Assessing the Role of Sensitive Attributes in Adversarial Debiasing

Diego Minatel[1], **Antonio R. S. Parmezan**[1], **Vinicius M. A. Souza**[2], **Solange O. Rezende**[1]

[1]Institute of Mathematics and Computer Science, University of São Paulo, São Carlos 13566-590, Brazil

[2]Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Curitiba 80215-901, Brazil

dminatel@usp.br, parmezan@usp.br, vinicius.mourao@pucpr.br, solange@icmc.usp.br

Abstract. *Fairness in machine learning refers to the development of models that do not systematically disadvantage individuals or groups based on sensitive attributes. One commonly adopted principle is fairness through unawareness, which holds that models should not explicitly incorporate sensitive attributes into the decision-making process. However, in some scenarios, such as healthcare, variables like sex and age are essential for accurate diagnoses and prognoses. Previous studies have assessed the influence of including or excluding such information during model training. Nonetheless, they have not considered Adversarial Debiasing, a classification algorithm specifically designed to promote equitable results. To address this gap, we propose a comprehensive empirical analysis to investigate the role of sensitive attributes in this algorithm. We experimentally evaluated Adversarial Debiasing across 20 settings using 23 datasets from varying domains, one predictive performance metric, three group fairness metrics, and a non-parametric statistical test. Our findings indicate that classifiers trained without including sensitive information in the input feature set produce more precise and fairer outcomes.*

1. Introduction

In recent years, it has become publicly acknowledged that machine learning models can make undesirable decisions from a societal perspective [Mavrogiorgos et al. 2024, Minatel et al. 2023b]. In response, researchers have proposed bias mitigation methods and adapted the learning process to incorporate fairness notions in order to address this concern [Minatel et al. 2025b, Minatel et al. 2023c, Zhang et al. 2018, Hardt et al. 2016, Kamiran and Calders 2012]. These efforts are essential for developing more impartial models, especially when they are trained on social data, where automated decisions may contribute to the spread of discrimination, inequality, misinformation, and defamation [Caton and Haas 2024, Barocas et al. 2023, Mehrabi et al. 2021].

One commonly adopted notion in the literature is *fairness through unawareness* [Grgic-Hlaca et al. 2016], which holds that a model should learn without incorporating sensitive information about individuals, such as race, gender, or age. This approach aims to prevent direct discrimination, since using such information may lead to

treating people with common traits adversely. However, avoiding direct discrimination does not guarantee the elimination of indirect discrimination [Žliobaitė and Custers 2016, Kamiran et al. 2010], also known as adverse impact, which may arise when model outcomes consistently disadvantage members of certain subpopulations.

Training models with sensitive attributes is justifiable in certain situations to avoid or mitigate adverse impacts. A well-known example is healthcare applications, where information such as age and sex is essential for accurate patient assessment [DSIT 2023]. Another example emerges in the criminal justice domain, where women have a significantly lower likelihood of recidivism than men [Corbett-Davies et al. 2023]. In such cases, it is important to consider the legal implications of utilizing these attributes in the country where the solution will be deployed [Zafar et al. 2017].

In legally regulated contexts, an open question remains as to whether the main fairness-aware methods proposed in the literature yield more accurate and fairer models with or without access to sensitive information. One of the most widely recognized algorithms for promoting group fairness (*i.e.*, the principle that individuals from different sociodemographic groups should receive similar outcomes) in classification tasks is Adversarial Debiasing [Zhang et al. 2018], which employs adversarial learning to produce outcomes invariant to the subpopulations represented in a given application. This algorithm has demonstrated strong performance in predictive accuracy and fairness metrics, especially in the healthcare domain [Zheng et al. 2025, Yang et al. 2023].

Previous studies have examined the role of sensitive attributes in building fairer models and argue that such attributes may be necessary depending on the characteristics of the data to achieve fair outcomes [Haeri and Zweig 2020, Žliobaitė and Custers 2016]. However, these studies often overlook the potential of fairness-aware algorithms, such as Adversarial Debiasing, to overcome inherent data limitations in developing more impartial classifiers. To the best of our knowledge, the literature still lacks comprehensive investigations into the influence of including or excluding sensitive attributes from the input feature set when training classifiers with Adversarial Debiasing.

