The use of Maximum Completeness to Estimate Bias in AI-based Recommendation Systems

Alessandro Simonetta[1],*, Maria Cristina Paoletti[1] and Alessio Venticinque[2]

[1] *Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy*

[2] *Department of Electrical and Information Engineering, University of Naples Federico II, Napoli, Italy*

Abstract

The use of AI based recommendation systems, based on data analysis using Machine Learning algorithms, is taking away people's full control over decision making. The presence of unbalanced and incomplete data can cause discrimination to religious, ethnic, and political minorities without this phenomenon being easily detectable. In this context, it becomes critically important to understand what are the potential risks associated with learning with such a dataset and what consequences it may have on the outcome of decision making using Machine Learning algorithms. In this paper, we tried to identify how to measure the group fairness of a prediction of a classification algorithm, to identify the quality features of the dataset that influence the learning process, and finally, to evaluate the relationships between the quality features and the fairness measures.

Keywords

Fairness, Clustering, Machine Learning, Completeness, ISO/IEC 25012, Maximum Completeness, Metrics, Bias, Classification

1. Introduction

In 2019, the Economist [1] stated that data is an important resource comparable to oil. Moreover, Forbes [2] defines data as the fuel of the information age, and Machine Learning (ML) as the engine that uses it. These technologies in addition to the evolution of networks [3, 4] is providing opportunities to develop new applications.

Many companies and organizations are investing, increasingly, in decision-making processes centered on AI based recommendation systems to offer a variety of services ranging from marketing [5, 6] to fault diagnosis [7]. Tools that make use of these types of algorithms relieve people from making decisions that may be influenced by moods, biases and subjective thoughts, they ensure fairness and repeatability at different times too. The reasons why ML algorithms arrived at a certain type of result may not be transparent or easily understood by users. For this reason, techniques have been introduced, such as *Explenable AI*, which allow analysts to understand how a given choice was arrived at, or *Reinforcement Learning*, which allows the decision-making process to be distributed across different levels in a counterbalance way [8, 9, 10, 11].

These decision systems are based on a data-driven approach and their results are strongly influenced by the quality of the information available (balance, completeness, ...). The correctness and authenticity of the data alone [12, 13, 14, 15, 16, 17] are not enough to guarantee their quality. Thus, the presence of poor quality in the data, or a low level of representativeness, can lead to biased learning.

A very striking example of discrimination is the result of the algorithm used in Florida on predicting the risk of re-offense, brought to light by the nonprofit organization ProPublica [18]. In fact, the algorithm, which assigned each person a score indicating the likelihood of reoffending, was trained using an unbalanced dataset, and as result, black people showed greater recidivism than other ethnicities [19].

In this paper, we will present a methodology that, starting from training data, allows us to estimate the risk of getting an unfair treatment in the prediction.

To do this, it is necessary to identify how to measure the fairness of a prediction of a classification algorithm, to discover the quality characteristics of the dataset that influence the goodness of learning, and, last but not least, to study the relationships that exist between quality characteristics and the fairness measures of the ML algorithm.

The study of the fairness of ML algorithms is widely debated topic in science, in [20], [21] many performance and fairness indices are studied, such as *False Positive* and *Equalized Odds.* These metrics could be used to enhance different performance characteristics, that could be in contrast each others and each of them fit best a particular class of objectives [22],[23]. For example, there are indices that prefer accuracy and others that prefer precision [20], the right trade-off must be found between the two, depending on the problem to be solved. Indeed, in the

SYSYEM 2022: 8th Scholar's Yearly Symposium of Technology, Engineering and Mathematics, Brunek, July 23, 2022

*Corresponding author. All the authors contributed equally.
alessandro.simonetta@gmail.com (A. Simonetta);
mariacristina.paoletti@gmail.com (M. C. Paoletti)
0000-0003-2002-9815 (A. Simonetta); 0000-0001-6850-1184 (M. C. Paoletti); 0000-0003-3286-3137 (A. Venticinque)

cancer detection we prefer recall rather than precision: better to plot a healthy patient as probably ill than not to screen one who actually is.

Other studies [22, 24, 25, 26] are aimed at identifying the relationship between sensitive attributes and the target one. These show dependencies between the number of incorrect predictions (e.g., ratio of predicted positive to real positive) and the features of the dataset. For example, if it get wrong advantageously with respect to sensitive attributes, that is, by attributing more positive outcomes to them, individuals belonging to this set are considered a privileged group. Conversely, if the algorithm get wrong negatively, associating more unfavorable outcomes than should normally be indicated, that group is considered an unprivileged.

Regarding data quality aspects, we identified the international standards ISO/IEC 25012 [27] and and ISO/IEC 25024 [28] as the models from which to draw the notion of completeness. This choice was also supported by the presence of studies [29],[30] that use these standards for dataset construction and maintenance of its quality over time. In particular, we identified the notion of maximum completeness [31] as satisfying the goal we had set for ourselves.

