Pre-processing Techniques to Mitigate Against Algorithmic Bias

Maliheh Heidarpour Shahrezaei
RSRC
DKIT
Dundalk, Ireland
Maliheh.heidarpour@dkit.ie

Róisín Loughran
RSRC
DKIT
Dundalk, Ireland
Roisin.Loughran@dkit.ie

Kevin Mc Daid
RSRC
DKIT
Dundalk, Ireland
kevin.mcdaid@dkit.ie

***Abstract*—A significant portion of current AI research is focused on ensuring that model decisions are fair and free of bias. Such research should consider not merely the algorithm but also the datasets, metrics and approaches used. In this paper, we work on several pre-processing techniques to achieve fair results for classification tasks by assigning weights, sampling and changing class labels. We used two well-known classifiers, Logistic Regression and Decision Tree, performing experiments on a popular data set in the fairness domain. This research aims to compare the effects of different pre-processing techniques on the resulting confusion matrix elements and the derived fairness metrics. We found that the Massaging technique with the Logistic regression classifier resulted in the Disparate Impact value that was closest to one. While, for the Decision Tree classifier, Reweighting and Uniform Sampling performed better than Massaging for all of our fairness metrics and both sensitive attributes.**

Keywords—Fairness, Bias, Metric, Pre-processing, Mitigation Techniques

I. Introduction

Bias can be intentionally or unintentionally introduced into Artificial Intelligence (AI) systems or algorithms, or it can develop when they are employed in a particular application. The presence of negative bias in such systems can result in unfair outcomes, weakened public trust, legal and ethical challenges, and diminished human dignity [1]. The goal of bias mitigation research is to develop algorithms and/or other approaches that produce predictions that are considered fair for both privileged and underprivileged groups, particularly in relation to sensitive attributes.

Debiasing algorithms can, in general, be divided into three categories. Pre-processing algorithms modify data before it is presented to a processing algorithm [2]. One of the advantages of these techniques is that they are employed early in the life cycle, so if the data incorporates biases, the classification model could learn this bias and possibly mitigate against it. In-processing algorithms change the procedure of the algorithm to make fairer results [2]. Post-processing algorithms alter the predictions made by algorithms to mitigate the anticipated impact of bias [3]. The effectiveness of these strategies has been investigated using several fairness metrics and datasets, determining that there is no single algorithm that performs better than all other algorithms with all fairness measures across all datasets [4]. Each mechanism has its benefits and drawbacks. Pre-processing mechanisms can be applied to most classification algorithms, but there is a great deal of uncertainty regarding the level of accuracy that will be attained. Additionally, they might challenge the results' explainability [1]. Although post-process techniques can be used with any classification algorithm, they often produce inferior results since they are added relatively late in the learning process. In-processing techniques may explicitly enforce the necessary trade-off between accuracy and fairness in the goal function, but the selection of methods varies with parameters. In this study, we conducted experiments on several pre-processing bias reduction strategies based on Kamiran and Calders' studies [5]–[7].

The next two sections look at the popular fairness metrics definitions with their formula and the key information about the Adult Income dataset. Section IV describes the methodology that has been used in this research. Section V discusses the results obtained and some conclusions and future work are proposed in Section VI.

II. Metrics

Most of the research into fairness in algorithms examines the amount of bias that is formed, influenced, or mitigated. In the case of the binary classification task, there are four possibilities based on the confusion matrix comparing model predictions with the actual results, True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). These values are calculated for both privileged and unprivileged groups and they can help to make measurements resulting from various definitions of fairness conceptualized in ML applications [8]. True Positive Rate (TPR) is the probability of an actual positive individual being correctly identified [9]. The False Positive Rate (FPR) is the probability of anegative event being wrongly categorized as positive. True Negative Rate (TNR) is the probability of actual negative correctly identified [9]. The most popular measurements for classifying ML fairness and their definitions based on core measurements are given below. In the following formulae, the number of members of dataset is denoted by N, the number of members of the unprivileged group is denoted by N_U and the number of members of the privileged group is denoted by N_P.

A. Disparate Impact

Disparate Impact (DI) gives a measure of how similar the percentage of positive predictions is among groups. The ideal value is one [10]. Equation (1) is used for calculating this measurement.

$$\mathrm{DI} = \frac{(\mathrm{TP_U} + \mathrm{FP_U})/\mathrm{N_U}}{(\mathrm{TP_P} + \mathrm{FP_P})/\mathrm{N_P}} \tag{1}$$

B. Demographic/Statistical Parity

Statistical Parity (SP) resembles DI but instead of using the ratio, the difference is taken [10]. Regardless of whether a person belongs to the sensitive category, the likelihood of a positive outcome should be the same. The ideal result is zero. Equation (2) shows the calculation of SP.