Motivated by this gap, we present a comprehensive empirical analysis of the role of sensitive attributes in training classifiers using the Adversarial Debiasing algorithm. The contributions of this paper are threefold: (i) it investigates the impact of sensitive attributes on a fairness-aware algorithm; (ii) introduces an experimental protocol to evaluate the outcomes of classifiers trained with and without sensitive attributes; and (iii) conducts a robust empirical study across 23 datasets from diverse domains, such as healthcare, finance, and demographics.

Our findings suggest that not providing sensitive information as input to the decision-making process of classifiers trained employing Adversarial Debiasing leads to fairer and more accurate outcomes. With this study, we aim to enhance the understanding of the role of sensitive attributes in Adversarial Debiasing—more specifically, their inclusion in the input feature set—to foster further discussion in this research area and to support researchers and practitioners in developing fairer machine learning solutions.

2. Related Work

Group fairness analysis evaluates model behavior concerning the different subpopulations defined by sensitive attributes, also known as protected attributes. The literature typically

categorizes the subpopulations into two groups: privileged and unprivileged. This type of analysis does not depend on whether the model was trained using these attributes; it only requires that such information be available during model induction [Barocas et al. 2023, Mehrabi et al. 2021].

The main fairness notions associated with this type of analysis are demographic parity, equal opportunity, and equalized odds [Caton and Haas 2024, Hardt et al. 2016, Dwork et al. 2012]. Achieving these notions requires attaining parity in outcomes between privileged and unprivileged groups based on the performance metric associated with each definition. For instance, equal opportunity demands that both groups have the same recall score. Since achieving perfect parity is often a challenge in practice, we convert these notions into measurable metrics by computing either the difference or the ratio between the groups' scores [Barocas et al. 2023].

These metrics help measure the effectiveness of incorporating the notion of fairness through unawareness into the learning process. As previously discussed, specific legal frameworks support this concept and prohibit the use of sensitive data in automated decision-making, leading to the development of approaches that exclude such attributes throughout the training process [Zhao et al. 2022]. However, some studies take the opposite stance and argue that these attributes should be employed during model training [Haeri and Zweig 2020, Žliobaitė and Custers 2016].

In [Žliobaitė and Custers 2016], the authors contend that sensitive attributes are necessary for building fairer regression models but should not be used after training—that is, the sensitive attribute should not serve as input to the model. They also discuss the regulatory conflict between restrictions on collecting and employing sensitive data and the fact that such information can lead to fairer regressors. In the same direction, [Haeri and Zweig 2020] starts from the premise that machines behave differently from humans and that fairness cannot be guaranteed simply by ignoring sensitive attributes. In certain scenarios, they demonstrate that incorporating these attributes during training leads to fairer classifiers and propose applying the Kolmogorov-Smirnov test to identify such cases, which occur when the data distribution varies across the groups.

These publications evaluated traditional classification algorithms, such as Random Forest or Multi-layer Perceptron, on a minimal set of benchmark datasets. However, they did not consider Adversarial Debiasing [Zhang et al. 2018], one of the most widely used fairness-aware algorithms for classification. Other studies have conducted experimental evaluations that investigated Adversarial Debiasing. One found that cross-validation with stratification by class and group helps identify hyperparameter values that lead to fairer classifiers [Minatel et al. 2025a, Minatel et al. 2023a]. Another showed that balancing the distribution of sensitive attributes is more effective for enabling adversarial learning to achieve fairness objectives [Beutel et al. 2017].

Our work differs from these studies by providing a robust evaluation of how including or excluding sensitive information from the input feature set affects Adversarial Debiasing, following the methodology detailed in Section 4. We consider over twenty datasets, along with predictive performance and multiple group fairness measures to assess this fairness-aware algorithm explicitly designed to promote less biased outcomes across groups.

3. Adversarial Debiasing

Adversarial Debiasing leverages adversarial learning to build fairer classifiers. This classification algorithm is trained on the tuple (X, Y, Z) to induce the classifier $\hat{Y} = f(X)$, where X is the feature matrix, Y represents the ground-truth labels, and Z corresponds to the protected attributes. As portrayed in Figure 1, the architecture of Adversarial Debiasing consists of two neural networks: the predictor and the adversary. The predictor learns to predict Y given X, and its output layer serves as the input to the adversary network. The adversary then attempts to predict Z, namely the value of the protected attribute associated with the given instance [Zhang et al. 2018].