This paper will start from the state of the art, Section II, comparing the advantages and disadvantages of different approaches to fairness and show alternative synthesis solutions that are useful in identifying critical issues in the input data. In section III, we will present our solution developed from what has been proposed in the literature and in particular we will decline it into two different versions defining its pros and cons. In section IV, we will point out to identify the limitations of the present work and what are the possible future developments. Finally, in section V, we will present concluding remarks.

2. State of Art

In general, a classification model is called fair if the mistakes are equally distributed across the different groups, identified within the sensitive attribute. The input features X of the values space are mapped onto a target variable R according to a function f(X). Where the values of R are represented by classes or a score, i.e., in a range of a scale of N different values, they can be mapped back to a binary value by defining a threshold. In order to train these classifiers, example data are used in which the input features X, are associated to truth variable Y with the real result. From the goodness of these examples derives the quality of the resulting classifier.

From now on, we will refer to A as the sensitive attribute, which can identify a minority. Although, these variables are treated individually, an underprivileged group could be identified through a combination of them too. One work including metrics for assessing fairness is [22] where *Disparate Impact* and *Demographic Parity* (*Statistical Parity*) are introduced. The first one use the ratio between $P(R=1|A=a_i)$ and $P(R=1|A=a_j)$, instead the second one is the difference between the two probabilities. Demography Parity is present in [32], too, and it is referred to as *Independence*, as it indicates the degree of independence of the target variable R compared to the sensitive attribute. Another measure of fairness reported in [22] is the *Equalized Odds* which is satisfied if the prediction is conditionally independent to the sensitive attribute, given the true value; it highlight the difference between true positive rate and false positive rate. In our work we call the latter index as *Separation*. The *Equal Opportuinity* [22], requires the true positive rates to be similar across groups. Other metrics for estimating faireness are the *Sufficiency* [32], similar to Equal Opportunity, but focuses on true values rather than predicted values, and the *Overall Accuracy Equality* [21], which tests the average error between predictions across groups.

Determine what fairness metrics are best for finding what is the right configuration of an algorithm to use in a decision support system depends on the purpose for which the it should be built and what discrimination risks it may be exposed to. Studies [23], [33] point out that it is not possible to maximize all metrics simultaneously and therefore one must choose among the features that these measures tend to enhance, such as accuracy and recall. In [25] the authors present a framework for comparing indices and highlighting when the maximization of one conflicts with that of another. Their work goes beyond analyzing individual metrics, and groups them according to their characteristics (fairness of treatment, fairness of opportunity, interest groups, sensitive attributes, etc.) and their usefulness with respect the target that the system has (i.e. support for film discovery on a streaming platform). These are clustered using a hierarchical algorithm applied to correlation between metrics to identify similar ones. The results are then diagrammed in a simplified manner through the use of Principal Component Analysis (PCA) to reduce the state space in two dimensions. In particular, the authors conclude that through the use of PCA it is possible to explain the relationships among different metrics by reducing the state space in a range from one to three component.

The idea of using balancing indices to predict the risk of discrimination can be found in [34]. In this work for the first time they use a measure of fairness applied to the sensitive attribute and not to the comparison among groups. In the next section we will start from this topic and then propose two different solutions for calculating a synthetic index related to the sensitive attribute.

3. Methodology

The first formal criterion introduced in [32] requires that the sensitive attribute A be statistically independent of the predicted value R. Assuming we use a dataset with a field A having cardinality m, $A = \{a_1, ..., a_m\}$, the random variable A is independent compared to R if and only if for each $i, j \in [1, m]$, with $i \neq j$ we have that:

$$P(R = 1|A = a_i) = P(R = 1|A = a_j) \quad (1)$$

To understand how far the two predictions deviate from the ideal case (zero difference), we can calculate the distance between two probabilities:

$$\mathfrak{U}(a_i, a_j) = |P(R = 1|A = a_i) - P(R = 1|A = a_j)| \quad (2)$$

To get a synthesis value of the non-independence between A and R [34], the arithmetic mean of the distances can be considered:

$$\mathfrak{U}(a_1, .., a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathfrak{U}(a_i, a_j) \quad (3)$$

Instead of using the equation 2, some authors [22] apply a different notion of independence:

$$\varepsilon_i = |P(R = 1|A = a_i) - P(R = 1|A \neq a_i)| \quad (4)$$

What we have seen so far fails to explain whether there are groups, within the sensitive attribute, that undergo the same mode of treatment nor the presence of discrimination among groups, the present work was born from this reflection. For explaining our idea we prefer to use an example previously mentioned.
The sensitive attibute A=*Race*, of *Compas* dataset, contains six different ethnicities shown in the first column of the table 1.

