AHP–Gaussian To Enhance Model Selection Based On Multiple Fairness Criteria

Diego Minatel[1], **Nícolas Roque dos Santos**[1], **Vinícius Ferreira**[1], **Mateus Modesto**[2]

[1]Institute of Mathematics and Computer Science
University of São Paulo, São Carlos, Brazil

[2]PECEGE Institute, Piracicaba, Brazil

{dminatel, nrsantos, vfsilva}@usp.br, profmateusmodesto@gmail.com

Abstract. *The challenge of developing impartial models that minimize the propagation of unfair predictions is directly linked to optimizing multiple fairness concepts. Therefore, identifying which model best combines these concepts is essential for promoting fairness in machine learning. The field of Multi-Criteria Decision Analysis addresses similar issues by developing techniques for choosing the best alternative in complex problems. One standout method is AHP–Gaussian, which, through the Gaussian factor, defines the relevance of each criterion used in decision-making. This eliminates any human factor in weighing the criteria's importance, making it an excellent alternative in the fairness-aware model selection task. To the extent of our knowledge, no study in the literature has proposed this approach before. This paper handles this gap and proposes applying AHP–Gaussian to select fairer models in classification tasks involving people. According to the results, AHP–Gaussian is more effective at selecting classifiers that balance predictive power and maximization of distinct fairness concepts than traditional multi-criteria methods.*

1. Introduction

Incorporating fairness notions into the machine learning (ML) process has proven to be adequate in preventing the propagation of discriminatory effects in society through ML-supported decision-making systems. In classification tasks, a group fairness analysis is commonly conducted to avoid disproportionate outcomes among different sociodemographic groups [Barocas et al. 2023, Caton and Haas 2020, Mehrabi et al. 2021].

This type of analysis assesses whether the predictive outcomes among the different groups satisfy group fairness notions such as demographic parity, equal opportunity, and equalized odds [Dwork et al. 2012, Hardt et al. 2016]. For example, positive prediction rates must be the same for all groups analyzed to fulfill demographic parity. This behavior is desirable, especially in applications like recruiting systems, to avoid disadvantaging a particular group in the recruiting process [Barocas et al. 2023].

Frequently, it is unclear which fairness notion best aligns with a specific application, and at times, various metrics may need to be optimized [Minatel et al. 2023b]. In such cases, a classifier must be evaluated according to different fairness metrics in addition to predictive performance measures. In this context, the model selection task

becomes much more complex, involving multiple criteria to evaluate and determine the best classifier [Black et al. 2022].

The field of research in Multi-Criteria Decision Analysis tackles this challenge by providing decision-makers with methods to make and justify their choices when faced with complex problems by evaluating different points of view and criteria [Aruldoss et al. 2013]. Therefore, methods developed in this area can be a promising path for fairness-aware model selection. The well-known Analytic Hierarchy Process (AHP) [Saaty 2008] stands out among the methods developed. It is widely applied in business, management, economics, ecology, and social studies, among many other areas [Adem Esmail and Geneletti 2018, Aruldoss et al. 2013, Darko et al. 2019, Podvezko 2009].

In the AHP method, the criteria weights are manually defined by a decision-maker using the Saaty scale [Saaty 2008]. This approach is well-suited when there is extensive knowledge of the problem, as it allows for precise assignment of the importance level to each characteristic selected for evaluation. However, in most cases, these weights are difficult to assign manually, leading to overestimation or underestimation of the importance of a specific criterion. This particularity makes AHP less suitable as a multi-criteria method for model selection.

To address this issue, [Dos Santos et al. 2021] proposed an evolution of the AHP method, called AHP–Gaussian, which solves the problem of obtaining weights manually. Specifically, this method defines the weights of the multiple criteria used in decision-making based on data calculated through a Gaussian factor. With the elimination of human influence, AHP-Gaussian becomes highly attractive for large-scale applications in fairness-aware model selection, bringing the expertise of the AHP method in choosing the best alternative for complex problems within machine learning. To the extent of our knowledge, this approach has not been explored in the literature before. Therefore, this paper handles this gap and proposes applying AHP–Gaussian to enhance model selection based on multiple fairness criteria.

