Enhancing Fairness in Classification Tasks with Multiple Variables: a Data- and Model-Agnostic Approach

Giordano d'Aloisio[0000−0001−7388−890X], Giovanni Stilo[0000−0002−2092−0213], Antinisca Di Marco[0000−0001−7214−9945], and Andrea D'Angelo[0000−0002−0577−2494]

Department of Engineering and Information Sciences and Mathematics, University of L'Aquila, Italy
giordano.daloisio@graduate.univaq.it,
{giovanni.stilo,antinisca.dimarco}@univaq.it
andrea.dangelo6@student.univaq.it

Abstract. Nowadays assuring that *search* and *recommendation* systems are fair and do not apply discrimination among any kind of population has become of paramount importance. Those systems typically rely on machine learning algorithms that solve the classification task. Although the problem of fairness has been widely addressed in binary classification, unfortunately, the fairness of multi-class classification problem needs to be further investigated lacking well-established solutions. For the aforementioned reasons, in this paper, we present the *Debiaser for Multiple Variables*, a novel approach able to enhance fairness in both binary and multi-class classification problems. The proposed method is compared, under several conditions, with the well-established baseline. We evaluate our method on a heterogeneous data set and prove how it overcomes the established algorithms in the multi-classification setting, while maintaining good performances in binary classification. Finally, we present some limitations and future improvements.

Keywords: Machine learning · Bias and fairness · Multi-class classification · Preprocessing algorithm.

1 Introduction

Bias impacts human beings as individuals or groups characterized by a set of legally-protected sensitive attributes (e.g., their race, gender, or religion). If not managed, the inequalities reinforced by search and recommendation algorithms can lead to *severe societal consequences*, such as discrimination and unfairness [14]. Both *search* and *recommendation* algorithms provide a user with ranked results that fit and match their needs and interests. Both tasks often convey and strengthen bias in terms of *imbalances* and *inequalities*, primarily if they rely on or encompass machine learning algorithms as those which solve classification problems. For this reason, assuring that search and recommendation

systems are fair and do not apply discrimination among any kind of population has become of paramount importance, mainly because they are pervasive in several domains (e.g., justice [26], health care [30], education [4], etc.).

Over the years, different methods have been proposed to mitigate bias at several levels of data processing. However, we notice that the multi-class classification problem is still not effectively addressed, even if it is widely adopted and constitutes a building block for personalization and search systems in several domains [21,29,16].

For this reason, in this paper, we present the *Debiaser for Multiple Variables (DEMV)*. This novel approach is a generalization of the *Sampling* algorithm proposed by Kamiran et al. in [17]. DEMV is model and data-agnostic, allowing its introduction in already existing systems without particular effort and without introducing structural changes. The DEMV enhances fairness both in binary and multi-class classification problems, handling any number of sensitive variables and with any classifier. We exhaustively show, with different datasets, that our method outperforms the state-of-the-art methods in the multi-class classification while achieving comparable performances in the binary one.

This paper is structured as follows: in Section 2, we recall some background knowledge used in our work and describe some bias mitigation methods in the context of multi-class classification problem; in Section 3, we describe in detail the proposed approach; Section 4 is dedicated to the experimental analysis that has been conducted both in binary and multi-classification problems; finally, Section 5 describes some points of improvement of our approach and concludes the paper.

2 Background Knowledge and Related Work

In the last ten years, the study of bias and fairness in machine learning acquired considerable relevance in literature. Many definitions and metrics have been proposed to address different kinds of bias and fairness [23]. In this section, we recall the definition of fairness we use in this paper and then, we describe the related work in the context of bias mitigation in multi-class classification problem.

2.1 Fairness definition

Demographic (Statistical) Parity (DP) [20,11] is one of the most used definitions of *group fairness* [23], which assumes the independence among the predicted positive label y_p and the sensitive variables $S_1, \ldots, S_n$.

Formally, let $\hat{Y}$ be the predicted value and S be a generic binary sensitive variable where $S = 1$ and $S = 0$ identify the privileged and unprivileged groups, respectively. A predictor is *fair* under DP if:

$$P(\hat{Y} = y_p | S = 1) = P(\hat{Y} = y_p | S = 0) \tag{1}$$

A different formulation for the DP is the *Disparate Impact* [12], which considers the ratio among the two probabilities:

$$0.8 \leq \frac{P(\hat{Y} = y_p|S = 1)}{P(\hat{Y} = y_p|S = 0)} \leq 1.2 \tag{2}$$

In this case, following the *80% rule* [12], the value must be between 0.8 and 1.2 in order to have *fairness*. DP falls into the *We Are Equal (WAE)* metrics family, which holds that all groups have similar abilities concerning the task (i.e., have the same probability of being classified in a certain way) [13].

