Improved Boosted Classification to Mitigate the Ethnicity and Age Group Unfairness

Ivona Colakovic and Sašo Karakatič[a]

Faculty of Electrical Engineering and Computer Science,
University of Maribor, Koroška cesta 46, Maribor, Slovenia

Keywords: Fairness, Classification, Boosting, Machine Learning.

Abstract: This paper deals with the group fairness issue that arises when classifying data, which contains socially induced biases for age and ethnicity. To tackle the unfair focus on certain age and ethnicity groups, we propose an adaptive boosting method that balances the fair treatment of all groups. The proposed approach builds upon the AdaBoost method but supplements it with the factor of fairness between the sensitive groups. The results show that the proposed method focuses more on the age and ethnicity groups, given less focus with traditional classification techniques. Thus the resulting classification model is more balanced, treating all of the sensitive groups more equally without sacrificing the overall quality of the classification.

1 INTRODUCTION

In recent years Machine Learning (ML) has been used to solve problems in different areas such as finance, healthcare, retail and logistics. Thus, outperforming humans and automating human tasks have led to the wide adoption of ML in a wide variety of fields. With a growing number of ML applications, many have started to raise the issues about the accountability when decision are made by ML, and especially, the *fairness* of those decisions. The often raised question about ML being fair is still not entirely researched. Many have proposed different approaches to mitigate algorithm bias, but do not consider the existing bias in data.

The ability of ML to discover patterns relevant to decision-making that humans can overlook is what makes ML useful. Discovering patterns in data is the power of ML, but the data given to ML is entirely a human product. In that way, ML can discover historical bias in data (Barocas et al., 2019), which is our social bias reflected in the collected data, which ML learns and thus becomes unfair itself. There are different types of fairness, but in this work, we consider *group fairness*, which is the equal chances of different groups to be positively classified (Binns, 2019; Mehrabi et al., 2021). Usually, the unfairness is learned by ML, when sensitive features are playing the role in the decision-making process (i.e., when they are included in the patterns in the classification model). To prevent the ML from building models with sensitive features, they are removed from the data, before the ML process commences. But, as other features may indirectly reflect the sensitive feature values, we have to control for the sensitive feature, i.e., we have to measure the quality of the ML model for every value of the sensitive feature.

1.1 Fair Machine Learning

Fairness, especially in societal problems, is defined differently by fairness researchers, thus making the fairness measure very hard to define (Barocas et al., 2019; Verma and Rubin, 2018). However, this does not mean that fairness can not be measured, just that the standard classification metric that everyone would agree upon does not exist or is not agreed upon yet. Therefore, various metrics were proposed, although most group fairness metrics are based on *statistical parity* – equal chances of each group to have a positive outcome (Binns, 2019).

While unfairness can appear in data that does not include people, it is more likely to have a bigger impact in applications where data represents individuals. Some ML applications with discrimination have already been identified, like Amazon's Prime system for determining who is eligible for advanced services, that turned out to be racially biased (Ingold and Soper,

[a] https://orcid.org/0000-0003-4441-9690

Colakovic, I. and Karakatič, S.
Improved Boosted Classification to Mitigate the Ethnicity and Age Group Unfairness.
DOI: 10.5220/0011287400003269
In *Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*, pages 432-437
ISBN: 978-989-758-583-8; ISSN: 2184-285X

2016). Analysis of COMPAS, the system used by judges, probation, and parole officers, used to assess the likelihood of criminal recidivating, showed that black people are more likely to be incorrectly classified as a higher risk to re-offend (Angwin et al., 2016b). In comparison, white people are more likely to be incorrectly classified as a lower risk to re-offend (Angwin et al., 2016a). Google's ad systems tend to show more ads for high-paid jobs to males than female users (Datta et al., 2015). These applications do not imply ML's inefficiency, but rather the need for further research on fair ML.

