The Inefficiency of Achieving Fairness with Protected Attribute Suppression

Lucas R. Aragão[1], Maria de Lourdes M. Silva[1], Javam C. Machado[1]

Laboratório de Sistemas e Bancos de Dados (LSBD)
Departamento de Computação
Universidade Federal do Ceará, Brazil

{lucas.aragao, malu.maia, javam.machado}@lsbd.ufc.br

Abstract. *In recent years, there has been an increase in the use of artificial intelligence for various tasks, including classifying individuals for purposes such as granting bank loans. Although this technology has enabled the automation of tasks, it has also raised social and ethical concerns due to the potential propagation of bias against historically discriminated groups. The attributes that contain these groups are known as protected attributes. This work suggests that simple methods of suppressing these attributes are insufficient to eliminate bias and achieve fairness in classification algorithms. We analyzed the correlation and independence between attributes and evaluated the impact of suppression on the classification task, considering both utility and fairness.*

1. Introduction

The benefits of using artificial intelligence (AI) technology are accompanied by ethical and social concerns regarding using Machine Learning (ML) models on personal data. One of the most persistent concerns in AI is the potential propagation or amplification of certain biases in the model's responses. The hiring process in companies [Langenkamp et al., 2020] and bank loans [Orji et al., 2022] are examples of personal data employed in ML for training and classification. However, individuals whose data are used in those tasks are sometimes targets of discrimination generated or propagated by the classifier. The area that studies methods for bias mitigation is called *fairness* [Kearns and Roth, 2019], which focuses on finding ethically responsible practices in the AI context. Algorithmic discrimination occurs due to biased data used to train ML models. The bias derives from historical discrimination against groups of individuals who are associated with protected attributes. One example was Amazon's recruitment model in the last decade, which penalized resumes that contained the word "women" [Kearns and Roth, 2019].

Some previous works use the suppression of protected attributes as a bias mitigation strategy [Dhar et al., 2021]. Our hypothesis is that suppressing the protected attribute is insufficient to mitigate bias. This work empirically confirms our hypothesis. We suggest that by creating and comparing ML models with and without the protected attribute on the training data. We justify the inefficiency of the isolated suppression by doing a correlation analysis between the attributes using well-known datasets in the fairness literature. Since our work is based on empirical results from a few experiments, and the scope limits the number and depth of those, we encourage a broader approach later.

We present the paper as follows: Section 2 explains fundamental fairness concepts and approaches for attribute correlation analysis. Section 3 discusses previous works in fairness. Section 4 explains the methodology and details the experiments. Section 5 discusses the experiments' results. Section 6 concludes the work and debates future works.

2. Theoretical foundation

This section describes the essential knowledge that was used to produce this work, including the group fairness concept and some techniques for correlation analysis.

2.1. *Group Fairness*

Group fairness is a fairness field that studies discrimination against groups of individuals who share some common feature. The aim is to ensure that different groups have the same opportunities for success. Bias mitigation techniques often modify the input, classifier, or output to satisfy fairness constraints [Pitoura et al., 2021]. Mathematically, given two groups partitioned by the protected attribute $A = \{0, 1\}$, the probability of the classifier $\hat{Y}$ to give a positive output must be the same for both groups, as shown in Equation 1. This concept is called independence [Barocas et al., 2023].

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

Disparate impact (DI) is a well-known metric that captures and measures bias in a dataset. It refers to the negative impact of practices on a specific protected group [The US EEOC].

2.2. Techniques for correlation and independence analysis

This subsection discusses two techniques for analyzing correlation and attribute independence. These investigations enable an understanding of how changes in one attribute are associated with changes in others. The first one is the Φ coefficient(ϕ_k), an association metric between two variables [Baak et al., 2020]. Defined as a Pearson correlation variation, the Φ correlation is helpful due to its flexibility since it measures the correlation between different types of variables. Bayesian network is another useful technique that captures the correlation among attributes [Pearl, 1988]. They are probabilistic models that illustrate the dependence relation between different attributes in a dataset. The network is a directed graph where the nodes represent the attributes, and the edges denote the dependence relations. Those models use ML algorithms for the network's structure estimation based on the conditional probability distribution associated with each node.