2.2 Related Works

Over the years, many methods have been proposed to mitigate bias at different levels of data processing [23,7]. In particular, we distinguish among **pre-processing** methods, which modify the data to remove the underlying bias; **in-processing** methods, which change the learning algorithm to remove discrimination during the model training process; **post-processing** methods, which re-calibrate an already trained model using a holdout set not used during the training phase. In general, the sooner a technique can be applied, the better because it can be chained with other bias mitigation methods in the later processing phases [31,1].

Among *pre-processing* methods, one widely adopted is the *Sampling* algorithm proposed by Kamiran et al. in [17]. Its method rebalances both privileged and unprivileged users in the case of binary classification with a single sensitive variable.

Formally, let be S the sensitive variable with $\{w, b\} \in S$ representing the privileged and unprivileged groups, respectively, and let be Y the target label with $\{+, -\} \in Y$ defining the positive and negative outcomes. The sampling algorithm first splits the original dataset into four groups:

- Deprived group with Positive label (DP): all instances with $S = b \wedge Y = +$;
- Deprived group with Negative label (DN): all instances with $S = b \wedge Y = -$;
- Favored group with Positive label (FP): all instances with $S = w \wedge Y = +$;
- Favored group with Negative label (FN): all instances with $S = w \wedge Y = -$.

Then, the algorithm balances the groups iteratively until their *observed* sizes are equal to their *expected* ones.

We like to note that very few methods are able to mitigate the bias in the multi-class classification problems. Among those are the two *post-processing* methods proposed by Krishnaswamy et al. [19]. They start from the definition of a set of deterministic classifiers H and a class of groups G made by elements in the form $g : N \to \{1, -1\}$, where $g(i) = 1$ iff item i is in group g, -1 otherwise. For each group g, they identify the best classifier h^*_g as the one assuring the highest accuracy for that group. Then, they define fairness as a constraint among the accuracy of each possible classifier and the optimal one for each group g. The core difference between the two proposed approaches lays in the definition

of G. The first one, named *Proportional Fairness (PF)*, considers each possible subset of the dataset, so there is no limitation on G. The second method, *BeFair (Best-effort Fair)*, considers instead each linearly separable group (i.e., if $g \in G$, then g and $N \backslash g$ are linearly separable). To solve the unfairness of the input classifier, both methods optimize a MinMax function.

An *in-processing* method that solves unfairness in multiple classification settings is the one presented by Agarwal et al.[2]. The algorithm addresses two definitions of fairness at once: *Demographic Parity* and *Equalized Odds* [15]. The authors formulate such definitions as linear constraints and then build an Exponentiated Gradient reduction algorithm that yields a randomized classifier with the lowest error subject to the desired fairness constraints. Also, in this case, the method follows a MinMax approach in which the players try to minimize the given constraint and maximize the classifier's score. Although the authors study their algorithm mainly in binary classification problems, they also show how it can be applied to regression and multi-classification problems.

Note that most of the methods in the literature are primarily designed for binary classification problems, and the minority of them apply during the *pre-processing* phase. On the contrary, our proposed method is able to work in the pre-processing stage. Since it can be used at the beginning of the processing phase, it can be possibly chained with other algorithms in later steps. Moreover, it natively supports multi-class classification and multiple sensitive variables in an affordable yet successful way, as is shown in the experimental section 4.

3 Debiaser for Multiple Variables (DEMV)

In this section, we describe in detail *Debiaser for Multiple Variables (DEMV)*, a bias mitigation method for multiple sensitive variables in the classification context.