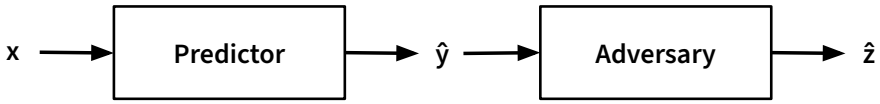

Figure 1. Simplified architecture of Adversarial Debiasing.

This formulation aims to optimize the predictor's ability to predict Y while minimizing the adversary's ability to predict Z. In doing so, the classifier learns to produce outputs invariant to the group associated with the protected attributes. Adversarial Debiasing can be applied to optimize fairness notions such as demographic parity, equal opportunity, and equalized odds. The definition presented in this section, illustrated in Figure 1, describes the Adversarial Debiasing approach for enforcing demographic parity. To apply other fairness definitions, the adversary must receive additional information. For instance, in the case of equalized odds, the adversary must be given both the predicted label $\hat{Y}$ and the actual label Y. Although the adversary accesses the protected attributes during the training step, the user decides whether or not to include these attributes in X.

4. Methodology

In this section, we present the methodology adopted in this study. We detail the experimental protocol conceived to evaluate the impact of including or excluding sensitive attributes from the input feature set during model training. We also describe the benchmark datasets, the evaluation metrics, and the rationale behind our experimental design.

4.1. Proposed Experimental Protocol

The primary goal of this study was to assess the effect of including protected attributes Z in the input feature set X during the training of models using the Adversarial Debiasing algorithm. For this purpose, we designed a novel experimental protocol, as displayed in Figure 2. In the initial step, we created two distinct versions of each preprocessed dataset: one where the protected attributes were retained in the input feature set ($Z \subseteq X$), and another where they were excluded ($Z \not\subseteq X$). Section 4.2 details the datasets and the protected attributes considered in this investigation.

For each dataset version, we performed a five-fold cross-validation stratified by both group and class, as suggested by [Minatel et al. 2025a], to reduce evaluation bias. We opted for five folds due to the limited number of examples in some of the selected datasets. Additionally, we ensured that the folds used in the versions with and without the protected attributes from the same original dataset contained the same examples. Within

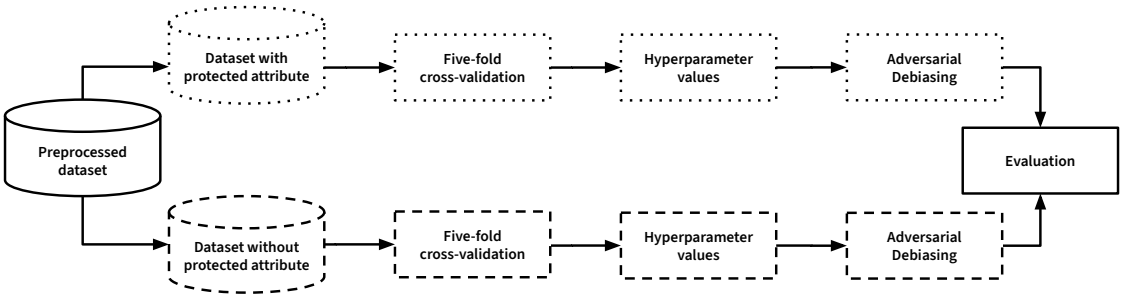

Figure 2. Overview of the experimental protocol proposed. First, we generated two versions of a dataset: one preserving the protected attributes and another excluding them. Next, we applied Adversarial Debiasing with diverse settings for both versions. Finally, we compared the performance of the resulting models to assess the impact of protected attributes in inducing fairer classifiers.

each fold, we trained twenty classifiers using the Adversarial Debiasing algorithm. To obtain these models, we varied the number of training epochs from 50 to 240 in increments of 10. All other hyperparameters—including the number of hidden units in the classifier, batch size, and adversary loss weight—followed the default settings provided by the AIF360 Python library[1]. Although tuning these hyperparameters for each dataset could improve performance, we kept them fixed across all experiments to ensure a consistent comparison. This choice also reduces computational costs and promotes reproducibility by limiting variability in the experimental setup.

