Exploration of Potential New Benchmark for Fairness Evaluation in Europe

Magali Legast[1], Lisa Koutsoviti-Koumeri[2], Yasaman Yousefi[3] and Axel Legay[1]

[1]*Université catholique de Louvain*
[2]*Hasselt University*
[3]*Università di Bologna*

Abstract

With the increase use of artificial intelligence systems and the associated concerns regarding automated discrimination, research in the field of fairness has increased in the past years. To evaluate their work in fair machine learning, researchers have often been using the same three datasets (Adult, COMPAS, and German credit) as benchmarks. However, those datasets each present serious limitations. In this work, we first explore what other datasets could potentially be used as replacement, specifically in a European context. We then use an experimental approach to compare Adult and COMPAS with a new candidate, Student Performance (a.k.a Student Alcohol Consumption). Our early results highlight the scarcity of easily accessible European datasets suitable as benchmarks for fairness evaluation of problems with positive or negative outcome, as well as the high influence dataset selection can have on experimental results.

Keywords

Fairness, Datasets, Machine learning, Bias mitigation, Classification, Fair Classification, Benchmark

1. Introduction

With the increase use of artificial intelligence systems, legitimate ethical and legal concerns have been growing, including the risk that some people may be treated more negatively than others, thus resulting in discrimination [18, 21, 22]. This problem is prevalent in machine learning, where the training data is of paramount importance, while usually retaining historical and social biases that are then learned by the prediction models [21].

Research in the field of fairness has been constantly growing in the past few years, with the problem of fair classification receiving the most attention [21]. Many fairness metrics and bias mitigation methods have been developed in that sub-field, but less attention has been given towards quality benchmark datasets, specially regarding European data. A subset of only three datasets has been surpassing all others in term of popularity [12, 17], namely Adult [3], COMPAS [10], and German credit [16]. Those popular datasets have nevertheless been shown to present serious limitations [12], such as, but not limited to, old age (Adult and German credit),

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany
magali.legast@uclouvain.be (M. Legast); lisa.koutsoviti@uhasselt.be (L. Koutsoviti-Koumeri); yasaman.yousefi3@unibo.it (Y. Yousefi)
0000-0003-4246-1158 (M. Legast); 0000-0002-9490-6035 (L. Koutsoviti-Koumeri); 0000-0003-1483-2978 (Y. Yousefi); 0000-0003-2287-8925 (A. Legay)

noisy data [2, 9, 14], label bias (COMPAS [1]), or even coding mistakes preventing retrieval of the sensitive attribute (German credit [14]).

Given these observations, the emerging consensus is that their use as benchmarks should be avoided, lest duly justified. Efforts have been made to provide alternatives (for instance a dataset suite to replace Adult [9]), to disseminate good practices (like in [1]), or to facilitate access to other datasets (such as a large collection of fairness datasets [12] along with its search engine [13]). Building on these works, we explore potential alternatives to the three popular yet flawed historical benchmarks, while focusing on the European context. We also compare, with an experimental approach, results stemming from the use of Adult and COMPAS as benchmarks with a potential replacement, Student Performance [7]. More specifically, we compare the results of a bias mitigation method and several popular fairness metrics for models trained with each of those three datasets.

Our results highlight the scarcity of open access and easily accessible datasets for fairness evaluation in a European context. Further, the analysis of those three examples confirms that different datasets may lead to different results when evaluating bias mitigation methods and fairness metrics. This stresses the importance of dataset selection in experiments. We also aim to expand this work with a more thorough search for datasets and further experiments encompassing more bias mitigation methods, fairness metrics and sensitive attributes.

2. European Datasets

As mentioned in the introduction, the datasets Adult, COMPAS and German credit have been widely used as benchmarks in fair classification studies, even though they are not as suitable for fair machine learning research as previously thought. All three of them contain real life tabular data about individuals with one or more attribute(s) recognized as sensitive/protected and a target label that is deemed positive or negative for the individual.

Those key characteristics for fair classification should be shared by potential replacement datasets. Being open access also increased their appeal. We thus aim to find a potential new benchmark that is also easily accessible, specially considering the fast past environment of computer science conferences where there is often little time dedicated to the selection of datasets. Lastly, we focus on the European context. Indeed, the worldwide influence of the United-States and difference in data privacy culture and legislation across countries have made datasets from the USA dominant in the field, while data from other areas, including Europe, remains less accessible. This can cause further bias given that models and results from a certain place aren't necessary applicable to other geographical locations [9].