979-8-3503-6021-9/23/$31.00 ©2023 IEEE

$$\mathrm{SP} = \left(\frac{\mathrm{TP_U} + \mathrm{FP_U}}{\mathrm{N_U}}\right) - \left(\frac{\mathrm{TP_P} + \mathrm{FP_P}}{\mathrm{N_P}}\right) \tag{2}$$

C. *Average Odds Difference*

Average Odds Difference (AOD) calculates the inequality between groups on the combination of the TPR and the FPR. In other words, the rates of true positives and false positives should be the same for the privileged and unprivileged groups [10]. The ideal result for this is zero. Equation (3) is used for calculating AOD.

$$\mathrm{AOD} = \frac{(\mathrm{FPR_U} - \mathrm{FPR_P}) + (\mathrm{TPR_U} - \mathrm{TPR_P})}{2} \tag{3}$$

D. *Equal Opportunity*

Equal Opportunity (EO) ensures everyone is treated similarly and satisfies the same requirements [11]. It mandates that the privileged and unprivileged groups should have similar TPR. The ideal result is zero. Equation (4) shows the calculation of EO.

$$\mathrm{EO} = \mathrm{TPR_U} - \mathrm{TPR_P} \tag{4}$$

The Accuracy (ACC) of an algorithm is the most common measure of performance [12]. ACC measures the number of correct predictions divided by the number of total predictions. In dealing with an imbalanced dataset Balanced Accuracy (BA) can be used [13]. BA calculates the average of TPR and TNR [13]. Equations (5) and (6) were used for calculating ACC and BA respectively.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right) \tag{6}$$

III. DATASET

The Adult Income dataset was extracted, pre-processed, and donated to the UCI Machine Learning Repository in 1996 from the census database by Ronny Kohavi and Barry Becker [14]. The target of the prediction task was a binary variable indicating whether the respondent's income exceeded \$50,000. The size of this dataset is 48,842. This dataset is imbalanced in terms of class labels. The total number of people who earn more than 50k is 11687 (24%) whereas the number of people who earn equal or less than 50k is 37155 (76%). Moreover, the Adult Income dataset suffered from representation bias [1]. This arises when populations are not fairly portrayed in comparison to reality. This is driven by incorrect sampling procedures that leave out portions of the population or by population changes [15]. Classifiers usually are configured to optimize accuracy which often reinforces bias patterns that are already present in the dataset [15]. The favorable outcome for this dataset is income greater than 50K. Table I shows the distribution of favrable and unfavorable each of the sensitive attributes. One of the sensitive attributes in the Adult Income dataset is sex, for which the privileged group is male, and the unprivileged group is female. The privileged group attains a favourable position in comparison to the unprivileged group [16]. The other sensitive attribute is race.

TABLE I. DISTRIBUTION OF CLASS LABELS IN SENSITIVE ATTRIBUTES

Class Label/ Sensitive Attribute		**$Y = 1$ income>50k**	**$Y = 0$ income<=50k**
sex	male	9,918	22,732
	female	1,769	14,423
race	white	10,607	31,155
	non-white	1,080	6,000

I. METHODOLOGY

This paper aimed to evaluate four pre-processing bias mitigation approaches, namely Reweighting, Uniform Sampling (US), Preferential Sampling (PS), and Massaging. All of these methods were implemented based on Kamiran and Calders' papers [5]–[7]. In this experiment, the outcome of these pre-processing techniques was compared with two baselines, LR and DT classifiers in terms of core metrics then popular fairness metrics. Two algorithms were employed to compare how pre-processing mitigation techniques impact on different models.

The original Adult Income dataset was split into three sets: 70% training set, 15% validation set, and 15% test dataset. First, a baseline experiment with a standard LR and DT on an Adult Income dataset was performed to compare and benchmark the results of the debiasing experiments. We trained the model on the training set and then through the validation set we found the optimal classification threshold which led to the best-balanced accuracy. As Adult Income datasets are imbalanced toward class labels by 76% class label 0 (income<50k) and 24% class label 1 (income >50k) using traditional accuracy as a performance metric can be misleading. Therefore, we aimed to maximize the balanced accuracy to let the model have a balanced view of a model's performance across different classes.

