Bursting the Burden Bubble?
An Assessment of Sharma et al.'s Counterfactual-Based Fairness Metric

Yochem van Rosmalen, Florian van der Steen, Sebastiaan Jans, and Daan van der Weijden

Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands
{y.m.vanrosmalen,f.a.vandersteen,s.j.j.jans,d.j.vanderweijden}@students.uu.nl

Abstract. Machine learning has seen an increase in negative publicity in recent years, due to biased, unfair, and uninterpretable models. There is a rising interest in making machine learning models more fair for unprivileged communities, such as women or people of color. Metrics are needed to evaluate the fairness of a model. A novel metric for evaluating fairness between groups is Burden, which uses counterfactuals to approximate the average distance of negatively classified individuals in a group to the decision boundary of the model. The goal of this study is to compare Burden to statistical parity, a well-known fairness metric, and discover Burden's advantages and disadvantages. We do this by calculating the Burden and statistical parity of a sensitive attribute in three datasets: two synthetic datasets are created to display differences between the two metrics, and one real-world dataset is used. We show that Burden can be more nuanced than statistical parity, but also that the metrics can disagree on which group is treated unfairly. We therefore conclude that Burden is a valuable metric to add to the existing group of fairness metrics, but should not be used on its own.

Keywords: Fairness metrics · Burden · Statistical parity · Decision boundary · Sensitive attributes · Unprivileged groups · Classification

1 Introduction

The use of algorithms in automated decision making has increased over the past years, due to the increased accuracy in which they can predict or classify desired outcomes. However, many of these algorithms are black boxes, whose decision process is not transparent to humans. This is undesirable since it could lead to the unfair treatment of certain groups, without being able to provide an explanation [1]. This has led to an increased demand of fair and explainable models, with many new frameworks for providing explanations being proposed [13, 15, 16], as well as new metrics to measure the fairness of a model [9, 11, 19].

Metrics for fair machine learning measure how well a particular model is towards different groups within a dataset. Although the definition and practical implementation of fairness varies between different metrics, their overarching goal is to provide insight into the level of fairness between different groups regarding sensitive attributes (e.g. age, gender, socioeconomic status).

One of the new frameworks is CERTIFAI [18], a framework that tests the robustness of a model, as well as providing explanations and a metric to measure fairness. It does so by generating counterfactuals for each datapoint in the dataset. This counterfactual is a synthetic datapoint, generated to have the other possible outcome, while being as close as possible to the original datapoint. The counterfactual provides insight as to what features should change to have the model classify the datapoint differently. Not only does this provide an explanation for why a certain classification was made, this also allows us to measure the – possibly unfair – difference in treatment for certain groups (e.g. male and female). By calculating the average distance for a group between original datapoints in the negative outcome class (e.g. loan application denied) and their generated counterfactuals (e.g. loan application approved), the *Burden* of a group can be calculated. The Burden of groups can be compared, in order to address which groups have a higher Burden and thus have more 'difficulty' converting from the negative to the positive predicted outcome class.

Sharma, Henderson, and Ghosh [18] claim that this metric for fairness, which they call Burden, is a more nuanced version of other fairness metrics, like *statistical parity* (SP), which is a formal non-discrimination criterion used to measure fairness [11]. It is calculated by measuring the ratio of the probability of receiving a positive outcome from a model between groups (See Sec. 2.1). However, this claim is not validated in their study. In this study, the claim is tested, by comparing the Burden metric to SP. We focus on SP because it, and Burden likewise, does not take the actual ground truth target value into account but rather the model's prediction. This is clear in the formula for SP (Eq. 2). This matter is more in-depth explained in Sec. 2.

In this study, we investigate situations where both metrics give different results to see if Burden can provide more nuance and if it is a good fairness metric in practice. This is tested on two synthetic datasets with hypothetical data and a real-world loan application dataset [20]. All three datasets have a *binary* outcome class: a favorable outcome and an unfavorable outcome. This means that the models used are also binary classification models.

2 Related Work

In this section, the fairness metrics SP (Sec. 2.1) and Burden (Sec. 2.2) are explained more in-depth to get a better theoretical understanding of how and why they work as fairness metric.