Merging the positive characteristics of the popular datasets with our context of interest, we formulate our dataset selection criteria as follows: An open access dataset with tabular data about European subjects that is no more than 25 years old and is adapted to the problem of fair classification leading either to a positive or negative outcome for the subject.

To find such dataset, we used the search engine for fairness datasets[1] presented in [13] as it is the most complete collection of datasets for fair machine learning to the best of our knowledge. Its database comprises over 200 datasets for diverse domains and fairness tasks.

[1]This search engine is available at http://fairnessdata.dei.unipd.it/

Filtering on tabular data and fair classification with a positive or negative outcome, we are left with 22 datasets (as of January 29 and February 2, 2024). Out of those, fourteen datasets contain data from the United-States, two from elsewhere in America, three have no mention of localization in their description, three are European, and only one is from Asia. There are thus no dataset referenced for Africa nor Oceania. Out of the three European datasets, only two contain data collected in this century, Dutch Census [6] and Student Performance [7] (a.k.a "Student" and "Student Alcohol Consumption"). Since Dutch Census is part of the IPUMS International collection [2] and requires approval to be accessed, only Student Performance fully fits our criteria. Presented in [8], this dataset contains social, gender and study information about students in two Portuguese schools for the core classes of Mathematics and Portuguese in secondary education (high school). It has been used in a few fairness studies and is referenced in the dataset survey [19].

3. Experiment

With this experiment[3], we compare different models trained with bias mitigation on Adult, COMPAS, and Student Performance. For Student Performance, we use data related to the Portuguese subject version of the dataset since it has the most instances (649). We consider the sensitive attributes *sex* and *age*, with students who are 18 or older as the protected group, as in [19]. We take the usual sensitive attributes *sex* and *race* for Adult and COMPAS. We did not include German credit, as it is impossible to retrieve its protected attribute (*sex*), making interpretation of results misleading.

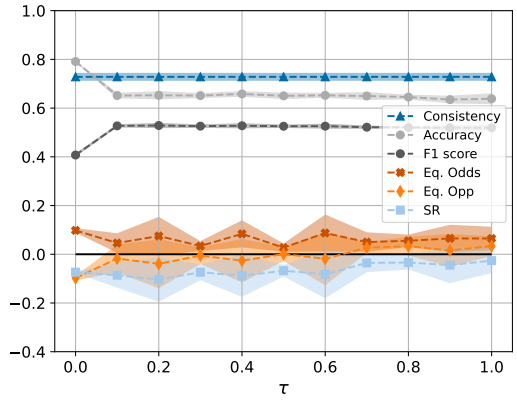
For each of the considered datasets, we compute several classifiers using the training and in-processing bias mitigation meta-algorithm presented in [5]. This algorithm considers a constrained optimization problem that is approximately solved with provable guarantees. The constraint enforces a minimal value, the fairness penalty parameter τ, for a chosen fairness metric. We use the AIF360 [4] implementation with Statistical Parity ratio as the fairness constraint and gradient descent. We train different models with τ value ranging from 0 (no bias mitigation) to 1 (constraint of perfect statistical parity). We then evaluate the performance and fairness of the resulting models using different metrics to assess their evolution with different constraint levels. The fairness metrics we use are Statistical Rate difference (SR) [11], the most used metric and based on prediction only, Equality of Opportunity (Eq. Opp) and Equalized Odds (Eq. Odds) [15], the most used metrics based both on predictions and ground truth, and Consistency [24], the most used metric based on similarity [23, 17]. We evaluate each metric for each classifier, then report the average over 10 folds and the corresponding confidence interval.

You can see in Figure 1 the results for the different models, each represented by the fairness penalty parameter τ it was trained with. Results for Adult and COMPAS with sensitive attribute *race* are close to those with *sex* and are not presented here due to space restriction.