Table 1
Probability for Sensitive Attribute Race

$A = a_i$	$P(R = 1\|A = a_i)$	Centroid
Caucasian	0.33	0.26
Hispanic	0.28	
Other	0.20	
Asian	0.23	
African-American	0.58	0.65
Native-American	0.73	

The study [19], has determined that African-Americans are the unprivileged group compared to the rest of the other ethnic groups (Native-American, Caucasian, Asian, Mexican, Other).
The synthetic index described by the equation 3 can show on average how much a sensitive attribute is at risk of discrimination, but might underestimate the inequity. With reference to the dataset *Compas*, in the table 1 it is observed that the values in the second column $P(R = 1|A = a_i)$ cluster around 0.26 and 0.65. Figure 1 shows graphically the values of the table 1, with evidence of the centroids identified through the K-Means algorithm. The difference between the value calculated using the equation 3 (P=0.24) and the value obtained by considering centroids (P=0.39) turns out to be 0.15 points. This demonstrates what was stated earlier with respect to the use of a central tendency index. However, the equation 3 can be used in the presence of more than two clusters.

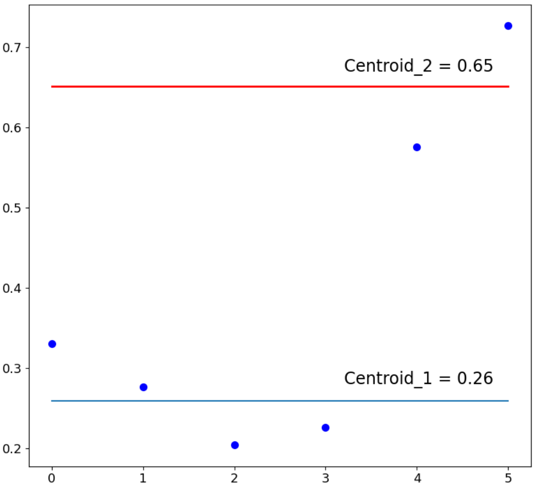
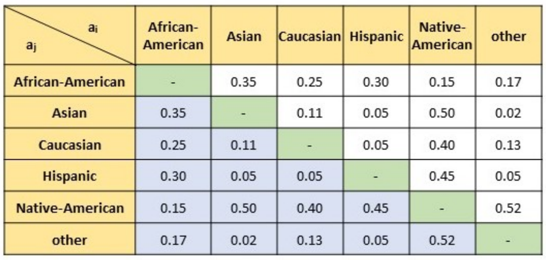


Figure 1: Scatter Plot: probability of A=Race and K-Means centroid



a_i a_j	African-American	Asian	Caucasian	Hispanic	Native-American	other
African-American	-	0.35	0.25	0.30	0.15	0.17
Asian	0.35	-	0.11	0.05	0.50	0.02
Caucasian	0.25	0.11	-	0.05	0.40	0.13
Hispanic	0.30	0.05	0.05	-	0.45	0.05
Native-American	0.15	0.50	0.40	0.45	-	0.52
other	0.17	0.02	0.13	0.05	0.52	-

Figure 2: Probability of $\mathfrak{U}(a_i, a_j)$ with A=Race for couples in Equation 2

In the following, we will illustrate alternative methods for calculating different synthetic fairness indices that allow for greater sensitivity to discriminatory situations. Next, we will try to identify the relationship that exists between the dateset completeness index and the identified fairness indices. This will allow us to anticipate the risks of bias arising from incomplete data.

3.1. Dataset

In this section are present the list of the dataset used for the sperimentation:

- COMPAS Recidivism Dataset [19];
- Recidivism in juvenile justice [35]
- UCI Statelog German Credit [36];
- default of credit card clients Data Set [37];
- Adult Data Set [38];
- Student Performance Data Set [39].

The sensitive attributes are listed in the following table 2

Table 2
Datasets and Sensitive Attributes

Dataset	Attribute	Cardinality
Compas	Race	6
	Sex	2
	Age	3
Juvenile	V3_nacionalitat	35
	V2_estranger	2
	V1_sexe	2
	V5_edat_fet_agrupat	3
	V4_nacionalitat_agrupat	5
	V8_edat_fet	5
UCI	Sex	2
	Education	7
Income	Education	16
	Race	5
	Sex	2
	Native country	41
Statelog	Status	4
	Sex	2
	foreignworker	2
Student	Sex	2
	Age	6
	mMather job	5
	Father job	5
	Mother Education	5
	Father Education	4

For each dataset we have calculated the predicted value using a classification model, but only *Compas Dataset* and *Recidivism in juvenile justice* have already this information, so we have used the original one. Different models for classification are present in literature and are implemented in software libraries. We chose the logistic regression [40] that offers categorical results and therefore can be trained to predict the membership of an item in a class.

