

# Fair Play for Individuals, Foul Play for Groups? Auditing Anonymization’s Impact on ML Fairness

Héber H. Arcolezi<sup>a</sup>, Mina Alishahi<sup>b,\*</sup>, Adda-Akram Bendoukha<sup>c</sup> and Nesrine Kaaniche<sup>c</sup>

<sup>a</sup>Inria Centre at the University Grenoble Alpes, France

<sup>b</sup>Open Universiteit, Netherlands

<sup>c</sup>Samovar, Télécom SudParis, Institut Polytechnique de Paris, France

**Abstract.** Machine learning (ML) algorithms are heavily based on the availability of training data, which, depending on the domain, often includes sensitive information about data providers. This raises critical privacy concerns. Anonymization techniques have emerged as a practical solution to address these issues by generalizing features or suppressing data to make it more difficult to accurately identify individuals. Although recent studies have shown that privacy-enhancing technologies can influence ML predictions across different subgroups, thus affecting fair decision-making, the specific effects of anonymization techniques, such as  $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness, on ML fairness remain largely unexplored. In this work, we systematically audit the impact of anonymization techniques on ML fairness, evaluating both individual and group fairness. Our quantitative study reveals that anonymization can degrade group fairness metrics by up to four orders of magnitude. Conversely, similarity-based individual fairness metrics tend to improve under stronger anonymization, largely as a result of increased input homogeneity. By analyzing varying levels of anonymization across diverse privacy settings and data distributions, this study provides critical insights into the trade-offs between privacy, fairness, and utility, offering actionable guidelines for responsible AI development. Our code is publicly available at: <https://github.com/hharcolezi/anonymity-impact-fairness>.

## 1 Introduction

As machine learning (ML) systems increasingly shape critical decision-making across domains such as healthcare, finance, and social services, concerns about *privacy* and *fairness* have gained significant attention. Privacy is essential to safeguard personal data, ensuring compliance with regulatory frameworks and protecting individuals from potential data misuse. Fairness, on the other hand, ensures that ML models provide unbiased and equitable outcomes across different demographic groups. Both aspects are fundamental to fostering trust and accountability in AI-driven decision-making.

The growing emphasis on privacy is reflected in strict regulatory frameworks such as the General Data Protection Regulation (GDPR) [32] and the California Consumer Privacy Act (CCPA) [25], which impose stringent requirements on data collection, storage, and sharing. Beyond privacy, the proposed European Union AI Act [33] extends regulatory concerns to fairness, requiring AI systems to be transparent, non-discriminatory, and aligned with ethical and social

values. These evolving legal frameworks highlight the dual imperative of privacy and fairness in the design of ML models.

Given the importance of balancing privacy and fairness in machine learning, a substantial body of work has investigated their interplay—primarily through the lens of differential privacy (DP) [40, 20]. Much of this literature focuses on central DP mechanisms [17], such as DP-SGD, which have been shown to exacerbate group fairness disparities in some settings [7], while other studies report more bounded effects [15]. On the other hand, few works have concluded that Local DP mechanisms positively impact group fairness metrics [5, 12, 28, 29] by removing the dependency between protected attributes and the target variable.

Despite growing research on the interplay between fairness and DP, the *fairness implications of anonymized datasets on ML remain unexplored*. Anonymization methods, such as  $k$ -anonymity [38, 39], provide privacy guarantees by generalizing specific attributes or suppressing data to prevent re-identification attacks [27, 26, 30]. These techniques are widely used due to their simplicity, interpretability [21, 19], and compatibility with existing privacy regulations such as GDPR, as noted by the Article 29 Working Party [6].

Ensuring fairness in ML involves mainly two key principles: (i) *group fairness*, which ensures that model predictions are consistent across demographic groups [13, 8], and (ii) *individual fairness*, which ensures that similar individuals receive similar treatment [18]. While anonymization techniques are effective in preserving privacy, they introduce transformations such as generalization or suppression, which may distort data distributions and induce unintended bias in ML models [3]. These alterations can inadvertently affect protected attributes, shift group distributions, and influence fairness metrics.

Existing research on the interplay between anonymization and fairness has primarily focused on *dataset-level fairness*, evaluating how anonymization techniques alter dataset properties [36]. Some studies have explored optimal parameter selection, such as determining the best value of  $t$  in  $t$ -closeness, to balance privacy and fairness trade-offs [22, 35]. However, *a significant gap remains in understanding the direct impact of anonymization on model fairness*, particularly how these techniques influence bias propagation and fairness metrics in ML models. Addressing this challenge is crucial for the development of privacy-preserving yet fair AI systems.

**Our contributions.** In this paper, we present the first in-depth, systematic audit of the impact of three widely-used anonymization methods, namely,  $k$ -anonymity [38, 39],  $\ell$ -diversity [27], and  $t$ -closeness [26] on fairness in ML. While anonymization is widely

\* Corresponding Author. Email: [mina.sheikhalishahi@ou.nl](mailto:mina.sheikhalishahi@ou.nl)

used to meet privacy regulations such as GDPR, its downstream effects on fairness remain poorly understood. We fill this gap by examining their effects on *group fairness metrics* (e.g., equalized odds [23]), *individual fairness metrics* (e.g., similarity fairness [18]), while also examining the trade-offs with *utility metrics* (e.g., F1-score). Our findings highlight the nuanced interplay between anonymity, fairness, and utility, offering valuable insights and actionable guidance for practitioners working with anonymized datasets. In summary, the key contributions of this paper are:

- We conduct a comprehensive audit on the effects of  $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness on both group and individual fairness metrics across diverse datasets and ML models. This dual focus offers a nuanced understanding of how anonymization techniques influence different aspects of fairness.
- We analyze how factors such as record suppression thresholds, dataset size, and target class balance modulate the fairness-utility trade-off. This enables a fine-grained understanding of how anonymization interacts with data characteristics in practice.
- Based on our findings, we provide practical guidelines for mitigating fairness risks in ML pipelines. These include strategies for balancing privacy and fairness under different anonymization configurations, empowering practitioners to make informed decisions when deploying privacy-preserving technologies.

**Outline.** The remainder of this paper is structured as follows. In Section 2, we present preliminary concepts. In Section 3, we formulate the problem, describe our research questions, and present our experimental setup. Afterward, in Section 4, we present the experimental results with an analysis and guidelines for practitioners, and we conclude the paper with a discussion for future work in Section 5.

## 2 Preliminaries

### 2.1 Anonymization Methods

Three main anonymization methods are commonly used:  $k$ -anonymity [38, 39],  $\ell$ -diversity [27], and  $t$ -closeness [26]. These methods aim to reduce the risk of re-identification by transforming the data while maintaining its utility for downstream tasks. In the context of anonymization, attributes in a dataset are typically categorized as follows: 1) *identifiers* that uniquely identify individuals, such as social security numbers, 2) *quasi-identifier* (QI) attributes that, when combined, may potentially identify an individual, such as zip codes, birth dates, and genders. While these attributes lack uniqueness in isolation, their conjunction often yields a one-to-one mapping with identifiers. And 3) *sensitive* attributes that adversaries are prohibited from discovering, such as a patient’s disease or an employer’s salary. An equivalence class, also known as an equivalence group, is a set of records in a dataset that share the same values for the quasi-identifier attributes. We now formally define each anonymization method.

**$k$ -anonymity:** A dataset  $D$  satisfies  $k$ -anonymity with respect to a set of  $QIs$ , if and only if every equivalence class  $E \subseteq D$  formed by the unique combinations of values in  $QI$  contains at least  $k$  records [38, 39]. Formally,  $\forall E \subseteq D, |E| \geq k$ , where  $E$  is an equivalence class of records that share the same generalized values across  $QI$ . The  $k$ -anonymity property ensures that any individual represented in the dataset cannot be distinguished from at least  $k - 1$  other individuals based on the quasi-identifier attributes, thus reducing the risk of re-identification. Achieving  $k$ -anonymity typically involves

generalization (replacing specific values with broader categories) or suppression of quasi-identifiers to create equivalence classes.

**$\ell$ -diversity:** A dataset  $D$  satisfies  $\ell$ -diversity with respect to a set of attributes  $QIs$  and a sensitive attribute  $S$  if, for every equivalence class  $E \subseteq D$  formed by unique combinations of values in  $QI$ , there are at least  $\ell$  well-represented distinct values of  $S$  [27]. Formally,  $\forall E \subseteq D, |\text{distinct}(S(E))| \geq \ell$ , where  $S(E)$  denotes the set of sensitive attribute values in the equivalence class  $E$ .  $\ell$ -diversity ensures that each equivalence class contains sufficient diversity in the sensitive attribute values, mitigating the risk of attribute disclosure.

**$t$ -closeness:** A dataset  $D$  satisfies  $t$ -closeness with respect to a set of attributes  $QI$  and a sensitive attribute  $S$  if, for every equivalence class  $E \subseteq D$ , the distribution of  $S$  within  $E$  is within a distance  $t$  from the distribution of  $S$  in the overall dataset  $D$  [26]. Formally,  $\forall E \subseteq D, d(S(E), S(D)) \leq t$ , where  $d$  is a distance metric, such as the Earth Mover’s Distance (EMD) [34], and  $S(E)$  and  $S(D)$  represent the distributions of  $S$  in  $E$  and  $D$ , respectively. The  $t$ -closeness property ensures that sensitive attribute distributions in each equivalence class closely resemble the overall distribution, reducing the risk of inference attacks. Generalization and suppression are applied to satisfy the  $t$ -closeness constraint.

### 2.2 ML Fairness Metrics

Machine learning fairness [8] refers to the principle that ML models should produce predictions or decisions that are impartial and equitable across individuals or groups, particularly when protected attributes such as race, gender, or socioeconomic status are involved. This study employs a comprehensive evaluation of fairness using two primary categories of metrics: group fairness metrics (Section 2.2.1) and individual fairness metrics (Section 2.2.2).

#### 2.2.1 Group fairness metrics.

Group fairness metrics assess the degree to which model outcomes are distributed equitably across demographic groups defined by protected attributes. These metrics aim to ensure that the treatment of groups is consistent and does not result in discrimination.

**Model Accuracy Difference (MAD)** quantifies the difference in model accuracy between two specific demographic groups defined by a protected attribute  $A$ . Specifically, MAD evaluates whether the model performs equally well for individuals in group  $A = 1$  compared to those in group  $A = 0$ . Formally:

$$\text{MAD} = \Pr [\hat{Y} = Y | A = 1] - \Pr [\hat{Y} = Y | A = 0] \quad (1)$$

where  $Y$  is the true label,  $\hat{Y}$  is the predicted label, and  $A$  is the binary protected attribute defining two groups ( $A = 1$  and  $A = 0$ ). A value of  $\text{MAD} = 0$  signifies equal model accuracy across the two groups.

