Evaluating Fairness Metrics[1]

Zahid Irfan, Fergal McCaffery and Róisín Loughran

Regulated Software Research Center,
Dundalk Institute of Technology,
Dundalk, Co Louth, Ireland
{Zahid.Irfan, Fergal.McCaffery , Roisin.Loughran}@dkit.ie

Abstract. Artificial Intelligence systems add significant value to decision-making. However, the systems must be fair because bias creeps into the system from sources like data and preprocessing algorithms. In this work, we explore fairness metrics discussing the shortfalls and benefits of each metric. The fairness metrics are demographic, statistical, and game theoretic. We find that the demographic fairness metrics are independent of the actual target value and hence have limited use. In contrast, the statistical fairness metrics can provide the thresholds to maximize fairness. The Minimax criterion was used to guide the search and help recommend the best model where the error among protected groups was minimum.

Keywords: Artificial Intelligence, Fairness, Bias, Game Theory, Minimax, Pareto Front

1 Introduction

Fairness becomes a key factor when machines use algorithms to make decisions. This is important when we have privileged groups due to social, demographic, economic, or other factors. Privileged groups can have an unfair advantage over unprivileged groups, also called protected groups. This leads to bias and sometimes severe harm to protected groups. It is possible that some protected groups can be mistreated if nothing is done to ensure fairness. In this context, we study the various fairness metrics and attempt to understand the limitations or benefits. We must gain the theoretical background to ensure we use the correct metrics. As is often the case, there is no one solution or fix, but if we understand the basics, it's a much more informed decision. Some work has been done to evaluate fairness metrics, but it is geared towards evaluating metrics from the perspective of data or algorithms [1].

Fairness definitions can be grouped as statistical and individual fairness definitions [2]. Statistical fairness definitions divide the protected groups to ensure equalized influence using some statistics like error rate etc. Individual fairness definitions seek to ensure that similar individuals are given similar treatment.

[1] This research was supported through the HEA's Technological University Transfer Fund (TUTF) and Dundalk Institute of Technology (DkIT).

In this study, we focus on three different fairness measures: demographic, statistical, and minimax. Metrics based on a dataset demographic include statistical parity differences and disparate impact [3]. The statistical distribution of the predictions determines the equal opportunity and receiver operator characteristic curve/ area under the curve metrics. The minimax fairness criteria provide a notion of fairness focusing on the groups and ensuring that each group is not worst off [4] [5].

This paper is organized as follows: Section 2 offers mathematical background and explains the metrics used. Section 3 details the experiment, Section 4 gives the results, Section 5 discusses the results, and Section 6 provides the conclusions and future work.

2 Mathematical Background

2.1 Binary Classification Problem

We consider a binary classification problem defined by features to understand fairness metrics. The population is classified into two classes based on one feature. The classes are sometimes referred to as positive and negative, depending on the domain context. For example, if we deal with credit approval, a positive class will be where the credit is approved, and a negative would be where the credit is not approved. An example scenario is shown in Figure 1. Here, the two classes are shown in orange and blue. For the sake of simplicity, the two classes are assumed to have Gaussian distribution. Classification algorithms are trained to predict which class an unseen sample of data is likely to belong to [6].

The true positive rate (TPR), or the hit rate, is the rate at which the classifier correctly predicts the positive class. The false positive rate (FPR) is the rate of incorrect positive classification by the classifier.

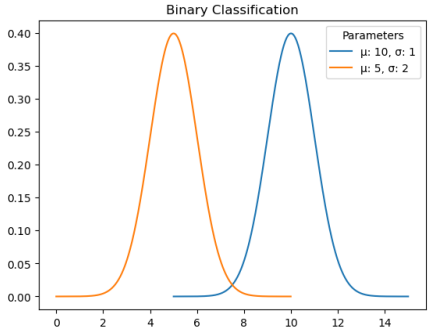


Fig. 1. Binary Classification

2.1.1 Receiver Operating Characteristic (ROC) Curve

The receiver operating characteristic (ROC) curve helps determine the classifier accuracy [6]. The ROC shows the values of TPR against FPR for different classification threshold values. As the threshold between the two separate classes' changes, the TPR and FPR change. Table 1 indicates the relationships between the limits of TPR and FPR