3. Related works

There is a vast literature on fairness in classification algorithms (Caton and Haas [2024]). Due to space limitation and the specific problem we attack in this paper, we focused our related work on some recent approaches to feature correlation for fairness classification algorithms. Still, we only provide a brief description of those approaches. Please refer to Caton and Haas [2024] for a detailed analysis. Martini [2023] analyses individuals' data from a hospital and creates a few ML models for classifying whether an individual had a seizure. The author conducted a correlation analysis using the Pearson correlation to examine the relationship between the attributes in the data and demonstrated the impact of each attribute on the outcome. Pedreshi et al. [2008] and Kamiran and Calders [2009] propose approaches to creating non-discriminatory environments. Both affirm that extracting the protected attribute is insufficient to ensure fairness. However, none of them goes deeper into the topic. Le Quy et al. [2022] surveys the most used datasets for fairness studies and examines dependence between different attributes. It creates a Bayesian network for each explored dataset, followed by the results' analyses, which detail the dependence relations.

4. Methodology

This section outlines the methodology used in this work, including details about the datasets, the approach to identifying the suppression of the protected attribute's inefficiency, and the evaluation techniques.

4.1. Datasets

The datasets used for this study are some of the most used in fairness literature. The Adult Census Dataset [Becker and Kohavi, 1996] is the 1996 American census. The ML task here is to classify an individual's yearly income, and the protected attribute is gender. The COMPAS dataset [Larson et al., 2016] uses data from a tool for classifying criminal recidivism of individuals. The assignment is to decide if an individual will commit crimes in the next two years, and the protected attribute is race.

The Law School dataset [Wightman, 1998] is about American law school admission. The task is to classify whether an individual will succeed in the exam, and the protected attribute is also race. The Bank Marketing dataset [Moro et al., 2014] gets data from a Portuguese bank marketing campaign, which classifies whether an individual entered the campaign or not, and the protected attribute is marital status. Table 1 shows a synthesis of the raw datasets.

To pre-process the datasets, we applied the following four criteria: (1) We removed features with explicit identifiers, such as names, and those with many null values. (2) In cases where multiple attributes had similar meanings, we eliminated redundant attributes. (3) We removed the instances with remaining null from the dataset. (4) Finally, for visualization purposes, we selected a maximum of seven features most relevant to the context of each dataset.

Table 1. Datasets description.

Datasets	#instances	#fields	Context	Protected attribute
Adult	48842	14	Annual income	Gender
COMPAS	7213	54	Ethical and criminal issues	Race
Law School	18692	12	Students admission	Race
Bank Marketing	45211	16	Marketing campaign	Marital status

4.2. Highlighting the inefficiency

To show the inefficiency of the protected attribute suppression, we compare the predictions of an ML model trained with and without the protected attribute. For this task, we select the Logistic Regression (LR) model as the predictor since it is a classic and widely used model in literature Berkson [1944]. We will refer to the strategies by their initials throughout the text. The first setting is the **L**ogistic **R**egression with **P**rotected attribute (LRP) and the second is the **L**ogistic **R**egression **W**ithout the **P**rotected attribute (LRWP). We evaluate if the predictions of both settings yield similar results about the fairness metric, then the suppression is insufficient to ensure fairness.

4.3. Evaluation techniques

We create a heatmap for each dataset based on the ϕ_k correlation. The tone of the cell highlights the correlation between two attributes; the darker the cell, the more significant the correlation of the related attributes. In addition, we create a Bayesian network for each dataset to demonstrate the dependence between attributes. We used an exact algorithm that enumerates all the exponential numbers of structures and finds the best [Koivisto and Sood, 2004]. The protected attribute and the label are highlighted in green and gray, respectively.

We evaluate the model's utility by using the AUC (Area Under the Curve) score based on the ROC (Receiver Operating Characteristics) curve [Fawcett, 2006]. The values vary from 0 to 1, so the closer the value is to one, the better the model can distinguish the two classification classes. Finally, we use the DI to evaluate the suppression effect on the classifier's prediction, considering comparing LRP and LRWP. We set the threshold based on the four-fifths rule [The US EEOC]. Above that value, the environment is considered discriminatory.