After training the models, we evaluated them using predictive performance metrics and group fairness measures, as described in Section 4.3. The reported scores correspond to the average across the five iterations of the cross-validation procedure. We compared the mean results between classifiers built with and without the inclusion of protected attributes. Finally, we applied the Wilcoxon signed-rank test, a non-parametric method for comparing paired samples, to assess whether the observed differences were statistically significant.

We implemented the source code in Python employing the following libraries: Pandas (dataset loading and data manipulation), Scikit-learn (cross-validation and predictive performance metrics), AIF360 (Adversarial Debiasing and group fairness metrics), and SciPy (Wilcoxon signed-rank test). The complete source code is available in the following public repository: `https://github.com/diegominatel/assessing-role-sensitive-attributes`.

4.2. Datasets with Sensitive Attributes

We selected 23 benchmark datasets from various domains commonly used for fairness evaluation in the machine learning research community. Table 1 summarizes their main characteristics, including the number of instances (#I), the number of attributes (#A), the protected attribute (#PA), the unprivileged group (#UG), the proportion of instances belonging to the positive class (#PPC), and the proportion of instances belonging to the unprivileged group (#PUG).

[1]Documentation available at `https://aif360.readthedocs.io/en/stable/modules/generated/aif360.sklearn.inprocessing.AdversarialDebiasing.html`.

Table 1. Characteristics of the benchmark datasets evaluated.

Dataset	#I	#A	#PA	#UG	#PPC	#PUG	Reference
Alcohol	1,885	11	Ethnicity	Non-White	98.19%	8.75%	[Kelly et al. 2017]
Amphet	1,885	11	Ethnicity	Non-White	48.22%	8.75%	[Kelly et al. 2017]
Arrhythmia	452	278	Sex	Female	43.57%	55.95%	[Kelly et al. 2017]
Bank Marketing	45,211	42	Age	Under 25	11.69%	2.95%	[Kelly et al. 2017]
Cannabis	1,885	11	Ethnicity	Non-White	78.09%	8.75%	[Kelly et al. 2017]
Census Income	48,842	76	Race and Sex	Non-White or Female	24.90%	37.35%	[Kelly et al. 2017]
Coke	1,885	11	Ethnicity	Non-White	55.06%	8.75%	[Kelly et al. 2017]
Contraceptive	1,473	10	Religion	Islam	55.48%	86.22%	[Kelly et al. 2017]
Crack	1,885	11	Ethnicity	Non-White	13.68%	8.75%	[Kelly et al. 2017]
Credit Card	30,000	24	Sex	Female	22.12%	60.37%	[Kelly et al. 2017]
Diabetes	45,715	50	Gender	Female	24.21%	54.73%	[Kelly et al. 2017]
Dutch Census	60,420	12	Sex	Female	47.60%	50.10%	[Van der Laan 2001]
Ecstasy	1,885	11	Ethnicity	Non-White	45.83%	8.75%	[Kelly et al. 2017]
German Credit	1,000	36	Sex	Female	70.00%	31.00%	[Kelly et al. 2017]
Heart	383	13	Age	Non-Middle-Aged	46.12%	41.41%	[Kelly et al. 2017]
Heroin	1,885	11	Ethnicity	Non-White	14.85%	8.75%	[Kelly et al. 2017]
LSD	1,885	11	Ethnicity	Non-White	43.28%	8.75%	[Kelly et al. 2017]
Nicotine	1,885	11	Ethnicity	Non-White	77.29%	8.75%	[Kelly et al. 2017]
Recid. Female	1,395	176	Race	Non-White	37.32%	53.36%	[Larson et al. 2016]
Recid. Male	5,819	375	Race	Non-White	49.69%	62.04%	[Larson et al. 2016]
Ricci	118	6	Race	Non-White	47.45%	42.37%	[Feldman et al. 2015]
Student	480	46	Sex	Female	73.84%	36.61%	[Amrieh et al. 2015]
Titanic	1,309	6	Sex	Male	40.07%	61.84%	[Vanschoren et al. 2014]