**Equalized Odds Difference (EOD)** [23] ensures that the model’s prediction is independent of the protected attribute, conditioned on the true label. It measures the disparity in true positive and false positive rates between two specific demographic groups defined by a protected attribute. Formally:

$$\text{EOD} = \Pr [\hat{Y} = 1 | Y = 1, A = 1] - \Pr [\hat{Y} = 1 | Y = 1, A = 0] \quad (2)$$

An  $\text{EOD} = 0$  indicates equality of true positive rates across groups.

**Statistical Parity Difference (SPD)** [18] evaluates whether the likelihood of a positive outcome is the same across two specific de-

mographic groups defined by a protected attribute. Formally:

$$\text{SPD} = \Pr[\hat{Y} = 1 \mid A = 1] - \Pr[\hat{Y} = 1 \mid A = 0] \quad (3)$$

An SPD = 0 indicates equal selection rates between the groups.

### 2.2.2 Individual fairness metrics.

Individual fairness metrics evaluate the consistency of a model’s predictions for similar individuals [18].

**Lipschitz Fairness (LF)** evaluates the maximum sensitivity of a model’s predictions to changes in its input features. It measures the largest rate of change in the model’s outputs relative to input variations, quantified by the Lipschitz constant. A lower LF value indicates better fairness, as it reflects reduced sensitivity and ensures that similar inputs yield similar predictions. Formally:

$$\text{LF} = \max_{i \neq j} \frac{\text{diff}(f(x_i), f(x_j))}{\text{dist}(x_i, x_j)}, \quad (4)$$

where  $f(x)$  is the model’s prediction for input  $x$ ,  $\text{dist}(x_i, x_j)$  is the distance between inputs  $x_i$  and  $x_j$ , and  $\text{diff}(f(x_i), f(x_j))$  is the difference between the model’s predictions for inputs  $x_i$  and  $x_j$ , often computed using measures such as entropy for classification tasks.

**Similarity Fairness (SF)** assesses the degree to which a model treats similar inputs consistently. It evaluates the average variation in predictions for input pairs that are deemed similar based on a defined similarity metric. A lower SF value indicates better fairness, as it ensures consistent treatment of similar individuals. Formally:

$$\text{SF} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}(x_i)|} \sum_{x_j \in \mathcal{N}(x_i)} |f(x_i) - f(x_j)| \cdot \text{sim}(x_i, x_j), \quad (5)$$

where  $\mathcal{N}(x_i)$  represents the neighborhood of  $x_i$ ,  $\text{sim}(x_i, x_j)$  is the similarity score derived from the inverse of the distance, and  $n$  is the total number of data points.

**Neighborhood Consistency Fairness (NCF)** evaluates the consistency of predictions within local neighborhoods, ensuring equitable treatment for instances with similar characteristics. A lower NCF value reflects better fairness, as it ensures that predictions are consistent for neighboring inputs. Formally:

$$\text{NCF} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}(x_i)|} \sum_{x_j \in \mathcal{N}(x_i)} \mathbb{I}(f(x_i) = f(x_j)), \quad (6)$$

where  $\mathcal{N}(x_i)$  is the neighborhood of  $x_i$ ,  $\mathbb{I}(\cdot)$  is the indicator function, which equals 1 if predictions for  $x_i$  and  $x_j$  are identical, and  $n$  is the total number of data points.

**Approximation via  $k$ -Nearest Neighbors (k-NN):** Computing LF, SF, and NCF requires evaluating pairwise distances between all data points, resulting in a computational complexity of  $O(n^2)$ , which is prohibitive for large datasets. To improve efficiency, we approximate these metrics using  $k$ -NN, reducing the complexity to  $O(kn)$ , where  $k \ll n$ , as also motivated in prior work [41]. Specifically, each sample is compared with its  $k$ -nearest neighbors instead of the entire dataset. We use Euclidean distance and set  $k = 100$ , which provides a balance between computational tractability and the ability to capture a meaningful local neighborhood. Throughout the paper, we refer to these approximated versions as Approximate LF (ALF) and Approximate SF (ASF).

## 3 Methodology and Experimental Setup

This section formulates the problem, introduces the research questions, and describes the experimental setup.

### 3.1 Problem Formulation

While anonymization aims to protect individual privacy by transforming the data before learning, its downstream impact on fairness in ML is poorly understood. We therefore aim to systematically audit how different anonymization methods influence group and individual fairness outcomes, as well as predictive performance, across a range of ML models and datasets.

To formalize this problem, we define the following components. Let  $\mathcal{D} = \{(a_i, x_i, y_i)\}_{i=1}^n$  be a dataset consisting of  $n$  i.i.d. samples drawn from an unknown joint distribution over  $\mathcal{A} \times \mathcal{X} \times \mathcal{Y}$ , where:

- $A \in \mathcal{A}$ : Protected attribute(s) (e.g., race, gender), often binary ( $A \in \{0, 1\}$ ), with  $A = 1$  denoting the *privileged group* and  $A = 0$  the *unprivileged group*.
- $X \in \mathcal{X}$ : Non-sensitive, non-protected features (e.g., employment status, credit score).
- $Y \in \{0, 1\}$ : Binary target variable (e.g., loan approval outcome), where  $Y = 1$  denotes a favorable decision.

Given this setting, a supervised ML model learns a predictive function  $f : \mathcal{A} \times \mathcal{X} \rightarrow [0, 1]$  such that  $\hat{Y} = f(A, X)$  approximates the true label  $Y$ . In practice, the model is trained on a dataset  $D'$ , which may differ from  $D$  due to privacy constraints. Specifically, to preserve privacy, we consider an anonymization mechanism  $\mathcal{A}$ , which transforms the raw dataset  $D$  into an anonymized dataset  $D' = \mathcal{A}(D)$ . In this work,  $\mathcal{A}$  represents one of the following methods:  $k$ -anonymity,  $\ell$ -diversity, or  $t$ -closeness, each of which enforces privacy by applying generalization and suppression over quasi-identifiers. We then train ML classifiers on both original ( $D$ ) and anonymized ( $D'$ ) datasets, and compare them using performance (e.g., F1-score) and fairness metrics (Section 2.2).

### 3.2 Research Questions

We structure our evaluation around the following research questions:

- **RQ1) How do different anonymization techniques and anonymity levels affect the fairness of ML models?** This research question explores the impact of three widely used anonymization techniques ( $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness) on the fairness of ML models. By adjusting their respective privacy parameters (i.e.,  $k$ ,  $\ell$ , and  $t$ ), we evaluate how these techniques influence fairness metrics and whether certain configurations disproportionately affect specific demographic groups. The experiments addressing this question are presented in Section 4.1.
- **RQ2) What is the impact of varying the record level suppression in anonymization on the fairness of ML models?** Because suppression often targets outlier data, this may disproportionately affect certain sub-populations, potentially exacerbating fairness disparities. This research question investigates how varying suppression thresholds (removing rows) impact fairness metrics. The experiments addressing this question are detailed in Section 4.2.
- **RQ3) What is the impact of varying target distributions on the fairness of ML models?** This research question examines how changes in the target distribution, specifically by varying the threshold used to binarize the *target variable* (see "Datasets" in

Section 3.3), influence fairness metrics. Varying the target distribution affects the balance between positive and negative outcomes in the dataset, which can, in turn, affect fairness metrics between demographic groups. The experiments addressing this question are presented in Section 4.3.

- **RQ4)** *How does dataset size influence the relationship between anonymization and fairness in ML models?* This question explores the role of dataset size in mediating the relationship between anonymization and fairness. By systematically varying the data fraction, we analyze how sample size influences the trade-offs between privacy, fairness, and utility. The experiments addressing this question are detailed in Section 4.4.
- **RQ5)** *To what extent are the fairness results obtained with XGBoost representative across different ML classifiers?* This research question investigates whether the fairness results observed in our default experiments using XGBoost [14] (**RQ1** – **RQ4**) generalize across other ML classifiers. By comparing the fairness metrics and predictive performance of multiple classifiers (*e.g.*, Random Forest, Neural Networks), we aim to assess whether the trends observed with XGBoost are consistent. The experiments addressing this question are presented in Section 4.5.

### 3.3 Experimental Setup

**Environment:** All algorithms are implemented in Python3 and executed on a local machine with 2.50GHz Intel Core i9 and 64GB RAM. The source code is publicly available in our *GitHub repository*: <https://github.com/hharcolez/anonymity-impact-fairness>.

**Datasets:** For our experiments, we used three widely-used benchmark datasets in the fairness literature: the *Adult* dataset from the UCI ML repository [9], the *Compas* dataset gathered by ProPublica [4], and the *ACSIncome* retrieved with the *folktables* [16] Python library. The datasets are randomly split into a training set (80%) and a testing set (20%). To simulate a worst-case privacy scenario, all attributes ( $X$  and  $A$ ) are treated as quasi-identifiers subject to generalization or suppression during anonymization. For fairness evaluation, we consider both *gender* and *race* as protected attributes across all datasets. A detailed description of each dataset is provided in Appendix A.

**Anonymization parameters:** The anonymization parameters are varied as follows:  $k \in \{1, 2, \dots, 9, 10, 25, 50, 75, 100\}$  for  $k$ -anonymity,  $\ell = 2$  for  $\ell$ -diversity (binary target), and  $t \in \{0.45, 0.50, 0.55\}$  for  $t$ -closeness<sup>1</sup>. Note that for  $k = 1$ , no anonymity is satisfied, which serves as the *non-private baseline*. Unless otherwise mentioned, the allowed record suppression level is fixed at `supp_level = 20%`<sup>2</sup>. The anonymization methods were implemented with Anjana [37]. The same generalization levels applied to the training set are replicated for the test set to ensure consistency and prevent discrepancies between training and evaluation.

**Model training:** For our experiments, we use XGBoost [14] as the default ML classifier. In Section 4.5, we further benchmark the performance and fairness of other state-of-the-art ML classifiers, including LightGBM (LGBM) [24], Random Forests [10], and Neural Networks (*i.e.*, Multi-Layer Perceptron – MLP). Classifiers are trained using their default hyperparameters to ensure consistency across experiments. All models are trained and evaluated on both the original

and anonymized datasets. Evaluation metrics are computed on predictions from the transformed test sets, enabling a comparative analysis of fairness and utility under different anonymization scenarios.