The main idea behind the proposed method is that all the possible combinations of the sensitive variables values and of the label's values for the definition of the sampling groups must be considered. We approach the problem by recursively identifying all the possible groups given by combining all the values of the sensible variables with the belonging label (class). For each group, we compute its expected and observed sizes, defined respectively as:

$$W_{exp} = \frac{|\{X \in D | S = s\}|}{|D|} * \frac{|\{X \in D | L = l\}|}{|D|} \tag{3}$$

$$W_{obs} = \frac{|\{X \in D | S = s \wedge L = l\}|}{|D|} \tag{4}$$

Where $S = s$ is a generic condition on the value of the sensitive variables[1] and $L = l$ is a condition on the label's value. If $W_{exp} \backslash W_{obs} = 1$ implies that the group is fully balanced. Otherwise, if the ratio is less than one, the group size is larger

[1] The variables can be binary, discrete or categorical ones

than expected, so we have to delete a random element from the considered group. Instead, if the ratio is greater than one, the group is smaller than expected, so we have to duplicate an item from the group by random sampling. For each group, we recursively repeat this operation until $W_{exp} \backslash W_{obs}$ converge to one.

The group-balancing operation is implemented by the SAMPLE function, whose pseudo-code is depicted in listing Algorithm 1. This function takes as input the group g and the expected (W_{exp}) and observed size (W_{obs}). The core of this algorithm is a while loop that checks if the value of $W_{exp} \backslash W_{obs}$ is different from 1. If so, the algorithm selects a random index in the range of $(0, length(g) - 1)$ and duplicates the corresponding item if $W_{exp} \backslash W_{obs} > 1$ or removes it if $W_{exp} \backslash W_{obs} < 1$. Finally, the algorithm returns the sampled group when the while condition becomes true.

Algorithm 1: Pseudo-code of SAMPLE

```
   Input: (Group g, Expected size W_exp, Observed size W_obs)
   Output: Balanced group g
1  while W_exp \ W_obs ! = 1 do
2      i = random value ∈ {0, ..., length(g) − 1}
3      if W_exp \ W_obs > 1 then
4          duplicate item at position i in g
5      else if W_exp \ W_obs < 1 then
6          remove item at position i from g
7      recompute W_obs
8  return g
```

The SAMPLE algorithm is invoked inside DEMV whose pseudo-code is showed in listing Algorithm 2. The main *DEMV* function takes as input the dataset D, the categorical sensitive variables $S_1, \ldots, S_n$, the label L and other parameters useful for the recursion: a counter i initially set to 0, an array G initially empty and a boolean *condition* initially set to *true*. Lines from 2 to 9 define the base condition of the function. They check if the counter i is equal to the number of sensitive variables. If so, the algorithm iterates the possible values of the label and creates the corresponding group g. Then, it computes the expected and observed sizes and it balances the group using the SAMPLE function (listing Algorithm 1). Finally, the approach adds the balanced group g_b to the array G (used to collect all the sampled groups) and returns it. Lines from 10 to 14 identify the recursive part of the function. In particular, if the value of i is not equal to the number of sensitive variables, the algorithm increments the value of i by one and appends to G the result of a series of recursive calls. These recursive invocations differ from each other only in the condition passed as input. In fact, the algorithm iterates for all the possible values of the sensitive variable S_i and, for each value s, it does a recursive call adding the condition of $S_i == s$ to the previous ones through an $\wedge$ connector. Finally, lines from 15 to

19 define the returning conditions of the function. In particular, the maximum number of samples obtainable from the combination of n sensitive variables plus the label is given by the product of all the sensitive variables' and label's values, that is:

$$\prod_{1,\dots,n}^{i} |S_i| * |L|$$

If the length of G is equal to this value, then the function has considered and balanced all the groups and it returns the final sampled dataset D_S. Otherwise, the function being in the middle of the recursive tree, returns G which will be again merged with the result of other recursive functions. DEMV algorithm can also be applied to binary classification problems; in that case, the number of sampling groups will be equal to:

$$\prod_{1,\dots,n}^{i} |S_i| * 2$$

Algorithm 2: Pseudo-code of DEMV

Input: (Dataset D, Sensitive variables $S_1, S_2, \dots, S_n$, Label L, $i = 0$, $G = []$, condition=$true$)
Output: Sampled dataset D_S

$n = length(\{S_1, S_2, \dots, S_n\})$
if $i == n$ **then**
  foreach $l \in L$ **do**
    $g = \{X \in D|$ condition $\wedge\ L == l\}$
    $W_{exp} = \frac{|\{X \in D|\text{condition}\}|}{|D|} * \frac{|\{X \in D|L == l\}|}{|D|}$
    $W_{obs} = \frac{|g|}{|D|}$
    $g_b = SAMPLE(g, W_{exp}, W_{obs})$
    add g_b to G
  return G
else
  $i = i + 1$
  foreach $s \in S_i$ **do**
    $G^{'} = DEMV(D, S_1, \dots, S_n, i, G, \text{condition} = condition \wedge S_i == s)$
    add $G^{'}$ to G
  if $length(G) == \prod_{1\dots n}^{i} |S_i| * |L|$ **then**
    $D_S =$ merge all $g \in G$
    return D_S
  else
    return G