The first method to reduce bias was Reweighting. Weights were assigned to every object of our training set based on the tuple of sensitive attributes and class labels according to one of these groups. For the sensitive attribute Sex:

- Privileged Positive (PP): (male, income>50k)
- Un-privileged Positive (UP): (female, income>50k)
- Privileged Negative (PN): (male, income <=50k)
- Un-privileged Negative (UN): (female, income <=50k).

The objects in UP were assigned higher weights than the objects in UN and the objects in PN were assigned higher weights than the objects in PP [5]. The classifier used these weights in the learning process. Table II displays the membership count for each group within the original training set (OTS) alongside the corresponding weight values calculated for each group (weights).

Sampling techniques were introduced as not all classifiers directly take in a weighting vector [5]. These techniques impacted the training data similar to Reweighting by changing the distribution of samples [7]. In the Sampling method, the dataset was modified by using replacement

TABLE II. ADJUSTMENT FOR EACH TECHNIQES

Parameter	**PP**	**UP**	**PN**	**UN**
OTS	6,881	1,253	15,986	10,069
weights	0.79	2.15	1.09	0.86
STS	5,440	2,694	17,427	8,628
M	-1441	+1441	+1441	-1441

samples to sample the objects based on their weights [5]. The PP and UN groups as their weights were less than one were undersampled (remove samples) while the PN and UP with weights more than one were oversampled (duplicate samples) [5]. Table II illustrates the new membership counts for each group in the sampling training set (STS).

There are numerous methods for oversampling (e.g. SMOTE) [17], undersampling (e.g. UECMS) [18] or hybrid sampling in a mix of both (e.g. CDSMOTE) [17] to reduce the effect of class imbalance. In this research, two methods, US and PS were used. In the case of US objects in each group had the same chance to be duplicated or removed [5]. On the other hand in PS a ranker algorithm was used to sort the training data based on their positive class probability [6]. In this study, we used LR as the ranker to sort the data objects based on the positive probability estimates. This ranker was used as the objects close to the decision boundary are more likely to be biased [6]. The objects in UP and PP were sorted in ascending order and UN and PN in descending order. Then for the sampling, the top objects of UP and PN were duplicated, and the top objects of PP and UN were skipped. For the sensitive attribute sex, this led to:

- (female, income>50k) with lowest probability to be predicted income>50k were duplicated
- (male, income>50k) with lowest probability to be predicted having income>50k were skipped.
- (female, income<=50k) with highest probability to be predicted having income>50k were skipped
- (male, income<=50k) with highest probability to be predicted having income>50k were duplicated.

The last technique that was used in this research is Massaging. In the Massaging method, the class label for limited objects of the privileged group was changed from positive to negative. The same number from the unprivileged object was changed from a negative label to a positive label [5]. The number of class labels that were changed (M) is provided in Table II. Equation (7) is used for calculating M.

$$M = \frac{|SP| * N_U * N_P}{N} \tag{7}$$

1441 males who earned more than 50k were changed to the males who earned less than 50k. Also, 1441 females who earned less than 50k were changed to females who earned more than 50k. Changing the label class of the objects was not done randomly. The same ranker in PS which sorted the training data based on their positive class probability was used. The training data was divided into two groups. The Promotion group was sorted descending, and the Demotion group was sorted ascending:

- Promotion: (female, income<=50)
- Demotion: (male, income>50k)

Selecting the top objects from each group which were the closest objects to the decision boundary can help to have less impact on the accuracy performance [7].

II. RESULTS AND DISCUSSION

The LR and DT classifiers for the unmitigated Adult Income dataset both resulted in an accuracy of 0.74 and a balanced accuracy of 0.74.

Table III shows core metrics for LR and all pre-processing techniques mitigating against bias. As we see, mitigation techniques led to a decrease in the difference between females' value and males' value in the positive predictions (TP and FP). The reason is that as the algorithm was biased against females, the pre-processing techniques through higher weights, adding more samples, or changing the class labels from negative to positive, tried to fit the model to be fairer toward females. Therefore, as an effect of pre-processing bias mitigation techniques, the number of positive predictions for the test dataset increased for the females and decreased for the males. On the other hand, the difference between females' value and males' value in the negative prediction (TN and FN) increased through the pre-processing mitigation techniques. This occurred as, during the mitigation techniques, we trained the model on fewer negative class (unfavourable) labels for the females and more negative class labels for the males. Therefore, the number of negative predictions was reduced for the females and increased for the males and, as a result, the gap between the males and females in these metrics rose.