1.2 Existing Literature

Some work has already been done on increasing the fairness of ML models. But with this paper, we build upon the existing literature, where the fairness is addressed with the ensemble of classifiers. Fair Forests (Raff et al., 2017) were proposed for fairness induction in decision trees. An alteration of how information gain is calculated with respect to sensitive feature is proposed. This approach improves fairness in decision trees, whereas we wanted an iterative process, where fairness is corrected in steps. Thus, instead of a forest (ensemble of trees) of uncorrelated classifiers, we opted to use the boosting technique in an ensemble of decision trees. This was already touched upon in a case study done by researchers from the University of Illinois (Fish et al., 2015), where the boosting technique increased fairness in the Census Income dataset. This approach relabels existing instances according to fairness rules. It focuses on improving individual fairness, while we wanted to focus on group unfairness. Next, AdaFair (Iosifidis and Ntoutsi, 2019) was proposed for boosting instances using cumulative fairness while also tackling the problem of class imbalance of used datasets. This approach changes how the weights are updated, so it considers the model's confidence score and equalized odds. On the other hand, our approach uses the maximum difference between two groups to calculate estimator error and fairness of each group to update weights, making the equal treatment of group the main priority of here proposed approach.

1.3 Contributions

From this, we present the boosting classification ensemble, which strives to optimize both, the group fairness and the overall model quality simultaneously.

Our proposed approach is used to address the common unfairness problem in the Drug benchmark dataset, notorious for its historical bias for age and ethnicity (Donini et al., 2020).

Thus, our main contributions are the following:

- We define fairness of sensitive feature group, which we use to alter the weights that the original AdaBoost calculated.
- We propose **Fair AdaBoost** classification ensemble, for balanced group results of sensitive feature, with achieving the same overall quality.

2 METHODOLOGY

AdaBoost was first introduced by Freund and Schapire in 1995 (Freund and Schapire, 1997). It is an adapting boosting algorithm in which weak learners are combined in order to create a strong one. The boosting technique enables a weak learner to learn from his own mistakes and boost his knowledge. Estimators are created iteratively, and each estimator at the end receives its weight corresponding to its accuracy. The final prediction is presented as a weighted sum of all estimators.

At the beginning of the algorithm, an equal weight is assigned to each instance. At the end of every iteration, weights of misclassified instances are increased, allowing the learner to focus on more challenging instances in the next iteration. Weights are adapted according to an error in estimation that suggests the importance of instances until a certain number of iterations or a perfect estimator with estimator error 0 is achieved.

2.1 Fair AdaBoost

We propose the Fair AdaBoost algorithm to expand the multi-class AdaBoost algorithm (Hastie et al., 2009) that considers fairness in training its classification model. As well as AdaBoost, Fair AdaBoost is based on a boosting technique, where each data instance gets a weight updated through iterations until an optimal result is achieved. Therefore, misclassified instances, get increased weight while the weight of correctly classified instances decreases. In addition to that, weights of the instances also contain the error rate for its sensitive feature group. At the end of each iteration, a model is created that, together with its weight, is defined according to the model performance, combining results for the final prediction. A few hundred iterations could be performed before the estimator is perfect, having estimator error 0 or before it starts to stagnate.

As presented, Fair AdaBoost takes into account fairness when updating instance weights and, in that

Algorithm 1: Fair AdaBoost weights boosting stage.

```
$w_0 = 1/S$                              ▷ $S$ is a number of instances
for $i = 1, ..., n$ do                   ▷ $n$ is a number of iterations
    $learn(data, w_{i-1})$
    $predict(X)$
    $calculate\ accuracy$
    $calculate\ fairness\ per\ group\ as\ in\ Equation\ 2$
    $calculate\ estimator\ error\ as\ in\ Equation\ 3$
    $update\ weights\ according\ to\ Equation\ 1$
end for
```

way, differs from AdaBoost. Fair AdaBoost procedure is shown in the form of pseudocode in Algorithm 1, where in the beginning, every instance has an equal weight. In every boosting iteration, the estimator is fitter with instance weights defined at the end of the previous iteration. When the prediction is given, the overall accuracy of that model (acc_{global}) is calculated, as well as the accuracy of each group of the sensitive feature. After, the fairness of each sensitive feature group is calculated as shown in Equation 2, with acc_{max} being the highest accuracy any group has achieved and acc_k the accuracy of the k group.

