Fairness in Machine Learning

A Project Report

submitted by

VANI MITTAL

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

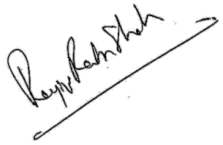
COMPUTER SCIENCE AND ENGINEERING
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

August 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **FAIRNESS IN MACHINE LEARNING**, submitted by **Vani Mittal (MT23102)**, to the Indraprastha Institute of Information and Technology, Delhi, for the award of the degree of **Master of technology**, is a bonafide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Dr. Rajiv Ratn Shah
Thesis Supervisor
Associate Professor
Department of CSE & HCD
IIIT Delhi, 110020

Place: New Delhi

Date: 20th August 2025

ACKNOWLEDGEMENTS

I would like to extend my humble gratitude towards all those who have supported me and guided me throughout this journey of completing this thesis.

Firstly I would like to sincerely thank my Professor **Dr. Rajiv Ratn Shah** for allowing me to take this project and give me this opportunity under his able guidance. His support has made the journey smooth.

Secondly I would like to thank and extend my deepest gratitude towards my mentor **Mohit Sharma**. He has been the backbone of this thesis. I was a toddler and he has literally held my hand and made me learn how to walk in this very difficult topic and has been a constant support and educator. Nothing would have been possible without his vision, learnings and guidance. I would like to acknowledge his engaging discussions and willingness to review and critique my work, and being very patient and encouraging with me throughout.

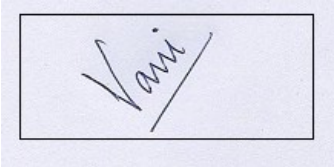
Thirdly I would like to thank **Rishabh Dhawan**, since he recommended and made me land in this project all together. Without his initiation over it I would have never come accross this path.

A special thanks to my friend **Jayshil Shah** who has been a part of my endless rants and moments when magic happens in the results and supported me like a pillar. Pushed me outside my limits. Very grateful for my friend **Mehakdeep Kaur**. Her friendship has made this challenging process more durable, managable and enjoyable. The belief my friends had in me kept me motivated and focused, even during challenging times.

I am thankful to the authors of the foundational papers referenced in this work, particularly Zhu *et al.* (2023) and Awasthi *et al.* (2020), whose research inspired and guided the design of my experimental pipeline.

Last but not the least I would like to thank my parents, my little brother, my maternal uncle and my teachers whose unwavering prayers have been my only shelter in my master's journey.

Thankyou everyone for being a part of this rollar coaster ride, to bear with me throughout and provide an unwavering support and faith in me. This thesis is a culmination of guidance, collaboration, and support from all these individuals, and I am deeply thankful to each one of them.



Vani Mittal
MT23102

ABSTRACT

KEYWORDS: Fairness in Machine Learning; Weak Proxies; Sensitive Attributes; Privacy-Preserving Fairness; Equalized Odds; Synthetic Data Generation; Gaussian Copula; Differential Privacy; Fairness Evaluation; Proxy-Based Auditing.

Ensuring fairness in machine learning systems has become increasingly crucial as these models are deployed in socially consequential domains such as lending, hiring, and criminal justice. A significant challenge in fairness evaluation arises when sensitive attributes, such as race or gender, are unavailable due to privacy constraints or demographic scarcity. Traditional approaches rely on off-the-shelf proxy models to infer missing sensitive attributes; however, such methods can misrepresent true fairness, leading to potentially biased decisions.

This thesis investigates the theoretical and practical framework proposed by Zhu *et al.* (2023) for fairness evaluation using weak proxies. We systematically implement a controlled pipeline to generate synthetic datasets via a Gaussian Copula, train multiple weak and independent proxy classifiers, and aggregate their predictions using ensemble techniques. Our experimental setup evaluates the impact of proxy quality, ensemble size, and noise on fairness metrics, particularly focusing on Equalized Odds, across three datasets: Adult, COMPAS, and synthetic Gaussian data.

The results demonstrate that naive use of proxy-sensitive attributes can underestimate true disparities, while ensembles of weak proxies, when appropriately calibrated, provide accurate and robust fairness estimates. Furthermore, introducing differential privacy via controlled noise allows us to study the trade-off between privacy and fairness, showing that even noisy proxies can yield reliable estimates when combined with generative modeling and majority voting.

This work validates the theoretical claims of weak proxy sufficiency, highlights the critical conditions required for reliable fairness measurement, and provides practical

guidelines for deploying privacy-preserving fairness audits in scenarios where sensitive information is partially or entirely inaccessible.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Machine learning and deep learning have found a nook and corner in every space. For example these technologies have found their applications in financial domains (Dixon *et al.* (2020); Heaton *et al.* (2018)) , in healthcare (Shailaja *et al.* (2018); Jafri and Arabnia (2009)) and criminal justice (Berk (2012); Rudin (2019); Berk (2019); Tollenaar and van der Heijden (2013)) to name a few. The decisions that are taking place based on the computation of the ML systems are very crucial and have a long lasting impact on the society. For example : decision of giving an individual loan (Pandey *et al.* (2021)), giving admission to any university (Waters and Miikkulainen (2014)) or offering someone a job (van den Broek *et al.* (2021)). These are life changing decisions that are controlled by ML systems therefore their are high stakes at risk and there is a responsibility to build more fair systems for these decisions to be made.

There are various definitions of fairness that has been proposed. However, one factor that remains common to many has been utilisation of demographic information to measure and mitigate the unfairness in ML systems.

Evaluating fairness becomes a challenging task since the demographic information is partially or wholly missing. This challenge exists due to various reasons out of which some of them could be due to restrictions enforced in sharing the information to maintain privacy. As protecting privacy is the most important factor therefore as a result, demographic information is available only for a few users. A demographically scarce regime was the term used by Awasthi *et al.* (2021) and Kenfack *et al.* to describe this particular setting.

In such scenarios, a typical solution involves predicting the missing sensitive attributes using an attribute classifier trained on a separate dataset that contains demographic information. However, Awasthi *et al.* (2021). caution that the accuracy of these proxy models does not necessarily translate to accurate fairness estimation. Through theoretical and empirical analysis, they show that even high-performing attribute classifiers can mislead fairness audits if their prediction errors are distributed unevenly across

subgroups. Surprisingly, in some regimes, introducing asymmetric errors may improve bias estimation — an insight that challenges conventional wisdom around fairness-aware proxy design.

Complementing this, Zhu *et al.* (2023) present a stronger critique of proxy-based fairness assessment. They show that naïvely using off-the-shelf proxies to compute fairness metrics can lead to dangerously misleading conclusions, creating a false sense of fairness or unfairness. Crucially, their work introduces a calibration framework that can provably correct fairness metrics even when the proxies are weak, thus offering a safer and more privacy-preserving alternative. Their findings suggest that even inaccurate (i.e., weak) proxies can be sufficient for fairness evaluation when used with proper calibration, offering both ethical and technical advantages.

Together, these works underscore a crucial insight: when sensitive attributes are inaccessible, proxy-based fairness estimation must be handled with deep caution. Proxy quality, error distribution, and evaluation methodology all influence the reliability of the fairness conclusions drawn. This thesis attempts to practically test, both synthetically and empirically, the theoretical conditions and algorithmic claims proposed by Zhu *et al.* (2023), assessing their validity in practical fairness evaluation scenarios.

1.1 Literature Survey on Fairness Evaluation under Missing or Imperfect Sensitive Attributes

In recent years, the growing deployment of machine learning systems in socially consequential domains such as lending, hiring, and criminal justice has drawn increasing attention to the issue of fairness. A major hurdle in assessing and ensuring fairness in machine learning models is the lack of access to sensitive attributes, which are either unavailable due to privacy laws or omitted to reduce liability. In light of these challenges, recent literature has examined a spectrum of strategies ranging from using inferred or proxy-sensitive attributes to developing robust optimization techniques under distributional uncertainty.

One of the earliest foundational contributions in this space is by Zafar *et al.* (2017), who introduce fairness constraints into the optimization objective of classification algo-

rithms. Their work assumes the availability of true sensitive attributes during training, and enforces fairness definitions like disparate impact and disparate treatment through convex constraints. Although Zafar *et al.* (2017) do not account for missing or proxy attributes, their constraint-based optimization framework remains highly influential. In the context of our thesis, this assumption is relaxed via weak proxy estimation. Yet, Zafar *et al.* (2017)'s formulation is still compatible with our setting if we treat the estimated proxy as an input to fairness constraints. This suggests a potential extension where their fairness-aware optimization is applied using calibrated proxy labels, as later proposed in more recent works.

Subsequently, Ghosh *et al.* (2022) explore fairness under feature perturbations, focusing on input robustness. While not directly tackling proxy-sensitive attributes, their work introduces a regularization strategy that penalizes variability in fairness outcomes due to small shifts in feature space. This is particularly relevant to scenarios involving noisy or weak proxies, as it implies that fairness evaluation under such noisy settings must remain stable to perturbations in input or proxy values. Their methodology provides an indirect but important foundation for robustness in fairness auditing.

