International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-025-00721-2

REGULAR PAPER

General procedure to measure fairness in regression problems

Juliett Suárez Ferreira[1] · Marija Slavkovik[2] · Jorge Casillas[1,3]

Received: 28 June 2024 / Accepted: 15 January 2025

Abstract
Fairness in artificial intelligence has emerged as a critical ethical concern, with most research focusing on classification tasks despite the prevalence of regression problems in real-world applications. We address this gap by presenting a general procedure for measuring fairness in regression problems, focusing on statistical parity as a fairness metric. Through extensive experimental analysis, we evaluate how different methodological choices, such as discretization methods, algorithm selection, and parameter optimization, impact fairness outcomes in regression tasks. Our primary contribution is a systematic framework that helps practitioners assess and compare fairness across various approaches to solving regression problems, providing clear guidelines for selecting appropriate strategies based on specific problem requirements. The results demonstrate the importance of carefully considering procedural decisions when evaluating fairness in regression contexts, as these choices influence both model performance and fairness outcomes.

Keywords Fair AI · Regression · Statistical parity · Fairness metrics · Bias in AI · Multiobjective optimization

1 Introduction

Artificial intelligence (AI) techniques are being widely applied, with individuals either interacting directly with AI systems or being affected by AI-embedded decision-making processes [1]. While many AI applications focus on classification tasks, there is a significant class of problems requiring continuous or ordinal outputs, regression problems, that present unique challenges for ensuring fairness and non-discrimination.

Ensuring fair and non-discriminatory AI decisions is crucial, as it represents one of the fundamental rights recognized by the European Union. Algorithmic decision systems (ADS) carry known risks of discrimination [2], which can manifest as either intentional or unintentional injustice resulting from human prejudice and stereotyping based on sensitive attributes [3]. Notable instances of AI applications failing to make fair decisions [4, 5] highlight the critical need for robust fairness measures and mitigation strategies.

Most fair AI research concentrates on fairness metrics for machine learning classification tasks. In classification tasks, the goal of machine learning is to correctly assign predefined categories or labels to input data, based on patterns and relationships identified from training examples. For example, the compas case [6], which predicts a recidivism risk score between 1 and 10, has been addressed as a binary classification problem: low risk vs. medium–high risk, with scores from 1 to 4 classified as low risk; this is an example of an ordinal classification task where the order of the labels matters. These kinds of problem have also been classified as ordinal regression. Some problems require a real value, rather than a label, to be assigned to the input data. These are called regression problems. For example, for the Law School Admission Council dataset [7], the prediction outcome is the Under-

Marija Slavkovik and Jorge Casillas have contributed equally to this work.

✉ Juliett Suárez Ferreira
juliettsuarez@correo.ugr.es

Marija Slavkovik
marija.slavkovik@uib.no

Jorge Casillas
casillas@decsai.ugr.es

[1] Data Science and Computational Intelligence Institute (DaSCI), University of Granada, Calle Periodista Daniel Saucedo Aranda, s/n, 18071 Granada, Spain

[2] Department of Information Science and Media Studies, University of Bergen, Fosswinckels gate 6, 5007 Bergen, Norway

[3] Department of Computer Science and Artificial Intelligence (DCSAI), Higher Technical School of Computer and Telecommunications Engineering, University of Granada, 18071 Granada, Spain

Published online: 13 February 2025

graduate Grade Point Average of law students, a continuous variable also called 'target variable.' A regression problem can be addressed as a classification problem by finding a good way to discretize the target variable.

Despite the abundance of regression problems in real-world applications, there exists no in-depth analysis comparing classification and regression fairness measures when applied to regression problems. The common practice of converting regression problems into classification tasks through discretization lacks a thorough analysis of its consequences for fairness. Additionally, the influence of methodological choices, such as discretization methods and model selection, on fairness outcomes remains understudied.

This paper addresses these critical gaps by introducing a *general procedure* for assessing fairness in regression problems. The proposed procedure is supported by an experimental analysis that focuses on evaluating fairness supported by multiple hypotheses.