This work contributes twofold. First, it transposes AHP–Gaussian concepts to the machine learning domain. Second, it introduces a novel model selection method based on multiple criteria. Our findings suggest that AHP–Gaussian is more effective in selecting classifiers that combine strong predictive performance with greater fairness in their decisions than the tested methods, as demonstrated by the results of several metrics. Consequently, by applying the AHP–Gaussian to model selection, we can promote fairer results that help to reduce the spread of discriminatory effects in our society.

2. Background and Related Work

This section presents the terminology, fundamental concepts, and related works required to comprehend our proposed method and the adopted experimental setup.

2.1. Group Fairness Analysis

The primary goal of group fairness analysis is to identify that a model does not produce asymmetric prediction results for different sociodemographic groups derived from protected attributes[1]. In the Fairness in Machine Learning literature, these sociodemographic

[1]Protected attributes are characteristics that encompass sensitive information, such as gender and race.

groups are commonly divided into two analysis groups: privileged and unprivileged. The unprivileged group consists of those who have historically received disadvantaged treatment, and its composition varies according to the case study [Barocas et al. 2023, Mehrabi et al. 2021].

There are three main group fairness notions adopted in binary classification tasks: demographic parity, equal opportunity, and equalized odds. A model satisfies the concept of demographic parity if it has equal positive prediction rates for both privileged and unprivileged groups [Dwork et al. 2012]. In contrast, equality of opportunity requires parity for these groups in the recall score [Hardt et al. 2016]. Finally, the classifier must achieve equivalence in both true positive and false positive rates across the analyzed groups to satisfy equalized odds [Hardt et al. 2016].

However, satisfying any of these fairness notions is very challenging, and according to [Chouldechova 2017], it is impossible to achieve all these concepts simultaneously. Therefore, a more practical way to assess whether a model is fair or unfair is to convert these concepts into group fairness measures, with each metric's score indicating how far the classifier is from achieving the associated concept.

To perform this conversion, we typically calculate the measure score associated with a specific fairness notion (*e.g.*, recall in the case of equal opportunity) for both the privileged and unprivileged groups. Then, we compute the ratio between these scores, ensuring that the higher score is placed in the denominator so that the result falls between 0 and 1, making it easier to interpret. A value of 1 indicates that the classifier has fully achieved the evaluated fairness concept, while a score close to 0 means that the model is further from satisfying this concept. For instance, a classifier with a recall score of 0.80 for the privileged group and 0.60 for the unprivileged has a metric score associated with equal opportunity equivalent to $\frac{0.60}{0.80} = 0.75$.

A key point is conducting a similar analysis for the metric used to assess the classifier's predictive performance. For example, in this work, where we applied the Macro F1-Score, it is also crucial to calculate the Macro F1-Score for both groups and then compute the ratio, as aforementioned. Table 1 presents the acronyms of the group fairness measures applied in this study.

Table 1. Group fairness measures: the ratio is calculated between the scores of the privileged and unprivileged groups, with the lower score always used as the denominator.

Acronym	Description	Value Range	Ideal Value
RDP	The ratio of scores relative to demographic parity	$[0, 1]$	1
REO	The ratio of scores relative to equal opportunity	$[0, 1]$	1
RDO	The ratio of scores relative to equalized odds	$[0, 1]$	1
RMF1	The ratio of Macro F1-Score	$[0, 1]$	1

Group fairness analysis is important for achieving more impartial results and must be integrated into the learning process when the model's decisions could impact people's lives. However, even if fairness concepts are satisfied, this does not guarantee that the classifier is fair, as contradictory as it may seem. It is equally necessary for the model to have good predictive power since a model with a high error rate, even if it is equal for both groups, is considered unfair. Thus, the main difficulty in building a classifier

is to balance good predictive performance with adequate scores in several fairness measures [Barocas et al. 2023].