Estimator error is then calculated as shown in Equation 3, where w_f is the fairness weight given as an input parameter, and acc_{diff} is the maximum difference between accuracy of any two sensitive feature groups. If the estimator error is 0 (the estimator returns perfect class predictions), the boosting is stopped.

$$w_{i,j} = w_{i,j(AB)} \times fairness_k \quad , \quad j \in K \tag{1}$$

$$fairness_k = \frac{acc_{max}}{acc_k} \tag{2}$$

New instance weights are calculated as in AdaBoost and then multiplied by $fairness_k$, as shown in Equation 1. Meaning, weight of j instance in i-th iteration ($w_{i,j}$) is weight of j instance in i-th iteration calculated by AdaBoost ($w_{i,j(AB)}$) multiplied by fairness of k group to which j instance belongs to regarding sensitive feature.

$$err = (1 - acc_{global}) \times (1 - w_f) + acc_{diff} \times w_f \tag{3}$$

3 EXPERIMENT

For evaluation of Fair AdaBoost, we conducted an experiment on the UCI Drug consumption dataset (Fehrman et al., 2016) using 5-fold cross-validation. For comparison, we used the AdaBoost algorithm and Decision Tree. AdaBoost contains 50 decision tree with a maximum depth of 1, the learning rate at 1.0, and the SAMME.R algorithm. Decision Tree has a random state set at 123, *gini* for criterion function, and no defined maximum depth. Boosting weights using AdaBoost and Fair AdaBoost is performed in 50 iterations.

Beside standard metrics for model evaluation, such as accuracy, F-score, TPR and TNR, for measuring fairness we observe accuracy and F-score by sensitive group as well as their maximum difference. We also include equalized odds difference that represents equal chances of same instances to be either positive or negative, and demographic parity difference that demonstrates difference between groups with largest and smallest selection rate (Bird et al., 2020).

3.1 Dataset

In the experiment, we use the Drug consumption dataset, as used in (Mehrabi et al., 2020; Donini et al., 2020). Data was collected in a survey (Fehrman et al., 2017) where participants had to state the frequency of various drugs consumption. It contains answers from 1885 people, where each participant is described with 12 personal attributes and 18 attributes that correspond to each drug in a survey. Dataset can be used for solving different problems, so we chose binary classification in which we predict heroin consumption. For binary classification purposes, we transform values of drug attributes from [”Never Used”, ”Used over a Decade Ago”, ”Used in Last Decade”, ”Used in Last Year”, ”Used in Last Month”, ”Used in Last Week”, and ”Used in Last Day”] to [”Used”, ”Not Used”]. The dataset contains several possible sensitive attributes, from which we chose and separately tested age and ethnicity.

Table 1: Drug consumption dataset description.

	Dataset	
Instances	1885	
Attributes	32	
Sensitive attribute	Age	Ethnicity
Class ratio (+:-)	1 : 5.73	
Positive class	Used	

3.2 Results

Figure 1 shows the results of classification with age as a sensitive attribute. Values of these metrics closer to 1 are better results. AdaBoost and Fair AdaBoost outperform Decision Tree in almost every metric. While AdaBoost and Fair AdaBoost achieve similar results, Fair AdaBoost better classify positive cases. Figure

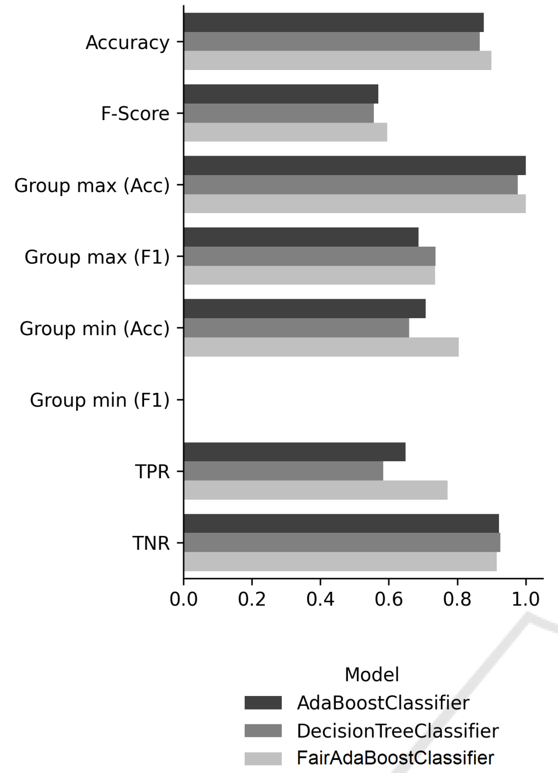

Figure 1: Evaluation of classification on Drug consumption dataset using attribute age as sensitive attribute (higher values represent better results).