- *H1*: Fairness is sensitive to different discretization methods/thresholds. We will investigate how the selection of different discretization methods or thresholds affects fairness when the original regression problem is transformed into a classification task.
- *H2*: Different fairness results can be obtained using different methods to solve the same problem. We will measure fairness using the definition of SP for classification and regression to check how the selection of different methods to solve the problem affects fairness.
- *H3*: hyper-parameter optimization can lead to different fairness values when using the same method to solve the problem. We present a study of hyper-parameter optimization based on the proposal made by Valdivia et al [8] specifically for regression problems.
- *H4*: Using measures designed for classification allows for the detection of unfairness in regression problems. We will explore the adaptation of classification metrics for regression scenarios, shedding light on the advantages and drawbacks.
- *H5*: The solution using regression techniques is better when continuous output is needed. We present a comparison of the solutions using classification and regression techniques, highlighting the differences in fairness and error.

Our experimental analysis is focused on statistical parity (SP). SP is a fairness metric that quantify whether different groups (e.g., based on gender or race) receive similar treatment in decision-making processes, evaluating the broad view of equality of opportunity [9]. We also contribute a state-of-the-art review of SP metrics for regression problems, offering practitioners clear guidelines for selecting appropriate fairness measures based on problem characteristics.

This paper is organized as follows. The related work, found in Sect. 2, provides a study of SP definitions, with a particular focus on definitions for regression problems, as well as guidelines for their application based on the characteristics of the problem. Section 3 introduces a methodological framework for evaluating fairness in regression tasks.

In Sect. 4, we propose a general procedure for quantifying SP in regression tasks and evaluate it by measuring SP in multiple datasets. The evaluation of *H1* is detailed in Sect. 4.2. Section 4.3 presents the analysis of the SP measure for classification tasks, also evaluating the hypotheses *H2* and *H4*. Section 4.4 provides an analysis of a definition of the SP measure specifically designed for regression tasks and compares the metric using different methods (*H2*). The study of hyper-parameter optimization to evaluate *H3* is covered in Sect. 4.5, which also highlights the selection of regression techniques *H5* when continuous output is required. The results are discussed in Sect. 5, and the conclusions and future work are presented in Sect. 6.

2 Related work

In this section, we present different definitions of fairness metrics specifically for statistical parity in classification and regression. We aim to provide a revision of the fairness metrics proposed for the regression setting, specifying which of them can be applied taking into account the problem specifications. We highlight some existing tools that implement these metrics and an approach in hyper-parameter optimization focused on fairness.

2.1 Fairness and statistical parity in classification

Several literature reviews have been carried out on different aspects of fair AI [3, 10]. Fairness notions have been divided mostly into two different groups: individual and group fairness. *Individual fairness* is the requirement that models assign similar predictions to similar individuals [11]. *Group fairness* is the requirement that models assign similar predictions to different groups. *Subgroup fairness* studied by Mehrabi et al [12] as another category that aims to obtain the best properties of the group and individual notions of fairness. Authors like Verma and Rubin [13] make a distinction about *counterfactual fairness* (defined by Kusner et al [14]) as a way of interpreting sources of bias using causal graphs.

We here focus on group fairness and we are using some concepts in this context that need to be defined:

- protected attribute (*PA*): an attribute that divides the population into groups (such as gender and race). The algorithm outcomes should preserve parity taking this

attribute into account. In the compas example, this attribute is the race.

- privileged value (*PV*): a value of a *PA*, indicating a group that has been in systematic advantage (historically). In the same way, unprivileged values imply the group in a systematic disadvantage (*UV*). The compas example considers the African-American group to be at historical disadvantage.
- favorable label (*FL*): a label for an outcome that provides an advantage to the recipient. Following the example of compas, the favorable label corresponds to attaining a low risk of recidivism score. The classification task involves only two distinct labels, whereas in the original problem, the values of a low risk range from 1 to 4, inclusively.

SP has been defined by Calders et al [15] as: The likelihood of a positive outcome (*FL*) should be the same regardless of whether the person is in the protected group (whether the person has a *PV* or a *UV* in *PA*). This measure represents the independence criteria defined by Barocas et al [9].