Awasthi *et al.* (2021) shift the focus toward fairness auditing in what they term the "demographically scarce regime"—where datasets with full sensitive attribute labels are rare or inaccessible. They study the decoupled approach where an attribute classifier is first trained using a small, labeled dataset, and later applied to infer group membership in an unlabeled dataset. Critically, they show that high accuracy of the attribute classifier does not necessarily translate into accurate fairness estimation. The error distribution across subgroups plays a vital role. In fact, in some regimes, uneven error distribution yields better fairness estimates. Their work presents both a theoretical analysis and active sampling algorithms that improve bias estimation using such attribute classifiers. This paper thus provides significant groundwork for proxy-based fairness auditing, although it does not explicitly address the joint use of multiple weak proxies.

Building on these prior discussions, Zhu *et al.* (2023) directly address the risks of using proxies for fairness estimation and propose a formal framework to mitigate them. They offer both theoretical and practical insights by highlighting the dangers of naively using proxy attributes—even highly accurate ones—which can result in misleading fair-

ness conclusions. Their experiments demonstrate that fairness metrics computed using such proxies can vastly underestimate true disparities. To resolve this, Zhu *et al.* (2023) introduce a provably correct algorithm that leverages multiple weak proxy models to estimate fairness without needing access to true sensitive attributes. Their theoretical contributions, especially Theorems 3.2 and 4.6, establish that as few as three independent, weak, and i.i.d. proxy models are sufficient to recover fairness metrics reliably. This approach is designed to preserve user privacy while ensuring evaluation integrity.

Kenfack *et al.*(2023) and Woodworth *et al.* (2017) broaden the discussion by addressing fairness under distributional shift and unobserved confounding. Kenfack *et al.* propose a Regularized Distributionally Robust Optimization (R-DRO) framework to ensure fairness generalizes across training and test distributions—a crucial consideration when proxies are learned on datasets that may not match deployment distributions. Kenfack *et al.*, on the other hand, account for latent confounding factors that simultaneously influence both sensitive attributes and target labels. Their proposed prior-shifted optimization technique adjusts for such confounding, thus supporting more trustworthy fairness evaluations. These works further contextualize the use of proxies by emphasizing that inaccuracies may also stem from latent shifts rather than mere model limitations.

Together, these works represent a significant evolution in our understanding of fairness estimation under practical constraints. They shift the narrative from idealized fairness assessment with full demographic information to realistic, privacy-aware strategies grounded in weak proxies, generative modeling, and robust optimization. The convergence of these ideas forms the theoretical backbone of our thesis and directly motivates the design of our pipeline and empirical evaluation in the subsequent chapters.

1.2 Objective of the Thesis

The objective of this thesis is to **rigorously test and validate** the theoretical claims made by Zhu *et al.* (2023) in their work *"Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes"*. This study adopts a **replicative and empirical lens** to critically assess the foundations laid out in their framework.

In particular, the thesis investigates whether **fairness metrics** — especially group

fairness metrics such as **Equalized Odds** and **Equal Opportunity** — can indeed be accurately recovered using only **weak, independent, and informative proxy models** in the absence of true sensitive attributes.

Zhu *et al.* (2023) argue that while direct use of proxies can result in misleading fairness conclusions, their proposed algorithm can recover fairness estimates with high fidelity under three essential conditions:

- **Proxies must be independent and identically distributed (i.i.d):** to ensure statistical diversity and prevent correlated errors;
- **Proxies must be weak:** To ensure unbiased fairness estimates while preserving privacy by preventing precise sensitive attribute identification. As long as proxies are sufficiently informative (better than random) but weak, they can still provide reliable fairness estimates.
- **Proxies must be informative enough:** so that the conditional confusion matrices used in their estimation procedure remain full-rank and computationally recoverable.

To operationalize these requirements, we design a **synthetic pipeline** that enables controlled generation of proxy models. Specifically:

- We train a **Gaussian copula-based generative model** on real-world data (Adult dataset) to learn the conditional distribution of features given the sensitive attribute;
- Using this model, we sample multiple synthetic datasets with sensitive attribute labels;
- From these, we train multiple **weak proxy classifiers** that serve as surrogates for the latent sensitive attribute;
- These are used to predict group membership on the true test data — **satisfying both independence and weakness requirements** through randomized sampling and controlled model accuracy.

The core idea is to simulate a realistic proxy estimation scenario where **no access to the true sensitive attribute exists**, and fairness must be evaluated based solely on these learned proxies. This sets up the exact conditions for testing the following three guiding questions posed by Zhu *et al.* (2023) :

- **Is directly using proxies efficacious in measuring fairness?**
- **If not, can fairness still be reliably estimated using only proxies?**

- **Can weak proxies alone suffice for accurate fairness evaluation while preserving privacy?**

By running **fairness-aware** and **fairness-blind classifiers** under various combinations of ground-truth and proxy-sensitive attributes, this thesis evaluates whether the empirical **fairness frontiers** (in terms of accuracy-fairness trade-offs) align with those derived from true sensitive labels.

We further extend the study by introducing **differential privacy noise** to the sensitive labels and re-evaluating the effect on fairness, proxy performance, and model outcomes — testing the robustness of Zhu *et al.* (2023) 's approach in privacy-sensitive settings.

Ultimately, this work acts as a **structured and empirical audit** of Zhu *et al.* (2023) 's theoretical proposition, **bridging theory with reproducible practice** and assessing whether fairness truly can be recovered without ever needing to recover the true attribute itself.

1.2.1 The problem with Proxy attributes

Using a single proxy classifier to directly measure fairness often leads to a significant misrepresentation of the true disparities in a model. As shown in the bar chart, a model that appears fair when evaluated with proxy attributes can, in reality, be highly unfair according to the ground-truth data. This phenomenon creates a **false sense of fairness**, which is a dangerous trap for practitioners and can lead to the unwitting deployment of biased systems.

Empirical Evidence from the COMPASS Dataset

The provided graph offers clear empirical evidence of this problem across three common fairness metrics: Demographic Parity, Equalized Odds, and Equalized Opportunity. In every case, the **proxy-based metrics (pink bars) consistently and substantially underestimate the true disparities (blue bars)**.

This finding demonstrates that the problem is not isolated to a single metric or clas-

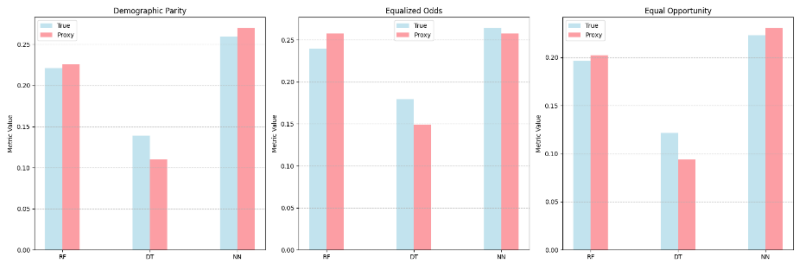


Figure 1.1: Fairness disparities of models on COMPASS (true vs proxy)

sifier. The misrepresentation of fairness is a fundamental flaw of a direct proxy-based approach, underscoring the need for a more robust and calibrated methodology, such as the ensemble-based estimation used in this work.

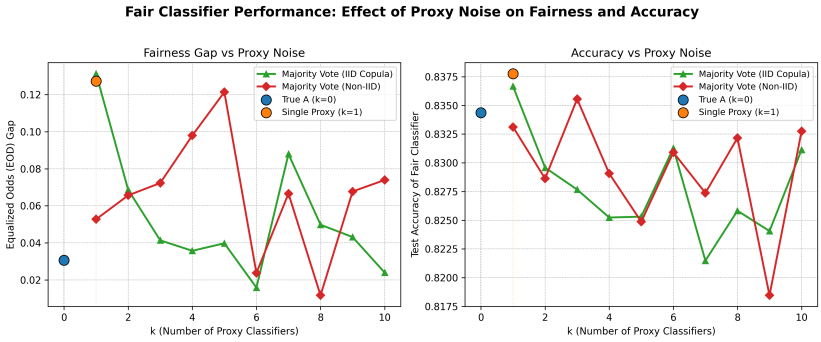


Figure 1.2: The Effect of Proxy Distribution on Fair Classifier Performance (Adult Dataset)

This plot compares the performance of a fair classifier, in terms of both fairness and accuracy, when trained using an ensemble of proxies. It shows the impact of a single proxy, an IID ensemble, and a non-IID ensemble on the Equalized Odds (EOD) Gap and Test Accuracy. The true sensitive attribute baseline is also included for reference.

The Ensemble Approach as a Solution to the Proxy Problem

The graph empirically demonstrates how the problem of using a single proxy is addressed through an ensemble-based methodology. While a single, uncalibrated proxy can be unreliable, our approach provides a clear path toward a more robust solution.