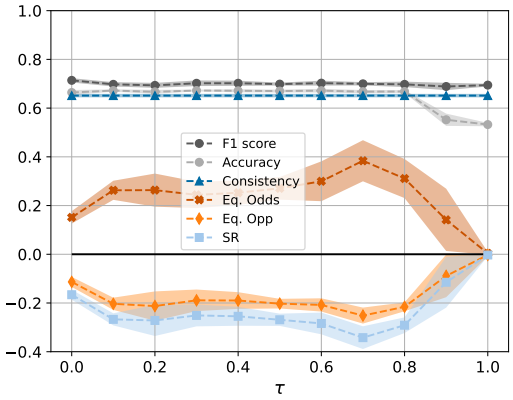
Let us first note that Consistency, an individual fairness metric, is not impacted by the Statistical Parity mitigation, which is based on group fairness. We thus focus only on the three

[2] See Harmonized International Census Data for Social Science and Health Research https://www.ipums.org/projects/ipums-international

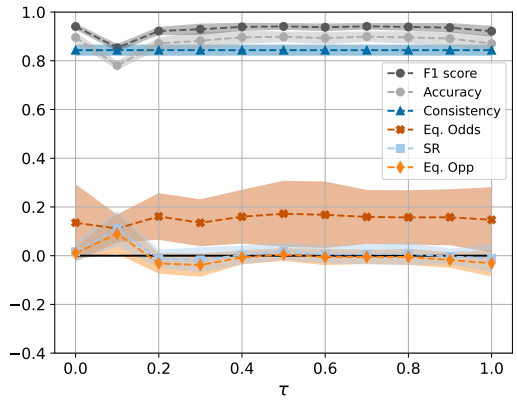
[3] The full code of the experiment is available at https://github.com/Magalii/AIF360/tree/EWAF2024



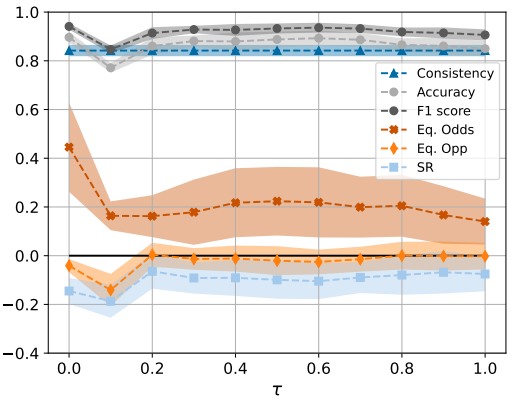
(a) Adult dataset with sensitive attribute *sex*
Women protected, Men privileged



(b) COMPAS dataset with sensitive attribute *sex*
Men protected, Women privileged



(c) Student dataset with sensitive attribute *sex*
Girls protected, Boys privileged



(d) Student dataset with sensitive attribute *age*
≥ 18 protected, < 18 privileged

Figure 1: Fairness and performance results for classifiers trained with different levels of fairness constraint

other metrics, related to group fairness, for the remainder of this section.

In Figure 1a, we see that the model trained on Adult without bias mitigation ($\tau = 0$) is already very close to fairness according to all metrics. Bias mitigation still induces a general improvement for all group fairness metrics. Eq. Opp indicates that the protected group (Women) becomes the one with a slightly higher true positive rate when the fairness constraint is greater than or equal to 0.7 . We also note a significant drop in accuracy and increase of the F1-score as soon as bias mitigation is introduced, but only marginal changes after that.

In Figure 1b, for models trained on COMPAS, there is first an overall decrease in fairness for all group metrics from $\tau = 0.1$ to $\tau = 0.7$. Fairness then increases again for higher values of τ, which is coincident with a significant drop in accuracy, while the F1-score remains steady.

In Figure 1c, the unconstrained model trained on Student Performance with *sex* as protected attribute is near perfect fairness according to SR and Eq. Opp. The constrained models never surpass these values. Eq. Odds shows more bias than both SR and Eq. Opp for all models but

one. This metric is less mitigated than for Adult, even though the original value was higher to start with.

In Figure 1d, the model trained on Student Performance with *age* as sensitive attribute and no bias mitigation shows significant bias according to SR. This bias is efficiently mitigated by the model, even with the lowest level of fairness constraint. Bias reported by SR and Eq. Opp is very low even before bias mitigation. Except for a brief increase when τ is 0.1, their bias level is reduced, with Eq. Opp extremely close to equality.

Overall, even though the same algorithm as been used for the training and bias mitigation of all of these models, the results vary significantly for each of the different datasets studied. Additionally, for Adult and COMPAS the overall tendencies are similar for the two sensitive attributes considered (*sex* and *race*), but we see a very significant difference when considering *sex* or *age* for Student Performance.