In the binary classification setting, where the objective is to predict a binary label (*FL* or not), SP can be computed taking into account the true positive[1] (TP) and the false positive[2] (FP) values of the predicted outcome. The mathematical definition of this metric can be found in Eq. 1 where *U* and *P* are the total number of outcomes for each group of *PA* while the values U_{tp} and P_{tp} are the TP for each group and U_{fp} and P_{fp} are the FP in each group of *PA*.

$$SP_C = \frac{(U_{tp} + U_{fp})}{U} - \frac{(P_{tp} + P_{fp})}{P} \tag{1}$$

2.2 Statistical parity in regression

In the regression setting, the objective is to predict an ordinal or continuous outcome. Therefore, to apply the definition SP_C to a regression problem, the outcome should be discretized to obtain a binary variable. This will transform the original problem into a binary classification task. The risk of trying to solve regression problems as classification tasks has been widely studied [16], highlighting the risks associated with the enforcement of labels or categories in continuous or ordinal variables. Therefore, to apply the fairness metrics designed for classification problems to regression problems, the favorable label *FL* and the unfavorable label are a set of values grouped by a label instead of a single value.

One way to calculate two groups is to divide the output based on a threshold. The phenomenon of incompatibility between threshold policies and fairness measures has been considered by CorbettDavies et al [17] as differences in the base rate between groups that directly affect fairness measures. The same phenomenon has been studied in the fairness literature as impossibility results [18] but also in statistics as the problem of inframarginality [17].

The issues encountered with thresholds also arise when, rather than binary protected attributes, the problem involves continuous protected attributes like age, which are frequently grouped to assess fairness. The interpretations of *PA* and *PV* are adapted to this context to consider a group of values delimited by thresholds.

Another approach to assess fairness in regression problems involves using measures specifically designed for this context. Numerous suggestions have been made to measure independence criteria, particularly with respect to SP in regression problems. A summary of the SP measures in the regression settings is provided in Table 1.

Table 1 Statistical parity definitions for regression problems

Approach	References	Cat	Cont
Pearson correlation	[19]		✓
Mann–Whitney U test	[20]	✓	
Kolmogorov–Smirnov test	[21, 22]	✓	
Distance of the average outcomes	[23–25]	✓	
Area under the curve	[25]	✓	
Mutual information	[26–28]	✓	✓
Pairwise	[29]	✓	✓
	[30]	✓	

The approach to implement the measure, the references in which the measure was proposed and whether the measure can handle categorical (Cat) and/or continuous (Cont) protected attributes

Bias control in linear regression has been studied by Calders et al [25]. The authors presented two measures to quantify bias based on mean predictions and area under the curve (AUC). The metrics: group fairness in expectation [24] and the definition of SP by Yan et al [23], can be considered similar to the mean prediction metric proposed by Calders et al [25] since they use the distance of the average outcomes to calculate the differences between groups. The mathematical definition of this metric can be observed in 2 where $\hat{y_u}$ and $\hat{y_p}$ are the predicted outcomes for the unprivileged and privileged groups and the values n_{y_u} and n_{y_p} represent the number of elements in each group.

$$SP_R = \frac{\sum(\hat{y_u})}{n_{y_u}} - \frac{\sum(\hat{y_p})}{n_{y_p}} \tag{2}$$

Statistic tests have been proposed to be used in fairness. [20] proposed a method for fair regression using the Mann–Whitney U statistic [31]. In [21], the authors use the Kolmogorov–Smirnov (K–S) statistic to compute SP. The

[1] **True positive** (TP): The predicted and the actual outcome are in a positive class.

[2] **False positive** (FP): A positive outcome is predicted as negative.

proposal of Chzhen et al [22] is similar to this one; they use the total variation distance to measure SP and change it to use the K–S distance as a constraint imposed to obtain the optimal fair regression function due to the computational cost. These definitions have a stronger statistical justification in terms of independence.