5. Experiments and results

5.1. Correlation study

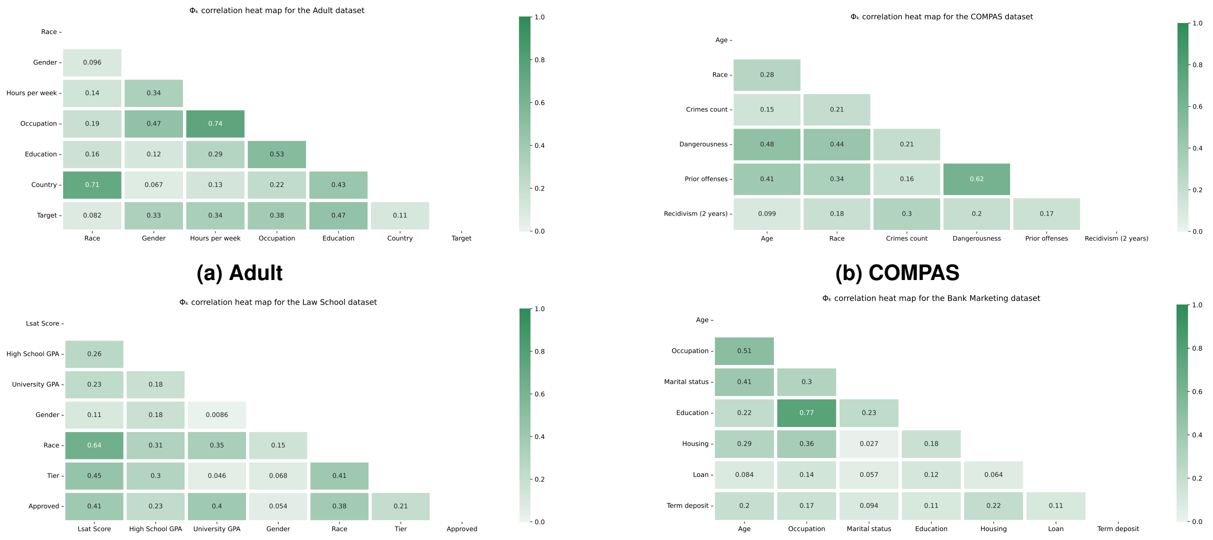
Figure 1 shows heatmaps. Expected values appear in the result, such as the high correlation between occupation and education in the adult and bank marketing datasets. This is intuitive since the higher an individual's education degree, the better their occupation. However, some correlations, like country and race, require attention as geographical location may provide clues about social information. The suppression of race from the dataset would not vanish such tips. In these cases, the discriminatory bias will remain there.

Similarly, the law school and bank marketing datasets present identical behavior regarding LSAT score and age, respectively. The correlation between those attributes and the protected ones is considered high in both cases. This happens in the law school dataset because of the imbalance between the number of people from each race, which only reflects the discrimination against these groups of people in those exams.

The protected attributes in the Adult and Law School datasets have a significant correlation value with the label. In the other datasets, those values are low, especially in Bank Marketing, where the value is below 0.1, so it may not directly affect the prediction. An accentuated correlation exists between race and dangerousness in the COMPAS heatmap. Such an elevated value can be better understood when the historical context of discrimination against the Afro-American population is known. In such a case, the dataset only reflects the racist behaviors of the society where the data was taken.

5.2. Bayesian networks

Figure 2 presents the Bayesian networks for each dataset, which contains the dependence relation of each dataset's attributes, highlighting the protected and target attributes. The relationship between attributes becomes more apparent when analyzing networks. For instance, in the adult dataset, the hours per week attribute depends on occupation, and we previously observed that they are highly correlated. Similarly, occupation depends on gender, while the protected attribute also influences education, hours per week, and



(a) Adult

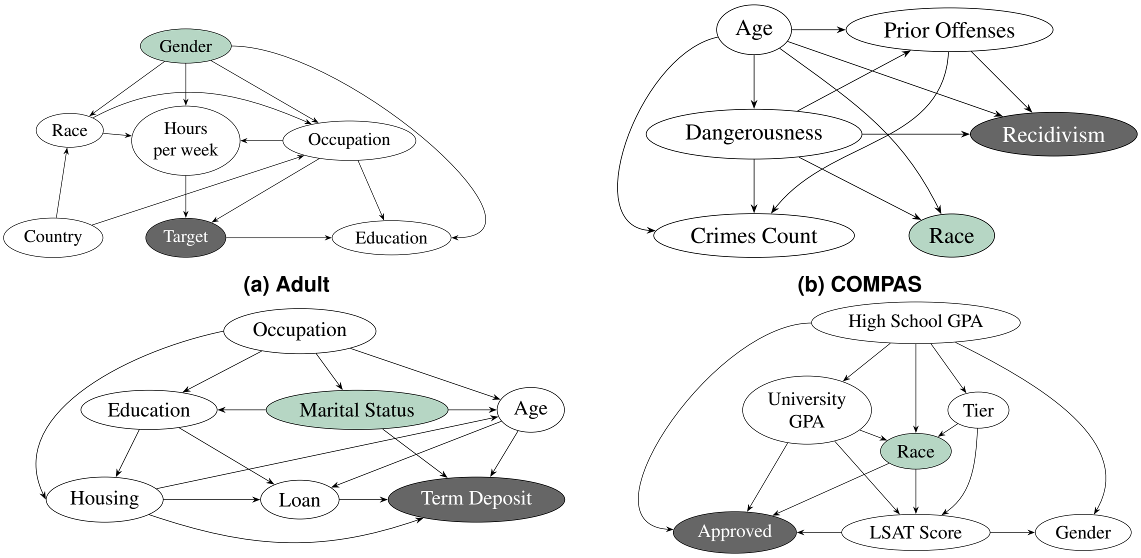
(b) COMPAS

(c) Law School

(d) Bank Marketing

Figure 1. Datasets heatmaps using ϕ_k correlation

race. This implies that simply suppressing the gender attribute is insufficient to prevent the inference of the protected attribute.