In order to evaluate the completeness of the dataset we examined the *Max Completeness*, introduced in [41]. Maximum completeness is an index measuring the percentage degree of completeness of the dataset (*Incomplete*=0, *Complete*=1) with respect to one or more categorical attributes, when the expected value is that in which the attributes considered have a number of replications equal to that of the predominant. Assuming we wish to calculate the completeness of a dataset on a categorical attribute A:

$$C_{MAX}\ (A) = \frac{N}{K_A \cdot M_{PC}} \qquad (5)$$

Where:

- N is the total number of instances in dataset;
- K_A is the number of classes of attribute A;
- M_{PC} is the maximum number of elements of a class of attribute A.

This index could be calculated on multiple categorical attributes by considering as K_A the number of possible combinations of the chosen categorical variables and as M_{PC} the maximum number of items grouped over the number of attributes considered. For example, considering the Compas dataset and the attributes *Race* and *Sex*, to have the maximum completeness $C_{MAX}\ (Race, Sex) = 1$, we must have a number of items for all combinations of race and sex equal to 2,626, that is the number of items of male and African-American ethnicity, which is the category most numerous. To do this the number of records within the dataset must increase to 31,512 from the current N=6,172 ($C_{MAX}\ (Race, Sex) = 0.19$).

3.2. Clustering Method

Starting from the consideration that the conditional probabilities of the random variable R with respect to membership in a sensitive attribute ($P(R = 1|A = a_i)$) may determine affinity in treatment equivalence classes, we thought of using unsupervised clustering algorithms to identify possible clusters in the probabilities. Among the ML algorithms that were analyzed, we chose *K-Means* and *DBSCAN* [42]. K-Means is a clustering algorithm that tends to separate samples into K groups with equal variance, minimizing the within-cluster sum-of-squares criterion. To use this method, it is necessary to know a priori the K number of clusters into which to divide the samples. Once the centroids have been calculated, it is possible to use the equation 2 or the 3 depending on the value of K to obtain the synthetic index. Compared with [34], bundling multiple instances of the sensitive

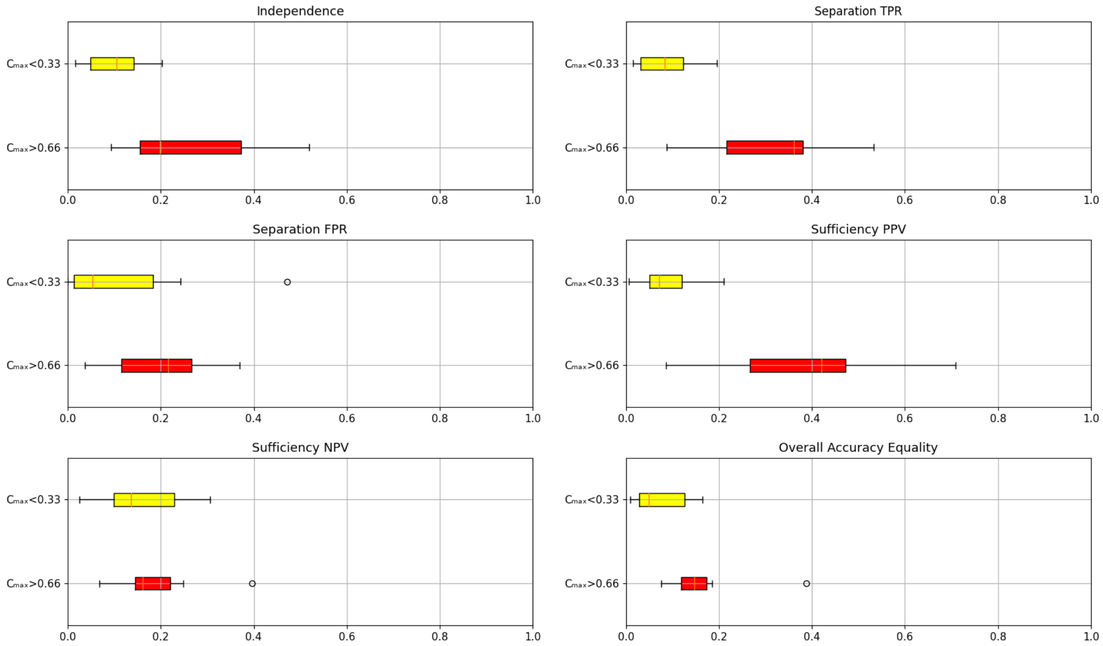


Figure 3: K-Means method, C_{MAX} minor of 0.33 and major of 0.66

attribute into groups results in a lower m-number, indeed, the term a_i is composed of all elements treated similarly. The critical issue of correctly identifying the K-number to be used for clustering led us to study other approaches and subsequent experimentation with DBSCAN. This sees clusters as areas of high density separated by areas of lower density. The application of such an algorithm needs only the parameters indicating the number of samples found in the area forming a cluster and the one indicating the density required to form a cluster (*eps*). One of its limitations is that some elements turn out not to belong to any cluster; it was decided to treat them as unitary clusters. Since the concept of a centroid does not exist for DBSCAN, the value of the fairness metric was calculated as follows:

- clusters: probability as per equation 3 to calculate cluster fairness;
- individual elements: average of the fairness indices of individual instances of the sensitive attribute.