Steinberg et al [26, 27] used mutual information (MI) to introduce approximations of the independence, separation, and sufficiency group fairness criteria for regression models. Moreover, SP defined by Yan et al [28] uses MI to calculate unfairness. The use of correlation is proposed by Komiyama et al [19] to reduce bias in regression problems; this approach is valid for handling continuous *PA*.

Different pairwise fairness metrics have been proposed by Narasimhan et al [29] to be used specifically for ranking problems (mainly used in recommendation problems) and have been extended to be used in regression problems. The metrics are defined on the basis of the comparison of pairs of examples. They proposed that a regressor satisfies SP if the distribution of the outcomes is equal for the privileged and unprivileged groups. The proposed measures handle binary and continuous *PA*. Following a similar approach, [30] has proposed a version of SP for ordinal regression tasks.

It is worth mentioning that along with SP the concept of disparate impact [15] also represents the independence criteria. In 2021, the 2021/144 law was approved to regulate locally with respect to automated employment decision tools in New York City [32]. The law requires a bias audit on an automated employment decision tool before using it. In the text, the proposed measures for regression problems are based on the average of the outcomes and the average of the outcomes below the median of the distribution. The authors [33] performed an extensive analysis of the metrics included in the law and proposed two metrics derived from the binarization of the outcome based on one or multiple thresholds. The first metric works by summarizing data across various proportional thresholds, using an approach that calculates AUC. The second measure is proposed as the likelihood of getting fair binary data when selecting a threshold at random. Both proposals correspond to the concept of disparate impact, also considered a measure of independence, so we do not include them in the table; nevertheless, it is important to note that they can be considered similar to the proposals measuring average results and AUC differences [23–25].

The selection of the measure to be applied can be made taking into account the type of *PA* in the problem to be solved, as well as the output of the problem itself. In our case, we will use the measure based on the distance of the average outcomes (Eq. 2) because we will evaluate binary *PA* and ordinal and continuous outputs. However, when faced with continuous *PA*, measures based on correlation, MI, and pairwise should be considered [19, 26–29]. In addition, when solving ranking problems, pairwise measures should be more suitable [29, 30].

2.3 Tools and hyper-parameter optimization

There are several tools that are mainly focused on classification problems. The *AI Fairness 360* (AIF360) toolkit proposed by Bellamy et al [34] presents more than 70 fairness metrics for classification tasks. The *Fairlearn* project [35] provides three algorithms to mitigate unfairness, two of which could be used to solve regression problems. There are also fairness auditing tools such as *Aequitas* [36] and *FairML* [37]. The lack of regression metrics in these tools is disadvantageous for practitioners who need to use metrics designed specifically for regression tasks.

After selecting a metric to serve as the fairness indicator for a model, the process of choosing the final model involves a careful consideration of both the algorithm and its hyper-parameters. This selection process inherently involves navigating the trade-offs between fairness and accuracy. The methodology proposed by Valdivia et al [8] offers a comprehensive approach to evaluating various models by fine-tuning hyper-parameters through multiobjective optimization in the classification setting. This approach aims to find a range of solutions that balance fairness and error, allowing practitioners to make informed decisions based on their specific requirements. While initially developed for classification tasks, this methodology is versatile and applicable to various machine learning contexts. In our work, we extend this approach to the regression setting, integrating it into a general procedure for model selection. This extension demonstrates the importance of considering fairness–accuracy trade-offs across different types of machine learning problem, not just in classification tasks.

3 General procedure to measure fairness in regression problems

In this section, we present a general procedure for measuring fairness in regression problems. Taking into account that regression problems have also been solved as classification problems and classification metrics have been applied to measure fairness, we propose the procedure in Fig. 1 to measure fairness in regression problems.

Before applying the methodology illustrated in Fig. 1, it is necessary to select the concept of fairness that should be used and the metrics defined to measure it. This methodology is applicable to different notions of fairness and, thus, is reusable. The results will be different solutions to the same problem that can be selected to be used in terms of fairness according to the most suitable metric value.