Table 1. TPR vs FPR

TPR	FPR	Comments
0	0	Every point classified as negative
0	1	Every negative point classified as positive, while positive as negative. (Simple class inversion makes this optimal)
1	0	Optimal point (not necessarily achievable)
1	1	Every point is classified as positive

Figure 1 displays the plot between the TPR and FPR for an example classifier which defines the ROC., The area under the curve (AUC) is a measure of accuracy of the given classifier. Ideally the area should be 1.0 (meaning a TPR=1, FPR=0).

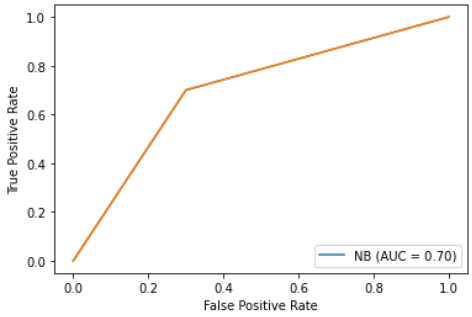


Fig. 2. Receiver Operating Curve ROC and Area Under Curve (AUC).

In the following sections, we demonstrate how the ROC and AUC provide the basis of fairness. The ROC helps in identifying equalized Odds and Equal Opportunity thresholds. While the AUC provides the overall accuracy of the classifier. If used for protected attributes, the AUC can lend an idea about the bias present in the system [7].

2.2 Demographic Fairness Metrics

2.2.1 Statistical Parity Difference

Statistical Parity Difference (SPD) measures the difference between the probability of the privileged and unprivileged classes receiving a favourable outcome. This measure must be equal to 0 to be fair.

$$SPD = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \quad (1)$$

Where $\hat{Y}$ is the model predictions, and A identifies the protected attribute (A=0 for unprivileged class, A=1 for privileged class).

2.2.2 Disparate Impact

Disparate Impact (DI) compares the proportion of individuals that receive a favourable outcome for two groups, a privileged group and an unprivileged group. This measure must be equal to 1 to be fair.

$$DI = P(\hat{Y} = 1 \mid A = 0) \,/\, P(\hat{Y} = 1 \mid A = 1) \quad (2)$$

Where $\hat{Y}$ is the model predictions, A identifies the protected attribute (A=0 for unprivileged class, A=1 for privileged class).

2.3 Statistical Fairness Metrics

2.3.1 Equalized Odds

The equalized odds definition, according to [8], is given by the following. Let A=1 and A=0 represent the privileged and unprivileged demographics, respectively.

$$P\left[\hat{Y} = 1 \middle| Y = y, A = 0\right] = P\left[\hat{Y} = 1 \middle| Y = y, A = 1\right], where\ y \in \{0,1\} \quad (3)$$

Considering the above equation for y=1, the equation shows TPR across privileged and unprivileged groups. While if we consider y=0, the equation represents the false positive rate (FPR) across privileged and unprivileged groups. This represents the threshold in ROC where both TPR and FPR are equal for privileged and unprivileged demographics.

2.3.2 Equal Opportunity

In this case, the equal opportunity fairness criteria are met when the TPR for both groups is the same. Regarding ROC, this means that the TPR is equal for both the privileged and unprivileged groups.

$$P\left[\hat{Y} = 1 \middle| Y = 1, A = 0\right] = P\left[\hat{Y} = 1 \middle| Y = 1, A = 1\right]. \quad (4)$$

2.4 Game Theoretic Fairness

2.4.1 Minimax Fairness Criteria

Equal Opportunity and Equalized Odds work well for groups; however, they only guarantee when we need individuals [9]. Game Theory is an economic framework that helps model economic problems as games [10]. Nash Equilibrium is the solution of the games when n- players engage in a non-cooperative zero-sum game [11]. Recent research has proposed to model learning with fairness as minimax group fairness [5, 4].