2.2. Related Work

In recent years, many studies have proposed different methods to incorporate fairness concepts into the machine learning process [Mavrogiorgos et al. 2024]. In [Celis et al. 2019, Narasimhan 2018], the authors introduced methods that apply one or more fairness concepts as constraints in optimizing the classification algorithm's objective function. Other works [Calmon et al. 2017, Minatel et al. 2023c, Minatel et al. 2023d] have applied preprocessing techniques to reduce discriminatory bias present in the data. Meanwhile, studies such as [Hardt et al. 2016, Mishler et al. 2021, Pleiss et al. 2017] presented postprocessing methods that adjust classifier predictions to make them more impartial.

In the context of fairness-aware model selection, [Minatel et al. 2023e] proposed a one-criterion method based on Differential Item Functioning. In this method, classifiers are modeled as test items, and selection is performed using the area method, where the classifier that produces the smallest ABC (Area delimited Between Classification characteristic curves) is considered the most impartial.

As discussed in Section 2.1, multi-criteria evaluation is often necessary, as selecting a fair classifier involves considering various fairness measures in addition to predictive performance metrics, making the model selection process more complex [Black et al. 2022]. The AHP method has been incorporated into a framework to determine which fairness criteria are most important to evaluate for a specific machine learning application [Zhang et al. 2020]. However, as discussed in Section 1, using AHP for model selection is impractical in most cases since the importance of the criteria is defined manually. In this context, the multi-criteria method proposed by [Parmezan et al. 2017], called MCPM, has been applied to select more impartial classifiers [Minatel et al. 2023d, Minatel et al. 2023e].

3. Proposed Method

This section presents our proposal for applying AHP–Gaussian [Dos Santos et al. 2021] in model selection based on multiple fairness criteria.

The first step in applying AHP–Gaussian to model selection is determining the decision matrix $\mathbf{X}$. In this paper's approach, matrix $\mathbf{X}$ comprises m trained models evaluated on n measures, where each value x_{ij} indicates the result of model i on measure j. This matrix serves as the foundation for choosing the best alternative. Table 2 presents an example of the decision matrix $\mathbf{X}$.

Table 2. Example of decision matrix $\mathbf{X}$, where the value x_{ij} represents the result of trained model i on measure j.

model	measure 1	measure 2	$\cdots$	measure n
1	x_{11}	x_{12}	$\cdots$	x_{1n}
2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	x_{m1}	x_{m2}	$\cdots$	x_{mn}

The score value for each measure in matrix $\mathbf{X}$ must range between 0 and 1. To ensure that all measures have the same ideal value, a value adjustment to metrics where

the ideal value is 0, such as the false positive rate, is necessary. Therefore, for a measure p with this characteristic, the adjusted value $\forall i$ is $x'_{ip} = 1 - x_{ip}$. After this step, all measures in the decision matrix have 1 as the ideal value.

Next, we compute the matrix $\tilde{\mathbf{X}}$, which represents the normalized version of matrix $\mathbf{X}$ with unit sum, where each value $\tilde{x}_{ij}$ is given by Equation 1:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sum_{k=1}^{m} x_{kj}} \tag{1}$$

The Gaussian factor determines the importance of each metric based on the data from matrix $\tilde{\mathbf{X}}$. Thus, we calculate each measure's Gaussian factor according to Equation 2, where $\bar{\tilde{x}}_j$ and σ_j are the mean and standard deviation, respectively, of the values of measure j in $\tilde{\mathbf{X}}$.

$$g_j = \frac{\sigma_j}{\bar{\tilde{x}}_j} \tag{2}$$

Finally, after determining the Gaussian factors and the matrix $\tilde{\mathbf{X}}$, we calculate the final score s_i of each model i using Equation 3.

$$s_i = \sum_{k=1}^{n} \tilde{x}_{ik} \times g_k \tag{3}$$

Our proposed method selects the model with the highest s_i. Probability normalization can be applied to the scores to enhance the selection process interpretability, providing a better indication of each model's likelihood of being chosen.