4. Discussion

On the one hand, the search for new potential benchmarks highlights the scarcity of European datasets for use in fair classification with a positive or negative outcome. Indeed, out of the over 200 datasets referenced in the search engine used [13], only Student Performance [7] fits our selection criteria. This dataset shows very little to no bias with regard to the sensitive attribute *sex*, at least regarding the most common fairness definitions, strongly reducing its interest when studying fairness related problems. The other most common sensitive attribute considered for this dataset is *age*. However, being an older student is most often a direct result of past failures, which is itself usually deemed an appropriate criterion to predict future exam results. It is thus questionable whether this attribute should be considered protected or not. Other attributes we have not explored here could also be of interest. For example, attributes related to alcohol consumption are studied in [20], which also extends the discussion to label bias.

So, despite the existing efforts to mitigate the collective data documentation debt and offer new alternatives to Adult, COMPAS and German credit, there is still a need to bring forward new European datasets, as well as data from other underrepresented continents. This may include improvement of the visibility and centralization of existing datasets or collection of new data adapted to fairness related questions.

On the other hand, our study illustrates that the same procedure applied to different datasets may lead to significantly different results, which is congruent with the results in [19]. This showcases the importance the choice of dataset can have on fairness evaluation and when presenting results. We thus recommend to use several different datasets since they may lead to varying results, to look beyond open access data if needed, and to avoid making broad claims based on few examples, which echoes some of the recommendations in [1] and [19]. We also encourage researchers to consider the selection of data not as a minor step, but as a meaningful part of the research, and to provide justification on the choices made in that regard.

Beyond this discussion, we aim to expand this work to study more datasets, include non-open access data, as well as more bias mitigation methods, fairness metrics and sensitive attributes.

References

[1] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. *ArXiv*, 2021. URL https://arxiv.org/abs/2106.05498.

[2] Matias Barenstein. Propublica's compas data revisited. *ArXiv*, 2019. URL http://arxiv.org/abs/1906.04711.

[3] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *ArXiv*, 2018. URL https://arxiv.org/abs/1810.01943.

[5] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019. URL https://doi.org/10.1145/3287560.3287586.

[6] Minnesota Population Center. Integrated public use microdata series, international: Version 6.4 2001 dutch census. Minneapolis: University of Minnesota, 2015. URL http://doi.org/10.18128/D020.V6.4.

[7] Paulo Cortez. Student performance. UCI Machine Learning Repository, 2008. URL https://archive.ics.uci.edu/dataset/320.

[8] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference, Porto*, pages 5–12. EUROSIS-ETI, 2008. URL https://repositorium.sdum.uminho.pt/handle/1822/8024.

[9] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *ArXiv*, 2022. URL http://arxiv.org/abs/2108.04884.

[10] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4, 2018. URL https://www.science.org/doi/10.1126/sciadv.aao5580.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226. Association for Computing Machinery, 2012. URL https://doi.org/10.1145/2090236.2090255.

[12] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6): 2074–2152, 2022. URL https://link.springer.com/10.1007/s10618-022-00854-z.

[13] Alessandro Fabris, Fabio Giachelle, Alberto Piva, Gianmaria Silvello, and Gian Antonio Susto. A search engine for algorithmic fairness datasets. In *Proceedings of the 2nd European Workshop on Algorithmic Fairness*. CEUR Workshop Proceedings, 2023. URL https://ceur-ws.org/Vol-3442/paper-08.pdf.

[14] Ulrike Grömping. South german credit data: Correcting a widely used data set. Technical report, Mathematics, Physics and Chemistry, Department II, Beuth University of Applied Sciences, Berlin, Germany, 2019.

[15] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[16] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. URL https://doi.org/10.24432/C5NC77.

[17] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. 1(2):11:1–11:52. URL https://doi.org/10.1145/3631326.

[18] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *EAE Papers and Proceedings*, volume 108, pages 22–27, 2018. URL https://pubs.aeaweb.org/doi/10.1257/pandp.20181018.

[19] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1452, 2022. URL https://doi.org/10.1002/widm.1452.

[20] Daphne Lenders and Toon Calders. Real-life performance of fairness interventions - introducing a new benchmarking dataset for fair ML. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 350–357. ACM, 2023. URL https://dl.acm.org/doi/10.1145/3555776.3577634.

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):115:1–115:35, 2021. URL https://doi.org/10.1145/3457607.

[22] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, 2021. URL https://doi.org/10.1146/annurev-statistics-042720-125902.

[23] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7. IEEE, ACM, 2018. URL https://dl.acm.org/doi/10.1145/3194770.3194776.

[24] Richard S. Zemel, Ledell Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*. PMLR, 2013. URL https://proceedings.mlr.press/v28/zemel13.html.