Let $(x_i, y_i)|_{i=1}^{N}$, where x_i is the feature vector divided into *K* groups $\{G_!, G_2, \ldots, G_K\}$. A class *H* of models map features to predicted labels, y_i. The minimax problem is with the following constraints defined with *L*, the loss function taking values in [0, 1].

$$h^* = \arg\min_{h \in \Delta H} \{\max_{1<k\leq K} \epsilon_k(h)\} \quad (5)$$

The average population error $\epsilon(h)$ and average group error $\epsilon_k(h)$ are defined as follows.

$$\epsilon(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i) \quad (6)$$

$$\epsilon_k(h) = \frac{1}{|G_k|}\sum_{(x,y)\in G_k} L(h(x), y) \quad (7)$$

The algorithm described in [4] is to minimize $\epsilon(h), h \in \Delta H$ subject to $\epsilon_k(h) \leq \gamma, k = 1, .., K$.

The algorithm iterates over the scenarios where two players, Learner and Regulator, are engaged in zero-sum games. At each iteration, the regulator determines a weighting over groups, and the learner responds by computing model h_t to minimize the weighted prediction error. The regulator updates the weights by using the Exponential Weights Algorithm. The algorithm converges to Nash Equilibrium [11], and the solution space iterates over Pareto Fronts [12], which means no group is worse off due to any change.

3 Experiment

3.1 Dataset

We use the German Credit Dataset [13] to perform the analysis. The original dataset has a large set of possible values. It is selected because it has a binary target (Good/ Bad Risk), and the dataset has two protected attributes i.e., age and sex. We use the reduced dataset attributes and values. For example, the original dataset had bank balance limits instead of little, moderate, rich, and quite rich. The dataset's attributes are as follows.

Table 1. Attributes of German Credit Dataset

Attribute	Possible Values

Age	Integer values
Sex	Male, Female
Job	Employed, unemployed
Housing	Own, Free, Rent
Saving accounts	Little, moderate, rich, quite rich
Checking accounts	Little, moderate, rich, quite rich
Credit Amount	Integer Values
Duration	Duration in month
Purpose	Business, car, domestic appliances, education, furniture/equipment, radio/TV, repairs, vacation/others
Risk	Good, bad

The risk attribute is the binary target attribute which is good or bad. We clean the data before using it and transform it into numeric attributes. The protected attributes are age and sex. The dataset's loss function provided in the dataset is the following. The loss function allows to optimize the training of the model. The loss function stipulates that a false positive is five times more damaging than a false negative.

Table 2. Loss Function of German Credit Dataset

Actual/Predicted	**Good**	**Bad**
Good	0	1
Bad	5	0

3.2 Fairness Metrics

Fairness metrics are used to measure the fairness of classification algorithm [14]. We use the following metrics to evaluate fairness. As described above in section 2, the fairness metrics are interpreted by their values. Here is a summary of all the metrics used in this experiment.

Table 3. Fairness Metrics and Criteria

Fairness Metric	**Criteria**
Statistical Parity Difference	0 means demographic fairness
Disparate Impact	1 means demographic fairness
Equalized Opportunity	TPR is the same for both groups
ROC	Closer to (1,1) is better
AUC	Higher is better and closer to 1.0
Minimax-Fairness	Models weights uniform distribution

3.3 Classifiers

In this experiment, we use the Support Vector Classifier (SVC), Gaussian Process Classifier (GPC), Gaussian Naïve Bayesian (GNB), and Linear Discriminant Analysis (LDA) Classifiers to do a comparative analysis. These classifiers have been chosen because they represent diverse types. The Support Vector Classifier is kernel-based, while Gaussian Process Classifier and Naïve Bayesian are probabilistic, and finally, Linear Discriminant Analysis is a dimensionality reduction technique. Thus, we have attempted to cover different types of classification.

4 Results

We ran a series of experiments described above on the German Credit Dataset. The results are detailed in this section. The objective of the experiments was to evaluate the fairness metrics to determine which metrics are most helpful in building fairness into the system. We use multiple classifiers on the same dataset. This approach enables the determination of the fairness metrics performance across classifiers.