Alcohol, Amphet, Cannabis, Coke, Crack, Ecstasy, Heroin, LSD, and Nicotine are subsets of the Drug Consumption dataset, which contains survey responses on legal and illegal drug use. Each subset defines a different target variable, indicating the consumption or non-consumption of the corresponding drug. We split the Recidivism dataset into two subsets: one containing male instances and the other containing female instances. We binarized the target classes in Arrhythmia (presence vs. absence of cardiac arrhythmia), Contraceptive (use vs. non-use of contraceptive methods), and Student (low vs. medium-high academic performance). All datasets underwent preprocessing, which included one-hot encoding of categorical features and standardization. For the Credit Card, Diabetes, Dutch Census, and Ricci datasets, we adopted the preprocessing procedures suggested in [Le Quy et al. 2022].

4.3. Evaluation Metrics

The evaluation metrics defined in the experimental protocol are macro-averaged F1-score for predictive performance (due to the class imbalance in the datasets) and demographic parity, equalized odds, and equal opportunity for group fairness analysis. Table 2 presents the details of each of these metrics.

Table 2. Description of the evaluation metrics.

Acronym	Description	Range value	Ideal value
Macro F1-Score	Macro-averaged F1-score	$[0, 1]$	1
RDP	Ratio of scores relative to demographic parity	$[0, 1]$	1
REO	Ratio of scores relative to equal opportunity	$[0, 1]$	1
RDO	Ratio of scores relative to equalized odds	$[0, 1]$	1

Previously, we discussed in Section 2 the need to transform group fairness notions into measurable metrics. For this purpose, we compute the ratio between the scores of the

privileged and unprivileged groups for the performance measures associated with these notions, which we refer to as RDP, REO, and RDO, as shown in Table 2. The computation of RDP, REO, and RDO is defined by Equations 1, 2, and 3, respectively, where the subscripted terms (*e.g.*, $\text{Recall}_{\text{pg}}$ and $\text{Recall}_{\text{ug}}$) denote the scores obtained for examples belonging to the privileged (PG) and unprivileged (UG) groups. In these equations, FPR stands for false positive rate, and PPR refers to the predicted positive rate.

$$\text{RDP} = \frac{\min(\text{PPR}_{\text{PG}}, \text{PPR}_{\text{UG}})}{\max(\text{PPR}_{\text{PG}}, \text{PPR}_{\text{UG}})} \tag{1}$$

$$\text{REO} = \frac{\min(\text{Recall}_{\text{PG}}, \text{Recall}_{\text{UG}})}{\max(\text{Recall}_{\text{PG}}, \text{Recall}_{\text{UG}})} \tag{2}$$

$$\text{RDO} = \frac{\min(\frac{\text{Recall}_{\text{PG}}+\text{FPR}_{\text{PG}}}{2}, \frac{\text{Recall}_{\text{UG}}+\text{FPR}_{\text{UG}}}{2})}{\max(\frac{\text{Recall}_{\text{PG}}+\text{FPR}_{\text{PG}}}{2}, \frac{\text{Recall}_{\text{UG}}+\text{FPR}_{\text{UG}}}{2})} \tag{3}$$

We placed the higher value in the denominator to ensure each score lies between 0 and 1. This normalization allowed us to assess how close the classifier was to the ideal value of 1 for each fairness concept, regardless of which group was favored or disadvantaged. By doing so, we treated disparities symmetrically, focusing on the magnitude of unfairness rather than its direction. For instance, consider the REO metric: if $\text{Recall}_{\text{UG}} = 0.60$ and $\text{Recall}_{\text{PG}} = 0.75$, then $\text{REO} = \frac{\min(0.75, 0.60)}{\max(0.75, 0.60)} = \frac{0.60}{0.75} = 0.80$.

5. Results and Discussion

In this section, we present and discuss the results obtained according to the experimental setup described in Section 4, which aims to evaluate the influence of using sensitive attributes in the induction of classifiers through the Adversarial Debiasing algorithm.

Table 3 presents the average scores for the Macro F1-Score, RDP, REO, and RDO metrics obtained from training 20 classifiers per dataset on both versions with and without the protected attributes. Bold values highlight the highest average score for each dataset-metric pair. We omitted standard deviations, as they did not differ substantially between the classifiers generated from these two versions.