Finally, because of the reduction in the difference between the TP predictions for the males and females, the DI metrics became close to one and SP close to zero. The results with the LR classifier are provided in Table III. It shows that Massaging gave the fairest results for DI and SP in this experiment with a final result of 1.06 and 0.02 respectively. However, the US performed the fairest in regard to AOD and EO. There is a decrease in balanced accuracy, albeit slight, when each of the mitigation methods is applied. Results for US and Reweighting were similar which is unsurprising as they are both implemented according to the calculated weights. A negative value for SP in Reweighting and US means this classifier was still slightly biased against the females. Positive values for AOD and EO, after applying all the techniques, show the improved fairness of the classifier towards the females when these techniques are employed.

Table IV provides the results for the DT classifier. It showed that the application of the pre-processing mitigation techniques generally had a similar effect as in the case of the LR classifier, with the differences between the number of positive predictions for the males and females decreasing and the differences between the negative predictions increasing. However, applying PS and Massaging techniques did not decrease the differences between females and males for TP and FP as much as they did in the LR classifier, and they did not increase the difference between females and males in TN

TABLE III. LR WITH SEX SENSITIVE ATTRIBUTE METRICS

Techniques	**Baseline (LR)**	**Reweighting**	**US**	**PS**	**Massaging**
TP	U=117 P=1217	U= 175 P=926	U= 171 P=926	U=215 P=920	U=197 P=924
FP	U=214 P=1239	U=416 P=648	U=403 P=648	U=641 P=637	U=609 P=646
TN	U=1898 P=2196	U=1696 P=2787	U=1709 P=2787	U=1356 P=2798	U=1503 P=2789
FN	U=155 P=291	U=97 P=582	U=101 P=582	U=57 P=588	U=75 P=584
DI	0.28	0.78	0.76	1.29	***1.06***
SP	-0.36	-0.07	-0.08	0.09	***0.02***
AOD	-0.32	0.02	***0.01***	0.18	0.11
EO	-0.38	0.03	***0.01***	0.18	0.11
ACC	0.74	0.76	0.76	0.72	0.74
BA	0.74	0.71	0.71	0.69	0.7

TABLE IV. DT WITH SEX SENSITIVE ATTRIBUTE METRICS

Techniques	**Baseline (DT)**	**Reweighting**	**US**	**PS**	**Massaging**
TP	U=117 P=1218	U=177 P=1010	U=175 P=939	U=157 P=1119	U=143 P=1119
FP	U=215 P=1254	U=489 P=785	U=416 P=696	U=347 P=1051	U=304 P=1051
TN	U=1897 P=2181	U=1623 P=2650	U=1696 P= 2739	U=1765 P=2384	U=1808 P=2384
FN	U=155 P=290	U=95 P=498	U=97 P=569	U=115 P=389	U=129 P=389
DI	0.28	***0.77***	0.75	0.48	0.43
SP	-0.36	-0.08	-0.08	-0.23	-0.25
AOD	-0.32	***-0.01***	***0.01***	-0.15	-0.19
EO	-0.38	***-0.02***	***0.02***	-0.16	-0.22
ACC	0.74	0.75	0.76	0.74	0.74
BA	0.74	0.72	0.71	0.73	0.73

and FN as much as they did in the LR classifier. Therefore, applying these two preprocessing techniques prior to running a DT classifier for this dataset did not have as significant an effect on the fairness measures as you might expect based on the LR classifier results. The values for PS and Massaging for all four metrics deviated from the ideal fairness ranges. For example, the value of DI for PS was 0.48 which for a metric with an ideal value of 1 can considered unfair due to the high deviation. Nevertheless, Reweighting and Sampling had almost the same positive effect on fairness metrics with the DT as they had using LR.

These experiments were repeated for the other sensitive attribute in the Adult Income dataset, race, but space restrictions prevent us from presenting these results here.

Our findings show Reweighting and Sampling have a great improvement for DI, ST, AOD and EO as the fairness metrics. For example, they enhanced DI as a metric with an ideal value of 1, from 0.28 to over 0.75 for all four combinations of classifier and sensitive attribute. Applying PS and Massaging did not show such constant improvement in all results, which is surprising as Kamiran and Calder assert in their paper that "PS always outperforms the Reweighing". This may be due to their use of different classifiers and rankers. Also, they performed their experiments on a random sample of 1/3 Adult Income dataset, whereas we used the LR as our ranker and applied techniques to the whole Adult Income dataset. Moreover, defining the optimised threshold to find the best-balanced accuracy, may have affected PS and Massaging techniques' performance as this can impact the model's behaviour, making it more sensitive to different groups' outcomes.