(a) Adult

(b) COMPAS

(c) Bank Marketing

(d) Law School

Figure 2. Bayesian networks generated

Such behavior is also seen in the Bank Marketing and Law School datasets, where the label and some other attributes directly depend on the protected attributes, Marital Status and Race, respectively. Finally, the COMPAS dataset is the only one without links between the protected attribute and label.

We can also observe a dependence relation between race and dangerousness. In this case, it is possible to infer an individual's race from some value of dangerousness. As discussed in the heatmaps, this can be a historical consequence of the society where the data was extracted. Such behavior generates an imbalance in the data, causing a race to have a much higher number of registers than the others.

5.3. ML models

Table 2 shows the results of the tests involving the ML models. In none of the cases the DI improvement is enough to reach the threshold, even less to achieve the perfect situation, in which the metric is 1. Such a situation indicates that the discriminatory bias is still there and can be perpetuated by ML models that train with those datasets.

Table 2. Models' results.

Dataset	Protected attribute	DI LRP	DI LRWP	AUC LRP	AUC LRWP
Adult	Gender	0.05	0.57	0.65	0.63
COMPAS	Race	0.67	0.68	0.66	0.66
Law School	Race	0.67	0.66	0.59	0.59
Bank Marketing	Marital Status	0.66	0.78	0.64	0.65

An explanation for the slight improvement of the disparate impact after suppressing the protected attribute is the dependence relation between the suppressed one and the others. For instance, in the adult dataset, the attributes hours per week and occupation depend on the suppressed attribute, gender. This means that an individual's gender could be de-identified based on the values of the remaining attributes.

The AUC score does not significantly change in any of the cases. Only one of them showed a minor improvement in the metric's value. The small change indicates that the protected attribute extraction did not alter the models' response quality. In this scenario, individuals who are unprivileged continue to face discrimination, but now indirectly. Even with the increase of the DI in some cases, we need more to ensure a fair environment since the values did not surpass 0.8. A possible solution for that is using well-known fairness methods, such as implemented in the IBM AIF360[1] Python library.

6. Conclusion

In this paper, we studied the impact of suppressing protected attributes to provide fairness in classification algorithms. We indicate the insufficiency of the protected attribute suppression technique by comparing the logistic regression trained with and without the protected attribute. The DI results confirmed that suppression was insufficient. Studying the relation between attributes was essential to identify possible reasons for the inefficiency. Therefore, it is possible to identify the need for a preliminary analysis of the datasets, not just the data cleaning but also the study of the metrics between attributes. A broader approach to the problem can be explored for future works, expanding the experiments for cases with more than one protected attribute and multiclass classification.

Acknowledgment

This research was partially funded by CAPES under grant number 88887.694847/2022-00 and by CNPq under grant number 316729/2021-3.

[1] https://aif360.res.ibm.com/

References

M. Baak, R. Koopman, H. Snoek, and S. Klous. A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043, 2020.

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

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996.

J. Berkson. Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365, 1944.

S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56 (7), apr 2024. ISSN 0360-0300.

P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa. Pass: protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15087–15096, 2021.

T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

M. Kearns and A. Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.

M. Koivisto and K. Sood. Exact bayesian structure discovery in bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

M. Langenkamp, A. Costa, and C. Cheung. Hiring fairly in the age of algorithms. *arXiv preprint arXiv:2004.07132*, 2020.

J. Larson, M. Roswell, and V. Atlidakis. Compas. `https://github.com/propublica/compas-analysis`, 2016. July 29, 2022.

T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

V. G. Martini. Análise de equidade em algoritmos de ia na área da saúde: um estudo sobre viés de dados, medidas de pós-processamento e correlações de atributos. 2023.

S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

U. E. Orji, C. H. Ugwuishiwu, J. C. Nguemaleu, and P. N. Ugwuanyi. Machine learning models for predicting bank loan eligibility. In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, pages 1–5. IEEE, 2022.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

E. Pitoura, K. Stefanidis, and G. Koutrika. Fairness in rankings and recommenders: Models, methods and research directions. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2358–2361. IEEE, 2021.

The US EEOC. *Uniform guidelines on employee selection procedures*, March 2, 1979.

L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.