4.1 Demographic Fairness Metrics

First, we consider the demographics within the data itself. If we look at the Statistical Parity Difference (SPD), the values are not equal to zero, as described in section 2.2 for fairness. The situation is better for protected attribute sex as compared to age. As the table 4 shows that the values for attribute sex is closer to zero compared to age. It means that unfairness is present with respect to attribute age.

However, the disparate impact (DI) is relatively high (the ideal is 1) for both protected attributes. Since these metrics don't consider the actual values but rely only on the predicted target values, this is a shortcoming because of reliance on the model's prediction.

Table 4. Demographic Fairness Metrics Results

Metric	**Age**	**Sex**
Statistical Parity Difference	-0.1285	-0.0748
Disparate Impact	0.8212	0.8965

4.2 Statistical Fairness Metrics

4.2.1 Equal Opportunity Analysis

Ideally, as we know that the TPR should be the same for both attributes. The equal opportunity fairness metric, shown in Table 5, shows that the Gaussian Process Classifier performs less fairly than the other three classifiers. Gaussian Process Classifier is

slightly fairer in relation to the protected attribute sex compared to the age. As we can see that the GPC is less fair and not accurate because of low values of TPR.

Table 5. Equal Opportunity for classifiers

Classifier	Age		Sex	
	Old	Young	Male	Female
Support Vector Classifier	0.97	1	1.0	0.92
Gaussian Process Classifier	0.26	0.071	0.24	0.095
Gaussian Naïve Bayesian	1	1	1.0	1
Linear Discriminant Analysis	0.91	0.85	0.90	0.88

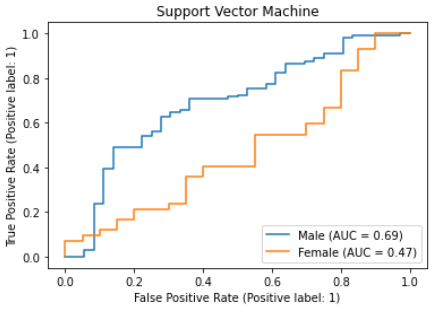
4.3 ROC /AUC Analysis

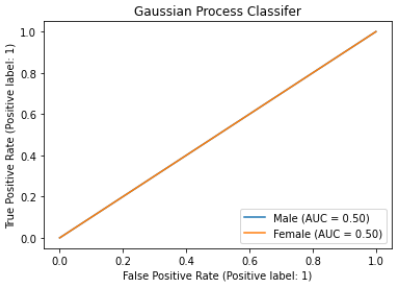
The ROC curve and AUC analysis were performed separately for each protected attribute. As discussed in section 3, the ROC and AUC identify the best classifiers [6] and find the equal opportunity fairness [8], respectively.

The ROC curves in Figures 3 and 4 show how the TPR and FPR change across the curve as the threshold changes. The AUC for each classifier calculated on both sensitive attributes is shown in Table 6. These values correspond to the ROCs shown in Figures 3 and 4. We can note that the AUC for GPC is 0.5, and the ROC is a diagonal line for both Age and Sex. This means that the GPC is only as good as a random guess. A similar conclusion was also observed in the Equal Opportunity metric. The GNB is the best among the classifiers, with an AUC value of 1.0, while LDA is better than SVC.

Table 6. AUC for each protected attribute for classifiers

Classifier	Age		Sex	
	Old	Young	Male	Female
Support Vector Classifier	0.59	0.47	0.69	0.42
Gaussian Process Classifier	0.5	0.5	0.50	0.50
Gaussian Naïve Bayesian	1.0	1.0	1.0	1.0
Linear Discriminant Analysis	0.76	0.87	0.82	0.72





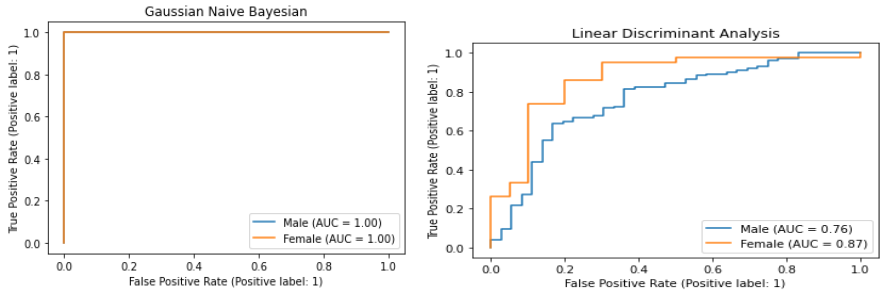


Fig. 3. ROC and AUC for the sensitive attribute sex.