As shown in Table 3, excluding sensitive attributes from the feature matrix yields higher average scores across all metrics for most datasets, with 'Without' outperforming 'With' in at least 17 datasets for each of these measures. These results are most significant for RDO, where models trained without sensitive attributes achieved the best results in 20 of the 23 datasets. Notably, the overall averages for RDP, REO, and RDO increase by at least 14 percentage points with sensitive attributes compared to those without, whereas the improvement in Macro F1-Score is eight percentage points.

We highlight three specific cases. The first is Arrhythmia, the only dataset in which 'With' outperformed 'Without' across all metrics, with a notable improvement of nearly seven percentage points in REO. The second is Alcohol, where 'Without' achieved scores near the ideal across all four metrics, while 'With' resulted in a drop of over fifty percentage points in those same indicators. The third involves the Bank Marketing, Census Income, and Recidivism Female datasets, where using sensitive attributes led to better

Table 3. Average scores for the Macro F1-Score, RDP, REO, and RDO.

Dataset	Macro F1-Score		RDP		REO		RDO	
	Without	With	Without	With	Without	With	Without	With
Alcohol	**0.9607**	0.4348	**0.9985**	0.4022	**0.9984**	0.4012	**0.9992**	0.4362
Amphet	**0.8563**	0.8035	**0.7928**	0.4232	**0.8228**	0.4675	**0.7420**	0.3874
Arrhythmia	0.8881	**0.8968**	0.6043	**0.6433**	0.7673	**0.8350**	0.6486	**0.6866**
Bank Marketing	**0.7895**	0.7752	**0.7050**	0.6900	0.6446	**0.7391**	0.6405	**0.6900**
Cannabis	**0.9067**	0.8206	**0.8754**	0.6263	**0.9596**	0.6752	**0.8382**	0.5193
Census Income	**0.9823**	0.9189	0.3427	**0.7945**	**0.8654**	0.6678	0.5803	**0.6927**
Coke	**0.8991**	0.8376	**0.7644**	0.6156	**0.7849**	0.5990	**0.7640**	0.5441
Contraceptive	**0.9494**	0.8871	**0.8251**	0.6915	**0.9221**	0.7533	**0.8776**	0.6903
Crack	**0.8512**	0.7994	**0.3304**	0.1839	0.1278	**0.1600**	**0.2088**	0.1210
Credit Card	**0.9911**	0.9852	**0.8454**	0.8431	**0.9601**	0.9394	**0.8983**	0.8787
Diabetes	0.9855	**0.9924**	**0.7674**	0.6195	**0.6723**	0.5396	**0.6922**	0.5917
Dutch Census	**0.9724**	0.9654	**0.6441**	0.5615	0.9599	**0.9695**	**0.9046**	0.7549
Ecstasy	**0.8312**	0.7423	**0.8029**	0.2844	**0.7760**	0.3013	**0.6414**	0.2479
German Credit	**0.9456**	0.9397	**0.8644**	0.8599	**0.8913**	0.8844	**0.8302**	0.8196
Heart	**0.8890**	0.8749	0.6612	**0.7412**	**0.8589**	0.8454	**0.6671**	0.6579
Heroin	**0.8132**	0.7727	**0.5554**	0.1763	**0.3283**	0.1316	**0.3287**	0.0929
LSD	**0.9195**	0.9037	**0.7115**	0.5610	**0.7922**	0.6881	**0.7257**	0.5746
Nicotine	**0.8863**	0.8092	**0.9363**	0.5000	**0.9663**	0.5388	**0.9212**	0.4534
Recidivism Female	**0.9012**	0.8870	0.8196	**0.8269**	0.8265	**0.8273**	**0.7917**	0.7826
Recidivism Male	0.9550	**0.9687**	**0.6948**	0.6810	**0.7851**	0.7786	**0.7430**	0.7259
Ricci	**0.7436**	0.2923	**0.5907**	0.3789	**0.6833**	0.5476	**0.4750**	0.3166
Student	0.8122	**0.8430**	0.6746	**0.6765**	0.7831	**0.7947**	**0.5801**	0.5322
Titanic	**0.8371**	0.5538	**0.6255**	0.3552	**0.8156**	0.2970	**0.6239**	0.2288
Average	**0.8942**	0.8132	**0.7145**	0.5711	**0.7823**	0.6253	**0.7010**	0.5402

scores in two metrics, while excluding them yielded higher scores in two others. In such scenarios, it is difficult to determine which approach is preferable, as the choice depends on the target application's specific fairness and performance priorities.