Applying the two clustering methods resulted in values that were higher on average than those calculated using only the arithmetic mean reported in the equation 3. Furthermore, this made it possible to identify groups that were similar in treatment type.

3.3. MinMax Method

Although both clustering methodologies gave good results to work on, another approach was explored starting from the definition found in [22] to calculate the value of the fairness metric trying to find the worst case. The process is described below referring to the Demographic Parity metric (equation 4) (Independence), but without loss of generality can be extended to all. This description places emphasis on the fact that a_i is considered as an unprivileged group and the set of all other elements as a privileged group. The algorithm, for all values of the sensitive attribute A, calculates the result of the equation 4 associated with each group by considering from time to time the element under observation as discriminated and all others as privileged. Considering, again, the field *Race*, if the element we are calculating for is *Asian*, this is a_i and all others constitute the other group in the equation. Once we have iterated this process for all values of the sensitive attribute we will go on to select the highest and lowest result. The former is the group for which the predicted variable R is most dependent on ethnicity, while the latter is the most independent. The difference between these two values indicates how large the inequality of treatment between the privileged and unprivileged group is, relative to the sensitive attribute considered.
Compared with the use of clustering, this methodology

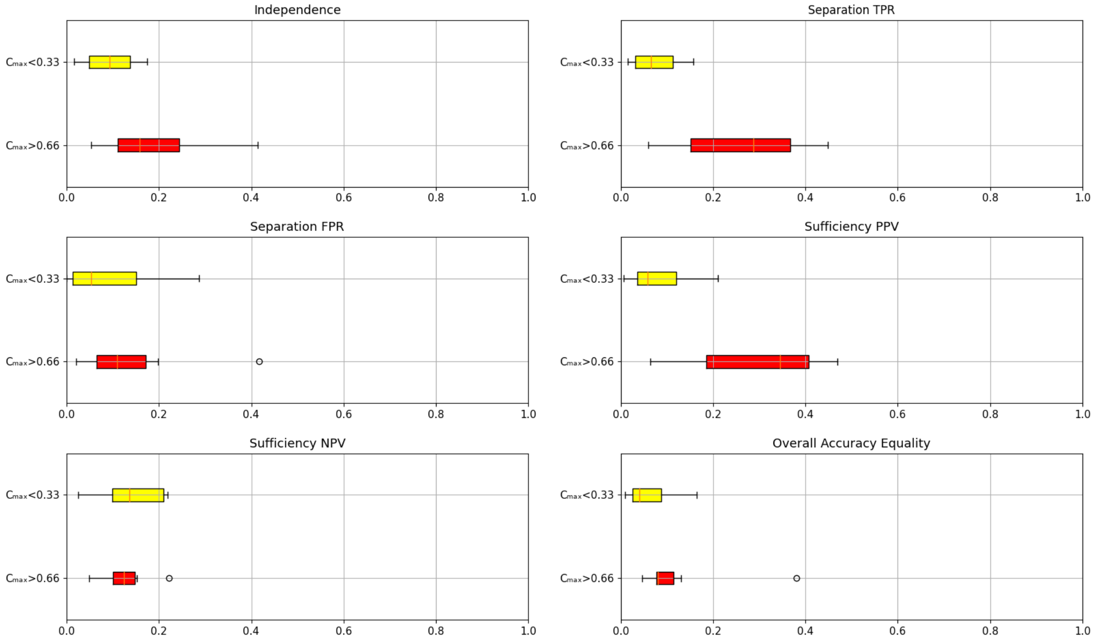


Figure 4: DBSCAN method, C_{MAX} minor of 0.33 and major of 0.66

arises in the worst case by considering as the index of treatment disequity the largest difference between the ε_i obtained by applying the equation 4 and reported in the table 3.

Table 3
Conditional Probabilities of Sensitive Attribute and ε_i by equation 4

$A = a_i$	$P(R = 1 \vert A = a_i)$	ε_i
Caucasian	0.33	0.22
Hispanic	0.28	0.18
Other	0.20	0.11
Asian	0.23	0.14
African-American	0.58	0.51
Native-American	0.73	0.64

ε_i=$|P(R = 1|A = a_i) - P(R = 1|A \neq a_i)|$

4. Discussion

During the experimental phase, the methods presented in the previous paragraphs were applied to sensitive attributes belonging to six known datasets. The fairness metrics used during the testing phase are: Independence, Separation, Sufficiency and Overall Accuracy Equality as defined in [32],[21]. Separation and Sufficiency were calculated by considering positive and negative predicted cases separately. The metrics resulting from this split are: *Separation TPR* (True Positive Rate), *Separation FPR* (False Positive Rate), *Sufficiency PPV* (Positive Predictive Value) and *Sufficiency NPV* (Negative Predictive Value). Once the results were evaluated for the six chosen fairness metrics, we related them to the maximum completeness balancing index.