The implementation of *DEMV* is available at the Territori Aperti RI

4 Experimental analysis

This section describes the experiments we conducted to evaluate the proposed method. We analyzed *DEMV* under heterogeneous conditions where a set of binary and multi-class datasets were employed. Our method was compared with *Exponentiated Gradient*[2] (whose adopted implementation is available on the Fairlearn library [5]). Following the documentation available online, we used as for the *Exponentiated Gradient (EG)*: the *Demographic Parity* for binary classifications and *Zero-one Loss* [10] for multi-class problems.

We used a Logistic Regression classifier and conducted *10-fold* cross-validation [27]. We decided to apply DEMV and EG only on the training set.

For all the experiments, we computed the following metrics on the testing set: *Statistical Parity (SP)*[20,11], *Disparate Impact (DI)*[12], *Zero-one Loss (Z.O. Loss)*[10], and *Accuracy (Acc.)*[28]. In addition, since DEMV has a stochastic behavior, for each training set, we ran DEMV and computed the corresponding metrics 30 times so that we can investigate how the removal or duplication of different samples can influence the *accuracy* and the *fairness* of our method. Since DI tends to show a reverse-bias situation more than SP and the other selected metrics, to highlight the maximum fairness point under DI better, we use the formulation proposed by Radovanovic et al. in [24]:

$$DI = min\left(\frac{p(\hat{y}=1|s=1)}{p(\hat{y}=1|s=0)}, \frac{p(\hat{y}=1|s=0)}{p(\hat{y}=1|s=1)}\right) \tag{5}$$

This metric computes the minimum among two formulations of DI wherein one the unprivileged group ($s = 0$) is at the numerator, and the other is at the denominator. The metric value is hence between 0 and 1, where 1 means complete fairness.

4.1 Employed datasets

The experiments have been conducted by employing eight well-known datasets (3 for the binary classification and 5 for the multi-class task), coming from the Bias and Fairness literature:

- **Adult Income (ADULT)** [18]: a binary dataset that comprises 30,940 items by 102 features (one-hot encoded). The goal is to predict if a person has an income higher than 50k a year. This information is represented by the `income` variable. The protected attributes are `sex`, and `race` and the unprivileged group is *black women* (items with `sex` and `race` equal to zero). The positive label is "*high income*".
- **ProPublica Recidivism (COMPAS)** [3]: This binary dataset is made of 6,167 samples by 399 attributes. The sensitive variables are `sex` and `race`. The goal is to predict if a person will recidivate in the next two years. The

favorable label, in this case, is *no*, and the privileged group is *Caucasian women* (items with `sex` and `race` equal to one).
- **German Credit (GERMAN)** [25]: This binary dataset classifies people described by a set of attributes as good or bad credit risks (`credit` variable). The dataset consists of 1,000 instances by 59 features (one-hot encoded). The sensitive variables are `sex`, and `age` and the unprivileged group is *women with less than 25 years*. The positive label is *low credit risk*.
- **Contraceptive Method Choice (CMC)** [22]: This multi-class dataset comprises 1,473 instances and ten columns about women's contraceptive method choice (*not-use*, *short-use*, and *long-use*). The sensitive variables are `religion` and `work`. The unprivileged group is *Islamic women who do not work* (both values equal one), and the positive label is *long-term use*.
- **Communities and Crime (CRIME)** [26]: This multi-class dataset is made of 1,994 instances by 100 attributes and contains information about the per-capita violent crimes in a community (variable `ViolentCrimesPerPop`). Since the label is continuous, we transformed it by grouping the values in 6 classes using equidistant quantiles. Following [6] the sensitive attribute is the percentage of the black population, but we also considered the ratio of the Hispanic population to have two sensitive variables. The unprivileged group is communities with a *high percentage of both black and Hispanic people* (both variables equal to 1), and the positive label is `100` (class of *low rate of crimes*).
- **Law School Admission (LAW)** [4]: This multi-class dataset comprises 20,694 samples by 14 attributes and contains information about the bar passage data of Law School students. We grouped the continuous label (`GPA`) in 3 groups using equidistant quantiles. The sensitive variables are `race` and `gender` and the unprivileged group are *black women* (both variables equal to one), and the positive label is 2 (class of high scores).
- **The Trump Effect in Europe (TRUMP)** [9]: This multi-class dataset is the result of a survey about political preference in Europe after Trump's presidential election. It is made of 7,951 features and 204 attributes. The label is `political view` and the sensitive variables are `gender` and `religion`. The unprivileged group is *non-catholic women* (both variables equal to 0), and the positive label is equal to 3.
- **Wine Quality (WINE)** [8]: This multi-class dataset comprises 6,438 instances and 13 attributes about wine quality. The sensitive attributes are the wine's color (`type` variable) and the alcohol percentage lower or higher than 10 (`alcohol` variable). The unprivileged group is *white wine with alcohol percentage* ≤ 10, and the positive label is 6 (high quality).