2.1 Statistical Parity

There are many metrics to measure how fair a model is, and there is no agreement on a best method, or even on the definition of fairness itself [3]. However, one of the most common and easy to implement fairness metrics is that of SP, or *demographic parity* [11]. In order to calculate SP, we have to calculate the acceptance rate (AR) for a specific group of S e.g. if S is a binary value the groups could be 0 and 1, which can be seen in Eq 1.

$$AR_{S=s} = P(\hat{Y} = 1 | S = s) \tag{1}$$

This means that the acceptance rate of the group where $S = s$ is defined as the probability (P) of the model predicting a positive outcome ($\hat{Y} = 1$)[1], given that $S = s$. To calculate the SP between two groups of a binary feature S, we look at the ratio of the acceptance rate of both groups, as seen in Eq 2.

$$SP_S = \frac{P(\hat{Y} = 1 | S = 0)}{P(\hat{Y} = 1 | S = 1)} \tag{2}$$

If the same percentage of individuals receives a positive score for each group, and thus the outcome of the ratio is 1, the two groups both have the same probability of receiving a positive outcome prediction from the model. This is seen as fair: if S is a sensitive attribute, e.g. age, there should be no difference in receiving a positive prediction. Perfect SP is almost never possible in practice, so often a maximum difference of 20% is used, to follow the 80% rule for disparate impact [8].

2.2 Burden

In 2020, a new method of measuring fairness is proposed in the CERTIFAI framework [18], where counterfactuals are used to measure different treatment of groups. The use of counterfactuals for explanatory purposes is not new [14], as is using counterfactuals to check if features are proxies for known sensitive features [12].

A counterfactual, in this context, is a datapoint calculated to be as similar to an original datapoint as possible, while receiving a different classification. A counterfactual datapoint can provide an individual with *recourse*: the counterfactual datapoint can show the individual which changes to the input features are necessary in order to change the qualification to the desired output class.

Sharma et al. [18] propose a genetic heuristic search for generating a counterfactual c for a datapoint x. The process is shown visually in Fig. 1. Starting from a randomly generated population of size N, the counterfactuals, $\mathbf{c}$, that are classified to be in the opposing class are selected. These are then mutated with probability P_m, which involves arbitrarily changing some feature values. Subsequently, crossover is applied with probability P_c, which involves randomly interchanging some feature values between individuals. Then, a top-k selection procedure is applied where only the most fit counterfactuals are selected. The fitness function $d(x, c)$ is a distance function calculated over the datapoint and

[1] The actual ground truth outcome, or target value, of a supervised dataset is denoted as Y, while the model's predicted outcome is denoted as $\hat{Y}$.

the counterfactual. The population is then filled back up to N by randomly generating new counterfactual points. This process is repeated for a predetermined maximum of generations. Finally, the fittest counterfactual $\mathbf{c}^*$ is selected.

With these fittest counterfactuals, the Burden of a group can be calculated. The Burden of a group with value s for feature S is defined as the mean of the distances between the datapoints in $S = s$ that had an unfavorable decision and their counterfactuals,

$$\text{Burden}_{S=s} = \mathbb{E}_{S=s}[d(\mathbf{x}, \mathbf{c}^*)], \tag{3}$$

where the distance function $d(\mathbf{x}, \mathbf{c}^*)$ can be chosen. Examples include the Manhattan (L_1) distance and the Euclidean (L_2) distance.

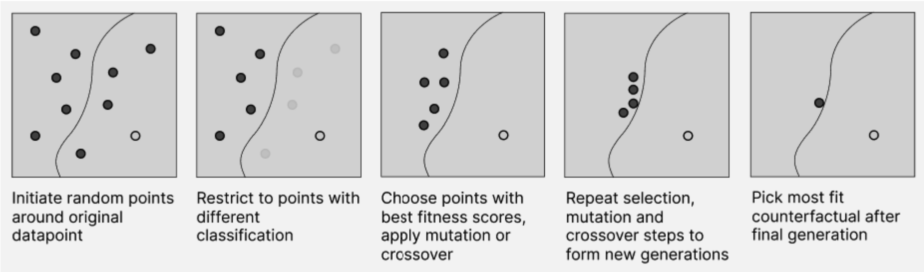

Initiate random points around original datapoint
Restrict to points with different classification
Choose points with best fitness scores, apply mutation or crcssover
Repeat selection, mutation and crossover steps to form new generations
Pick most fit counterfactual after final generation

Fig. 1: Visual representation of the counterfactual generation process of CERTIFAI. Adopted from [18].