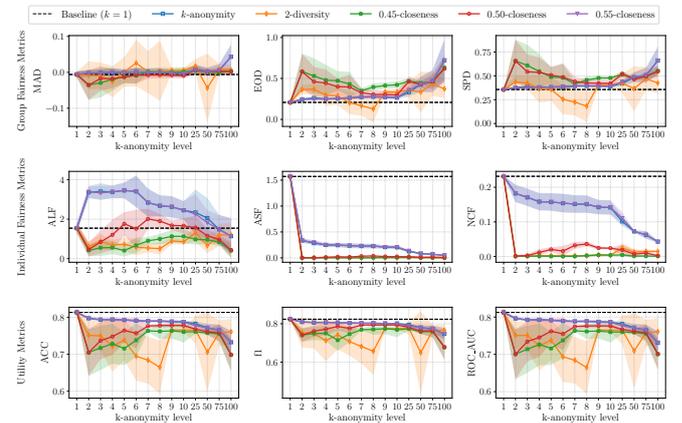
**Metrics:** We evaluate the performance of ML models trained on the original data (*i.e.*, baseline  $k = 1$ ) and anonymized data on utility and fairness. First, for *utility*, we use accuracy (ACC), F1-score (F1), and the area under the receiver operating characteristic curve (ROC AUC). Second, for *fairness*, we assess group and individual fairness metrics as defined in Section 2.2 (*i.e.*, MAD, SPD, EOD, ALF, ASF, and NCF). To address the randomness in train-test splitting and ML algorithms, all experiments are repeated over 40 runs, with the results reported as averages alongside their standard deviations.

## 4 Results and Analysis

Following the methodology and experimental setup described in the previous Section 3, this section presents an analysis of the impact of anonymization methods on ML fairness. Due to space limitations, we present in the main paper the results obtained using the *Adult* dataset with *gender* as the protected attribute. Additional details and discussions, including results with *race* as the protected attribute and those for both *Compas* and *ACSIncome* datasets (evaluating both *gender* and *race* as protected attributes), are provided in Appendix B. However, it is important to emphasize that the discussions in this section are broadly applicable to the findings across both datasets and all protected attributes.

### 4.1 Impact of Anonymization on Fairness in ML

To answer **RQ1** from Section 3.2, we consistently analyze the impact of anonymity methods on fairness in ML. Specifically, Figure 1 illustrates the impact of three anonymization techniques (*i.e.*,  $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (ACC, F1, ROC AUC) metrics, for the *Adult* dataset with *gender* as the protected attribute. The results are aggregated across different privacy parameter levels to provide a comprehensive overview of the trade-offs among anonymity, fairness, and utility in ML models.



**Figure 1.** Impact of anonymity methods ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness metrics (MAD, EOD, SPD), individual fairness metrics (ALF, ASF, NCF), and utility metrics (Accuracy, F1-score, ROC AUC) in ML. The results are based on the *Adult* dataset, with *gender* as the protected attribute for fairness evaluation.

From Figure 1, it is evident that *anonymization techniques negatively affect group fairness metrics* such as SPD and EOD. For instance, as the parameter  $k$  increases in  $k$ -anonymity, these fairness metrics tend to degrade, suggesting that stricter privacy constraints may amplify fairness disparities between demographic

<sup>1</sup> The values of  $t$  specify upper bounds on the Earth Mover’s Distance between the distribution of the sensitive attribute in the entire dataset and its distribution within each equivalence class.

<sup>2</sup> Our attempts with lower suppression levels (0%-15%) did not achieve anonymization for high  $k$  values because we operated under a *worst-case* assumption where all attributes except the target were treated as QIDs.

groups. Specifically, at  $k = 100$ , the EOD metric increases from  $\text{EOD} = 0.2$  (baseline  $k = 1$ ) to nearly  $\text{EOD} = 0.8$  for both  $k$ -anonymity and  $t$ -closeness, representing an approximately **four-fold increase**. Similarly, the SPD metric rises from  $\text{SPD} = 0.38$  (baseline  $k = 1$ ) to almost  $\text{SPD} = 0.68$  (for  $k$ -anonymity and  $t$ -closeness). A similar trend is observed in Figure 6 for the **Adult** dataset with `race` as the protected attribute, where both SPD and EOD exhibit slight increases for some values of  $k$ . Furthermore, as shown in Figures 7–10, at least two group fairness metrics consistently degrade across the **Compas** and **ACSIIncome** datasets, regardless of whether `gender` or `race` is used as the protected attribute. These findings highlight the broader implications of anonymization on fairness metrics across different datasets and protected attributes.

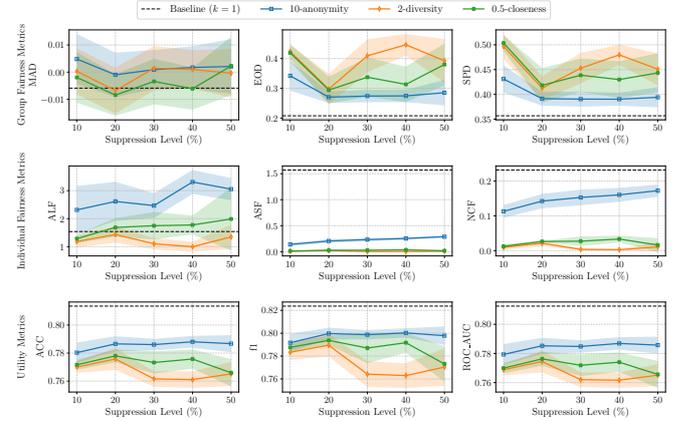
Regarding individual fairness, Figure 1 reveals distinct trends across the three evaluated metrics: ALF, ASF, and NCF. ALF exhibits mixed behavior depending on the anonymization technique. For  $k$ -anonymity (and 0.55-closeness), the metric tends to degrade for  $2 \leq k \leq 50$ , indicating that stricter anonymity levels increase the sensitivity of model predictions to input changes. This suggests reduced stability in preserving Lipschitz fairness under  $k$ -anonymity. Conversely, for  $\ell$ -diversity and 0.45-closeness, ALF shows improvement, demonstrating their potential to balance privacy and prediction stability for similar inputs. A similar trend is observed in Figure 6 for the **Adult** dataset with `race` as the protected attribute. In contrast, ASF generally worsens across all three anonymization methods in experiments with the **ACSIIncome** dataset, irrespective of the protected attribute (*i.e.*, see Figures 9 and 10).

These observations highlight the varied impact of anonymization on individual fairness metrics. While ALF demonstrates mixed trends, our results with all three datasets and protected attributes demonstrate that **anonymization techniques positively affect both ASF and NCF individual fairness metrics** (*i.e.*, see Figures 1, 6–10). Specifically, ASF consistently improves with higher levels of anonymization across all three techniques. This improvement indicates that anonymization reduces variations in predictions for similar individuals, fostering more equitable treatment at the individual level. Notably, for  $k \geq 10$ , the ASF metric stabilizes at nearly zero, reflecting a substantial alignment in predictions for similar data points. Similarly, the NCF metric also benefits from anonymization, showing a clear decline in its values as  $k$  increases. Lower NCF values imply fewer inconsistencies in predictions within local neighborhoods of similar inputs. For instance,  $k$ -anonymity demonstrates a significant drop in NCF at smaller  $k$ -values, with marginal improvements at higher levels of  $k$ . These trends can be attributed to the inherent differences in how ASF and NCF measure individual fairness. ASF captures the magnitude of prediction variations for similar inputs, penalizing even minor differences, whereas NCF focuses on exact consistency within local neighborhoods, relying on a binary indicator (see Eq. (6)). As anonymization increases, generalized attributes cause predictions to align more rapidly, leading to a faster decrease in ASF. In contrast, NCF, being less sensitive to small inconsistencies, decreases more gradually. Importantly, this smoothing effect and the resulting improvements in individual fairness metrics align with similar phenomena observed in differential privacy [18].

While privacy-preserving transformations reduce the risk of re-identification, they also degrade model utility [11], consistent with the well-known privacy-utility trade-off. Metrics such as accuracy, F1-score, and ROC AUC consistently decline as anonymization parameters are tightened. Notably, the lowest utility results are observed at  $k = 2$ , as  $\ell$ -diversity and  $t$ -closeness enforce the presence of at least two distinct classes within equivalence classes, leading to reduced predictive performance. However, utility metrics begin to recover as  $k$  increases, as there might be a majority class decision even under  $\ell$ -diversity or  $t$ -closeness. Despite this recovery, utility metrics remain below the baseline ( $k = 1$ ) and below the levels achieved under  $k$ -anonymity, underscoring the persistent trade-offs between privacy and utility.

## 4.2 Detailed Analysis #1: Record Suppression Levels and Their Effect on Fairness in ML

Following the findings of Section 4.1, to answer **RQ2** from Section 3.2, we now investigate the impact of the allowed record suppression level (`supp_level`) on fairness and utility in ML. Record suppression, which involves removing rows during anonymization, can exclude outliers from the dataset, potentially worsening fairness metrics as certain demographic groups may be disproportionately impacted. To understand this effect, we vary the suppression level (`supp_level`  $\in \{10, 20, 30, 40, 50\}$ ) and evaluate its impact on both group and individual fairness metrics, as well as utility. For the remainder of this analysis, the anonymity parameters are fixed at  $k = 10$  for  $k$ -anonymity and  $t = 0.5$  for  $t$ -closeness. The choice of  $k = 10$  represents a moderate level of anonymity, balancing privacy protection and data utility. Similarly,  $t = 0.5$  ensures a reasonable level of distributional closeness under  $t$ -closeness, reflecting practical settings often used in real-world applications. Fixing these parameters allows for a focused evaluation of how varying suppression levels impact fairness and utility, independent of additional variability introduced by the anonymity parameters.



**Figure 2.** Effect of allowed record suppression level (`supp_level`  $\in \{10, 20, 30, 40, 50\}$ ) in anonymization techniques (10-anonymity, 2-diversity, 0.5-closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. The results are derived from the **Adult** dataset, using `gender` as the protected attribute for fairness evaluation.

From the results for both datasets and protected attributes presented in Figures 2, 11–15, the impact of suppression levels on fairness and utility metrics reveals nuanced patterns. For instance, the effect of suppression on group fairness metrics such as SPD, EOD, and MAD is more mixed. In some cases, increasing suppression levels slightly improve these metrics, likely due to the exclusion of outliers that disproportionately skew fairness. For example, SPD and EOD in the **Adult** dataset with `race` as the protected attribute (Figure 11) exhibit slight improvements at moderate suppression levels. However, in other cases, particularly for the **ACSIIncome** dataset (Figures 14 and 15), the metrics remain stable or even worsen with higher suppression levels, suggesting that the removal of data points introduces disparities between demographic groups.