Figure 3 shows a heatmap where rows represent datasets, and columns denote evaluation metrics. The color of each cell reflects the result of 'Without' − 'With', with green indicating positive differences and red indicating negative ones. The color intensity is proportional to the magnitude of the difference. Each cell also includes one of three symbols ('↑', '↓', '≃') to indicate the outcome of the Wilcoxon signed-rank test at a 5% significance level. The symbol '↑' denotes a statistically significant difference favoring 'Without'; '↓' indicates a significant difference favoring 'With'; and '≃' means no statistically significant difference was found between these results.

The results of the Wilcoxon test, which performs pairwise comparisons, align with the average scores presented in Table 3, as the colors and symbols in each cell reflect the same trend in results. In 15 datasets, a statistically significant difference was observed across all analyzed metrics. Only German Credit showed no statistically significant difference in any metric, indicating that the sensitive attribute had a minimal impact on this dataset.

Our experimental results suggest that training classifiers with Adversarial Debiasing without including sensitive attributes in the feature matrix tends to produce fairer outcomes. Thus, we recommend applying Adversarial Debiasing without using sensitive data as input features, in alignment with the principle of fairness through unawareness. However, this recommendation should be carefully considered in the healthcare domain, where results were inconclusive: Diabetes performed better without sensitive attributes, Arrhythmia with them, and Heart showed similar performance in both cases.

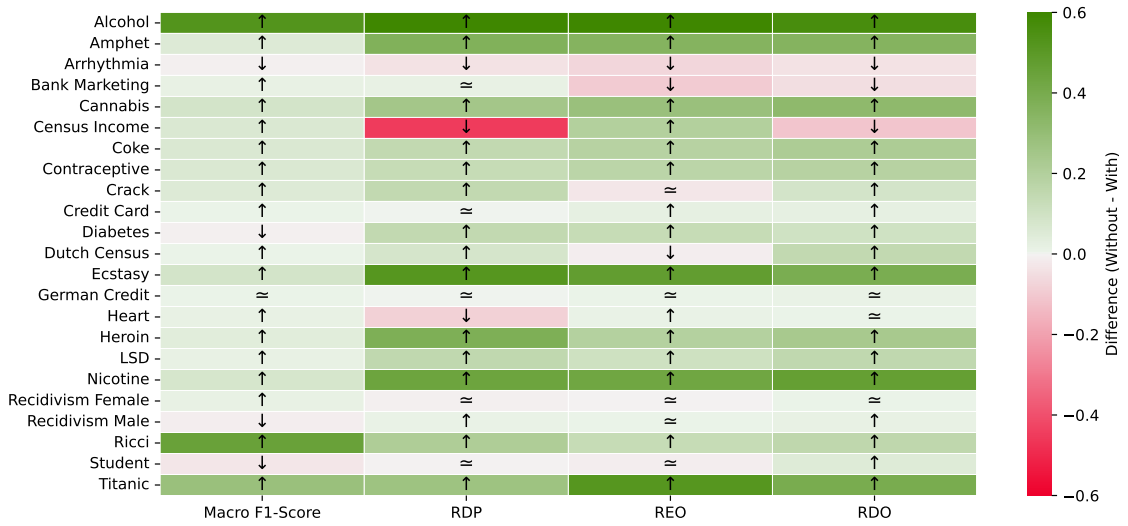

Figure 3. Heatmap of 'With' and 'Without' results, where the colors indicate the value of 'Without' minus 'With', and the symbols within each cell denote the outcome of the non-parametric Wilcoxon test. Symbols '↑' and '↓' indicate a statistically significant difference between the results, with the arrow direction showing the performance trend, while '≃' indicates no statistically significant difference.