For a binary sensitive attribute, the average distances can be divided by each other to give a ratio. If the ratio is larger than 1, the group in the numerator has a greater Burden, and if the ratio is smaller than 1, the group in the denominator has a greater Burden. An important difference between the two fairness metrics is that with the acceptance rate, *higher* is better, but with Burden, *lower* is better.

Remark that it is hard to verify that any difference in Burden between groups is not a measurement error, e.g. because of badly generated counterfactuals, due to the dimensionality of the data or complexity of the model. In this paper, however, we assume that any found difference in Burden is a true difference and it correctly implies that members in the unprivileged group have greater difficulty changing their outcome.

3 Methods

In this section, we describe the methodology of this study. The methodology is broken down in four parts: the creation of the two synthetic datasets, the description of the 'Default of Credit Card Clients' dataset, the classifier models and lastly CERTIFAI's counterfactuals and Burden. The Python code (Jupyter Notebook), saved models and generated data is available on GitHub[2].

3.1 Synthetic Datasets

We created two datasets to demonstrate two types of disagreements between Burden and SP that are theoretically possible. One synthetic dataset, D_A, shows that Burden disagrees with SP about if *there is* unfairness, and the other dataset, D_B, shows that Burden disagrees with SP about *which group* is treated unfairly. Both synthetic datasets consist of 80 datapoints.

Each datapoint in the two datasets consists of three features and a label. The legitimate (non-sensitive and non-proxying)[3] features X_1 and X_2, the sensitive attribute S (0 is unprivileged, 1 is privileged), and the target label Y (0 is unfavorable, 1 is favorable). This means that datapoint i is given by $D^{(i)} = (x_1, x_2, s, y)$. The legitimate features X_1 and X_2 are both sampled from normal distributions ($X \sim \mathcal{N}(\mu, \sigma^2)$) with different means μ. All samples have a standard deviation σ^2 of 1. The sensitive attribute S is selected (not sampled from a random distribution), as is the target label Y. The true underlying function between the legitimate features and the outcome can be derived from Table 1.

Dataset on Presence of Unfairness $\boldsymbol{D_A}$ Burden takes the average distance of a group to their counterfactuals into account, while SP does not. Therefore, the synthetic data needs to satisfy two properties: Firstly, it needs to satisfy SP, so for each group, the same number of datapoints needs to be predicted positive. Secondly, the average distance of the negatively predicted datapoints to their counterfactuals needs to differ between the two groups to show how Burden can find this unfairness.

As described in Section 3.1, the legitimate features X_1 and X_2 are sampled from a normal distribution and S and Y are selected. Sensitive attribute groups $S = 0$ and $S = 1$ both have a probability of 0.5 on the favorable outcome, which is considered fair according to the SP metric, i.e. $P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1) = 0.5$. The same is the case for target label Y. The distribution of dataset D_A is shown in Table 1. The datapoints of synthetic dataset D_A are plotted in Fig. 2a, along with their counterfactuals.

Dataset on Direction of Unfairness $\boldsymbol{D_B}$ This dataset should let Burden and SP disagree on which group is treated unfairly. This means that SP has to label one group of the sensitive attribute as unprivileged, and Burden has to label the other group as unprivileged. SP labels a group as unprivileged if the group has

[2] `https://github.com/yochem/bursting-the-burden-bubble`

[3] A proxying feature can reveal sensitive information. E.g. someones address, although not a sensitive feature, can reveal someones socioeconomic status because of their neighborhood.

less positively predicted outcomes as the other group ($P(\hat{Y}=1|S=0) \neq P(\hat{Y}=1|S=1)$). When we have a perfect classifier (accuracy of 1), $\hat{Y}=Y$. Using the definition of SP, an unprivileged group can be formed by having relatively fewer datapoints where $\hat{Y}=1$. For Burden to disagree with SP, the other S-group (i.e. the group that SP sees as privileged) needs to have a greater distance to their counterfactuals at the decision boundary, as illustrated in Fig. 1.