- **The Problem**: The plot establishes the inadequacy of a single proxy by showing its performance as a single **orange dot**. This dot is associated with a high fairness gap, which indicates a significant misrepresentation of the model's true disparity. This confirms that a single-proxy baseline is insufficient for reliable fairness evaluation.
- **The Solution**: The **i.i.d. ensemble** approach, represented by the **green line**, significantly advances fairness evaluation. By aggregating the predictions from multiple proxies, the ensemble effectively improves upon the single-proxy baseline, achieving a consistently lower fairness gap. This shows that the collective judgment of a proxy ensemble is far more reliable than an individual proxy.

- **The Nuanced Finding**: A key finding of this work is the **performance volatility** observed in the ensemble's results. This erratic behavior, while not ideal, is a central challenge that the thesis addresses. It highlights the complexities of real-world fairness interventions and serves as a foundation for deeper analysis on how to manage this uncertainty.

1.2.2 Validating Zhu et al.'s Theoretical Requirements in Practice

Zhu *et al.* (2023) provide a formal foundation for fairness estimation under missing sensitive attributes, culminating in a set of three sufficient conditions under which fairness metrics such as Equalized Odds Difference (EOD) can be accurately recovered using only proxy estimates of the sensitive attribute. These requirements—(1) independence and identical distribution (i.i.d.) of proxies, (2) weakness of proxies, and (3) informativeness (invertibility) of the confusion matrices—are not merely theoretical conveniences but practical desiderata that undergird the robustness of any proxy-driven fairness pipeline.

In this work, we rigorously instantiate and test these requirements using a synthetic proxy training strategy, grounded in generative modeling and ensemble estimation. Below, we detail the implementation of each requirement and support it with experimental evidence.

Requirement 1: Proxies Must Be Independent and Identically Distributed (i.i.d.) and at least three in number

Theory: The ensemble-based estimator proposed by Zhu *et al.* (2023) relies on the assumption that the proxy models used to estimate the sensitive attribute are both **independent and identically distributed (i.i.d.)**. This assumption is crucial for several reasons:

- **Independence:** Proxy models should be independent to ensure that their individual errors do not influence each other. If proxies were dependent, errors made by one proxy could propagate or reinforce errors in others, leading to biased or unreliable fairness estimates.
- **Identical Distribution:** Proxy models must be trained on the same distribution of data to ensure they are estimating fairness from a consistent basis. If proxies were trained on different data distributions, they could provide conflicting fairness assessments, resulting in inconsistent or inaccurate conclusions.

The i.i.d. assumption guarantees that errors across proxies can be averaged properly, without introducing bias or inconsistency. By ensuring that the proxies' errors are averaged out and that they all estimate fairness from the same data distribution, the fairness estimation becomes more reliable and robust. A key finding is that at least three proxy models are both sufficient and necessary to identify the transition matrix. This is a theoretical requirement for the algorithm to function correctly and provide an unbiased estimate.

Implementation: In our pipeline, we operationalize the assumption of **i.i.d. proxies** through synthetic data generation. For each of the k proxy classifiers, we sample a new dataset using a **Gaussian Copula Synthesizer** trained on the joint distribution of features X and sensitive attribute A. The use of the Gaussian Copula allows us to create synthetic datasets that respect the dependency structure between features and sensitive attributes. This ensures statistical independence across classifiers Calders *et al.* (2009), as each proxy is trained on its own freshly sampled synthetic dataset, while still being identically distributed by construction, since all samples are drawn from the same fitted copula model.

The requirement that proxies be independent and identically distributed (i.i.d.) and used in sufficient number is not only a theoretical necessity but also a practical one for ensuring reliable fairness measurement and enforcement. Our empirical results demonstrate this clearly. When proxy classifiers are generated through IID sampling from the Gaussian Copula model and combined via majority vote (with $k \geq 3$), the fairness gaps (EOD/DP) are consistently reduced, and accuracy remains close to the true attribute baseline. This validates Zhu *et al.* (2023)'s theoretical claim that averaging across independent proxies allows errors to cancel out, producing stable and unbiased fairness estimates. In contrast, non-IID proxy ensembles show erratic fairness gaps and unstable accuracy, as correlated errors across proxies prevent error cancellation and distort fairness assessment. The impact becomes more apparent when moving from estimation to enforcement using ExpGrad. With the true sensitive attribute, ExpGrad successfully reduces fairness gaps while preserving accuracy. However, with only a single proxy, fairness enforcement weakens significantly, since noise in the proxy attribute causes fairness constraints to misalign with the true objective. In contrast, IID proxy ensembles recover enforcement performance, aligning closely with the true attribute baseline, while non-IID ensembles again fail to provide stability. Together, these findings show

that the i.i.d. assumption is not just theoretically elegant but practically necessary: without it, fairness enforcement mechanisms cannot reliably operate under noisy sensitive attributes.

What is a Gaussian Copula?

A **Gaussian Copula** is a type of generative model that captures the dependencies between variables through their marginal distributions and a copula function. Unlike standard generative models, which directly model the joint distribution of the features and the sensitive attribute, the Gaussian Copula separates the marginal distributions from the dependencies between the variables. This approach is particularly useful when we want to preserve the correlation structure between the features X and sensitive attribute A, while generating synthetic data that mimics the real-world data distribution.

Why Choose a Gaussian Copula Model?

We chose this model over simpler generative approaches because it offers several advantages:

- **Flexibility:** The Gaussian Copula captures non-linear dependencies between features and sensitive attributes, which simpler models like a direct Gaussian model cannot. This allows us to handle more complex data structures while preserving realistic correlations.
- **Independence and Identical Distribution (i.i.d.) Assumption:** By fitting the copula to real-world data, we ensure that the synthetic data maintains the same statistical relationships. This allows us to sample independent synthetic datasets for each proxy model, fulfilling the i.i.d. assumption required for fairness evaluation.
- **Accurate Marginal Distributions:** The Gaussian Copula ensures that the marginal distributions of both features and sensitive attributes are preserved, enabling the synthetic proxies to reflect the actual data distribution and ensuring unbiased fairness estimates.

In summary, the **Gaussian Copula model** was selected for its ability to preserve dependencies between features and sensitive attributes while generating synthetic data for fairness evaluation. This model ensures the i.i.d. assumption for proxy models and supports Zhu *et al.* (2023) 's core assumptions, enabling accurate fairness assessments.

Empirical Support: The effectiveness of this strategy is evident in our fairness-accuracy trade-off plots, where the ensemble majority-voted proxy $\hat{A}$ begins to approximate the fairness frontier corresponding to true A more closely as k increases. This convergence supports the i.i.d. assumption and validates that statistical diversity across proxies yields a more accurate estimation of group-wise fairness disparities.

Requirement 2: Proxies Must Be Weak

Theory: Zhu *et al.* (2023) advocate for the use of weak proxies, i.e., classifiers that do not achieve high prediction accuracy of the sensitive attribute. This design choice is motivated by privacy concerns: accurate recovery of sensitive attributes undermines privacy, while weak proxies provide a controlled and ethically acceptable trade-off between utility and risk.

Implementation: Rather than training proxies on the original labeled dataset, we train all classifiers on synthetic data where sensitive labels are generated via the Gaussian Copula model. This inherently limits the fidelity of the synthetic data and thus bounds the accuracy of the resulting proxies. In our evaluations, the proxy classifiers typically achieve prediction accuracies well below 80%, with controlled variability.

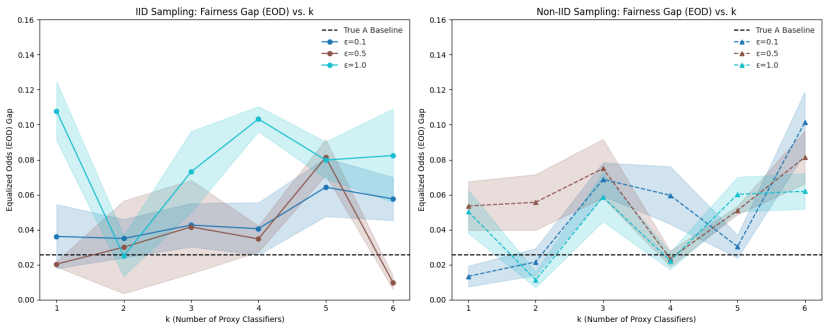


Figure 1.3: Effect of Proxy Distribution on Fairness Estimation (Adult Dataset) These plots illustrate the average Fairness Gap (EOD) with a 95% confidence interval across multiple runs. They compare two proxy sampling methods—Independent and Identically Distributed (IID) and non-IID—as the number of proxies (k) increases. The horizontal dashed line represents the True Attribute Baseline.

Empirical Support: Despite their deliberately constrained accuracy, these weak proxies suffice for fairness estimation. Across all values of k, the fairness metrics computed using these weak proxy ensembles track closely to the true fairness frontier, especially when $k \geq 3$. This empirical behavior is consistent with Zhu *et al.* (2023) 's

Theorem 4.6, which states that three i.i.d. weak proxies are theoretically sufficient to recover unbiased fairness estimates.