2 shows the results of classification with ethnicity as a sensitive attribute. Performance of measured algorithms is comparable to results when age is a sensitive attribute, while Fair AdaBoost is again better in classifying positive cases.

Results of classification are shown on Figure 3 and Figure 4 with age and ethnicity as sensitive attribute, respectively. Metrics on these graph measure fairness where lower results represent better results, since values indicate differences between groups. Fair AdaBoost achieves better results than Decision Tree and AdaBoost with different sensitive attributes.

We also evaluated model performance on different groups which sensitive attribute contains. Fair AdaBoost achieves better accuracy and F-score of elderly groups while maintaining good results in younger groups of people. When ethnicity is used as a sensitive attribute, Fair AdaBoost also achieves better-balanced results. Better accuracy is primarily achieved in Black-Asian ethnicity, as this group is often wrong classified by AdaBoost and DTR, while Fair AdaBoost achieves the best accuracy in this group. Figure 5 show the accuracy and F-score of used models by groups when a sensitive attribute is age, and Figure 6 when a sensitive attribute is ethnicity.

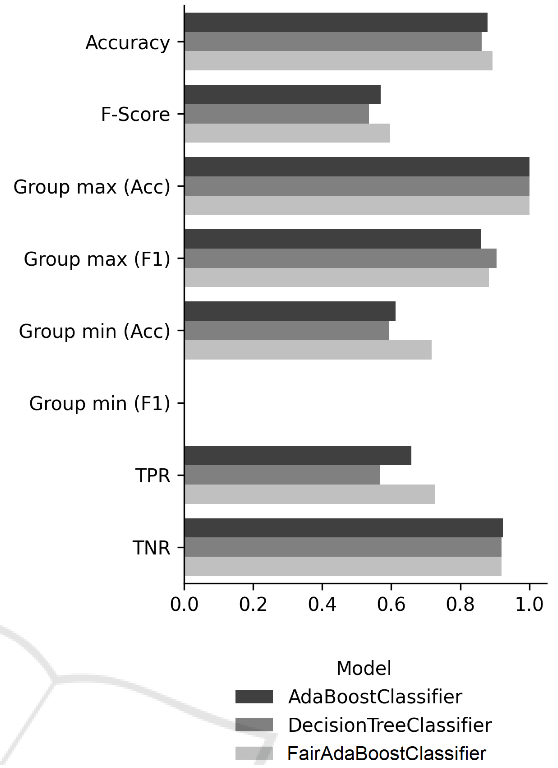

Figure 2: Evaluation of classification on Drug consumption dataset using attribute ethnicity as sensitive attribute (higher values represent better results).

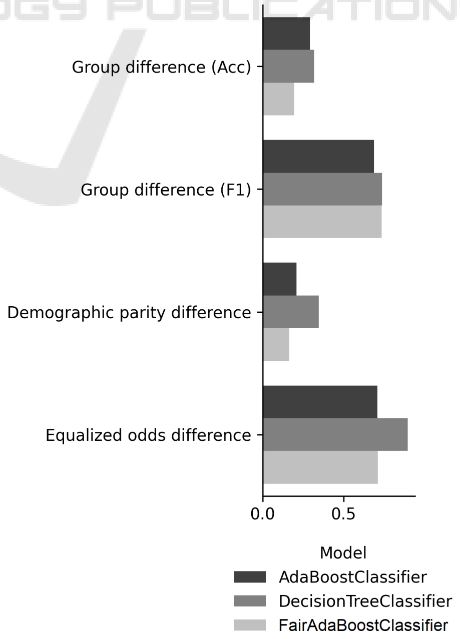

Figure 3: Evaluation of classification on Drug consumption dataset using attribute age as sensitive attribute (lower values represent better results).