Each diagram consists of two box plots containing the values of sensitive attributes divided in this way: Maximum Completeness values less than 0.33 (low risk in yellow) and greater than 0.66 (high risk in red). Intermediate values are not reported because it is more difficult for them to determine whether they are fair or not. We remarks that Fairness metrics, as defined, take value in the range [0,1] (*Fair*=0, *Unfair*=1).

Figure 3 shows the case where the six fairness metrics are calculated using the K-means technique. Note that all boxplots tails overlap, while for the body remains a clear separation for Separation TPR and Sufficiency PPV. The worst case is for the Sufficiency NPV metric where there is total overlap.

Figure 4 shows the results of applying the DBSCAN technique. In this case the values obtained are similar to the K-means, although there are worst results for Overall Accuracy Equality and NPV Sufficiency. Such plots were not optimal in 3 too.

Finally, in figure 5 the method of MinMax as for equation 4 is applied instead of the clustering algorithms. In these

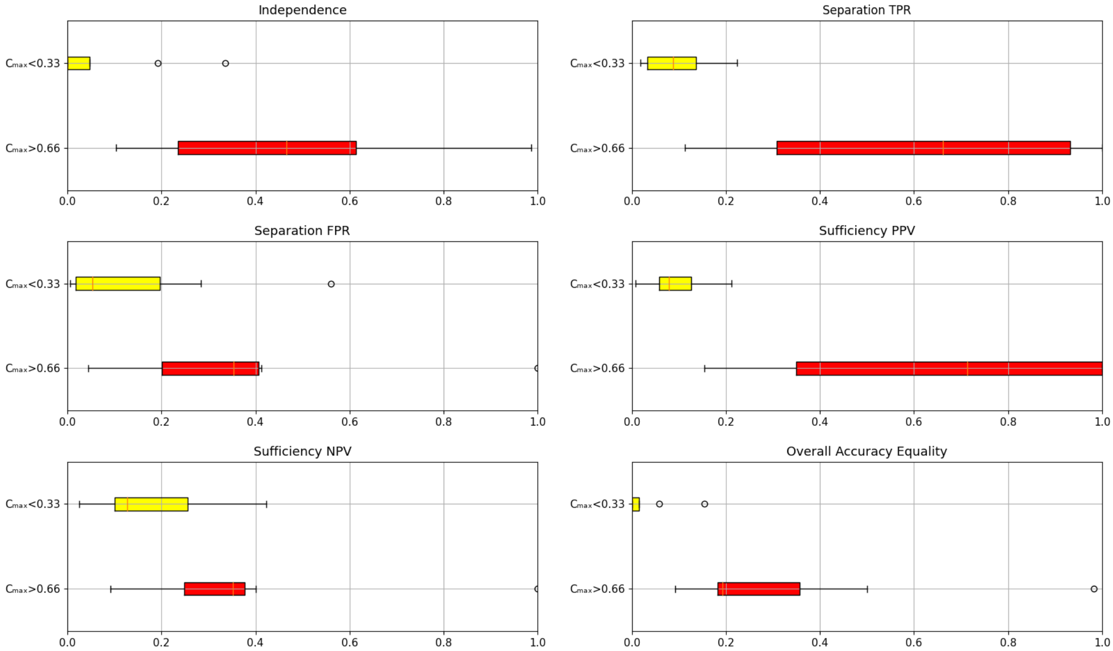


Figure 5: MINMAX method, C_{MAX} minor of 0.33 and major of 0.66

diagrams, one can see a lengthening of the boxplots related to the risk cases and a sharper separation between the two boxplots for all diagrams. In the Independence and OAE cases, there is no more overlap between the tails. In the PPV Sufficiency the high risk cases tend to one and almost full separation between values is achieved. For sensitive attributes that have a C_{MAX} greater than 0.66, there are values of fairness metrics greater than 0.2. In conclusion, the MinMax method yielded better results compared to clustering, but conversely, using methods such as K-Means and DBSCAN can help to better define treatment similarities among groups.

5. Current Limits and Future Works

The experimentation carried out has yielded encouraging results for what is the separation of high and low risk fairness forecast with respect to C_{MAX}. Although the result obtained shows values that are sometimes in a range that is not very wide, improvements have already been identified that can be investigated in future work. The first task is the choice of clustering method and parameters that affect the number of clusters. Having few clusters brings us closer to the worst case, and it becomes more easy to identify the correct meaning to give to each clusters, depending on their distance and their shape. However, establishing a number of clusters that forces such a division could join groups that are not actually treated equally. Another point we will investigate is the possibility of changing the algorithm to identify the synthetic value of the cluster for the metric under consideration. One solution, we will explore, is to integrate the calculation of the difference between maximum and minimum in a clustering algorithm that does not have a predefined number K of clusters, such as the already presented DBSCAN. Related to this algorithm, alternatives on how to treat elements that are not associated with any cluster will be explored. The question is: should these cases discarded because they are outliers or do they represent borderline cases, e.g., a highly discriminated minority?