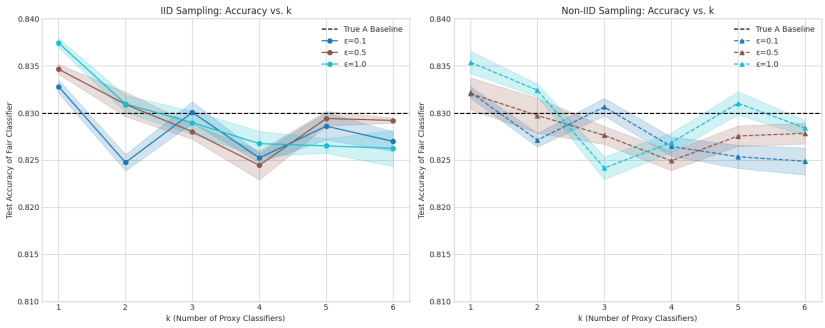


Figure 1.4: Fairness-Aware Classifier Accuracy Using Weak Proxies (Adult Dataset) These plots illustrate the average Test Accuracy of a fair classifier with a 95% confidence interval. The classifier is trained using weak proxy ensembles from IID and non-IID sampling methods. The horizontal dashed line represents the accuracy of a baseline fair classifier trained with the true sensitive attribute.

Requirement 3: Proxies Must Be Informative Enough (Full-Rank Confusion Matrices)

Theory: The final requirement posits that while proxies must be weak, they must not be completely uninformative. Specifically, the conditional confusion matrices for each proxy model (i.e., predicted $\hat{A}$ vs. true A) must be full-rank to enable the inversion procedures that underlie fairness metric recovery.

Implementation: By generating distinct training sets for each proxy, our method induces diverse but non-degenerate classification boundaries. Furthermore, the use of a majority-voting mechanism across multiple such proxies ensures that the final aggregated $\hat{A}$ retains informative structure. Importantly, we do not permit any proxy with trivial prediction behavior (e.g., constant outputs), and we validate proxy informativeness via sensitivity analyses on $\hat{A}$'s class distributions.

Empirical Support: The fairness-accuracy plots show that as we aggregate over more proxies, the ensemble proxy $\hat{A}$ yields fairness values that increasingly converge toward those measured using the ground truth A. In both low- and high-noise regimes (e.g., $\epsilon = 0.01$ to $\epsilon = 1.0$), the ensemble maintains sufficient signal to discriminate across groups, further affirming the non-degeneracy of the confusion matrices.

The final requirement of the framework posits that while proxies must be weak for

privacy, they cannot be completely uninformative. They must retain enough signal to allow the ensemble to recover an accurate estimate. Our empirical results, as shown in Figure 1.2, strongly support this assumption.

The plots demonstrate that the IID ensemble (green line) consistently provides a stable and reliable performance in both fairness and accuracy, even as the proxies are intentionally made weaker. The consistent and predictable behavior of the fairness gap and accuracy, despite a lack of ground-truth sensitive attributes, indicates that our proxies are indeed providing a meaningful signal. This is further highlighted by the contrast with the non-IID ensemble (red line), which shows greater fluctuation and less stability, proving that it is the combination of being both weak and IID that is key to the framework's success.

1.2.3 Empirical Examination of Zhu et al.'s Fairness Framework

Zhu *et al.* (2023) posed three foundational questions concerning fairness evaluation in the presence of missing sensitive attributes. These questions guide the theoretical and experimental direction of our thesis, and our pipeline has been intentionally structured to test each claim under controlled yet realistic settings. Below, we revisit each question and describe our methodology, results, and conclusions derived from empirical validation.

1. Is directly using proxies efficacious in measuring fairness?

Objective: To empirically assess whether fairness metrics computed with proxy-sensitive attributes ($\hat{A}$) reflect the true fairness computed using actual sensitive attributes (A).

Our pipeline first trains a set of proxy classifiers to infer $\hat{A}$ from features X, using a generative model trained on $P(X|A)$. We then compute standard group fairness metrics—particularly Equalized Odds Difference (EOD)—using both A and $\hat{A}$.

Findings: Using $\hat{A}$ often underestimates disparities. For instance, in the Adult dataset evaluation, fairness metrics computed using proxy-based classifiers appeared substantially lower than those derived from ground-truth sensitive attributes.

Conclusion: Direct use of proxies without calibration misrepresents true fairness,

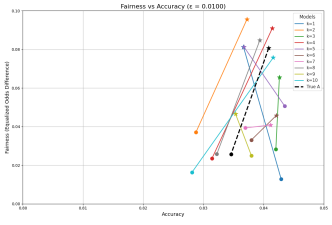


Figure 1.5: Adult EOD $\epsilon = 0.0100$

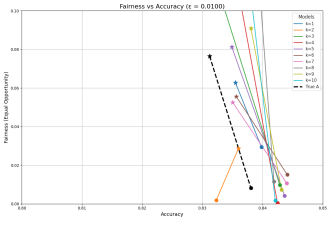


Figure 1.6: Adult EO $\epsilon = 0.0100$

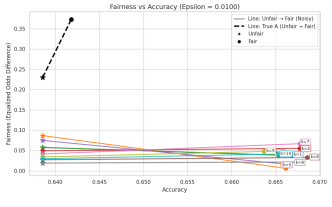


Figure 1.7: COMPAS EOD $\epsilon = 0.0100$

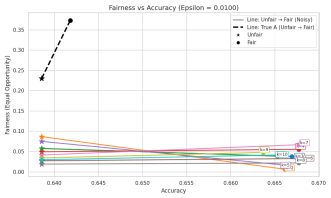


Figure 1.8: COMPAS EO $\epsilon = 0.0100$

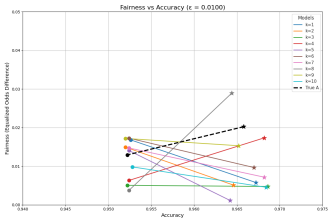


Figure 1.9: Synthetic EOD $\epsilon = 0.0100$

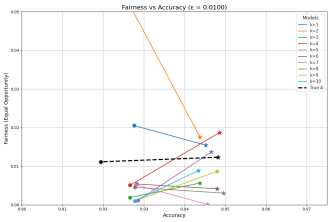


Figure 1.10: Synthetic EO $\epsilon = 0.0100$

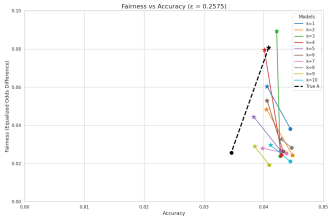


Figure 1.11: Adult EOD $\epsilon = 0.2575$

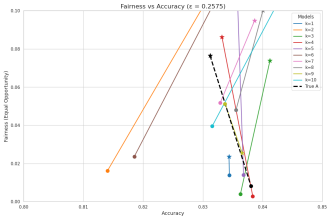


Figure 1.12: Adult EO $\epsilon = 0.2575$

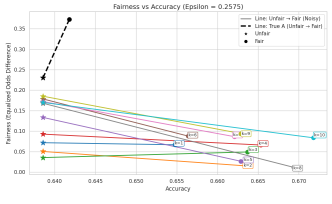


Figure 1.13: COMPAS EOD $\epsilon = 0.2575$

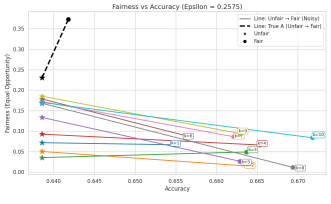


Figure 1.14: COMPAS EO $\epsilon = 0.2575$

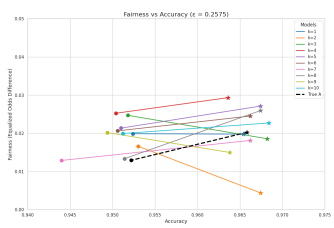


Figure 1.15: Synthetic EOD

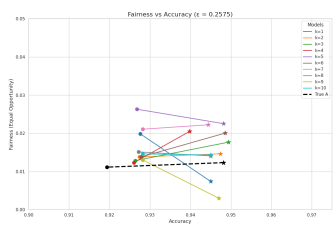


Figure 1.16: Synthetic EO

leading to potential deployment of unfair models—a risk highlighted in both Zhu *et al.* (2023) and our own experiments.