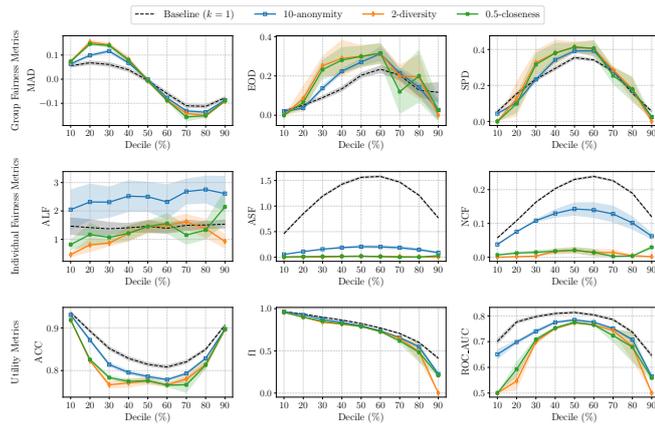
In addition, as the suppression level increases, individual fairness metrics such as ALF, ASF, and NCF show some degradation across all anonymization techniques and datasets. Higher suppression levels remove more rows, which disproportionately affects local neighborhoods and similar instances, introducing inconsistencies in predictions. For instance, in the **Adult** dataset with `gender` as the protected attribute (Figure 2), NCF values increase steadily as the suppression level rises, signaling worsening prediction consistency. Similarly, both ALF and ASF metrics are also worsened, highlighting increased variations in predictions for similar inputs. This

trend suggests that higher suppression levels may hinder the ability of anonymization methods to preserve fairness at the individual level.

In terms of utility, record suppression shows mixed impacts on utility metrics such as ACC, F1-score, and ROC AUC. In most cases, utility metrics remain stable or even improve slightly as the suppression level increases, likely due to the removal of noisy or less representative data. However, in a few instances (for  $\ell$ -diversity and  $t$ -closeness), utility decreases, particularly at higher suppression levels, reflecting the trade-off between privacy and maintaining a dataset that is representative of the original population. These observations suggest that while suppression can enhance privacy, *its effect on utility is nuanced and may vary depending on the dataset and the anonymization method applied.*

### 4.3 Detailed Analysis #2: ML Fairness Across Target Distribution Variations

To answer **RQ3** from Section 3.2, we now investigate the impact of target distribution variations on fairness and utility in ML models. With both **Adult** and **ACSIIncome** datasets, we modify the distribution of the income target variable by thresholding it at deciles ranging from 10% to 90%, simulating shifts in the balance between positive and negative classes. Similarly, for the **Compas** dataset, we modify the distribution of the COMPAS risk score target variable by thresholding it from scores 1 to 9. Figure 3 presents the results of these experiments focusing on the **Adult** dataset with gender as the protected attribute. Additional results for the **Adult** dataset with race, and both gender and race in the **ACSIIncome** and **Compas** datasets are shown in Figures 16–20.



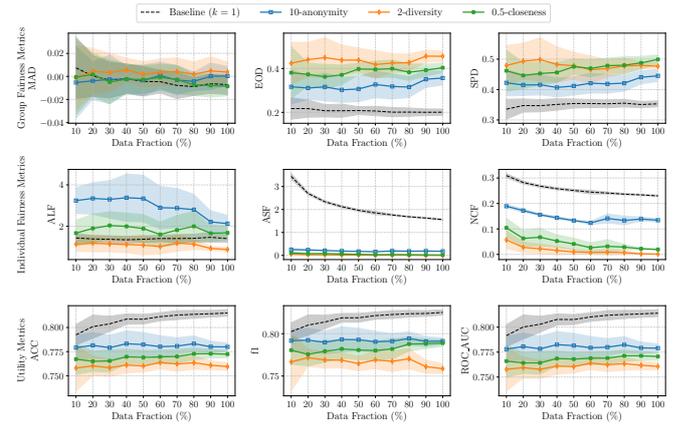
**Figure 3.** Effect of target distribution variation (*i.e.*, thresholded at deciles ranging from 10% to 90%) on anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) regarding group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are presented for the **Adult** dataset, with gender serving as the protected attribute for fairness evaluation.

The results across all datasets and protected attributes (Figures 3, 16–20) reveal consistent trends regarding the impact of anonymization techniques on fairness and utility metrics. Notably, the anonymized models tend to follow the same general shape as the baseline across varying deciles, indicating that the overall distributional patterns of fairness and utility metrics are preserved under anonymization. However, the magnitude of these metrics differs: anonymization consistently worsens group fairness metrics such as MAD, SPD, and EOD, reflecting an amplification of disparities between demographic groups. Conversely, individual fairness metrics, particularly ASF and NCF, generally benefit from anonymization, showing improved consistency in predictions for similar inputs. On the other hand, utility metrics such as ACC, F1-score, and ROC AUC are negatively impacted, with performance often falling below the

baseline. These findings align with the conclusions drawn for **RQ1** in Section 4.1, further underscoring the trade-offs between privacy, fairness, and utility introduced by anonymization techniques.

### 4.4 Detailed Analysis #3: Impact of Data Size on Fairness in ML

To address whether data fraction impacts fairness as formulated in **RQ3** in Section 3.2, we analyze the behavior of anonymization techniques across subsampled data fractions. Specifically, we vary the data fraction from 10% to 100% of the original dataset, subsampling the data randomly at each fraction level. Figure 4 illustrates the results for the **Adult** dataset with gender as the protected attribute; additional results for other datasets and protected attributes are provided in Figures 21–25 in Appendix B.



**Figure 4.** Effect of varying data fraction on the performance of anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) in terms of group fairness metrics (MAD, EOD, SPD), individual fairness metrics (ALF, ASF, NCF), and utility metrics (Accuracy, F1-score, ROC AUC) in ML. This analysis is performed using the **Adult** dataset, considering gender as the protected attribute for fairness evaluation.

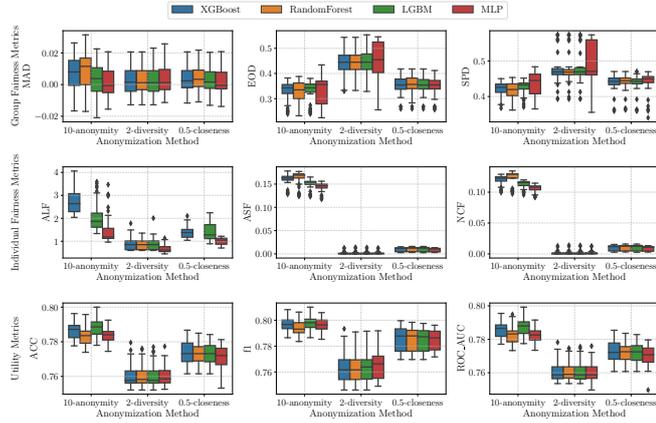
Figures 4, 21–25 reveal that performance and fairness metrics under anonymization techniques follow a relatively stable pattern across different data fraction levels sampled from 10% to 100%. For instance, anonymization introduces consistent increases in group fairness disparities compared to the baseline, regardless of the dataset size. Similarly, we observe a positive effect of anonymization on individual fairness that remains stable across data sizes. Performance metrics improve gradually with increasing data, as expected. In conclusion, these findings suggest that the trade-offs between privacy, fairness, and utility are primarily driven by the anonymization techniques and their parameter settings, rather than by the scale of the dataset. The random sampling of fractions introduces minor variability but does not fundamentally alter the patterns observed.

### 4.5 Detailed Analysis #4: Comparison of ML Classifiers in Fairness Under Anonymization

Finally, to address **RQ5** in Section 3.2, we compare the impact of anonymization techniques on fairness and utility across multiple state-of-the-art ML classifiers, including LGBM [24], Random Forest [10], Neural Networks (MLP), and XGBoost [14]. Figure 5 illustrates the results using the **Adult** dataset with gender as the protected attribute; additional comparisons are shown in Figures 26–30.

From Figure 5, 26–30, results show that the trends observed with XGBoost in previous sections (**RQ1–RQ4**) are consistent across other classifiers. Specifically, anonymization continues to negatively affect group fairness and to improve individual fairness. Moreover,

as expected, utility degrades under anonymization, regardless of the classifier. However, the magnitude of degradation varies slightly. XGBoost and LGBM tend to retain higher performance than Random Forest and MLP, especially in terms of F1-score and ROC AUC. Although minor differences exist, such as XGBoost’s marginal advantage in utility and fairness stability, the median values for all fairness and utility metrics are nearly identical across classifiers. This visual and statistical consistency reinforces that our findings are not specific to XGBoost. Hence, the conclusions drawn from XGBoost experiments can be considered broadly representative, supporting the generalizability of our results across different learning algorithms.



**Figure 5.** Comparison of the impact of different state-of-the-art ML classifiers on anonymized dataset ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) and relation to group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are based on the Adult dataset, with `gender` as the protected attribute for fairness evaluation.

#### 4.6 General Findings and Practical Guidelines

This section summarizes the key findings from Sections 4.1–4.5 and provides actionable guidelines for practitioners.

**Anonymity negatively impacts group fairness in ML:** Anonymization methods generally degraded group fairness metrics (*i.e.*, MAD, EOD, and SPD) as privacy parameters tightened. By coarsening quasi-identifiers and suppressing records, these anonymity-based methods distort the true joint distribution of  $(A, X, Y)$ , skew subgroup prevalences and error rates, and thereby amplify accuracy and positive-rate gaps between privileged and unprivileged groups.

**Anonymity positively impacts individual fairness in ML:** Anonymization techniques improve similarity-based individual fairness metrics such as ASF and NCF by inducing input homogeneity through generalization, which leads to more consistent predictions for similar individuals. A similar effect has been observed with differential privacy mechanisms [18]. However, since these improvements emerge from changes in data structure rather than targeted fairness interventions, they should be interpreted carefully to avoid overclaiming fairness benefits (see [1, 2, 31]).

**Higher record suppression levels negatively impacts individual fairness in ML:** Our findings reveal that higher suppression thresholds decrease the density of local equivalence classes, undermining the model’s capacity to generate consistent predictions for nearby instances. In contrast, group fairness metrics show mixed behavior, with minimal or no consistent degradation across suppression levels.

**Target distribution variations amplify group fairness disparities but stabilize individual fairness:** Adjusting the distribution of the target variable (*e.g.*, varying thresholds for binarization) has a significant impact on group fairness metrics, with disparities (*e.g.*, EOD and SPD) peaking at middle deciles. However, individual fairness metrics such as ASF and NCF remain relatively stable across target variations, benefiting from anonymization-induced generalization.

**Data size variations have minimal impact on fairness trends:** Across subsampled data fractions (10% to 100%), fairness and utility metrics follow consistent trends. Group fairness metrics remain negatively impacted by anonymization, while individual fairness metrics improve consistently. Utility metrics exhibit minor degradation at smaller fractions but generally remain stable.

**XGBoost findings generalize across ML classifiers:** Fairness and utility trends observed with XGBoost are consistent across other classifiers, such as Random Forest and Neural Networks (MLP). While XGBoost often achieves slightly better utility, the broader patterns—negative group fairness impacts, positive individual fairness outcomes, and utility trade-offs—remain similar. This consistency supports the generalizability of conclusions drawn from XGBoost experiments.