Table 1 summarizes the key datasets' information.

4.2 Experimental results

In this section, we present the results in binary and multi-class classification. For both experiments, we report charts showing the mean and the standard deviation

Table 1. Datasets information

	Adult	**Compas**	**German**	**CMC**	**Crime**	**Law**	**Trump**	**Wine**
Scope	Social	Justice	Social	Social	Justice	Education	Social	Food
Instances	30,940	6,167	1,000	1473	1,994	20,427	7,951	6,438
Features	102	399	59	10	100	14	204	13
Type	binary	binary	binary	multi	multi	multi	multi	multi
Sensitive variables	`sex` `race`	`sex` `race`	`sex` `age`	`work` `religion`	`black` `hisp`	`gender` `race`	`religion` `gender`	`type` `alcohol`
Percentage of sensitive group	5.02%	54.71%	10.50%	64.83%	23.62%	8.42%	30.71%	11.40%

for each of the metrics described in section 4 (y-axis) at each DEMV iteration (x-axis). At iteration zero, the graphics report the metrics of the original biased dataset, while at the end of the curves they show the metrics computed, on the whole, balanced dataset. On the right part of the plots are reported using bigger points, the same metrics obtained by the EG algorithm.

Binary classification. The results for binary classification are shown in figure 1 where the performance of DEMV at each iteration (x-axis) is shown.

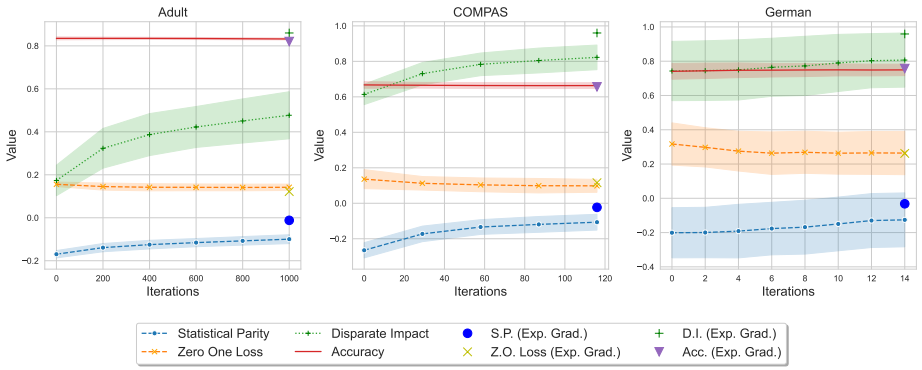

Fig. 1. Comparison of DEMV at each sampling iteration with EG for binary classification datasets.

The EG method can find the best trade-off between fairness and accuracy in the binary classification case. Instead, our approach has more difficulty improving fairness, especially when the bias is very high (see Adult dataset). However, our method can keep a high accuracy level when the dataset is fully balanced. In all the analyses, we can see that the complete balancing of the dataset leads to the best fairness of the classifier.

Multi-class classification. Similarly, the results for multi-class classification are shown in figure 2 where the performance of DEMV is shown at each iteration (x-axis).