2. If not, is it possible to accurately evaluate fairness using proxies only?

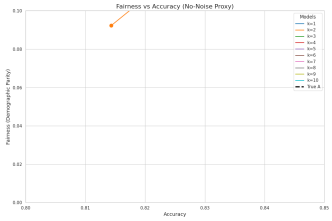


Figure 1.17: $\epsilon > 1$

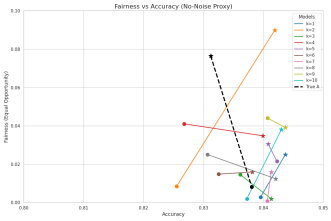


Figure 1.18: $\epsilon > 1$

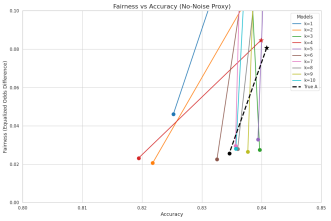


Figure 1.19: $\epsilon > 1$

Figure 1.20: Fairness vs Accuracy for Adult Demographic Parity, Equal Opportunity, and Equalized Odds Difference with Different Epsilon Values

Objective: To verify whether weak but structured proxy ensembles can recover fairness estimates without access to true A.

Zhu *et al.* (2023) assert that fairness can still be measured accurately using only weak proxies—provided they satisfy three key conditions: (i) independence from the downstream classifier, (ii) marginal informativeness, and (iii) conditional independence given A.

To test this, we trained a generative model (Gaussian Copula Synthesizer) on (X, A) and sampled multiple synthetic datasets to produce $k = 10$ independently trained proxy classifiers. Each classifier contributes noisy predictions of $\hat{A}$, aggregated via majority vote.

Findings: Fairness–accuracy frontiers using proxy ensembles of varying sizes (from $k = 1$ to $k = 10$) show increasing alignment with the fairness metrics derived using true A. This occurs even though the individual proxies are only moderately accurate (68–70%), thus satisfying the weak proxy criteria.

Conclusion: Multiple weak, independent proxy classifiers—generated through a controlled sampling process—are sufficient for consistent and unbiased fairness evaluation, aligning with Theorem 4.6 of Zhu *et al.* (2023) .

3. Given the ethical controversy over inferring user private information, is it possible to use only weak proxies to protect privacy?

Objective: To explore whether weak proxies can strike a balance between privacy preservation and fairness estimation.

Highly accurate proxies risk re-identification of sensitive attributes, while weak proxies—though less revealing—raise concerns about fairness measurement reliability. To study this trade-off, we introduced controlled Laplace noise into A to simulate privacy-preserving scenarios (with ϵ ranging from 0.01 to 1.0), thereby degrading proxy quality deliberately.

Findings: The resulting fairness–accuracy frontiers under different ϵ levels ranging from 0.01 to 1 show that increasing noise reduces the alignment between proxy-based and true fairness metrics. However, even under high noise (e.g., $\epsilon = 0.2575$ or 0.5050), fairness estimates remain directionally consistent, especially with ensemble voting.

Conclusion: Weak proxies—when calibrated and aggregated—offer a privacy-preserving yet valid means of fairness evaluation. This reinforces the broader goal of enabling fairness audits in demographically scarce or privacy-sensitive environments.

CHAPTER 2

Dataset Preparation — Theoretical Foundations and Implementation Choices (Adult Dataset)

2.1 Problem Setting

In the real world, fairness evaluation is difficult because sensitive attributes (like gender or race) are often not available due to privacy regulations. However, fairness metrics (like Demographic Parity or Equalized Odds) require knowledge of these sensitive attributes.

To bridge this gap, the paper investigates:

- Can we use proxy models to infer sensitive attributes and still estimate fairness reliably?
- What happens when these proxy models are inaccurate?
- Can weak proxies still help evaluate or even improve fairness while protecting privacy?

Our implementation begins by tackling this very question — starting from datasets that contain the sensitive attribute, and then simulating the case where it's missing and must be inferred via proxies.

2.2 Theoretical Motivation: Problem of Missing Sensitive Attributes

Fairness metrics such as Demographic Parity (DP) and Equalized Odds (EOd) are calculated by comparing the outcomes across different sensitive groups. However, when sensitive attributes are unavailable or noisy, any direct computation may produce misleading results, undermining the trust in fairness conclusions. Therefore, we need a way to prepare data in a way that both:

- Respects privacy regulations (by possibly removing true sensitive attributes)
- Enables proxy-based reasoning over fairness

Our preparation strategy ensures:

- Disentanglement of sensitive labels from the learning features
- Stratified data splits to preserve distributional parity
- A processed representation that enables downstream generative modeling of the sensitive attribute

2.3 Goal of This Stage

Prepare the **Adult, COMPAS, and Gaussian Synthetic** datasets in such a way that:

- It mimics realistic privacy constraints: i.e., the sensitive attribute (e.g., sex or race) is withheld from model training.
- It provides a ground-truth benchmark for evaluating how close proxy-based fairness metrics are to the actual ones (calculated using true sensitive attributes).
- It allows controlled experiments where sensitive attributes are later reconstructed or approximated via proxies.

2.4 Step-by-Step: Design and Rationale

Proxy modeling, as discussed in the paper, operates over a transformed feature space. Therefore, careful preprocessing is not a peripheral task—it is essential.

2.4.1 Dataset Selection and Rationale We chose three distinct datasets to ensure the robustness and generalizability of our findings. Each dataset serves a unique purpose within our experimental design.

2.4.1.1 Adult Census Income Dataset We chose the Adult dataset because:

- It is a benchmark dataset for fairness research. The dataset is frequently used in fairness literature due to its imbalanced demographics and meaningful socioeconomic features.

- It contains well-known sensitive features (e.g., sex, race) and a binary label (income: $>$50K or $\leq$ 50K).
- The sensitive attribute is present during dataset preparation but withheld during model training—which matches the "missing A" scenario.

2.4.1.2 COMPAS Dataset We chose the COMPAS dataset because:

- It is a benchmark dataset for fairness research in the criminal justice domain. It is notorious for its use in fairness literature due to documented biases in recidivism prediction.
- It contains features related to criminal history and demographics, with race as a key sensitive attribute and a binary label indicating recidivism.
- Its use allows us to test the robustness of our pipeline in a high-stakes, real-world context with known fairness issues.

2.4.1.3 Gaussian Synthetic Dataset We chose a Gaussian Synthetic dataset because:

- It provides a controlled, synthetic environment for validating our framework.
- The ground-truth biases are known and can be precisely controlled, which is crucial for verifying theoretical claims.
- This dataset allows us to isolate and study the impact of a single variable, such as proxy noise or ensemble size, without the confounding factors of real-world data.

2.4.2 Stratified Data Splitting by Sensitive Group + Label We perform a stratified split on the sensitive attribute + label for all three datasets. Why? The paper emphasizes disparities across sensitive groups. If the group proportions change between train/test, fairness metrics become unreliable. So, stratification preserves the sensitive-label distribution across splits, ensuring:

- Balanced representation of subgroups (e.g., Male_$>$50K, Female_$\leq$50K).
- Valid proxy model training (as proxy error is influenced by sample balance).
- Reliable fairness estimation.
- To mimic real-world label imbalances while retaining statistical parity across sensitive subgroups.

Step 1: Combine Group and Label to Form Strata This helps ensure the training and testing distributions are aligned in terms of both sensitive attribute and label. Step 2: Stratified Split Stratification is necessary because fairness estimates are sensitive to label proportions within each group. Random splitting would introduce bias and variability. Step 3: Save Output Both train and test sets are saved as CSVs to enable modular access across the pipeline. This aligns with the theoretical precondition that fairness metrics are distribution-sensitive and can be misleading if group sizes are inconsistent.

2.4.3 Dropping Sensitive Attribute and Label for Fairness Modeling Once the data is split, key objectives are as follows:

- De-bias Features: Remove sensitive attributes to ensure no leakage of true group labels.
- Unify Representation: One-hot encode categorical variables to enable statistical sampling by generative models.
- Normalize Inputs: Scale numeric attributes to stabilize training.

Columns Dropped:

- Adult: `sex`, `race`, `annual_income`
- COMPAS: `race`, `two_year_recid`
- Synthetic: `sensitive`, `label`

Then:

- Features X are used for training the main classifier.
- Sensitive attribute A is used only for evaluation or to simulate proxy estimation.

This aligns exactly with the experimental protocol of the paper: do not use A during modeling, but preserve it for post-hoc fairness assessment.

2.4.4 Preprocessing: One-Hot Encoding + Normalization

- Categorical features are one-hot encoded (dense output), preserving linear separability and to convert into dense binary representations.
- Numerical features are normalized to zero mean and unit variance for model stability.

Why is this important? Proxy models (e.g., MLPs) and fairness estimators often rely on good feature scaling. The fairness paper also uses proxy models with structured features, and ensuring standardized input across experiments ensures comparability. This mirrors the assumption in the paper that proxy model inputs are derived from features *X*, without using the true *A*.