4. Experiments

We designed the experimental protocol to evaluate different multi-criteria methods in fairness-aware model selection. Figure 1 provides an overview of this protocol. Firstly, a preprocessed dataset is divided into training and testing sets. Next, for a given classification algorithm and a range of hyperparameter values (detailed in Section 4.2), we performed five-fold cross-validation. In this study, we chose $k = 5$ in cross-validation because some datasets have few examples (Section 4.1). Moreover, due to the data imbalance, we employed stratified sampling by class and group both in the train-test split and during cross-validation [Minatel et al. 2023a].

After cross-validation, we computed the average of the following measures: Macro F1-Score, RMF1, RDP, REO, and RDO. We chose Macro F1-Score to assess classifier performance and RMF1 to evaluate the disparity in the Macro F1-Score between privileged and unprivileged groups. RDP, REO, and RDO were selected because they are the main fairness group metrics. We employed the average values of these five measures as input for the following multi-criteria methods:

- **Sum of All Criteria (SAC):** is a simple method used as the baseline for the results. We calculated the SAC score by summing the values of each input criterion, ensuring that each criterion has the same weight in the final score;

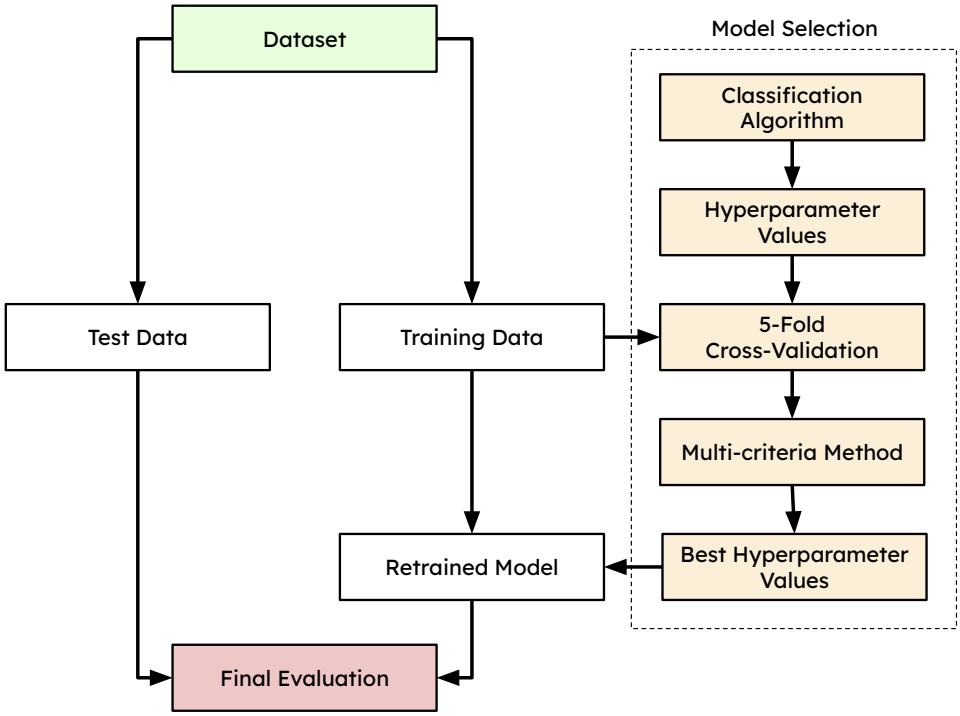

Figure 1. Overview of the adopted experimental protocol. For each tuple {*training data, classification algorithm, multi-criteria method*}, we retrained the selected model using the entire training data and evaluated it on the test data.

- **Multi-Criteria Performance Measure (MCPM):** is the method proposed by [Parmezan et al. 2017], where each metric represents a distinct axis. MCPM calculates the area of the triangle formed by each pair of measures, and its final score is obtained by summing these areas.
- **AHP–Gaussian**: is our proposed method in Section 3.

The five measures used as input for the multi-criteria methods have an ideal score of 1. Therefore, the AHP–Gaussian, MCPM, and SAC methods select the classifier with the highest score in their respective results.

At the end of the model selection stage, we selected one classifier for each tuple {*training data, classification algorithm, multi-criteria method*} for evaluation on the test set. The evaluation is conducted individually for each measure incorporated into the multi-criteria methods. To statistically compare the results, we applied Friedman's non-parametric hypothesis test for paired data, followed by multiple comparisons using the Nemenyi *post-hoc* test, with a significance level of 5% [Demšar 2006].