**Guidelines.** Based on our empirical findings, we propose the following recommendations for practitioners working with anonymized datasets in ML pipelines:

- **Use moderate privacy parameters** (*e.g.*,  $k = 10$ ,  $t = 0.5$ ) to balance privacy, fairness, and utility. Stricter settings can severely degrade group fairness and predictive performance.
- **Handle record suppression carefully** as high suppression thresholds may disproportionately remove minority or outlier samples, harming both individual and group fairness. When possible, combine suppression with imputation or domain-specific filtering to minimize bias.
- **Avoid median splits when binarizing continuous target variable**, as they tend to maximize group disparities. Evaluate multiple cutpoints across deciles. We found that thresholds below the 30th or above the 70th percentile yields lower group disparities; whereas median splits (40–60%) tend to exacerbate them.
- **Interpret individual fairness improvements cautiously.** Improvements in individual-fairness scores (ASF, NCF) are primarily due to feature homogenization and may not reflect “*genuine improvements in equitable treatment*”. Practitioners should therefore exercise caution and perform targeted case audits to avoid the risk of “fair washing/hacking” [1, 2, 31], for example.

## 5 Conclusion and Perspectives

This study systematically audits the tradeoff between anonymization techniques and fairness in ML. Through a comprehensive analysis, we evaluated the effects of three well-known anonymization methods ( $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness) on group and individual fairness metrics, as well as utility metrics, across multiple datasets and ML classifiers. By addressing five key research questions, our findings highlight the inherent trade-offs between privacy, fairness, and utility in anonymized ML models. Overall, our results show that anonymization tends to negatively affect group fairness metrics, often exacerbating disparities between demographic groups as privacy constraints increase. In contrast, similarity-based individual fairness metrics tend to improve under stronger anonymization, driven by increased data homogeneity. While this effect aligns with phenomena observed in differential privacy [18], it stems from data structural smoothing rather than fairness-aware optimization, and must therefore be interpreted with caution. Importantly, we demonstrate

that these findings are robust across different datasets, protected attributes, target distributions, data scales, suppression levels, and learning algorithms, confirming the generalizability of our conclusions. This work opens several promising research directions. A theoretical characterization of how anonymization influences fairness, *e.g.*, through the lenses of causal modeling or information-theoretic frameworks, remains an open challenge. Future efforts could also focus on designing privacy mechanisms that jointly optimize for fairness constraints, or on adapting anonymization strategies to richer settings such as multi-class classification or regression.

## Acknowledgements

This work has been partially supported by the French National Research Agency (ANR), under contracts “ANR-24-CE23-6239” JCJC project AI-PULSE, “ANR-19-P3IA-0003” MIAI @ Grenoble Alpes, “ANR-22-CE39-0002” JCJC project EQUIHID, and “ANR 22-PECY-0002” IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR.

## References

- [1] U. Aivodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp. Fairwashing: the risk of rationalization. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 161–170. PMLR, 09–15 Jun 2019.
- [2] U. Aivodji, H. Arai, S. Gambs, and S. Hara. Characterizing the risk of fairwashing. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14822–14834. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/7caf5e22ea3eb8175ab518429c8589a4-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/7caf5e22ea3eb8175ab518429c8589a4-Paper.pdf).
- [3] M. Alishahi and N. Zannone. Not a free lunch, but a cheap one: On classifiers performance on anonymized datasets. In *Data and Applications Security and Privacy*, pages 237–258. Springer International Publishing, 2021.
- [4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.
- [5] H. H. Arcolezi, K. Makhlof, and C. Palamidessi. (local) differential privacy has no disparate impact on fairness. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 3–21. Springer, 2023.
- [6] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymization techniques, 2014. [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf).
- [7] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [8] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- [9] B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: <https://doi.org/10.24432/C5XW20>.
- [10] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [11] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’08, page 70–78. Association for Computing Machinery, 2008. ISBN 9781605581934.
- [12] A. N. Carey, K. Bhaila, and X. Wu. Randomized response has no disparate impact on model accuracy. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5460–5465, 2023.
- [13] S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), 2024. ISSN 0360-0300.
- [14] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. KDD ’16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785.
- [15] A. S. de Oliveira, C. Kaplan, K. Mallat, and T. Chakraborty. An empirical analysis of fairness notions under differential privacy. *arXiv preprint arXiv:2302.02910*, 2023.
- [16] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.
- [17] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.
- [18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, Jan. 2012.
- [19] A. Fathalizadeh, V. Moghtadaiee, and M. Alishahi. On the privacy protection of indoor location dataset using anonymization. *Computers & Security*, 117:102665, 2022. ISSN 0167-4048.
- [20] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In *International Joint Conference on Artificial Intelligence*, IJCAI-2022, page 5470–5477. International Joint Conferences on Artificial Intelligence Organization, 2022.
- [21] S. A. Gaballah, L. Abdullah, M. Alishahi, T. H. L. Nguyen, E. Zimmer, M. Mühlhäuser, and K. Marky. Anonify: Decentralized dual-level k-anonymity for medical data donation. *Proc. Priv. Enhancing Technol.*, 2024(3):94–108, 2024.
- [22] S. Hajian and J. Domingo-Ferrer. A study on the impact of data anonymization on anti-discrimination. In *IEEE International Conference on Data Mining Workshops*, pages 352–359, 2012.
- [23] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [25] C. S. Legislature. California consumer privacy act of 2018, 2018. Assembly Bill No. 375, Chapter 55.
- [26] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *International Conference on Data Engineering*, pages 106–115, 2007.
- [27] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), 2007. ISSN 1556-4681.
- [28] K. Makhlof, H. H. Arcolezi, S. Zhidou, G. B. Brahim, and C. Palamidessi. On the impact of multi-dimensional local differential privacy on fairness. *Data Mining and Knowledge Discovery*, 38(4): 2252–2275, 2024.
- [29] K. Makhlof, T. Stefanović, H. H. Arcolezi, and C. Palamidessi. A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results. In *2024 IEEE 37th Computer Security Foundations Symposium (CSF)*, pages 1–16, 2024. doi: 10.1109/CSF61375.2024.00039.
- [30] F. Martinelli and M. SheikhAlishahi. Distributed data anonymization. In *IEEE Intl Conf on Dependable, Autonomic and Secure Computing (DASC)*, pages 580–586, 2019.
- [31] K. Meding and T. Hagendorff. Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology*, 37(1): 4, 2024.
- [32] E. Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679>.
- [33] E. Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act) (text with eea relevance), 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- [34] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000.
- [35] S. Ruggieri. Data anonymity meets non-discrimination. In *International Conference on Data Mining Workshops*, pages 875–882, 2013.
- [36] S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. *Trans. Data Privacy*, 7(2):99–129, 2014. ISSN 1888-5063.
- [37] J. Sáinz-Pardo Díaz and Á. López García. An open source python library for anonymizing sensitive data. *Scientific Data*, 11(1):1289, 2024.
- [38] P. Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–

1027, 2001.

- [39] L. Sweeney.  $k$ -anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10 (05):557–570, Oct. 2002.
- [40] K. Yao and M. Juarez. Sok: What makes private learning unfair? *arXiv preprint arXiv:2501.14414*, 2025.
- [41] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

## A Description of Datasets

- Adult Dataset:** The Adult dataset, obtained from the UCI ML repository [9] originates from the 1994 U.S. Census Bureau database. For this study, we use the Reconstructed Adult dataset as described in [16], where the target variable, *income*, is represented as a discrete value. This variable can be binarized into  $> \text{ThresholdIncome}$  or  $\leq \text{ThresholdIncome}$ , where *ThresholdIncome* serves as a configurable threshold. By default, the threshold is set to the median income value; however, in Section 4.3, we vary this threshold to analyze the impact of target distribution shifts on fairness and utility. After preprocessing and data cleaning, the dataset consists of  $n = 45,849$  samples and 10 discrete and categorical attributes. The *protected attributes* used for fairness evaluation are gender and race, while the *target attribute* is income.
- Compas Dataset:** The Compas dataset, curated by ProPublica [4], contains defendants from Broward County, Florida, screened between 2013 and 2014. We restrict to Black and White individuals who received a COMPAS risk score within 30 days of arrest. After preprocessing, the dataset comprises  $n = 5,278$  and five attributes: *race*, *sex*, *age*, *priors\_count*, and *days\_b\_screening\_arrest*. The target variable is the COMPAS risk score (*v\_decile\_score*), which consists of a rating of 1 – 10; the higher the score, the more likely the defendant is to re-offend. This target variable can also be binarized into  $> \text{RiskScore}$  or  $\leq \text{RiskScore}$ , where *RiskScore* serves as a configurable threshold. By default, the threshold is set to the median score value (*i.e.*, 3); however, as in Section 4.3, we vary this threshold (1–10) to analyze the impact of target distribution shifts on fairness and utility. The *protected attributes* used for fairness evaluation are gender and race.
- ACSIncome Dataset:** The ACSIncome dataset, sourced from the U.S. Census Bureau’s American Community Survey (ACS), represents a geographically distributed sample of individuals across U.S. states. Similar to the Adult dataset, the target variable *income* is categorized as  $> \text{ThresholdIncome}$  or  $\leq \text{ThresholdIncome}$ , with *ThresholdIncome* configurable to several representative income levels (median by default). For this study, we use the 2018 1-Year ACS Public Use Microdata Sample (*survey\_year="2018"* and *horizon="1-Year"*) across all U.S. states. The whole dataset contains  $n = 1,599,229$  data points with 10 discrete and categorical attributes. To reduce computational resource consumption, we randomly sample 10% of the data. The *protected attributes* for *fairness evaluation* are gender and race (*i.e.*, *SEX*, *RAC1P*), while the *target attribute* is income.

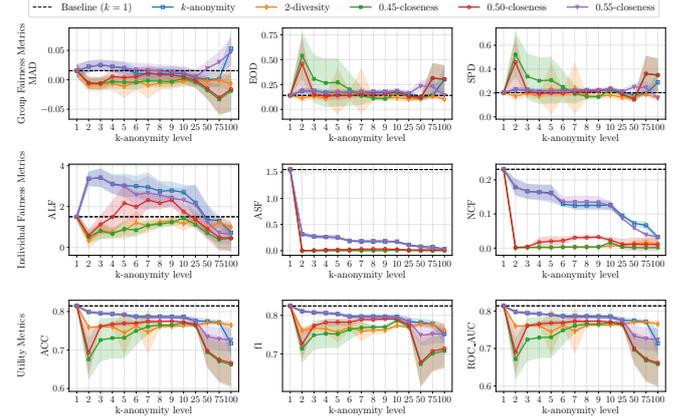
## B Additional Results

This section presents additional analyses answering the same research questions presented in Section 3.2 for Adult dataset with *race* as the protected attribute, on the Compas dataset with gender and *race* as protected attributes, and on the ACSIncome dataset with *SEX* and *RAC1P* as protected attributes.