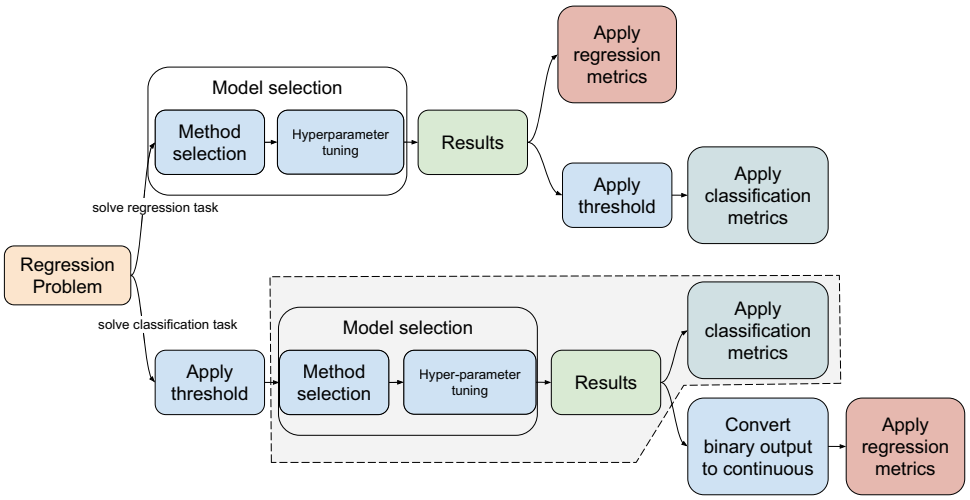


Fig. 1 General procedure to measure fairness in regression problems

The methodology aims to enhance the existing workflow for model selection in machine learning, particularly in terms of fairness and especially for regression problems. The methodology entails two main branches in which a regression problem can be solved as is or as a classification task. The gray area with dashed line borders in Fig. 1 corresponds to the procedure for computing fairness in classification problems.

The first step in deciding to solve a regression problem as a binary classification task is discretizing of the output. This is an important step that will be discussed in Sect. 4.2 since it could lead to solving a different problem, and the outputs cannot always be attributed back to the continuous output; this is not always necessary depending on the problem, but must be considered when making a decision about the method to be used to solve it.

The process of model selection encompasses the steps of method selection and hyper-parameter tuning. In the conventional machine learning process of selecting a model, it is common practice to experiment with various methods to identify the one that yields superior solutions based on a specific metric. Similarly, the hyper-parameters' optimization of the chosen method is a standard practice aligned with the same objective.

Algorithm selection can consider the desired interpretability of solutions, but it is also crucial to consider the trade-off between the error and the fairness measure values. In this sense, a multiobjective optimization approach is useful since it provides a range of different solutions, and the decision of which one will be used is completely in the hands of practitioners; this approach is applied in the hyper-parameter tuning step.

Finally, several comparisons can be made between the classification and regression metrics. Those comparisons enhance the opportunity for practitioners to understand the fairness in their models and choose the model with the best trade-off between error and fairness or the one more suitable in terms of fairness.

4 Experimental analysis

The experimental analysis presented in this section uses 11 datasets, different methods, the definitions of SP presented in Eq. 1 and Eq. 2 to compute fairness in classification and regression solutions, respectively, as well as the hyper-parameter optimization procedure to analyze the hypothesis presented in Sect. 1 and instantiate the general procedure proposed in this work.

Section 4.1.1 details the datasets used in the experiments. The algorithms and metrics used are described in Sect. 4.1.2. The remaining experimental analysis aims to support the hypothesis outlined in the introduction of this paper. Table 2 provides a detailed overview of the subsequent sections and the specific hypotheses each addresses along with an overview of the methodology used. The complete methodology is elaborated in each corresponding section.

Section 4.2 discusses the importance of choosing the correct discretization threshold or method to convert the continuous or ordinal output to binary in terms of fairness. This section focuses on *H1* and implements the *Apply threshold* step of the proposed procedure.