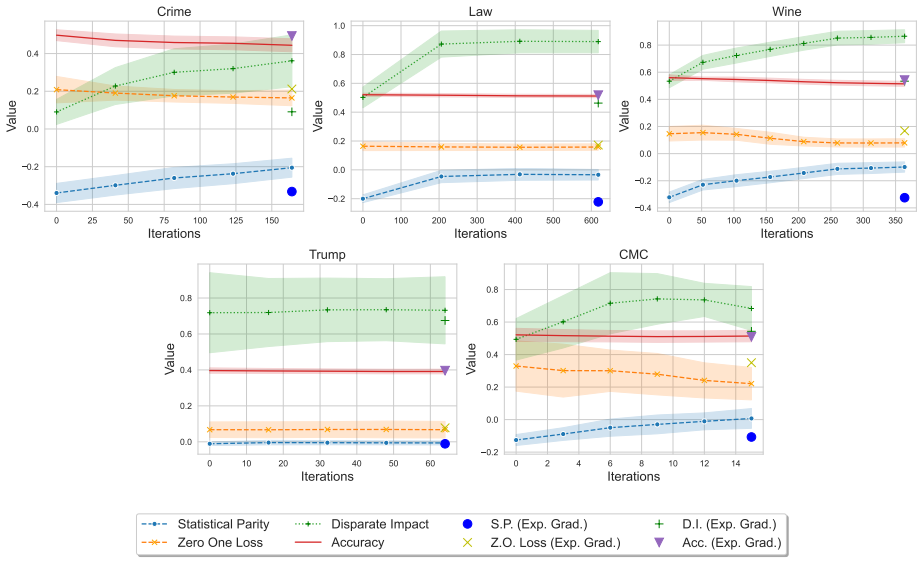

Fig. 2. Comparison of DEMV at each iteration with EG for multi-class classification datasets.

As it is possible to notice from the figure, DEMV outperforms EG in all metrics for what concerns multi-class classification problems. As said at the beginning of the section 4, Zero-One Loss is adopted as a minimization constraint for EG in the multi-class task. Moreover is it worth to note that not always the complete balancing of the dataset leads to the best fairness of the classifier. In fact, in two datasets, namely Trump and CMC, we observed that the best fairness under DI is achieved respectively with 15 and 9 iterations of the sampling algorithm. The intrinsic characteristics of such datasets can justify this behavior: in CMC, the size of the unprivileged group is about 65% of the total population, while Trump has a very shallow bias. In such situations, we observed that it is not convenient to fully balance the datasets' groups. Finally, we observe that DI has a higher variance than SP, especially in datasets with a shallow bias like Trump.

Discussion. From the above analyses, we can draw the following considerations about DEMV. Our method can constantly improve the fairness of the classifier both in binary and multi-class classification, with respect to the initial biased classifier, keeping the accuracy almost unchanged (up to 0.05 points in case of CRIME). Moreover, DEMV algorithm has the advantage of being data and model agnostic, meaning that it can be applied to any dataset with any number

of sensitive variables and any number of label's values, and it can be used with any classifier.

Concerning binary classification, DEMV little improves fairness, especially when the bias is very high; while other existing methods may perform better in these cases. For multi-classification setting, instead, our method outperforms the baseline, improving the fairness significantly (up to 0.4 points for DI in the case of the Law dataset). In addition, we observed that, in some particular circumstances (like CMC and Trump), achieving a complete balance of the sensitive groups does not lead to the best possible fairness. In such situations, a partial sampling of the groups is preferable.

5 Conclusion and Future Work

In this paper, we extended the work of [17] to present the *Debiaser for Multiple Variables*, a novel approach to enhance fairness in multi-class classification problems with any number of sensitive variables. We exhaustively compared it with the baseline method described in [2] performing both binary and multi-class classification.

We can summarise the following take away outcomes:

- DEMV is a novel approach, primarily defined for the under explored multi-class classification;
- DEMV is a better strategy to adopt than EG in the multi-class task;
- performing a complete balancing is not always the optimal solution for all the datasets;
- we used DEMV also in binary classification, observing an improvement for all metrics. However, as expected, other specifically designed methods perform better in such cases.

In the future, we like to investigate further which are the characteristics of the dataset that lead to optimal performance before a complete balance within the groups. Furthermore, we want to determine the impact of adopting a higher or lower number of sensible variables and if the method remains consistent both in terms of *accuracy* and *fairness*. Given the independence among the predicted positive label, DP may not be the best metric to use in case of multi-class classification, hence we will explore other metrics for evaluation. Finally, we will widely evaluate our approach with respect to other existing multi-class bias mitigation methods, also considering a more extensive set of datasets covering different domains and having distinct, not overlapping characteristics.

6 Acknowledgments

This work is partially supported by Territori Aperti a project funded by Fondo Territori Lavoro e Conoscenza CGIL CISL UIL, by SoBigData-PlusPlus H2020-INFRAIA-2019-1 EU project, contract number 871042 and by "FAIR-EDU: Promote FAIRness in EDUcation institutions" a project founded by the University of L'Aquila.