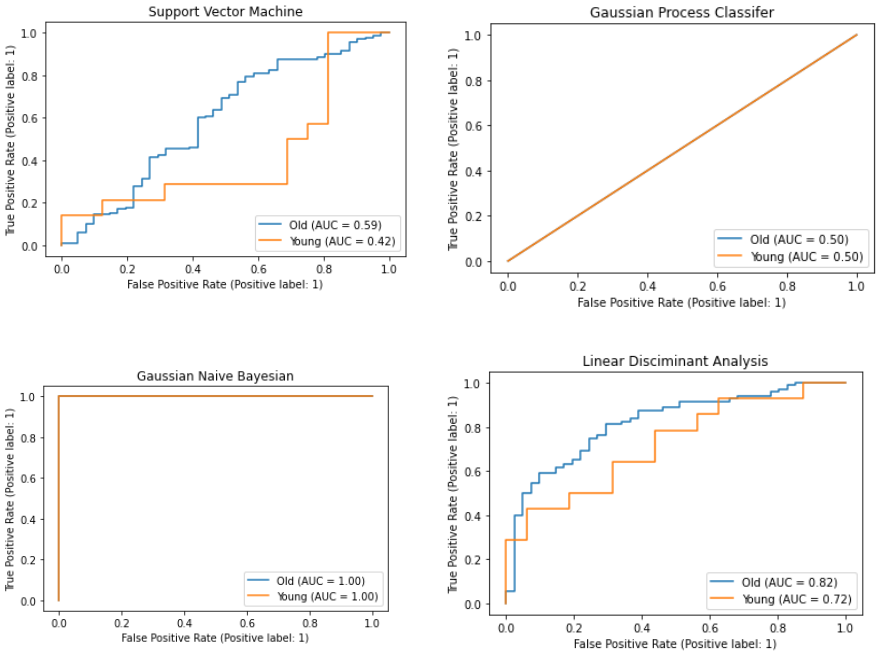


Fig. 4. ROC and AUC for the sensitive attribute age.

4.4 Minimax Fairness

A Minimax Fairness criterion was implemented as described in [3]. The scheme was used to find optimal weights for a classifier such that the discrepancy among prediction accuracies of the different groups is minimized along with the minimization of the overall prediction error.

Logistic Regression was used in these three classifiers, and the results for the two protected attributes show the errors converging in both cases. The sample weights are updated such that the average population and group errors defined by equations 6 & 7 are minimized. The plots in figure 5 show that as the algorithm proceeds, the values of errors for protected attributes age and sex are minimized, and their difference is also reduced.

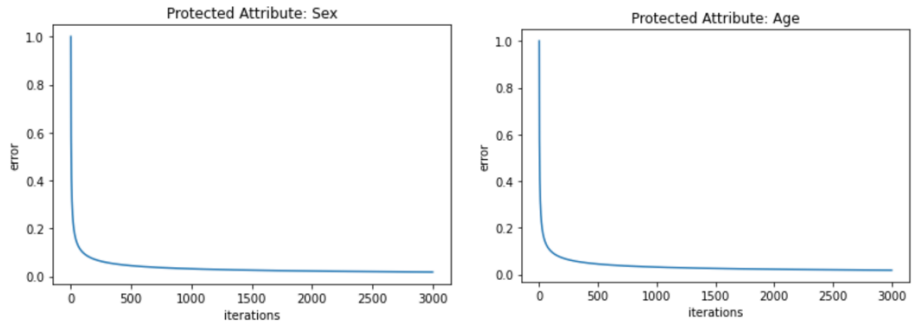


Fig. 5. Minimax errors for the sensitive attributes sex and age .