In Sects. 4.3 and 4.4, we investigate the metrics SP_C and SP_R. This includes a comparison of different methods to solve the same problem, highlighting the differences in fairness when using these metrics. We also explore *H2* and the step of *method selection*. Section 4.3 also explores *H4*, using the SP_C metric to determine whether there is a difference in fairness when solving a problem as a regression task versus a classification task. This is done by converting the output into a binary variable using a threshold. This experiment will com-

Table 2 Mapping of hypotheses to experimental methodologies and the corresponding sections in which they are analyzed

Hypothesis	Description	Section	Methodology
H1	Fairness sensitivity to discretization methods	4.2	Comparison of SP values for different discretization methods
H2	Fairness differences using different methods	4.3, 4.4	Comparison of SP measures for classification and regression methods
H3	Impact of hyper-parameter optimization on fairness	4.5	Multiobjective optimization of fairness and error
H4	Effectiveness of classification measures for regression problems	4.3	Comparison of SP_C for classification and regression solutions
H5	Superiority of regression techniques for continuous output	4.5	Comparison of error and fairness trade-offs for classification and regression techniques

pare the results of the *Apply the classification metrics* steps of the proposed procedure.

Section 4.4 provides a detailed analysis of the metric SP_R, which is specifically designed for regression tasks. It compares the results of the *apply regression metrics* step for the regression branch and contrasts it with the classification metric SP_C. This comparison highlights the differences, advantages, and limitations of the regression metric studied.

Section 4.5 discusses the impact of parameter tuning on the fairness computation. It studies the *H3* and the *hyper-parameter tuning* step of the general procedure in terms of fairness. In addition, this analysis allows us to explore the advantages of using a metric designed for the regression setting (*H5*) and to compare the results of both branches for the *apply classification metrics* and *apply regression metrics* steps.

4.1 Experimental setup

This section provides an overview of the datasets and algorithms used in the experiments. We begin by discussing the 11 datasets used, with five representing ordinal regression problems and the remaining six representing continuous regression problems. The same problem will be examined from both a binary classification and regression perspective. We also detail the characteristics of each dataset and the factors considered when discretizing the output into a binary problem. Lastly, we introduce the methods and metrics used in the experiments.

4.1.1 Datasets

Various datasets were collected from research articles that focus on fairness in regression, considering the analysis in [38] which enumerates various alternatives for each type of problem. Table 3 presents the available datasets for regression problems together with their respective references and main characteristics. These characteristics include binary protected attributes (*PA*), the variable to predict, privileged and unprivileged values (*PV*/*UV*), and *FL* favorable labels. The privileged values (*PV*) are highlighted in bold for each dataset.

These datasets encompass various regression tasks with ordinal and continuous output (Output Type). The column Variable to Predict (VtP in Table 3) contains the name of the target variable; we also specify the range of the values of this variable in the column Range. Additionally, we provide a threshold column (T) that contains a value used to transform these problems into binary classification tasks; these values were taken from previous studies using the same problems. For a more detailed description of each dataset and the preprocessing steps applied to them, see appendix in Sect. C; this explanation also includes a compilation of previous works using the same datasets.

For each dataset listed in Table 3, we employ regression techniques to predict ordinal or continuous output. Furthermore, we have transformed each problem into a binary classification task using the threshold column in Table 3. We will also solve this classification task by predicting the binary output.

In the regression problem, the target variable is the original numerical value (found in the column Variable to Predict). In the binary classification task, the target variable assumes two values for each dataset. One value corresponds to the favorable label, as described in the column *FL* in Table 3. The other value represents the complementary set of these favorable label values.

Equation 3 illustrates the conversion process from the variable to predict (VtP) to the binary output using a threshold (column T in Table 3). Section 4.2 is dedicated to a discussion of the implications of selecting different thresholds or discretization methods in terms of fairness.

$$\text{binary output} = \begin{cases} 1 \; if \;\; \text{VtP} \geq \text{T} \\ 0 \; if \;\; \text{VtP} < \text{T} \end{cases} \tag{3}$$

Table 3 Datasets description: the Output Type, the name used for identify the dataset (Dataset), the protected attribute (PA), the privileged (**PV**) and unprivileged values (UV) of the PA, the variable to predict (VtP), the range of that variable in the dataset (Range), the values of the favorable label (FL), the threshold used to obtain a binary classification problem (T), and the reference of each dataset (Ref)