6. Conclusion

The spread of ML algorithms for constructing decision systems make the data used in their construction increasingly important. Imbalances or biases that may be present within the information can affect the results of such systems, causing discrimination toward certain groups.
The use of the fairness metrics that have been presented becomes important to predict the impact related to such biases and go to act accordingly on both algorithms and input data in line with the objectives.

In this work we tried to provide a methodology to identify similar clusters by treatment type and to calculate a synthetic index that could predict how at risk the system is with respect to sensitive attributes. The experimentation carried out provided good results both in terms of identifying agglomerations that undergo similar treatments and in calculating a parameter that would give a conservative assessment of the metric.

The relationship between the maximum completeness index and the fairness indices calculated by the showed methods provided a guideline in order to recognize high-risk and lower-risk sensitive attributes. This will give to the analysts the information to better configure classification algorithms.

The results of this work lay the groundwork for future developments aimed at improving the identification of groups within sensitive attributes and researching alternative synthetic indices that will have a greater precision.

References

[1] The Economist, The world's most valuable resource is no longer oil, but data, The Economist, USA (6th May 2019).

[2] B. Marr, The 5 biggest data science trends in 2022, Oct 2021. URL: https://www.forbes.com/sites/bernardmarr/2021/10/04/the-5-biggest-data-science-trends-in-2022/?sh=22f5fc1d40d3(AccessedMay,2022).

[3] R. Giuliano, The next generation network in 2030: Applications, services, and enabling technologies, in: 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2021, pp. 294–298. doi:10.23919/EECSI53397.2021.9624241.

[4] C. Napoli, G. Pappalardo, E. Tramontana, A hybrid neuro-wavelet predictor for qos control and stability, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8249 LNAI (2013) 527–538. doi:10.1007/978-3-319-03524-6_45.

[5] S. Verma, R. Sharma, S. Deb, D. Maitra, Artificial intelligence in marketing: Systematic review and future research direction, International Journal of Information Management Data Insights 1 (2021) 100002. doi:https://doi.org/10.1016/j.jjimei.2020.100002.

[6] G. Capizzi, G. Lo Sciuto, C. Napoli, E. Tramontana, An advanced neural network based solution to enforce dispatch continuity in smart grids, Applied Soft Computing Journal 62 (2018) 768 – 775.

[7] T. Dhomad, A. Jaber, Bearing fault diagnosis using motor current signature analysis and the artificial neural network 10 (2020) 70–79.

[8] M. Matta, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, A. Nannarelli, M. Re, S. Spanò, A reinforcement learning-based qam/psk symbol synchronizer, IEEE Access 7 (2019) 124147–124157. doi:10.1109/ACCESS.2019.2938390.

[9] R. Brociek, G. Magistris, F. Cardia, F. Coppa, S. Russo, Contagion prevention of covid-19 by means of touch detection for retail stores, in: CEUR Workshop Proceedings, volume 3092, CEUR-WS, 2021, pp. 89–94.

[10] L. Canese, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, S. Spanò, Multi-agent reinforcement learning: A review of challenges and applications, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/11/4948. doi:10.3390/app11114948.

[11] N. Brandizzi, S. Russo, R. Brociek, A. Wajda, First studies to apply the theory of mind theory to green and smart mobility by using gaussian area clustering, volume 3118, CEUR-WS, 2021, pp. 71–76.

[12] F. Fallucchi, M. Gerardi, M. Petito, E. De Luca, Blockchain framework in digital government for the certification of authenticity, timestamping and data property, 2021. doi:10.24251/HICSS.2021.282.

[13] N. Brandizzi, V. Bianco, G. Castro, S. Russo, A. Wajda, Automatic rgb inference based on facial emotion recognition, in: CEUR Workshop Proceedings, volume 3092, CEUR-WS, 2021, pp. 66–74.

[14] B. Nowak, R. Nowicki, M. Woźniak, C. Napoli, Multi-class nearest neighbour classifier for incomplete data handling, in: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), volume 9119, Springer Verlag, 2015, pp. 469–480. doi:10.1007/978-3-319-19324-3_42.

[15] D. Połap, M. Woźniak, C. Napoli, E. Tramontana, R. Damaševičius, Is the colony of ants able to recognize graphic objects?, Communications in Computer and Information Science 538 (2015) 376–387. doi:10.1007/978-3-319-24770-0_33.

[16] S. Illari, S. Russo, R. Avanzato, C. Napoli, A cloud-oriented architecture for the remote assessment and follow-up of hospitalized patients, in: CEUR Workshop Proceedings, volume 2694, CEUR-WS, 2020, pp. 29–35.