4.5 Discussion

The Ethics Guide for Trustworthy AI builds a framework to create trustworthy systems using AI. It stipulates the development of lawful, robust, and ethical systems to achieve the purposes [15]. Systems that achieve fairness are the cornerstone of this endeavor. One aspect of Ethical AI is to ensure it acts fairly without detrimental bias against certain individuals and groups. In this paper, we evaluate the fairness criteria to identify if bias is present.

We performed a preliminary fairness analysis using demographic fairness metrics on the German credit dataset. The SPD and DI indicated that protected attribute sex was better than attribute age. It is imperative to understand that due to the lack of inclusion of actual target values, the metrics rely solely on the model's predictions. This means these metrics cannot be used to determine if the bias is because of the model or data.

The second set of fairness metrics involved equal opportunity and ROC/AUC. Since equal opportunity is calculated from the TPR and FPR, these are very good in giving the behavior of the classifier and the dataset. We used only one dataset in our experiments using four classification algorithms. We found that GNB was the best, while GPC was the worst.

The minimax fairness criteria provided weights for a model for adequate fairness. The minimax criteria ensured that, in the end, we had a set of weights that guaranteed the lowest maximum error. The results show that the errors in the protected attributes

of age and sex groups were minimized after the algorithm ended. In addition, the minimax resulted in a decrease in the difference between protected group errors. Hence it can be used as a fairness-enhancing metric among protected groups.

5 Conclusions and Future Work

In this work, we considered a series of fairness metrics applied to the German credit dataset. We calculated the SPD, DI, Equal Opportunity, ROC/AUC, and minimax metrics and analyzed the results. We have analyzed the fairness metrics to determine the biases against potentially protected groups.

In future work, we aspire to provide methods for mitigating bias and improving fairness metrics by comparing different datasets and algorithms.

Future work will also explore new fairness metrics from literature, evaluate other datasets and compare the results. The minimax fairness criterion is relatively new, and apart from the original algorithm proposed in [5], two algorithms have been proposed by [4]. We plan to further investigate this, among other new metrics for fairness, to reduce or eliminate detrimental biases in classification and machine learning systems. It would be an excellent exercise to use the minimax fairness criteria to compare its performance with datasets from different types of computational problems and classifiers.

References

1. Hinnefeld, J. Henry, et al. "Evaluating fairness metrics in the presence of dataset bias." arXiv preprint arXiv:1809.09245 (2018).
2. M. Keans, A. Roth and Sharifi-Malvaj, "Average Individual Fairness: Algorithms, Generalization and Experiments," Proceedings of the 33rd International Conference on Neural Information Processing Systems, p. 8242–8251, 2019.
3. L. E. Celis, V. Keswani, O. Yidiz and N. K. Vishnoi, "Fair distributions from biased samples: A maximum entropy optimization framework," CoRR,, vol. vol. abs/1906.02164, 2019.
4. E. Diana, W. Gill, M. Kearns, K. Kenthapadi and A. Roth, "Minimax Group Fairness: Algorithms and Experiments," Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 66-76, 2021.
5. N. Martinez, M. Bertran and G. Sapiro, "Minimax pareto fairness: A multi objective perspective," in International Conference on Machine Learning, 2020.
6. R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd Ed, John Wiley & Sons, 2001.
7. H. Fong, V. Kumar, A. Mehrotra and N. K. Vishnoi, "Fairness for AUC via Feature Augmentation," arXiv preprint arXiv:2111.12823, 2021.
8. M. Hardt, E. Price, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning," in Advances in Neural Information Processing Systems, 2016.

9. M. Kearns, S. Neel, A. Roth and Z. Wu, "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness," Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 2564-2572, 2018.
10. M. Maschler, S. Zamir and E. Solan, Game theory, Cambridge University Press, 2020.
11. J. F. Nash Jr, "Equilibrium points in n-person games," Proceedings of the national academy of sciences, vol. 36, no. 1, 1950.
12. L. Amoroso, " “Vilfredo Pareto.”," Econometrica , vol. 1, pp. 1-21, 1938.
13. "German Credit Dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).
14. A. Agarwal, H. Agarwal and N. Agarwal, "Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems," AI and Ethics, p. 1–13, 2022.
15. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1-35, 6 July 2022.