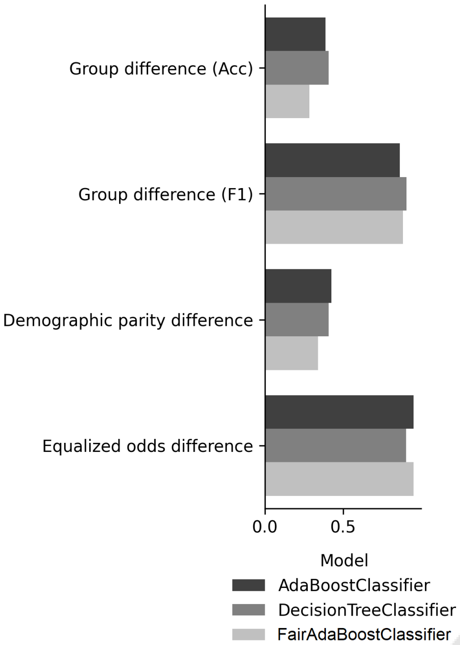

Figure 4: Evaluation of classification on Drug consumption dataset using attribute ethnicity as sensitive attribute (lower values represent better results).

4 CONCLUSION

In this work, we tackle the issue of unfair classification among sensitive groups, namely age and ethnicity of the individuals. For this, we propose Fair AdaBoost method, which is a boosting approach based on the AdaBoost algorithm that takes fairness into consideration during instance weights adaptation. We evaluate this approach with binary classification of Drug consumption dataset, which included the age and ethnicity of the participants, which shouldn't be taken into consideration in the classification process.

The results show that Fair AdaBoost improves fairness so that the overall accuracy and macro-averaged F-score are comparable to original AdaBoost, while fairness metrics calculated by sensitive feature improve. The biggest difference can be seen in TPR where Fair AdaBoost achieves the highest results. The equalized odds difference is the same as achieved with AdaBoost suggesting that the difference between groups from which instances are not the highest likely to be classified as positive or negative. However, demographic parity difference is lower than AdaBoost and individual decision tree, meaning that Fair AdaBoost has the most similar groups of sensitive feature in terms of equal rates to be classified as positive instance.

When considering age as sensitive attribute, we can observe that classification quality of different groups is more balanced than in AdaBoost and individual decision tree. This is especially evident in much fairer classification of the elder groups, even though they appear in small number.

Results from evaluation using ethnicity as sensitive feature show that Fair AdaBoost achieved notably better accuracy in Mixed-Black/Asian group. Results of other groups are comparable to the ones achieved by other approaches, suggesting that Fair AdaBoost does consider fairness in classification.

The proposed approach was evaluated on one dataset, but it should be tested on different datasets for more conceivable results. While, this experiment included Fair AdaBoost with CART decision tree classifier as base estimator, future work could examine the influence of different estimators in the ensemble. From the results, we concluded that boosting technique has the impact on fairness and in the future it would be interesting to apply fairness to different boosting algorithms such as XGBoost. And finally, here proposed Fair AdaBoost has to be thoroughly compared with other competing fairness ensuring ensemble classifiers, ideally on multiple datasets.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057).

REFERENCES

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016a). How We Analyzed the COMPAS Recidivism Algorithm.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016b). Machine Bias.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Binns, R. (2019). On the Apparent Conflict Between Individual and Group Fairness. *arXiv:1912.06883 [cs, stat]*. arXiv: 1912.06883.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *arXiv:1408.6491 [cs]*. arXiv: 1408.6491.