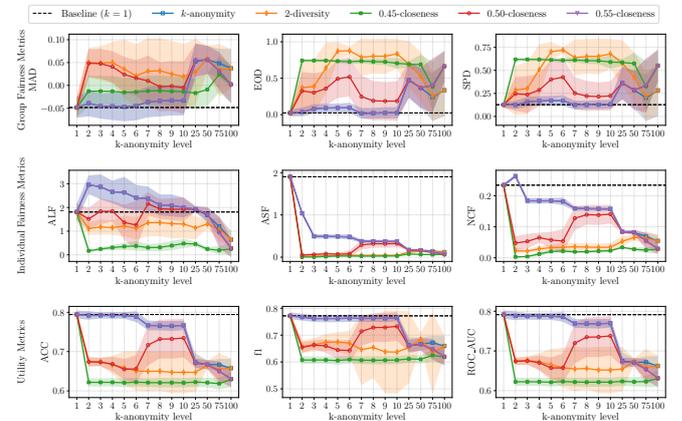
### B.1 Impact of Anonymization on Fairness in ML

Figures 6, 7, 8, 9, and 10 show the impact of  $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness group and individual fairness across different datasets. Specifically, they present results for the Adult dataset with *race* as the protected attribute and for both Compas and ACSIncome datasets, considering gender and *race* as protected attributes. These experiments extend the findings discussed in Section 4.1.

The results confirm that anonymization negatively affects group fairness, with the impact becoming more pronounced as privacy constraints become stricter. Conversely, while the trends in Approximate Lipschitz Fairness (ALF) vary across different settings, anonymization generally has a positive effect on individual fairness.



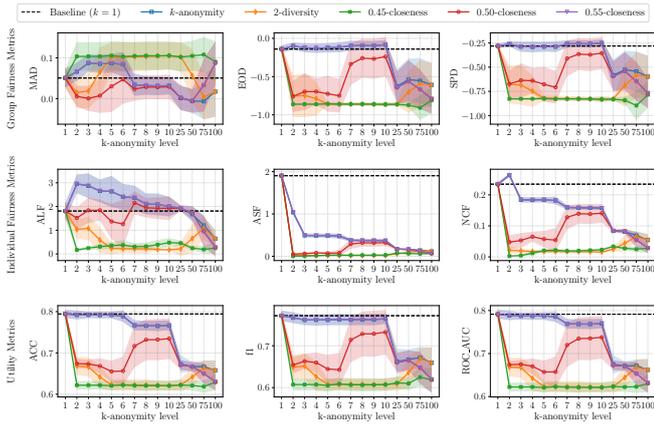
**Figure 6.** Impact of anonymity methods ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the Adult dataset with *race* as the protected attribute for fairness evaluation.



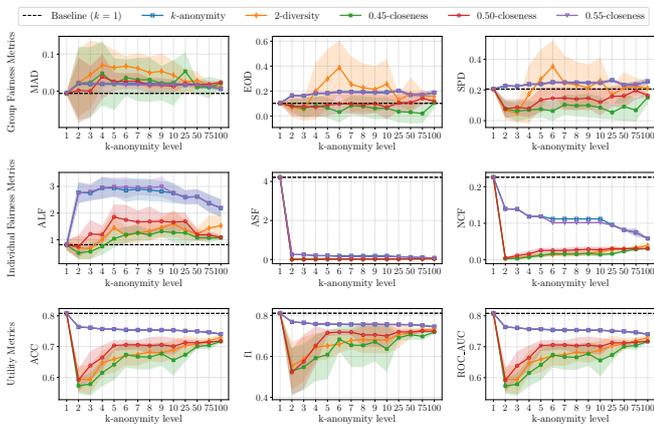
**Figure 7.** Impact of anonymity methods ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the Compas dataset with *gender* as the protected attribute for fairness evaluation.

### B.2 Impact of Record Suppression Levels on Fairness in ML

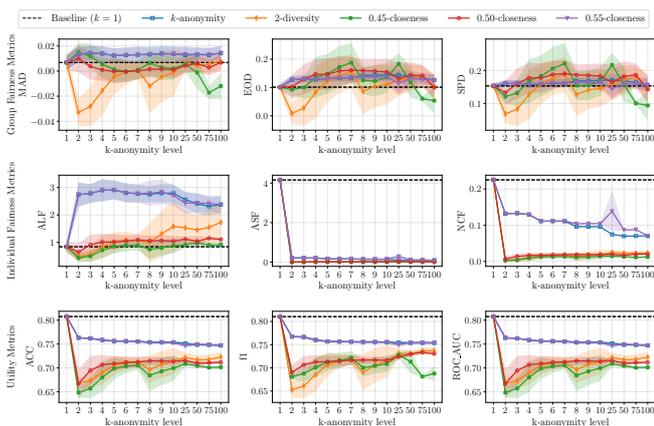
Figures 11, 12, 13, 14, and 15 demonstrate the impact of suppression levels on ML fairness for the Adult dataset with *race* as the protected attribute and for both Compas and ACSIncome datasets,



**Figure 8.** Impact of anonymity methods ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the Compas dataset with race as the protected attribute for fairness evaluation.



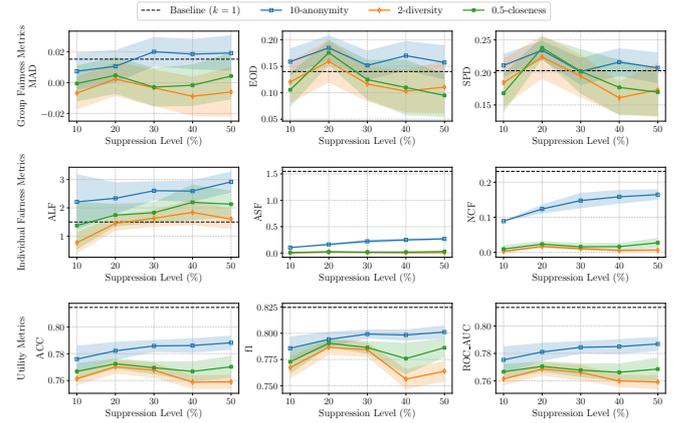
**Figure 9.** Impact of anonymity methods ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the ACSIncome dataset with gender as the protected attribute for fairness evaluation.



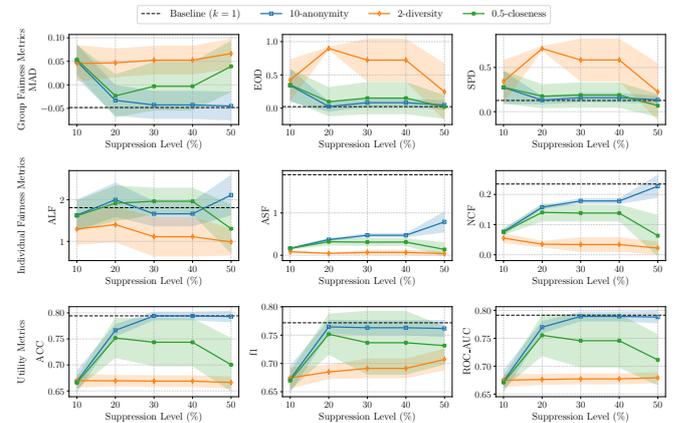
**Figure 10.** Impact of anonymity methods ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the ACSIncome dataset with race as the protected attribute for fairness evaluation.

considering gender and race as protected attributes, respectively. These results extend the experiments presented in Section 4.2.

It can be observed that the impact of suppression levels on fairness and utility metrics exhibits complex patterns across datasets and protected attributes. For group fairness, metrics such as SPD, EOD, and MAD show mixed trends—moderate suppression levels sometimes lead to slight improvements by removing outliers, while in other cases, especially in the ACSIncome dataset, fairness metrics remain stable or worsen due to the introduction of demographic disparities. In contrast, individual fairness metrics (ALF, ASF, and NCF) consistently degrade as suppression increases, as the removal of records disrupts local consistency in predictions. Regarding utility, suppression generally has a neutral or slightly positive effect by eliminating noisy data, but at high levels, it can reduce model performance, particularly for  $\ell$ -diversity and  $t$ -closeness.



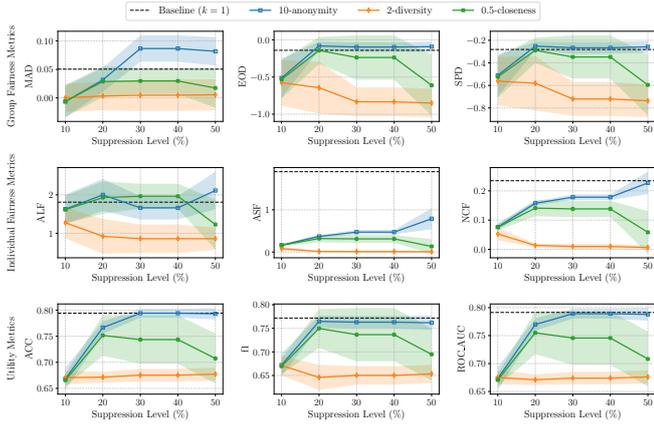
**Figure 11.** Impact of suppression level in anonymization ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the Adult dataset with gender as the protected attribute for fairness evaluation.



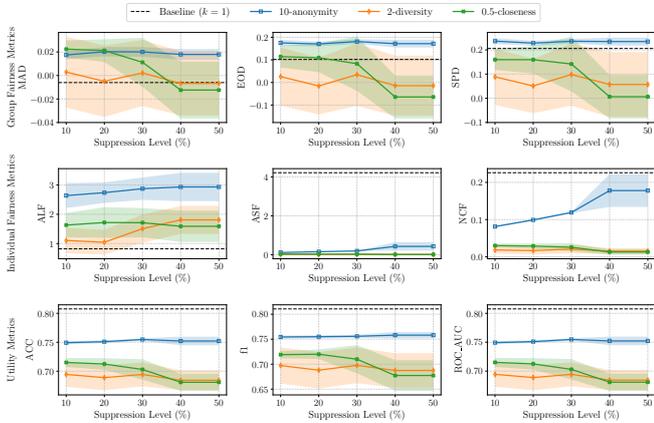
**Figure 12.** Impact of suppression level in anonymization ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the Compas dataset with gender as the protected attribute for fairness evaluation.