6. Concluding Remarks

This paper presented an in-depth comparative empirical analysis of the Adversarial Debiasing algorithm, evaluating classifiers trained with and without protected attributes to assess their impact on predictive performance and group fairness. Our experimental results indicate that using Adversarial Debiasing without incorporating sensitive information leads to fairer models with better predictive power than when these attributes are included in the feature matrix. However, these findings should be interpreted cautiously in the healthcare domain, as the results did not exhibit a consistent trend in this area.

In future work, we plan to expand the experimental setup by testing additional hyperparameters of the Adversarial Debiasing algorithm, such as the adversary loss weight. We also intend to compare the best-performing configurations for each dataset, both with and without protected attributes in the input feature set. This comparison may help identify which scenario has greater potential for producing fairer classifiers. Additionally, we aim to incorporate more healthcare datasets to investigate the Adversarial Debiasing performance in this domain, as well as datasets from other application areas where the use of protected attributes is applicable and ethically acceptable. Furthermore, we intend to broaden this analysis to include other fairness-aware methods and algorithms, thereby deepening our understanding of the role of sensitive attributes in building fairer machine learning models.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Numbers #2022/02176-8 and #2024/14211-8.

References

Amrieh, E. A., Hamtini, T., and Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. In *AEECT*, pages 1–5. IEEE.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7):166:1–166:38.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., and Goel, S. (2023). The measure and mismeasure of fairness. *J. Mach. Learn. Res.*, 24(1):14730–14846.

DSIT (2023). *Capabilities and Risks from Frontier AI: A Discussion Paper on the Need for Further Research Into AI Risk*. Department for Science, Innovation & Technology.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *ITCS*, pages 214–226. ACM.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 11. Curran Associates Inc.

Haeri, M. A. and Zweig, K. A. (2020). The crucial role of sensitive attributes in fair classification. In *SSCI*, pages 2993–3002. IEEE.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*, pages 3323–3331. Curran Associates Inc.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33.

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *ICDM*, pages 869–874. IEEE.

Kelly, M., Longjohn, R., and Nottingham, K. (2017). The UCI machine learning repository.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. ProPublica.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.*, 12(3):e1452.

Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., and Kyriazis, D. (2024). Bias in machine learning: A literature review. *Appl. Sci.*, 14(19):8860.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):1–35.

Minatel, D., da Silva, A. C. M., dos Santos, N. R., Cúri, M., Marcacini, R. M., and de Andrade Lopes, A. (2023a). Data stratification analysis on the propagation of discriminatory effects in binary classification. In *KDMILE*, pages 73–80. SBC.

Minatel, D., dos Santos, N. R., da Silva, A. C. M., Cúri, M., Marcacini, R. M., and de Andrade Lopes, A. (2025a). Influence of data stratification criteria on fairer classifications. *J. Inf. Data Manag.*, 16(1):161–169.

Minatel, D., dos Santos, N. R., da Silva, A. C. M., Cúri, M., Marcacini, R. M., and Lopes, A. d. A. (2023b). Unfairness in machine learning for web systems applications. In *WebMedia*, pages 144–153. ACM.

Minatel, D., Parmezan, A. R., Cúri, M., and Lopes, A. D. A. (2023c). Fairness-aware model selection using differential item functioning. In *ICMLA*, pages 1971–1978. IEEE.

Minatel, D., Parmezan, A. R. S., Roque dos Santos, N., Cúri, M., and Lopes, A. (2025b). A dif-driven threshold tuning method for improving group fairness. In *SAC*, pages 890–898. ACM.

Van der Laan, P. (2001). *The 2001 Census in the Netherlands: Integration of Registers and Surveys*, pages 39–52.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2014). OpenML: Networked science in machine learning. *SIGKDD Explor. Newsl.*, 15(2):49–60.

Yang, J., Soltan, A. A., Eyre, D. W., Yang, Y., and Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *npj Digit. Med.*, 6(1):55.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *AISTATS*, pages 962–970. PMLR.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *AIES*, pages 335–340. ACM.

Zhao, T., Dai, E., Shu, K., and Wang, S. (2022). Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *WSDM*, pages 1433–1442. ACM.

Zheng, G., Jacobs, M. A., Braverman, V., and Parekh, V. S. (2025). Towards fair medical ai: Adversarial debiasing of 3d ct foundation embeddings. *arXiv preprint arXiv:2502.04386*.

Žliobaitė, I. and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artif. Intell. Law*, 24:183–201.