[17] N. Dat, V. Ponzi, S. Russo, F. Vincelli, Supporting impaired people with a following robotic assistant by means of end-to-end visual target navigation and reinforcement learning approaches, in: CEUR Workshop Proceedings, volume 3118, CEUR-WS, 2021, pp. 51–63.

[18] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine

bias : There's software used across the country to predict future criminals. and it's biased against blacks., https://www.propublica.org/ (2016).

[19] J. Larson, S. Mattu, L. Kirchner, J. Angwin, Compas recidivism dataset, 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm/(AccessedOct,2021).

[20] J. Lee, Analysis of precision and accuracy in a simple model of machine learning, Journal of the Korean Physical Society, 2017, p. 866–870. doi:10.3938/jkps.71.866.

[21] A. Carey, X. Wu, The statistical fairness field guide: perspectives from social and formal sciences, AI and Ethics (2022) 1–23. doi:10.1007/s43681-022-00183-3.

[22] D. Pessach, E. Shmueli, Algorithmic fairness, 2020. URL: https://arxiv.org/abs/2001.09784. doi:10.48550/ARXIV.2001.09784.

[23] S. Prince, Bias and fairness in ai, 2019. URL: https://www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai/(accessedMar, 2022).

[24] G. Capizzi, G. Lo Sciuto, C. Napoli, M. Woźniak, G. Susi, A spiking neural network-based long-term prediction system for biogas production, Neural Networks 129 (2020) 271 – 279.

[25] M. Miron, S. Tolan, E. Gómez, C. Castillo, Addressing multiple metrics of group fairness in data-driven decision making, 2020. URL: https://arxiv.org/abs/2003.04794. doi:10.48550/ARXIV.2003.04794.

[26] G. Capizzi, G. Lo Sciuto, C. Napoli, R. Shikler, M. Wozniak, Optimizing the organic solar cell manufacturing process by means of afm measurements and neural networks, Energies 11 (2018).

[27] International Organization for Standardization, "ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model", 2008. URL: https://www.iso.org/standard/35736.html(accessedJan,2021).

[28] International Organization for Standardization, "ISO/IEC 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality", 2015. URL: https://www.iso.org/standard/35749.html(accessedJan,2022).

[29] J. Calabrese, S. Esponda, P. M. Pesado, Framework for Data Quality Evaluation Based on ISO/IEC 25012 and ISO/IEC 25024, in: VIII Conference on Cloud Computing, Big Data & Emerging Topics, 2020. URL: http://sedici.unlp.edu.ar/handle/10915/104778.

[30] F. Gualo, M. Rodriguez, J. Verdugo, I. Caballero, M. Piattini, Data quality certification using iso/iec 25012: Industrial experiences, Journal of Systems and Software 176 (2021) 110938. URL: https://www.sciencedirect.com/science/article/pii/S0164121221000352. doi:https://doi.org/10.1016/j.jss.2021.110938.

[31] A. Simonetta, A. Trenta, M. C. Paoletti, A. Vetrò, Metrics for identifying bias in datasets, SYSTEM (2021).

[32] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning, 2020. URL: https://fairmlbook.org/(AccessedSept,2021), chapter: Classification.

[33] K. Burkholder, K. Kwock, Y. Xu, J. Liu, C. Chen, S. Xie, Certification and trade-off of multiple fairness criteria in graph-based spam detection, Association for Computing Machinery, 2021, p. 130–139. doi:10.1145/3459637.3482325.

[34] A. Vetrò, M. Torchiano, M. Mecati, A data quality approach to the identification of discrimination risk in automated decision making systems, Government Information Quarterly 38 (2021) 101619. URL: https://www.sciencedirect.com/science/article/pii/S0740624X21000551. doi:https://doi.org/10.1016/j.giq.2021.101619.

[35] Department of Justice, Recidivism in juvenile justice, 2016. URL: https://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html(AccessedOct,2021).

[36] D. H. Hofmann, Uci statelog german credit, 1994. URL: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)(AccessedOct,2021).

[37] I.-C. Yeh, default of credit card clients data set, 2016. URL: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)(AccessedOct, 2021.

[38] R. Kohavi, B. Becker, Adult data set, 1996. URL: https://archive.ics.uci.edu/ml/datasets/adult(AccessedOct,2021).

[39] P. Cortez, Student performance data sett, 2014. URL: https://archive.ics.uci.edu/ml/datasets/student+performance(AccessedOct,2021).

[40] scikit-learn developers, Logistic regression, 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html(AccessedMay,2022).

[41] A. Simonetta, A. Vetrò, M. C. Paoletti, M. Torchiano, Integrating square data quality model with iso 31000 risk management to measure and mitigate software bias, CEUR Workshop Proceedings (2021) pp. 17–22.

[42] scikit-learn developers, Algorithmic fairness, 2022. URL: https://scikit-learn.org/stable/modules/clustering.html(AccessedMay,2022).