### B.3 ML Fairness Across Target Distribution Variations

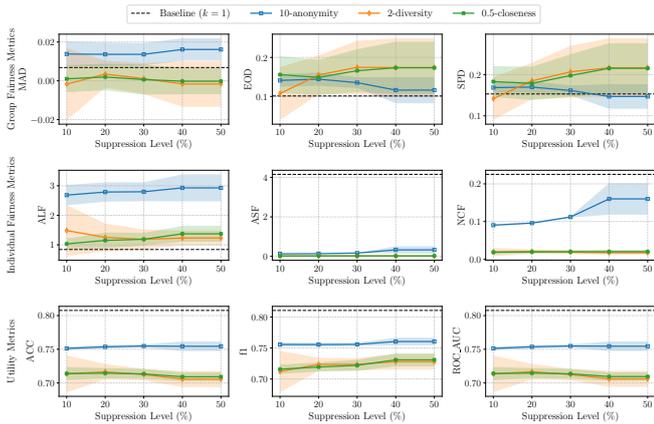
Figures 16, 17, 18, 19, and 20 show the ML fairness across target distribution for the Adult dataset with race as the protected at-



**Figure 13.** Impact of suppression level in anonymization ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the Compas dataset with `race` as the protected attribute for fairness evaluation.



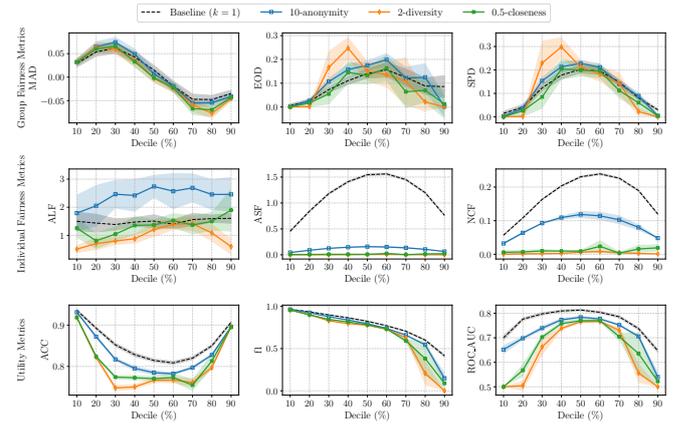
**Figure 14.** Impact of suppression level in anonymization ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) in ML. Results with the ACSIncome dataset with `gender` as the protected attribute for fairness evaluation.



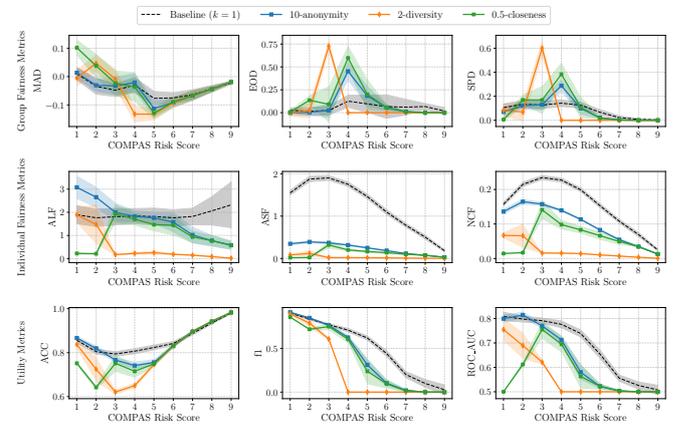
**Figure 15.** Impact of suppression level in anonymization ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) on group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results with the ACSIncome dataset with `race` as the protected attribute and for both Compas and ACSIncome datasets, consider-

ing gender and race as protected attributes, respectively. These results extend the experiments presented in Section 4.3.

These results reveal that anonymization techniques generally preserve the overall distributional patterns of fairness and utility metrics compared to the baseline. However, they significantly affect metric magnitudes: group fairness metrics (MAD, SPD, and EOD) deteriorate under anonymization, amplifying disparities between demographic groups, whereas individual fairness metrics (ASF and NCF) tend to improve, indicating greater consistency in model predictions. Meanwhile, utility metrics (ACC, F1-score, and ROC AUC) consistently decline, highlighting the trade-off between privacy and predictive performance.



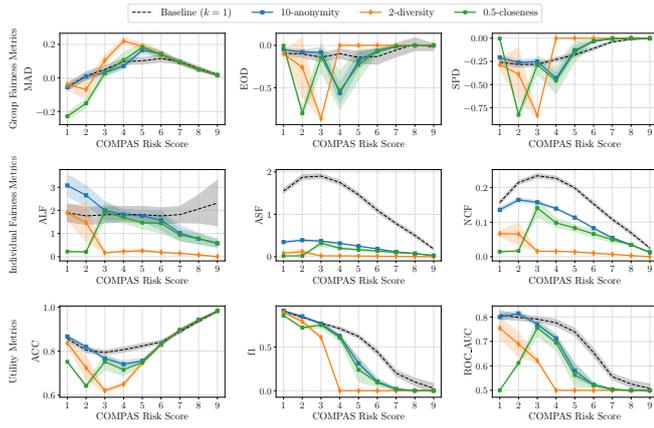
**Figure 16.** Effect of target distribution changes (*i.e.*, thresholded at deciles ranging from 10% to 90%) on anonymity techniques ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) regarding group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are presented for the Adult dataset, with `race` serving as the protected attribute for fairness evaluation.



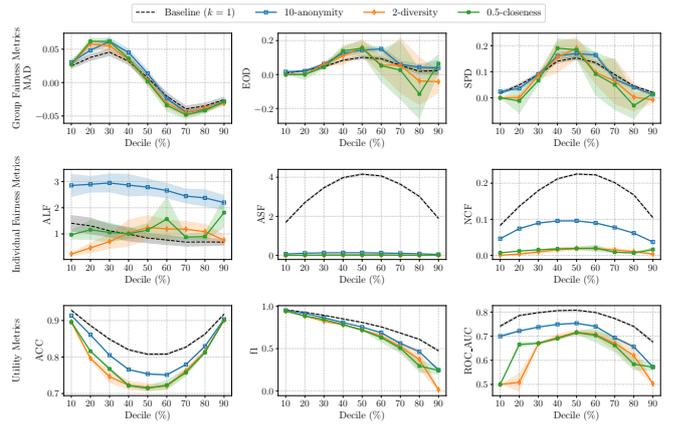
**Figure 17.** Effect of target distribution changes (*i.e.*, thresholded at COMPAS Risk Score ranging from 1 to 9) on anonymity techniques ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) regarding group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are presented for the Compas dataset, with `gender` serving as the protected attribute for fairness evaluation.

#### B.4 Impact of Data Size on Fairness in ML

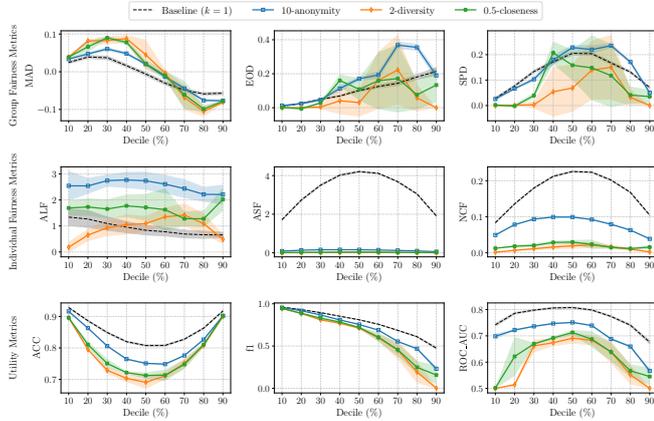
Figures 21, 22, 23, 24, and 25 show the effect of dataset sizes on ML fairness for the Adult dataset with `race` as the protected attribute and for both Compas and ACSIncome datasets, considering



**Figure 18.** Effect of target distribution changes (*i.e.*, thresholded at COMPAS Risk Score ranging from 1 to 9) on anonymity techniques ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) regarding group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are presented for the **Compas** dataset, with **race** serving as the protected attribute for fairness evaluation.



**Figure 20.** Effect of target distribution changes (*i.e.*, thresholded at deciles ranging from 10% to 90%) on anonymity techniques ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) regarding group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are presented for the **ACSIncome** dataset, with **race** serving as the protected attribute for fairness evaluation.



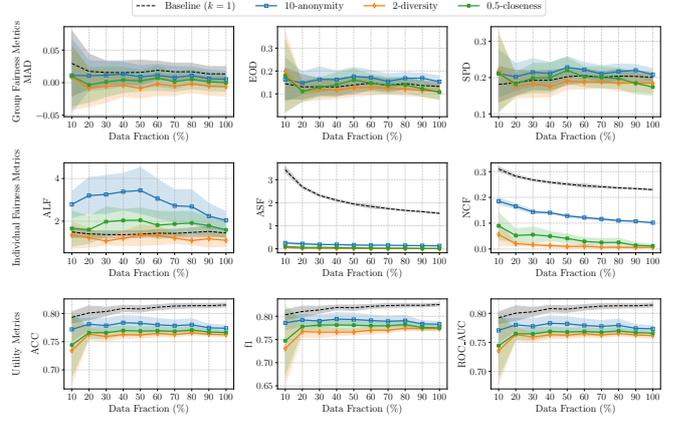
**Figure 19.** Effect of target distribution changes (*i.e.*, thresholded at deciles ranging from 10% to 90%) on anonymity techniques ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) regarding group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are presented for the **ACSIncome** dataset, with **gender** serving as the protected attribute for fairness evaluation.

gender and race as protected attributes, respectively. These results extend the experiments presented in Section 4.4.

These experiments demonstrate that fairness and performance metrics under anonymization remain relatively stable across varying data fractions, ranging from 10% to 100% of the dataset. These results indicate that the trade-offs between privacy, fairness, and utility are predominantly influenced by the choice of anonymization techniques and their parameter configurations, rather than by the dataset size. While random sampling introduces slight variations, it does not fundamentally alter the observed trends, reinforcing the robustness of the identified patterns.

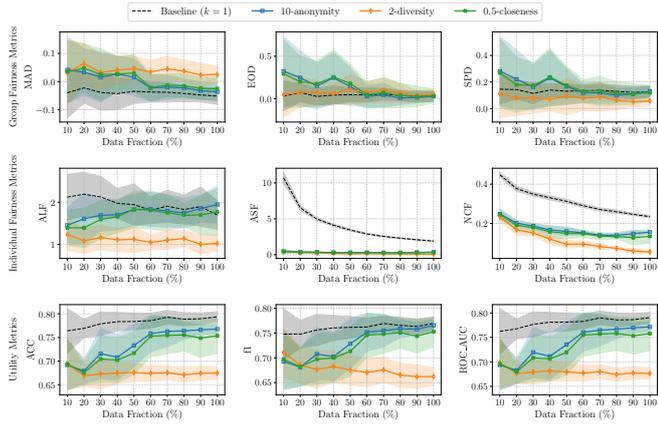
### B.5 Comparison of ML Classifiers in Fairness Under Anonymization

Figures 26, 27, 28, 29, and 30 show the ML fairness using different ML models on anonymized datasets for the **Adult** dataset with **race** as the protected attribute and for the **ACSIncome** dataset, considering **SEX** and **RACE** as protected attributes, respectively. These results extend the experiments presented in Section 4.5.