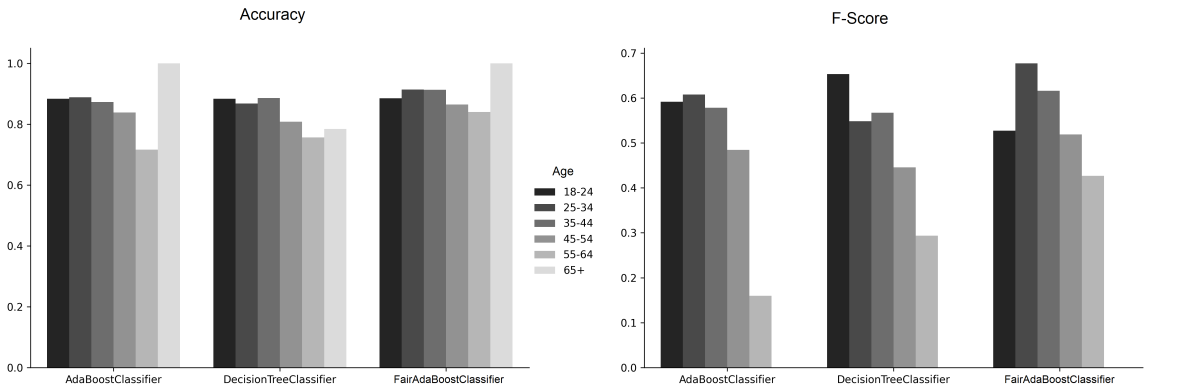

Figure 5: Accuracy (left) and F-score (right) of used algorithms for classification of Drug consumption dataset using age as sensitive attribute.

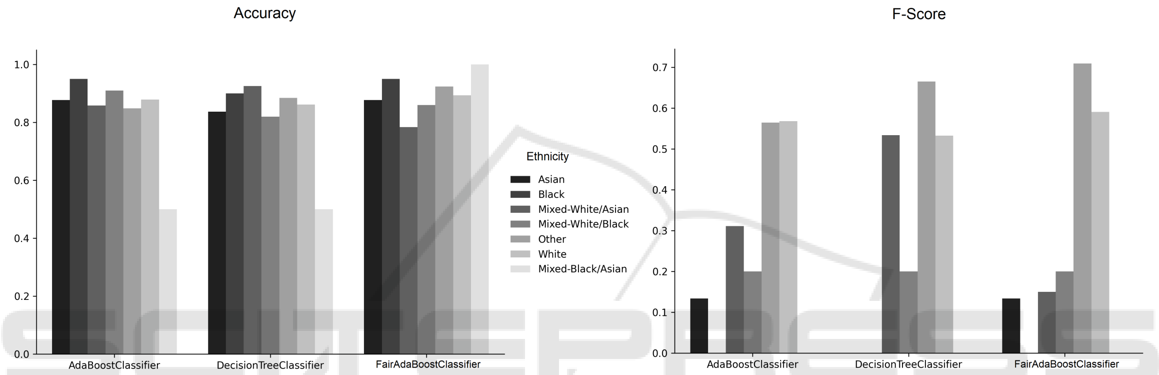

Figure 6: Accuracy (left) and F-score (right) of used algorithms for classification of Drug consumption dataset using ethnicity as sensitive attribute.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. (2020). Empirical Risk Minimization under Fairness Constraints. *arXiv:1802.08626 [cs, stat]*. arXiv: 1802.08626.

Fehrman, E., Egan, V., and Mirkes, E. (2016). UCI Machine Learning Repository.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N. (2017). The Five Factor Model of personality and evaluation of drug consumption risk. *arXiv:1506.06297 [stat]*. arXiv: 1506.06297.

Fish, B., Kun, J., and Lelkes, A. D. (2015). Fair Boosting: a Case Study. page 5.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360.

Ingold, D. and Soper, S. (2016). Amazon Doesn't Consider the Race of Its Customers. Should It?

Iosifidis, V. and Ntoutsi, E. (2019). AdaFair: Cumulative Fairness Adaptive Boosting. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 781–790. arXiv: 1909.08982.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35.

Mehrabi, N., Naveed, M., Morstatter, F., and Galstyan, A. (2020). Exacerbating Algorithmic Bias through Fairness Attacks. *arXiv:2012.08723 [cs]*. arXiv: 2012.08723.

Raff, E., Sylvester, J., and Mills, S. (2017). Fair Forests: Regularized Tree Induction to Minimize Model Bias. *arXiv:1712.08197 [cs, stat]*. arXiv: 1712.08197.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden. ACM.