To create a synthetic dataset with these properties, the distributions of proposed dataset D_B is shown in Table 1. Note that $P(\hat{Y}=1|S=0) \approx 0.57$ and $P(\hat{Y}=1|S=1) \approx 0.67$, and therefore has an imbalance in the positive outcome rates between the two groups, if the classifier has perfect accuracy. The average distance to the counterfactuals and thus the Burden however is higher for group $S=1$. The datapoints of synthetic dataset D_B are shown in Fig. 2b, along with their counterfactuals.

Table 1: The distribution of values per feature for both datasets D_A and D_B. The count is the number of datapoints sampled from the normal distributions for X_1, and X_2, with shown mean μ for their normal distribution $\mathcal{N}(\mu, 1)$.

Dataset D_A					Dataset D_B				
μ_{X_1}	μ_{X_2}	S	Y	count	μ_{X_1}	μ_{X_2}	S	Y	count
1	9	0	0	20	1	9	1	0	15
3.5	5	1	0	20	3.5	5	0	0	15
9	1	0	1	20	9	1	1	1	30
9	1	1	1	20	9	1	0	1	20

3.2 Default of Credit Card Clients Dataset

To explore the claim of Burden being more nuanced, the metric is also compared to SP on real-world data. This is done on a subset of the Default of Credit Card Clients dataset, also known as the Taiwan loan dataset, from [20] (from now on called Taiwan dataset). This is a dataset of credit card users, which records whether the individual defaults on a loan. Since defaulting on a loan is a negative outcome, the favorable label in this dataset is 0: 'did not default'. The unfavorable label is 1: 'default'. The sensitive attributes are gender, education, marriage,, and age [5]. In our experiments we treat gender as the sensitive attribute. Monthly payments were tracked for the other features, such as history of past payment, amount of bill statement, amount of previous payment,, and amount of given credit. More information on these features can be found in [20]. Datapoints with incorrect values were removed. The dataset contains 30,000 instances. The computational cost for generating counterfactuals is large because the genetic algorithm iteratively goes over large population sizes for many generations. Therefore, this study is limited to a random sample of 1000 instances from the Taiwan dataset.

3.3 Classifier

A binary classification model (classifier) is needed for calculating Burden, and SP. The choice of classifier is not very important for this study, because the counterfactuals can be learned independent of the model's internals. On all three datasets, a Logistic Regression model [6] was trained. This model was chosen because it is complex enough to learn a perfect decision boundary for the synthetic datasets, and it is traditionally used for this task [7].

The classifiers were trained on the datasets without their sensitive features. No hyper-parameter optimization was performed, and the datasets were not partitioned into separate tests for training, and evaluation. This was done to remove unnecessary complexity: our interest lies in evaluating fairness, not model performance. The Logistic Regression model was implemented in PyTorch for reasons concerning compatibility with the CERTIFAI framework. It used the binary cross-entropy loss function [6], and the stochastic gradient descent optimizer [17]. The learning rate was 0.001, and was trained using 2000 iterations. The number of input dimensions for each classifier was the number of legitimate features, i.e. X_1 and X_2 for the synthetic datasets, and 19 features for the Taiwan dataset relating to past payments, bill statements, and credit features.

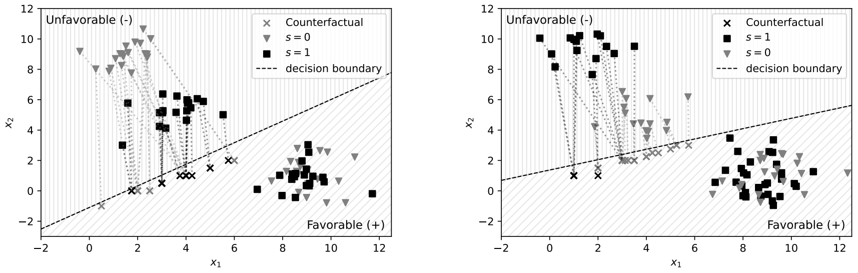
3.4 CERTIFAI's Burden

Using the `CERTIFAI.fit()` method, counterfactuals were generated given the model. The hyperparameters were the following: 10 generations of populations with size 60,000, of which maximal 10,000 retain after selection, of which maximal 5,000 retain for the next generation, and unconstrained generating of counterfactual features. The probabilities for crossover and mutation were adopted from [18]. For calculating the Burden, CERTIFAI's `check_fairness` method was used with as argument a mapping containing 1) the sensitive attribute and its value (e.g. `s: 0`), and 2) that it should be calculated over the unfavorable class (i.e. `favorable: 0`).