References

1. AI Fairness 360 - Resources, https://aif360.mybluemix.net/resources#guidance
2. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A Reductions Approach to Fair Classification. arXiv:1803.02453 [cs] (Jul 2018), arXiv: 1803.02453
3. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica, May **23**(2016), 139–159 (2016)
4. Austin, K.A., Christopher, C.M., Dickerson, D.: Will i pass the bar exam: Predicting student success using lsat scores and law school performance. HofstrA l. rev. **45**, 753 (2016)
5. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (May 2020), https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/
6. Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling Attribute Effect in Linear Regression. In: 2013 IEEE 13th International Conference on Data Mining. pp. 71–80 (Dec 2013). https://doi.org/10.1109/ICDM.2013.114, iSSN: 2374-8486
7. Caton, S., Haas, C.: Fairness in Machine Learning: A Survey. arXiv:2010.04053 [cs, stat] (Oct 2020), arXiv: 2010.04053
8. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decision support systems **47**(4), 547–553 (2009)
9. DaliaResearch: The Trump Effect in Europe (2017), https://www.kaggle.com/daliaresearch/trump-effect
10. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. Machine learning **29**(2), 103–130 (1997)
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ITCS '12, Association for Computing Machinery, New York, NY, USA (Jan 2012). https://doi.org/10.1145/2090236.2090255
12. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and Removing Disparate Impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268. ACM, Sydney NSW Australia (Aug 2015). https://doi.org/10.1145/2783258.2783311
13. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016)
14. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: From discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 2125–2126. ACM (2016)
15. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29**, 3315–3323 (2016)
16. Jiang, C., Liu, Y., Ding, Y., Liang, K., Duan, R.: Capturing helpful reviews from social media for product quality improvement: a multi-class classification approach. International Journal of Production Research **55**(12), 3528–3541 (2017)

17. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (Oct 2012). https://doi.org/10.1007/s10115-011-0463-8
18. Kohavi, R., et al.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Kdd. vol. 96, pp. 202–207 (1996)
19. Krishnaswamy, A., Jiang, Z., Wang, K., Cheng, Y., Munagala, K.: Fair for all: Best-effort fairness guarantees for classification. In: International Conference on Artificial Intelligence and Statistics. pp. 3259–3267. PMLR (2021)
20. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual Fairness. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
21. Le, Y., He, C., Chen, M., Wu, Y., He, X., Zhou, B.: Learning to predict charges for legal judgment via self-attentive capsule network. In: ECAI 2020, pp. 1802–1809. IOS Press (2020)
22. Lim, T.S., Loh, W.Y., Shih, Y.S.: A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine learning **40**(3), 203–228 (2000)
23. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys **54**(6), 1–35 (Jul 2021). https://doi.org/10.1145/3457607
24. Radovanović, S., Petrović, A., Delibašić, B., Suknović, M.: A fair classifier chain for multi-label bank marketing strategy classification. International Transactions in Operational Research (2021). https://doi.org/10.1111/itor.13059, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/itor.13059
25. Ratanamahatana, C.A., Gunopulos, D.: Scaling up the naive bayesian classifier: Using decision trees for feature selection (2002)
26. Redmond, M., Baveja, A.: A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research **141**(3), 660–678 (2002)
27. Refaeilzadeh, P., Tang, L., Liu, H.: Cross-Validation, pp. 1–7. Springer New York, New York, NY (2016). https://doi.org/10.1007/978-1-4899-7993-3_565-2
28. Rosenfield, G., Fitzpatrick-Lins, K.: A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing **52**(2), 223–227 (1986), http://pubs.er.usgs.gov/publication/70014667
29. Sánchez-Morillo, D., López-Gordo, M., León, A.: Novel multiclass classification for home-based diagnosis of sleep apnea hypopnea syndrome. Expert Systems with Applications **41**(4), 1654–1662 (2014)
30. Street, W.N., Wolberg, W.H., Mangasarian, O.L.: Nuclear feature extraction for breast tumor diagnosis. In: Biomedical image processing and biomedical visualization. vol. 1905, pp. 861–870. International Society for Optics and Photonics (1993)
31. Wolpert, D.H.: What does dinner cost?, http://www.no-free-lunch.org/coev.pdf