We have made the source code for this study in Python, along with benchmark datasets, results, and analyses, available in a public repository[2].

4.1. Datasets

We selected ten relevant benchmark datasets for binary classification, commonly used by the research community. Table 1 provides a summary of these datasets, including the

[2]Available at: https://github.com/diegominatel/ahp-gaussian-model-selection

total number of examples (#E), number of attributes (#A), proportion of positive class examples (#PC), protected attributes, unprivileged groups, and source reference.

Table 3. Dataset information: #E represents the total number of examples, #PC denotes the proportion of positive class examples, and #A indicates the number of attributes.

Dataset	#E	#A	#PC	Protected Attribute	Unprivileged Group	Reference
Arrhythmia	452	278	43.57%	Sex	Female	[Kelly et al. 2017]
Bank Marketing	45,211	42	11.69%	Age	Under 25	[Kelly et al. 2017]
Census Income	48,842	76	24.90%	Race and sex	Non-white/female	[Kelly et al. 2017]
Contraceptive	1,473	10	55.48%	Religion	Islam	[Kelly et al. 2017]
German Credit	1,000	36	70.00%	Sex	Female	[Kelly et al. 2017]
Heart	383	13	46.12%	Age	Non-middle-aged	[Kelly et al. 2017]
Recidivism Female	1,395	176	37.32%	Race	Non-white	[Larson et al. 2016]
Recidivism Male	5,819	375	49.69%	Race	Non-white	[Larson et al. 2016]
Student	480	46	73.84%	Sex	Female	[Amrieh et al. 2015]
Titanic	1,309	6	40.07%	Sex	Male	[Vanschoren et al. 2013]

We used the protected attributes 'gender' and 'age' to train the models on datasets Arrhythmia and Heart as they play critical roles in disease prediction. The Recidivism dataset was divided into Recidivism Female (female examples) and Recidivism Male (male examples). Finally, we binarized the class labels for datasets Arrhythmia (absence or presence of cardiac arrhythmia), Contraceptive (use or non-use of contraceptive methods), and Student (low-performance and medium-high-performance).

4.2. Classification Algorithms and Hyperparameter Settings

In this experiment, we applied the following traditional classification algorithms: Decision Tree (DT), k-Nearest Neighbors (k-NN), Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Machines (SVM), and eXtreme Gradient Boosting (XGB). In addition, we also used Adversarial Debiasing (AD), a well-known algorithm for mitigating discriminatory biases [Zhang et al. 2018]. Table 4 presents each classification algorithm and the range of numerical values for its hyperparameters[3]. We tested 32 hyperparameter settings per classification algorithm.

Table 4. Algorithms and their hyperparameter value ranges. Numerical variations in the format $(i : f : p)$ indicate that i and f represent the initial and final values, respectively, with p denoting the increment used.

Algorithm	Hyperparameter	Value Variation
Adversarial Debiasing	Number of epochs	$(50 : 330 : 9)$
Decision Tree	Minimum samples to be at a leaf node	$(1 : 33 : 2)$
	Minimum samples to split a node	$(4 : 5 : 1)$
k-Nearest Neighbors	Number of neighbors	$(1 : 33 : 2)$
	Power parameter for the Minkowski metric	$(1 : 2 : 1)$
Multilayer Perceptron	Number of neurons in the hidden layer	$(5 : 37 : 1)$
Random Forest	Number of trees	$(30 : 500 : 15)$
Support Vector Machines	Gamma	$(0.0025 : 0.02 : 0.0025)$
	Regularization	$(0.98 : 1.01 : 0.01)$
XGBoost	Number of trees	$(30 : 500 : 15)$

[3]For hyperparameters not specified, we used the default values provided by the algorithms in the following Python libraries: scikit-learn (DT, k-NN, MLP, RF, and SVM), aif360 (AD), and xgboost (XGB).

4.3. Experimental Results

This section presents the experimental results according to the previously described experimental setup.