4 Results

In the first two experiments, logistic regression models were trained on D_A and D_B respectively and both got an accuracy of 1.00. In Fig. 2, this is shown by the decision boundaries laying perfectly between both groups. After the models were trained, the counterfactuals for the unfavorable class were generated by CERTIFAI, also shown in Fig. 2. After this, the SP and Burden were calculated. The results of the metrics on both experiments are listed in Table 2.

For the experiment on D_A we see that the SP is met, thus having a ratio of 1 between groups. The Burden of group $S = 0$ is higher (Burden of 11.6) than of group $S = 1$ (Burden of 4.65). This difference in Burden can also be eyeballed using Fig. 2a, where the $S = 0$ group is further away from their counterfactuals than the $S = 1$ group.



(a) D_A, where Burden and SP disagree on the presence of unfairness.

(b) D_B, where Burden and SP disagree on the direction of unfairness.

Fig. 2: The synthetic datapoints (▼,■) for datasets D_A and D_B. The counterfactuals (×) for datapoints from the unfavorable outcome class are also included and are connected using a dotted line. The decision boundary (---) of the classifier is also shown.

For the experiment on D_B we see that the ratio in SP is 0.857. The Burden of the two groups are 3.31 and 11.0 for $S = 0$ and $S = 1$, respectively. The datapoints are plotted in Fig. 2b.

The results of the last experiment, on the Taiwan dataset, are also listed in Table 2. The model trained on this dataset achieved an accuracy of 0.78. The results are the following: SP is nearly met, with a value of 1.02 (0.967 over 0.948). Burden however shows that females have almost 1.5x higher Burden than males, respectively 1.38 and 0.940.

Table 2: SP and Burden for the three datasets. The acceptance rate and Burden are given per group ($S = 0$ and $S = 1$ for the synthetic datasets correspond to gender=female and gender=male respectively for the Taiwan dataset), as well as the Burden ratio and statistical parity (SP) between the two groups, in bold.

	Acceptance Rate		SP	Burden		
Dataset	$S = 0$	$S = 1$	0/1	$S = 0$	$S = 1$	0/1
Taiwan	0.967	0.948	**1.02**	1.38	0.940	**1.47**
D_A	0.500	0.500	**1.00**	11.6	4.65	**2.49**
D_B	0.571	0.667	**0.857**	3.31	11.0	**0.302**

5 Discussion

In this section, the results are discussed as well as the limitations of this study and directions for future work.

5.1 Discussion on Experimental Results

In the first experiment with dataset D_A, the results show that even though SP was met (ratio of 1.00), Burden shows that the model is unfair towards group $S = 0$, since their Burden is higher. This means that Burden can show unfairness between groups when SP can not. This is a positive result for the claim of Sharma et al. [18] that Burden provides more nuance than SP in this situation. The distance to the counterfactuals near the decision boundary is important here to actually find the unfairness.

In the second experiment on dataset D_B, the SP shows unfairness towards group $S = 0$ since SP is less than 1. Burden however tells us that group $S = 1$ is being treated unfairly, because the Burden of this group (11.0) is higher than the Burden of the other group (3.31). This means that SP and Burden can disagree on which group is treated unfairly.

As a third experiment, on real-world Taiwan data, the results show the same effect as the first experiment. Although the difference is smaller than with D_A, the results show that with SP the model is more fair than with Burden. Burden can thus add more nuance than SP.

5.2 Limitations and Future Work

Future work could find real-world examples of the synthetic dataset D_B, and if the disagreement between Burden and SP found in our synthetic experiment occurs in other situations.