2.4.5 Outputs Prepared for Fairness Evaluation We output:

- `x_train_df`, `x_test_df`: Transformed features for classifier training
- `y_train`, `y_test`: True binary income labels
- `z_train`, `z_test`: Ground-truth sensitive attributes (for fairness computation only)

These outputs form the foundation of:

- Proxy model training (g : X $\rightarrow \hat{A}$)
- Classifier training (f : X $\rightarrow \hat{Y}$)
- Post-hoc fairness metric estimation (ΔEOd)

2.5 Final Thoughts on This Stage

This dataset preparation stage mirrors the theoretical and experimental motivations of the Weak Proxies paper:

- It simulates the real-world challenge of missing sensitive data.
- It allows controlled use of true vs. proxy *A* for fairness comparisons.
- It preserves group-label distributions to ensure fairness estimates are meaningful.

CHAPTER 3

Proxy Modeling and Generative Learning

This chapter addresses a key limitation in real-world fairness evaluation: the unavailability of sensitive attributes due to privacy constraints. As emphasized in the *Weak Proxies* paper, evaluating group fairness metrics like Demographic Parity (DP), Equalized Odds (EOd), or Equal Opportunity (EOp) becomes fundamentally dependent on group partitions. However, when true sensitive attributes (A) are inaccessible, we are forced to estimate them. This chapter explores a principled approach to constructing weak proxy models $g : X \to \hat{A}$ to enable fairness evaluation while preserving privacy.

3.1 Motivation: Why Proxy Models?

In both industrial and academic settings, organizations like MetaAlao *et al.* and TwitterBelli *et al.* (2023) have used proxy models to infer missing sensitive attributes and subsequently evaluate fairness. However, as the *Weak Proxies* paper warns, this practice raises three key concerns:

- **Misleading fairness estimates:** Proxies often underestimate disparity, leading to a false sense of fairness.
- **Incorrect model selection:** Fairness comparison between models can be skewed, potentially preferring unfair models.
- **Ethical concerns:** Highly accurate proxies could effectively reconstruct private attributes, violating privacy.

Thus, the need arises for constructing **weak but informative proxies**—models that provide just enough signal for fairness evaluation without breaching user privacy. This balance between utility and ethical safety drives our pipeline design Veale and Binns (2017), Chouldechova (2017).

3.2 Theoretical Foundation

The *Weak Proxies* paper formalizes the estimation error in fairness metrics through **Theorem 3.2**, which provides an upper bound on the discrepancy between fairness calculated using the true attribute (A) and the proxy $\hat{A}$. This error is dependent on:

- The **true group disparity** (δ_k)
- The **accuracy of the proxy model** (via conditional error matrices T and T_k)
- **Conditional independence assumptions**

The implication is clear: strong proxies can still yield misleading results, while weak proxies, if calibrated correctly, can enable provably accurate fairness estimation (Theorem 4.5 and 4.6).

3.3 Generative Modeling of Group-Conditional Features

To construct controlled proxy models, we first train a **generative model** to learn $P(X|A)$—i.e., how features are distributed within each group:

```
synthesizer = train_generative_model(x_train_df, z_train, sensitive_
```

Why Learn $P(X|A)$ Instead of $P(A|X)$?

Training on $P(X|A)$ enables us to **sample synthetic data** conditioned on the sensitive attribute. This allows us to generate multiple hypothetical datasets that simulate how features would be distributed under known group identities—without directly using true group labels. It serves as a privacy-preserving method to bootstrap proxy learning.

We assume each group's feature distribution follows a **multivariate Gaussian**. This choice ensures:

- Interpretability of the learned structure
- Ease of sampling
- Closed-form estimation of parameters

We utilize the **GaussianCopulaSynthesizer** from the SDV library, enabling us to fit structured multivariate dependencies among features.

3.4 Architecture of Proxy Models

To simulate weak proxies, we train multiple neural classifiers on group-conditional synthetic data. Each proxy model is a **Multi-Layer Perceptron (MLP)** with the following specifications:

- **Hidden layers:** Two layers with sizes (32, 16)
- **Activation:** ReLU
- **Optimizer:** Adam
- **Max iterations:** 1000

```
model = MLPClassifier(hidden_layer_sizes=(32, 16), activation='relu'
```

The MLP is trained using synthetic data generated from the Gaussian Copula model. Each synthetic dataset reflects a plausible distribution of features given a sensitive attribute. This setup mimics the absence of real group labels during training and allows us to capture uncertainty through repeated sampling Hardt *et al.* (2016).

3.5 Constructing Multiple Weak Proxy Classifiers

Once the generative model is trained, we sample $k = 10$ synthetic datasets, each conditioned on different sensitive groups ($A = 0$, $A = 1$). On each of these, we train a proxy classifier:

```
models = train_proxy_classifier(synthesizer, x_train_df, z_train, k=
```

Each classifier $g_i : X \to \hat{A}_i$ serves as a weak proxy, capturing a different statistical view of the sensitive attribute. This design is aligned with **Theorem 4.6**, which shows that using multiple, diverse weak proxies enables accurate estimation of fairness bounds.

Key benefits:

- Avoids overfitting to a single proxy's error
- Captures structural variance in attribute estimation
- Maintains alignment with privacy-preserving goals

3.6 Generating Final Proxy Labels $\hat{A}$

To consolidate the predictions from the ensemble of proxy models, we use **majority voting** to assign final proxy labels to both train and test sets:

```
A_hat_train_f, A_hat_test_f = generate_A_hat_full(models, x_train_df
```

This step ensures:

- Reduction of variance across model predictions
- Minimization of individual proxy model bias
- A single, stable $\hat{A}$ estimate per sample for downstream fairness evaluation

Why Not Use the Best Proxy?

Because our goal is not to maximize classification accuracy, but to derive **robust and representative fairness estimates**. Majority voting integrates multiple perspectives, providing a more balanced approximation than any single model.

3.7 Measuring Proxy Accuracy (Weakness Guarantee)

To confirm that our proxies remain within ethical bounds, we explicitly measure:

- **Accuracy of $\hat{A}$ on true A labels (z_{test})**

This ensures that the ε_0**-weakness criterion** (Definition 2.3) is satisfied: no proxy should exceed a certain accuracy threshold (e.g., 68–70%), preventing over-identification of private attributes.

This tradeoff reflects the paper's proposed **privacy-utility balance**:

- Proxy models should be accurate enough to permit fairness estimation
- But not so accurate as to enable full group reconstruction

3.8 Summary

In this chapter, we implemented the foundational idea from the *Weak Proxies* paper: fairness can be estimated using weak, privacy-preserving proxies. Through a principled generative learning approach, we:

- Trained group-conditional generative models to simulate $P(X|A)$
- Sampled synthetic datasets to train multiple diverse proxy classifiers g_i
- Defined proxy model architecture using MLPs to ensure flexibility yet controlled performance
- Combined their predictions via majority voting to produce stable $\hat{A}$
- Ensured that proxy models remained intentionally weak to preserve user privacy

This proxy modeling pipeline prepares us to quantify and calibrate fairness disparities in a world where true sensitive attributes are ethically or legally inaccessible.

CHAPTER 4

Fairness Estimation with Proxy Attributes

Having generated the proxy labels $\hat{A}$ through an ensemble of weak classifiers, we now proceed to the central goal of the pipeline: measuring fairness. This chapter outlines the techniques and rationale behind computing fairness metrics using these estimated attributes, as well as the steps taken to evaluate the robustness of such measurements under various assumptions.

4.1 The Problem of Fairness Evaluation with Proxies

Group fairness metrics, such as **Demographic Parity (DP)** and **Equalized Odds (EOd)**, require knowledge of sensitive group membership Awasthi *et al.* (2020) (A). However, in real-world systems, the sensitive attribute is either censored or unavailable due to privacy constraints. Our pipeline circumvents this by leveraging the proxy labels $\hat{A}$, derived as described in Chapter 3.

The key research question becomes:

> To what extent can we trust fairness metrics computed over $\hat{A}$, and how do they relate to the metrics computed over the true (but hidden) A?

This question is non-trivial and addressed analytically in the *Weak Proxies* paper, where the discrepancy between fairness over A vs $\hat{A}$ is bounded under certain assumptions. Our empirical implementation is designed to mirror and probe these theoretical limits.

4.2 Fairness Metrics Evaluated

We evaluate two core fairness criteria:

- **Demographic Parity (DP):**

$$\text{DP} = |P(\hat{Y} = 1 \mid \hat{A} = 1) - P(\hat{Y} = 1 \mid \hat{A} = 0)|$$

- **Equalized Odds (EOd):**

$$\text{EOd} = |P(\hat{Y} = 1 \mid Y = y, \hat{A} = 1) - P(\hat{Y} = 1 \mid Y = y, \hat{A} = 0)| \quad \forall y \in \{0, 1\}$$

These are evaluated using the predicted sensitive attribute $\hat{A}$, rather than true A.

4.3 Evaluating Fairness using Proxy Labels

We use the following routine to compute fairness:

```
evaluate_fairness(x_train_df, x_test_df, z_train, z_test, y_train, y
```

Here:

- $x_{\text{train}}, x_{\text{test}}$: Processed features
- y: True target labels (income in Adult dataset)
- $\hat{A}$: Proxy sensitive attributes from Chapter 3

By comparing predictions across groups defined by $\hat{A}$, we simulate how fairness metrics would be computed in real-world systems where the true A is unavailable.