Table 5 shows the average scores for metrics Macro F1-Score, RMF1, RDP, REO, and RDO calculated on the test set. We organized the results by classification algorithms and multi-criteria methods. The values in bold indicate the best average score per classification algorithm, and gray cells highlight the top average result for each metric analyzed.

Table 5. Average classification algorithm results (%), with standard deviation in parentheses. Bold values indicate the best average result per algorithm, and gray cells highlight the top average result for each metric analyzed.

Algorithm	Selection	Macro F1	RMF1	RDP	REO	RDO
AD	SAC	71.66 (7.96)	89.20 (12.98)	75.78 (18.21)	84.04 (17.74)	73.94 (14.57)
	MCPM	71.45 (8.07)	88.92 (12.75)	75.45 (17.97)	83.58 (17.51)	73.71 (14.28)
	AHP–Gaussian	**71.85 (7.60)**	**89.31 (11.64)**	**76.48 (18.80)**	**84.58 (15.90)**	**74.88 (13.82)**
DT	SAC	69.88 (9.07)	90.08 (7.23)	68.58 (25.87)	**84.04 (14.05)**	69.83 (14.76)
	MCPM	**70.41 (8.55)**	**90.52 (7.18)**	67.67 (25.18)	83.16 (13.33)	68.91 (13.81)
	AHP–Gaussian	69.89 (9.08)	90.22 (7.24)	**69.08 (25.63)**	83.99 (14.07)	**70.28 (14.58)**
k-NN	SAC	66.92 (10.06)	94.04 (6.32)	**68.31 (21.81)**	**85.50 (10.69)**	**70.81 (15.15)**
	MCPM	66.92 (10.06)	94.04 (6.32)	**68.31 (21.81)**	**85.50 (10.69)**	**70.81 (15.15)**
	AHP–Gaussian	**67.03 (10.02)**	**94.22 (6.40)**	68.19 (21.75)	85.46 (10.55)	70.69 (15.02)
MLP	SAC	**73.61 (8.12)**	92.93 (4.67)	**66.44 (28.74)**	83.86 (16.61)	67.02 (19.94)
	MCPM	**73.61 (8.12)**	92.89 (4.69)	65.06 (27.70)	84.66 (16.75)	68.12 (21.20)
	AHP–Gaussian	73.50 (8.01)	**93.05 (5.08)**	65.17 (27.37)	**84.96 (17.48)**	**68.48 (21.01)**
RF	SAC	74.61 (8.81)	92.79 (4.12)	61.45 (24.88)	85.07 (17.01)	67.02 (17.17)
	MCPM	74.61 (8.81)	92.79 (4.12)	61.45 (24.88)	85.07 (17.01)	67.02 (17.17)
	AHP–Gaussian	**74.62 (8.92)**	**93.71 (3.64)**	**62.32 (24.97)**	**86.40 (14.61)**	**68.63 (16.67)**
SVM	SAC	67.88 (14.21)	93.89 (4.90)	63.56 (31.21)	**79.27 (29.18)**	65.40 (29.09)
	MCPM	67.88 (14.21)	93.89 (4.90)	63.56 (31.21)	**79.27 (29.18)**	65.40 (29.09)
	AHP–Gaussian	**68.88 (11.60)**	**94.19 (4.29)**	**65.13 (30.66)**	77.89 (28.70)	**68.06 (28.41)**
XGB	SAC	75.32 (8.37)	94.31 (3.79)	62.59 (25.29)	85.80 (11.25)	**71.09 (18.52)**
	MCPM	**75.37 (8.40)**	94.19 (3.63)	62.48 (25.47)	85.80 (11.25)	70.99 (18.71)
	AHP–Gaussian	75.16 (8.24)	**94.67 (3.72)**	**62.81 (25.87)**	**86.59 (11.66)**	69.90 (17.32)

As shown in Table 5, the AHP–Gaussian method achieved the best average for all metrics in AD and RF and had the best average in at least three of the five metrics for MLP, SVM, and XGB. Performing the best average results in AD is highly significant, as it is a classification algorithm specifically designed to reduce prediction biases. In contrast, the MCPM and SAC methods obtained the best averages for k-NN, while no method stood out over the others in DT. Additionally, AHP–Gaussian had the best overall average in four of the five analyzed metrics.