Furthermore, it is important to note that the computational complexity of the Burden metric is extremely high in comparison to a metric like SP. While SP is a simple calculation of a ratio between two percentages, the calculation of a single counterfactual for the Burden metric can take minutes. For large datasets it might thus be necessary to compute this metric for a representative sample of the dataset.

To conclude the discussion, which of SP and Burden is the better metric, and for which use-case, is a matter of subjective debate. It is in general a matter of subjective debate how fairness should be expressed in a number, or in fact even whether that is really possible or desirable at all [4]. Different stakeholders value different aspects of fairness differently. In fact, popular fairness metrics such as SP and equalized odds [10] are mutually exclusive under reasonable assumptions [2], which shows the extent of differences between metrics.

6 Conclusion

In this study we assessed the fairness metric introduced in Sharma et al. [18], using three experiments. The first experiment, using synthetic dataset D_A, shows that Burden can pick up unfairness when SP can not. The second experiment, using synthetic dataset D_B, shows that Burden and SP can even disagree on which group is treated unfairly. The last experiment, using the Taiwan dataset, shows that Burden is more nuanced than SP on a real-world dataset. The three experiments show that Burden can provide more information than SP, but this information may not be in line with SP.

We therefore conclude that Burden should not replace SP or any other fairness metric, because they measure different aspects of fairness. The advantage of using Burden as fairness metric is taking distance into account, which can provide more information about underlying unfairness between groups. The biggest disadvantage of Burden is the computational complexity. Therefore, considering its additional costly information about fairness, on large datasets, Burden can be applied on a random sample that reflects the global statistics of the original dataset.

Acknowledgments. We thank dr. Dong Nguyen, Yupei Du, and dr. Heysem Kaya for their great help.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Martin, K. (ed.) Ethics of Data and Analytics, chap. 6.1, pp. 254–264. CRC Press, Boca Raton, FL, 1st edn. (May 2022). https://doi.org/10.1201/9781003278290
2. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
3. Chouldechova, A., Roth, A.: A Snapshot of the Frontiers of Fairness in Machine Learning. Communications of the ACM **63**(5), 82—-89 (April 2020). https://doi.org/10.1145/3376898
4. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)
5. Council Regulation (EC): on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (2016), 02016R0679
6. Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958)
7. Crook, J.N., Edelman, D.B., Thomas, L.C.: Recent developments in consumer credit risk assessment. European Journal of Operational Research **183**(3), 1447–1465 (2007)
8. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and Removing Disparate Impact. International Conference on Knowledge Discovery and Data Mining (KDD) **21**(1), 259—-268 (August 2015). https://doi.org/10.1145/2783258.2783311
9. Hardt, M., Price, E., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems (NIPS) **29**(1) (2016). https://doi.org/10.5555/3157382.3157469
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
11. Kamiran, F., Calders, T.: Classifying without Discriminating. In: International Conference on Computer, Control and Communication (I4C). IEEE (February 2009). https://doi.org/10.1109/IC4.2009.4909197

12. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf
13. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. International Conference on Neural Information Processing Systems (NIPS) **31**, 4768—-4777 (2017). https://doi.org/10.5555/3295222.3295230
14. Mothilal, R.K., Sharma, A., Tan, C.: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 607—-617. No. 2 in FAccT, Association for Computing Machinery, New York, NY (2020). https://doi.org/10.1145/3351095.3372850
15. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. International Conference on Knowledge Discovery and Data Mining (KDD) **22**, 1135—1144 (2016). https://doi.org/10.1145/2939672.2939778
16. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations. AAAI Conference on Artificial Intelligence **32**(1) (2018). https://doi.org/10.1609/aaai.v32i1.11491
17. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)
18. Sharma, S., Henderson, J., Ghosh, J.: CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). pp. 166—-172. Association for Computing Machinery, New York, NY (2020). https://doi.org/10.1145/3375627.3375812
19. Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning Non-Discriminatory Predictors. In: Kale, S., Shamir, O. (eds.) Proceedings of the 2017 Conference on Learning Theory. pp. 1920–1953. No. 65 in PLMR (July 2017)
20. Yeh, I.C., Lien, C.h.: The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. Expert Systems with Applications **36**(2), 2473—-2480 (March 2009). https://doi.org/10.1016/j.eswa.2007.12.020