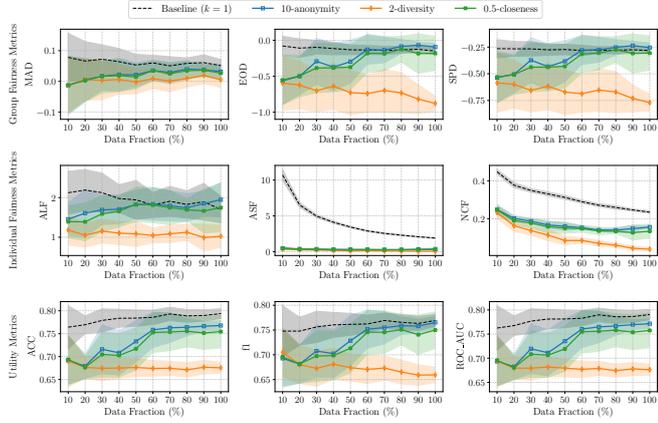


**Figure 21.** Effect of varying data fraction on the performance of anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) in terms of group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. This analysis is performed using the **Adult** dataset, considering **race** as the protected attribute for fairness evaluation.

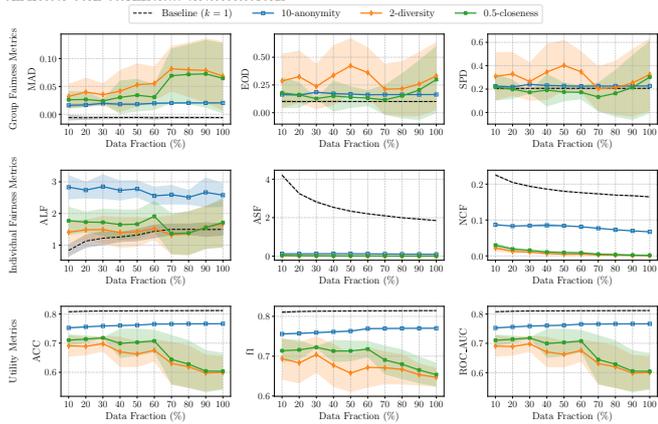
The results confirm that the trends observed with XGBoost remain consistent across different classifiers. While minor variations exist, such as XGBoost exhibiting slightly better utility and fairness stability, the overall patterns persist. Specifically, anonymization continues to negatively affect group fairness, improve individual fairness, and introduce trade-offs in utility across classifiers. These findings indicate that the insights gained from XGBoost-based experiments are broadly applicable to other models, reinforcing the generalizability of our study's conclusions.



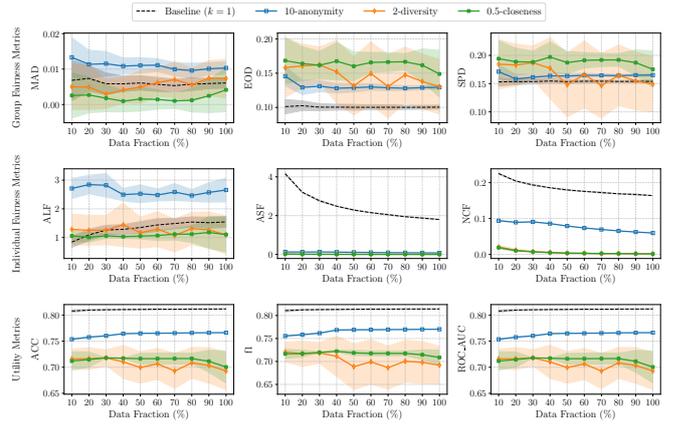
**Figure 22.** Effect of varying data fraction on the performance of anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) in terms of group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. This analysis is performed using the Compas dataset, considering gender as the protected attribute for fairness evaluation.



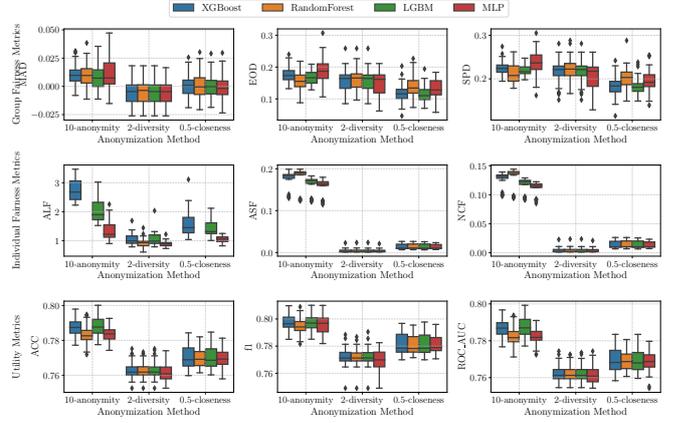
**Figure 23.** Effect of varying data fraction on the performance of anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) in terms of group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. This analysis is performed using the Compas dataset, considering race as the protected attribute for fairness evaluation.



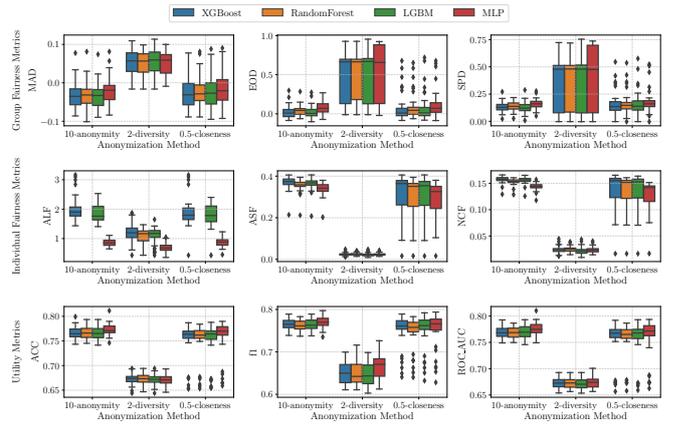
**Figure 24.** Effect of varying data fraction on the performance of anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) in terms of group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. This analysis is performed using the ACSIncome dataset, considering gender as the protected attribute for fairness evaluation.



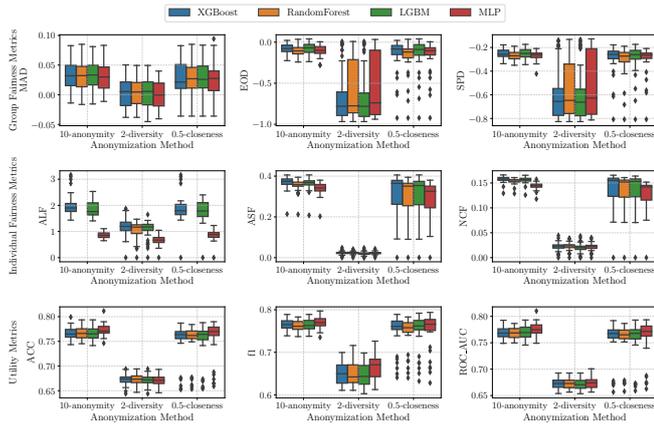
**Figure 25.** Effect of varying data fraction on the performance of anonymity techniques (10-anonymity, 2-diversity, 0.5-closeness) in terms of group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. This analysis is performed using the ACSIncome dataset, considering race as the protected attribute for fairness evaluation.



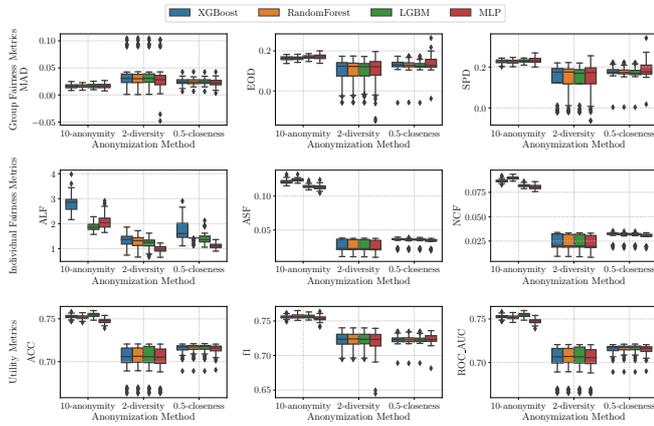
**Figure 26.** Comparison of the impact of different state-of-the-art ML classifiers on anonymized dataset ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) and relation to group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are based on the Adult dataset, with race as the protected attribute for fairness evaluation.



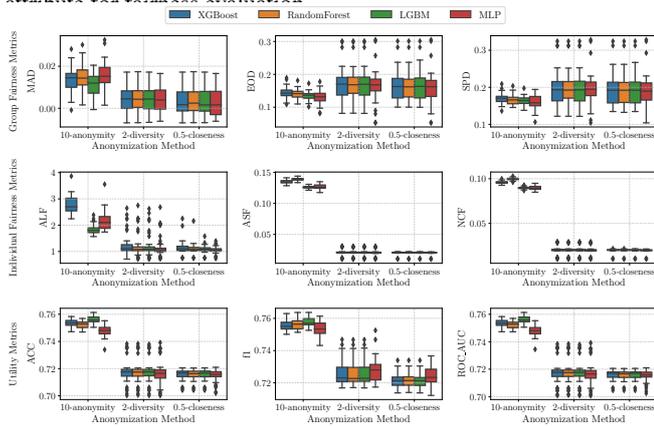
**Figure 27.** Comparison of the impact of different state-of-the-art ML classifiers on anonymized dataset ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) and relation to group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are based on the Compas dataset, with gender as the protected attribute for fairness evaluation.



**Figure 28.** Comparison of the impact of different state-of-the-art ML classifiers on anonymized dataset ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) and relation to group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are based on the Compas dataset, with `race` as the protected attribute for fairness evaluation.



**Figure 29.** Comparison of the impact of different state-of-the-art ML classifiers on anonymized dataset ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) and relation to group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are based on the ACSIncome dataset, with `gender` as the protected attribute for fairness evaluation.



**Figure 30.** Comparison of the impact of different state-of-the-art ML classifiers on anonymized dataset ( $k$ -anonymity,  $\ell$ -diversity,  $t$ -closeness) and relation to group fairness (MAD, EOD, SPD), individual fairness (ALF, ASF, NCF), and utility (Accuracy, F1-score, ROC AUC) metrics in ML. Results are based on the ACSIncome dataset, with `race` as the protected attribute for fairness evaluation.