The main characteristic to be evaluated in these multi-criteria methods applied to model selection is their ability to identify which classifier excels across all or most of its input criteria. In this regard, AHP–Gaussian stands out, as it achieved the best average results in five of the seven classification algorithms, demonstrating balanced performance in all the analyzed measures, whether performance or fairness metrics.

Table 6 presents the average scores obtained using the multi-criteria method applied to model selection. Each result is calculated as the average of the 70 models (7 classification algorithms $\times$ 10 datasets) selected by each method. Bold results indicate the best average score for each metric.

Table 6. Average results (%) on the test set, with the standard deviation in parentheses. The best result for each measure is highlighted in bold.

Multi-criteria	Macro F1	RMF1	RDP	REO	RDO
SAC	71.41 (9.79)	92.46 (6.90)	66.67 (24.74)	83.94 (16.94)	69.30 (18.41)
MCPM	71.46 (9.73)	92.46 (6.82)	66.28 (24.47)	83.86 (16.86)	69.27 (18.47)
AHP–Gaussian	**71.56 (9.20)**	**92.76 (6.56)**	**67.02 (24.55)**	**84.26 (16.54)**	**70.13 (18.02)**

The AHP–Gaussian method had the best overall average in all five measures used as selection criteria and the lowest standard deviation values. Despite having different selection strategies, the SAC and MCPM methods yielded similar results, both in the overall average and in algorithms analysis, as seen in the case of k-NN, where both selected the same classifiers.

To complement the analysis, Figures 2, 3, 4, 5, and 6 show the CD diagram derived from the Nemenyi *post-hoc* test for the results of Macro F1-Score, RMF1, RDP, REO, and RDO, respectively. We considered each tuple $\{$*dataset, classification algorithm, multi-criteria method*$\}$ in the Nemenyi *post-hoc* test. At the top of each diagram, we can observe the Critical Difference (CD), and the horizontal axis represents the average ranks of the model selection strategies, with the best-ranked data stratification criterion on the left. A black line connects criteria when no significant difference is detected between them.

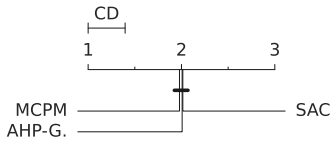

Figure 2. Macro F1

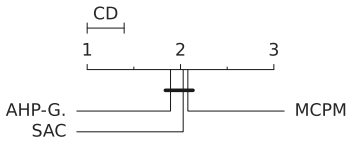

Figure 3. RMF1

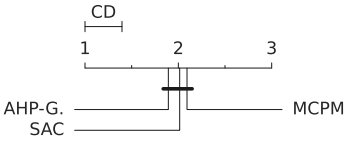

Figure 4. RDP

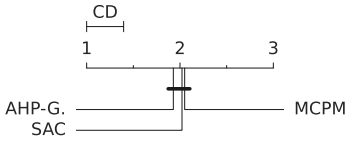

Figure 5. REO

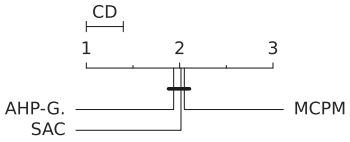

Figure 6. RDO

None of the CD diagrams showed statistically significant differences, but the results were consistent with those in Table 6. Specifically, AHP–Gaussian was ranked first in four of five measures. Only in Macro F1-Score was MCPM ranked first, with AHP in second place. However, MCPM was ranked last in all other measures.

The results presented in Tables 5 and 6, along with the CD diagrams, demonstrate the superior ability of the AHP–Gaussian method in selecting classifiers that combine good predictive performance with optimization across various group fairness measures compared to the multi-criteria methods tested in this experiment. These findings suggest that AHP–Gaussian is a simple but promising approach for scenarios where balancing predictive performance and fairness criteria is crucial.