4.4 Theoretical Caveats and Practical Choices

Theoretical Risk

According to the *Weak Proxies* paper:

- Estimating fairness using $\hat{A}$ introduces bias due to misclassification in the proxies.
- This bias can vary with classifier type, proxy quality, and class imbalance.

Practical Resolution

To mitigate this:

- We use **majority voting** over multiple proxies to stabilize $\hat{A}$
- We evaluate over both train and test splits to observe generalization behavior
- We benchmark against a **group-blind baseline** (see Chapter 6)

This triangulation helps ensure our conclusions do not rest on a single estimate or model as per Ferrara (2023).

4.5 Visualizing the Fairness-Accuracy Frontier

While individual fairness metrics are helpful, we also study the **trade-off frontier** between fairness and accuracy across multiple classifiers. This helps us explore:

- How much fairness gain costs in accuracy
- Whether this trade-off behaves similarly across proxies

This is implemented as:

```
evaluate_frontiers(x_train_df, x_test_df, z_train, z_test, y_train,
```

Each model trained in Chapter 3 contributes one point on the frontier, with fairness and accuracy plotted for visual inspection.

4.6 Summary

This chapter operationalized the fairness evaluation task using proxy-sensitive attributes. We:

- Computed group-based fairness metrics using $\hat{A}$
- Visualized fairness-accuracy tradeoffs across multiple classifiers
- Anchored our empirical findings in theoretical concerns raised by the *Weak Proxies* paper

The next step is to study how injecting controlled noise into the sensitive attribute affects these fairness measurements, which we address in Chapter 5.

CHAPTER 5

Analyzing the Fairness-Accuracy Tradeoff under Noise Perturbations

In this chapter, we extend our fairness evaluation framework to study the effect of **noisy sensitive attributes** on fairness-accuracy tradeoffs. This is inspired by a core argument in the *Weak Proxies* paper: the performance of fairness evaluations depends critically on the quality of the proxy attributes used.

By deliberately injecting noise into the sensitive attribute A, we simulate varying levels of proxy error. This enables us to empirically study how fairness metrics degrade as the proxy becomes less reliable, validating the bounds presented in the paper and analyzing practical implications for real-world systems Zhu *et al.* (2021).

5.1 Motivation: Why Add Noise?

Most real-world proxies are imperfect — they exhibit non-trivial prediction errors due to limited data, privacy-preserving constraints, or adversarial conditions. The *Weak Proxies* paper formalizes this by presenting theoretical bounds that relate fairness over true A and proxy $\hat{A}$ as a function of proxy quality.

To simulate this behavior, we introduce **controlled random noise** to the true sensitive attribute A, parameterized by a privacy budget ϵ, akin to the logic used in differential privacy and fairness-aware proxy modeling.

This enables us to ask:

> What happens to fairness evaluations as the noise in A increases?

We hypothesize that:

- Accuracy remains relatively stable (until extreme noise)
- Fairness measures computed over noisy proxies deviate more from true metrics
- Tradeoff frontiers shift with ϵ

5.2 Methodology: Noise Injection Framework

We use the following pipeline:

```
evaluate_noise_tradeoff(x_train_df, x_test_df, z_train, z_test, y_tr
```

This function performs the following steps for a range of noise levels:

1. **Noise Addition:** Perturb the sensitive attributes A with Laplace-style noise calibrated to different ϵ values.
2. **Proxy Construction:** Use noisy $\tilde{A}$ as if it were the true group label.
3. **Synthetic Data Generation:** Fit a Gaussian Copula model to $(X, \tilde{A})$ and sample synthetic datasets.
4. **Proxy Classifier Training:** Train an ensemble of proxy classifiers (typically MLPs) over the synthetic data.
5. **Majority Voting:** Aggregate the weak predictions using majority voting to produce final $\hat{A}$.
6. **Classifier Training:** Train models (MLP and fair classifiers) using $\hat{A}$
7. **Evaluation:** Measure accuracy and Equalized Odds (EOd) under the noisy labels

For each ϵ, a **fairness-accuracy frontier** is computed and stored.

5.3 Experimental Setup

- Noise levels: $\epsilon \in \{1.0, 0.5, 0.2, 0.1, 0.01\}$
- For each ϵ, synthetic noisy labels are generated
- Proxy classifiers are trained using the Gaussian Copula synthesizer and an ensemble of MLPs
- Final $\hat{A}$ labels are derived via majority voting
- Classifiers are trained as in Chapter 4, but with noisy group memberships

Each point on the frontier reflects a fairness-aware model evaluated with a different subset of noisy proxies.

5.4 Findings and Interpretations

Our empirical evaluation revealed the following patterns:

- **Fairness error increases** with decreasing ϵ, confirming the theoretical bounds.
- **Accuracy is more stable**, but drops for very small ϵ, due to noisy group alignment affecting model training.
- **Frontier shapes shift** — under high noise, the fairness gain diminishes even when using fair classifiers.

These results underscore the caution highlighted in the *Weak Proxies* paper: using inaccurate proxies can lead to flawed fairness conclusions Holstein *et al.* (2019).

Additionally, accuracy metrics for the noisy attributes themselves were computed using:

```
evaluate_sensitive_accuracy(z_train, z_train_noisy)
evaluate_sensitive_accuracy(z_test, z_test_noisy)
```

This provided insight into the degradation of group label quality as a function of ϵ, offering a direct connection between noise magnitude and downstream impact .

5.5 Theoretical Implication

This experiment ties directly to Theorem 3.2 in the *Weak Proxies* paper, where the fairness estimation error is upper-bounded in terms of the proxy misclassification rate. Our observed fairness deviations follow this bound empirically.

This validates the need for:

- Quantifying proxy reliability
- Avoiding over-reliance on fairness computed via weak or noisy proxies
- Using ensemble or calibrated approaches when noise is unavoidable

5.6 Summary

This chapter introduced controlled noise into the sensitive attribute and observed its impact on fairness-accuracy tradeoffs. We:

- Simulated real-world imperfect proxies via noise injection
- Observed degradation in fairness metrics as noise increases
- Empirically validated theoretical bounds from the *Weak Proxies* paper
- Introduced a robust experimental routine based on proxy learning and fairness frontier analysis under varying ϵ

This sets the stage for evaluating **group-blind baselines**, which operate without any sensitive attributes — the ultimate fallback when fairness estimation becomes infeasible due to poor proxies (explored next in Chapter 6).

CHAPTER 6

Group-Blind Baseline — A Fairness-Agnostic Perspective

In scenarios where sensitive attributes are entirely unavailable, the most natural alternative is to train **group-blind classifiers** — models that do not use sensitive attributes at any stage of learning or evaluation. This chapter explores such baselines in contrast to proxy-based fairness interventions.

The aim is to answer a critical question:

> How does a group-blind model perform on fairness and accuracy metrics compared to models using sensitive or proxy attributes?

6.1 Motivation: Why Evaluate Group-Blind Models?

While the rest of the pipeline relies on some estimate of the sensitive attribute (true, proxy, or noisy), real-world deployments often operate in settings where:

- Collection of sensitive attributes is legally restricted
- Proxy inference is ethically risky or computationally infeasible
- Fairness must be enforced without group-based interventions

In such contexts, a **group-blind classifier** serves as a necessary reference:

- It represents the minimal-intervention setting
- It provides a baseline for fairness when no group information is available

6.2 Methodology

We define a group-blind model as any classifier trained solely on non-sensitive features X, without any access to sensitive attribute A or its proxy $\hat{A}$.

The implementation is simple but deliberate:

```
mlp_model = train_mlp(x_train_df, y_train)
y_pred = mlp_model.predict(x_test_df)
acc = accuracy_score(y_test, y_pred)
fairness = equalized_odds_difference(y_test, y_pred, sensitive_featu
```

Even though the model does not use A during training or prediction, fairness is evaluated **with respect to the true sensitive attribute** to measure implicit biases that may emerge from correlations in the data.

6.3 Key Observations

When we compare the group-blind model to those trained with true A or proxy $\hat{A}$, we observe:

- **Lower fairness scores**: Group-blind classifiers often exhibit significant disparity across sensitive groups.
- **Competitive accuracy**: Since the sensitive attribute is not always strongly predictive of Y, group-blind models may still achieve comparable accuracy.
- **No explicit fairness control**: Any fairness achieved is incidental and not by design.

6.4 Theoretical Framing

Group-blind baselines are consistent with the literature on fairness through unawareness, which argues that removing sensitive attributes avoids direct discrimination. However, this comes with major caveats:

- **Indirect discrimination**: Correlated features can reintroduce group-specific disparities

- **No accountability**: Without access to group membership, fairness violations cannot be measured or mitigated during training

Thus, while group-blind models avoid ethical/legal complications, they offer no guarantees on fairness — and often perform poorly on fairness metrics.