III. CONCLUSION AND FUTURE WORK

In this paper, we examined four preprocessing bias mitigation techniques, namely Reweighting, US, PS and Massaging which we applied to LR and DT classifiers. This work was conducted on the Adult Income dataset, and we examined both the confusion matrix measure and some popular fairness measurements. Our results showed that with the LR classifier, the Massaging technique achieved the highest DI. However, for the DT classifier, Reweighting and US preformed better than Massaging in all of the presented fairness metrics. For our future work, we are keen to conduct experiments to uncover the underlying reasons for this discrepancy in our results not aligning with Kamiran's and Calder's studies. Moreover, we plan to investigate the impact of the choice of ranker on PS and Massaging as bias preprocessing techniques. It is also planned to explore new Sampling techniques to deliver improved fairness measures.

ACKNOWLEDGEMENT

This research is funded through the HEA's Technological University Transformation Found, co-funded by Dundalk Institute of Technology.

REFERENCES

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, Jul. 2021, doi: 10.1145/3457607.

[2] T. P. Pagano *et al.*, "Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods," *Big Data Cogn. Comput.*, vol. 7, no. 1, pp. 1–31, 2023, doi: 10.3390/bdcc7010015.

[3] P. K. Lohia, K. Natesan Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias Mitigation Post-processing for Individual and Group Fairness," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 2847–2851, May 2019, doi: 10.1109/ICASSP.2019.8682620.

[4] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy Artificial Intelligence: A Review," *ACM Comput. Surv.*, vol. 55, no. 2, 2023, doi: 10.1145/3491209.

[5] F. Kamiran and T. Calders, *Data preprocessing techniques for classification without discrimination*, vol. 33, no. 1. Springer, 2012. doi: 10.1007/s10115-011-0463-8.

[6] F. Kamiran and T. Calders, "Classification with No Discrimination by Preferential Sampling," *Informal Proc. 19th Annu. Mach. Learn. Conf. Belgium Netherlands*, pp. 1–6, 2010.

[7] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," *ICDM Work. 2009 - IEEE Int. Conf. Data Min.*, pp. 13–18, 2009, doi: 10.1109/ICDMW.2009.83.

[8] A. Narayanan, "Translation tutorial : 21 fairness definitions and their politics," p. 21, Accessed: Aug. 01, 2023. [Online]. Available: https://www.youtube.com/watch?v=jIXIuYdnyyk&t=113s

[9] H. Wang and H. Zheng, "True Positive Rate," *Encycl. Syst. Biol.*, pp. 2302–2303, 2013, doi: 10.1007/978-1-4419-9863-7_255.

[10] S. Verma and J. Rubin, "Fairness Definition Explained," *ACM*, vol. 18, pp. 1–7, May 2018, doi: 10.1145/3194770.3194776.

[11] M. Hardt, E. Price, and N. N. Srebro, "Equality of Opportunity in Supervised Learning," *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 3323–3331, Oct. 2016, doi: 10.48550/arxiv.1610.02413.

[12] A. Liang, J. A. Y. Lu, and X. Mu, "Algorithm design: a fairness-accuracy frontier," 2023.

[13] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3125–3128, 2010, doi: 10.1109/ICPR.2010.764.

[14] "UCI Machine Learning Repository: Adult Data Set." https://archive.ics.uci.edu/ml/datasets/adult (accessed Feb. 03, 2023).

[15] U. Gohar, S. Biswas, and H. Rajan, "Towards Understanding Fairness and its Composition in Ensemble Machine Learning," 2022, [Online]. Available: http://arxiv.org/abs/2212.04593

[16] R. Zandee, "A comparative study of bias mitigation methods applied on a multitude of classification algorithms," no. June, 2021.

[17] E. Elyan, C. F. Moreno-Garcia, and C. Jayne, "CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2839–2851, 2021, doi: 10.1007/s00521-020-05130-z.

[18] B. Yao and L. Wang, "An Improved Under-sampling Imbalanced Classification Algorithm," *Proc. - 2021 13th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2021*, pp. 775–779, Jan. 2021, doi: 10.1109/ICMTMA52658.2021.00178.