5. Concluding Remarks

This paper introduced the application of the AHP–Gaussian method for selecting fairer models considering multiple evaluation criteria. This method's advantage is that it calculates a Gaussian factor based on the data, which determines the importance of each input criterion in selecting the best classifier. Our experimental results indicate that AHP–Gaussian is the most effective at selecting classifiers that combine high predictive power with fairness-awareness among the multi-criteria methods tested, especially for the Adversarial Debiasing, Random Forest, and Support Vector Machines classification algorithms.

Although we designed the experimental setup to focus on binary classification and group fairness analysis, AHP–Gaussian is not limited to this type of analysis. We can apply it to select multiclass classifiers and tasks that assume other selection criteria. In future work, we intend to expand the experimental setup by increasing the number of classification algorithms, hyperparameter values, datasets, and metrics analyzed and applying AHP–Gaussian to multiclass classification. We also intend to apply other methods from the Multi-Criteria Decision Analysis field, such as Promethee and Thor.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Adem Esmail, B. and Geneletti, D. (2018). Multi-criteria decision analysis for nature conservation: A review of 20 years of applications. *Methods in Ecology and Evolution*, 9(1):42–53.

Amrieh, E. A., Hamtini, T., and Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. In *IEEE AEECT*, pages 1–5. IEEE.

Aruldoss, M., Lakshmi, T. M., and Venkatesan, V. P. (2013). A survey on multi criteria decision making methods and its applications. *American Journal of Information Systems*, 1(1):31–43.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Black, E., Raghavan, M., and Barocas, S. (2022). Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 30:3992–4001.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *ACM Computing Surveys*.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Darko, A., Chan, A. P. C., Ameyaw, E. E., Owusu, E. K., Pärn, E., and Edwards, D. J. (2019). Review of application of analytic hierarchy process (ahp) in construction. *International journal of construction management*, 19(5):436–452.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Dos Santos, M., de Araújo Costa, I. P., and Gomes, C. F. S. (2021). Multicriteria decision-making in the selection of warships: a new approach to the ahp method. *International Journal of the Analytic Hierarchy Process*, 13(1).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Kelly, M., Longjohn, R., and Nottingham, K. (2017). The UCI machine learning repository.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm.

Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., and Kyriazis, D. (2024). Bias in machine learning: A literature review. *Applied Sciences*, 14(19).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Minatel, D., da Silva, A. C. M., dos Santos, N. R., Curi, M., Marcacini, R. M., and de Andrade Lopes, A. (2023a). Data stratification analysis on the propagation of discriminatory effects in binary classification. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*, pages 73–80. SBC.

Minatel, D., dos Santos, N. R., da Silva, A. C. M., Cúri, M., Marcacini, R. M., and Lopes, A. d. A. (2023b). Unfairness in machine learning for web systems applications. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, pages 144–153.

Minatel, D., dos Santos, N. R., da Silva, V. F., Cúri, M., and de Andrade Lopes, A. (2023c). Item response theory in sample reweighting to build fairer classifiers. In *Annual International Conference on Information Management and Big Data*, pages 184–198. Springer.

Minatel, D., Parmezan, A. R., Cúri, M., and de A. Lopes, A. (2023d). Dif-sr: A differential item functioning-based sample reweighting method. In *Iberoamerican Congress on Pattern Recognition*, pages 630–645. Springer.

Minatel, D., Parmezan, A. R., Cúri, M., and Lopes, A. D. A. (2023e). Fairness-aware model selection using differential item functioning. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1971–1978. IEEE.

Mishler, A., Kennedy, E. H., and Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400.

Narasimhan, H. (2018). Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654.

Parmezan, A. R. S., Lee, H. D., and Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Systems with Applications*, 75:1–24.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, 30:5680–5689.

Podvezko, V. (2009). Application of ahp technique. *Journal of Business Economics and management*, (2):181–189.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International journal of services sciences*, 1(1):83–98.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). OpenML: Networked science in machine learning. *SIGKDD Explor.*, 15(2):49–60.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhang, Y., Bellamy, R., and Varshney, K. (2020). Joint optimization of ai fairness and utility: a human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 400–406.