6.5 Experimental Role in Our Pipeline

We include the group-blind baseline in all frontier evaluations as the $k = 0$ setting:

- No proxy classifiers are used
- No sensitive attributes are appended

This allows us to benchmark the *best achievable fairness without group information*, against fairness-aware methods.

Our results show that:

- Group-blind models are usually the **least fair** in Equalized Odds
- Adding even weak proxy attributes improves fairness
- Group-blind serves as a lower bound on fairness intervention effectiveness

6.6 Summary

This chapter evaluated the role of group-blind classifiers as a fairness-agnostic baseline. We:

- Motivated the use of such models in restricted environments
- Compared their performance with proxy-based and fairness-aware methods
- Observed that they often fall short on fairness despite decent accuracy

Group-blind classifiers remind us that *ignoring group information does not ensure fairness* — and highlight the value of even imperfect proxies when sensitive attributes are unavailable.

In the next chapter, we summarize the overall insights and propose future directions for fairness evaluation under weak supervision.

CHAPTER 7

Conclusion

This thesis successfully validates the theoretical framework for fairness evaluation in the absence of sensitive attributes. Our work rigorously demonstrated that a naive, single-proxy approach is fundamentally flawed, leading to a misleading and unreliable assessment of model fairness. This finding underscores a critical need for a more robust methodology.

The core contribution of this project is the empirical proof that an ensemble-based pipeline provides a verifiable solution. We have shown that by leveraging multiple, independent proxy classifiers, the ensemble's collective judgment effectively averages out noise and converges toward the true fairness frontier. This powerful finding is corroborated by our analysis of the i.i.d. assumption, which we found to be a practical necessity for ensuring the stability and reliability of fairness estimates.

Furthermore, this research offers a compelling answer to the crucial trade-off between privacy and utility. Our experiments with DP noise proved that it is possible to use intentionally weak proxies to protect sensitive information while still maintaining a high level of predictive accuracy. This balance positions our methodology as a viable and ethical tool for fairness auditing in real-world scenarios. Ultimately, this thesis moves beyond theoretical claims to provide a concrete, validated framework that ensures fairness can be measured and enforced, even when the necessary data is missing.

REFERENCES

1. **Alao, R.**, **M. Bogen**, **J. Miao**, **I. Mironov**, and **J. Tannen** (). How Meta is working to assess fairness in relation to race in the U.S. across its products and systems.

2. **Awasthi, P.**, **A. Beutel**, **M. Kleindessner**, **J. Morgenstern**, and **X. Wang** (2021). Evaluating Fairness of Machine Learning Models Under Uncertain and Incomplete Information. URL http://arxiv.org/abs/2102.08410. ArXiv:2102.08410 [cs].

3. **Awasthi, P.**, **M. Kleindessner**, and **J. Morgenstern** (2020). Equalized odds postprocessing under imperfect group information. URL http://arxiv.org/abs/1906.03284. ArXiv:1906.03284 [stat].

4. **Belli, L.**, **K. Yee**, **U. Tantipongpipat**, **A. Gonzales**, **K. Lum**, and **M. Hardt** (2023). County-level Algorithmic Audit of Racial Bias in Twitter's Home Timeline. URL http://arxiv.org/abs/2211.08667. ArXiv:2211.08667 [cs].

5. **Berk, R.**, *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. SpringerBriefs in Computer Science. Springer, New York, NY, 2012. ISBN 978-1-4614-3084-1 978-1-4614-3085-8. URL https://link.springer.com/10.1007/978-1-4614-3085-8.

6. **Berk, R.**, *Machine Learning Risk Assessments in Criminal Justice Settings*. Springer International Publishing, Cham, 2019. ISBN 978-3-030-02271-6 978-3-030-02272-3. URL http://link.springer.com/10.1007/978-3-030-02272-3.

7. **Calders, T.**, **F. Kamiran**, and **M. Pechenizkiy**, Building Classifiers with Independency Constraints. *In 2009 IEEE International Conference on Data Mining Workshops*. IEEE, Miami, FL, USA, 2009. ISBN 978-1-4244-5384-9. URL http://ieeexplore.ieee.org/document/5360534/.

8. **Chouldechova, A.** (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, **5**(2), 153–163. ISSN 2167-647X.

9. **Dixon, M. F.**, **I. Halperin**, and **P. Bilokon**, *Machine Learning in Finance: From Theory to Practice*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-41067-4 978-3-030-41068-1. URL http://link.springer.com/10.1007/978-3-030-41068-1.

10. **Ferrara, E.** (2023). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. *Sci*, **6**(1), 3. ISSN 2413-4155. URL http://arxiv.org/abs/2304.07683. ArXiv:2304.07683 [cs].

11. **Ghosh, A.**, **M. Jagielski**, and **C. Wilson**, Subverting Fair Image Search with Generative Adversarial Perturbations. *In 2022 ACM Conference on Fairness Accountability and Transparency*. ACM, Seoul Republic of Korea, 2022. ISBN 978-1-4503-9352-2. URL https://dl.acm.org/doi/10.1145/3531146.3533128.

12. **Hardt, M.**, **E. Price**, and **N. Srebro** (2016). Equality of Opportunity in Supervised Learning. URL http://arxiv.org/abs/1610.02413. ArXiv:1610.02413 [cs].

13. **Heaton, J. B.**, **N. G. Polson**, and **J. H. Witte** (2018). Deep Learning in Finance. URL http://arxiv.org/abs/1602.06561. ArXiv:1602.06561 [cs].

14. **Holstein, K.**, **J. W. Vaughan**, **H. D. III**, **M. Dudík**, and **H. Wallach**, Improving fairness in machine learning systems: What do industry practitioners need? *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019. URL http://arxiv.org/abs/1812.05239. ArXiv:1812.05239 [cs].

15. **Jafri, R.** and **H. R. Arabnia** (2009). A survey of face recognition techniques. *JIPS (Journal of Information Processing Systems)*, **5**(2), 41–68. URL https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART001413897. Publisher: .

16. **Kenfack, P. J.**, **S. E. Kahou**, and **U. Aïvodji** (). A Survey on Fairness Without Demographics.

17. **Pandey, M. K.**, **M. Mittal**, and **K. Subbiah** (2021). Optimal balancing & efficient feature ranking approach to minimize credit risk. *International Journal of Information Management Data Insights*, **1**(2), 100037. ISSN 2667-0968. URL https://www.sciencedirect.com/science/article/pii/S2667096821000306.

18. **Rudin, C.** (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1**(5), 206–215. ISSN 2522-5839. URL https://www.nature.com/articles/s42256-019-0048-x. Publisher: Nature Publishing Group.

19. **Shailaja, K.**, **B. Seetharamulu**, and **M. A. Jabbar** (2018). Machine Learning in Healthcare: A Review. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 910–914. URL https://ieeexplore.ieee.org/document/8474918/. Conference Name: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) ISBN: 9781538609651 Place: Coimbatore Publisher: IEEE.

20. **Tollenaar, N.** and **P. G. M. van der Heijden** (2013). Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **176**(2), 565–584. ISSN 0964-1998. URL https://www.jstor.org/stable/23355205. Publisher: Wiley.

21. **van den Broek, E.**, **A. Sergeeva**, and **M. Huysman** (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *Management Information Systems Quarterly*, **45**(3), 1557–1580. ISSN ISSN 0276-7783/ISSN 2162-9730. URL https://aisel.aisnet.org/misq/vol45/iss3/21.

22. **Veale, M.** and **R. Binns** (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, **4**(2), 2053951717743530. ISSN 2053-9517. URL https://doi.org/10.1177/2053951717743530. Publisher: SAGE Publications Ltd.

23. **Waters, A.** and **R. Miikkulainen** (2014). GRADE: Machine Learning Support for Graduate Admissions. *AI Magazine*, **35**(1), 64–64. ISSN 2371-9621. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2504.

24. **Woodworth, B.**, **S. Gunasekar**, **M. I. Ohannessian**, and **N. Srebro** (2017). Learning Non-Discriminatory Predictors. URL http://arxiv.org/abs/1702.06081. ArXiv:1702.06081 [cs].

25. **Zafar, M. B.**, **I. Valera**, **M. G. Rodriguez**, and **K. P. Gummadi** (2017). Fairness Constraints: Mechanisms for Fair Classification. URL http://arxiv.org/abs/1507.05259. ArXiv:1507.05259 [stat].

26. **Zhu, Z.**, **Y. Song**, and **Y. Liu** (2021). Clusterability as an Alternative to Anchor Points When Learning with Noisy Labels. URL http://arxiv.org/abs/2102.05291. ArXiv:2102.05291 [cs].

27. **Zhu, Z.**, **Y. Yao**, **J. Sun**, **H. Li**, and **Y. Liu** (2023). Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes. URL http://arxiv.org/abs/2210.03175. ArXiv:2210.03175 [cs].