Output type	Dataset	PA	**PV** / UV	VtP	Range	FL	T	Refs
Ordinal	wine	Color	**Red - 1** / White - 0	Quality	{3, ..., 8}	{6, 7, 8}	6	[39]
	compas	Race	**White - 1** / AA - 0	Recidivism	{1, ..., 10}	{1, 2, 3, 4}	5	[6]
	singles	Gender	**Male - 1** / Female - 0	Income	{1, ..., 9}	{5, 6, 7, 8, 9}	5	[40]
	obesity	Gender	**Male - 1** / Female - 0	Obesity	{0, ..., 5}	{0, 1, 2, 3}	4	[41]
	drugs	Gender	**Female - 1** / Male - 0	Coke Recency	{0, ..., 6}	{0, 1, 2}	3	[42]
Continuous	insurance	Gender	**Male - 1** / Female - 0	Charges	{11K, ..., 64K}	$\{\leq 40K\}$	40K	[43]
	parkinson	Gender	**Male - 1** / Female - 0	UPDRS score	{7, ..., 55}	$\{< 17.1\}$	17.1	[44]
	older-adults	Gender	**Male - 1** / Female - 0	Mistakes	{1, ..., 27}	$\{< 8\}$	8	[45]
	crime	Race	**White - 1** / Other - 0	% Crimes	{0, ..., 1}	$\{< 0.15\}$	0.15	[46]
	lsac	Race	**White - 1** / Other - 0	ugpa[1]	{1, ..., 5}	$\{\geq 3.2\}$	3.2	[7]
	student	Gender	**Female - 1** / Male - 0	Final Grade	{0, ..., 19}	$\{\geq 12\}$	12	[47]

[1] Undergraduate Grade Score Average

4.1.2 Methods and metrics

To solve the classification and regression tasks, we use several algorithms. We use the algorithm *Logistic regression* (LogR) implemented in *Scikit-learn* [48] to predict binary outcomes. We implemented a variant of the ordinal classification method (OrdR) proposed by Frank and Hall [49] to predict ordinal variables where the order of the values holds significance (the tasks with output type equal to ordinal in Table 3). For problems with continuous output, we employed *Linear Regression* (LinR), also implemented in *Scikit-learn*. We will call this set of methods *Linear*.

We are also applying another set of tree-based methods to solve classification and regression problems. Specifically, *Classification Tree* and *Regression Tree* (CT and RT, respectively) are also implemented in *Scikit-learn*. We call this set of methods *Tree*.

Fairness measurement is made taking into account the definitions of SP_C and SP_R that can be found in Eqs. 1 and 2 respectively. The results will present the mean value of the metric derived from a tenfold cross-validation experiment for each dataset, unless otherwise specified. An exception is made for the older-adults dataset, due to its size, which employs a threefold cross-validation procedure.

The methods and metrics used in this study are not part of the contribution of the article. They are well known and we are using them to instantiate the methodological choices of the general procedure to measure fairness in regression problems proposed in Sect. 3 and to support the study of the hypothesis presented in Sect. 1.

4.2 Analysis of binary discretization methods in regression problems in terms of fairness

To solve a problem with continuous or ordinal output, such as a binary classification problem, a threshold or method is used to convert the output. The use of a threshold is problem-specific and is sometimes already known (the examination is approved when the score is higher than a certain value), while it can be calculated or suggested by experts in other cases.

Table 4 shows the differences in SP when different thresholds are selected for the same problem. The result corresponds to a single run of a CT for each dataset. We have selected the mean, the median, the division made using the *K-means* algorithm implemented in *Scikit-learn* with two groups, and the threshold used in our experimentation. The selection of the threshold for our experimentation is taken from studies that have used the same dataset; an explanation of these sources can be found in Appendix C.

From the results of Table 4 we can observe that, in several cases, the K-means threshold deviates significantly from the others. For instance, the value in compas, older-adults, drugs, lsac, and parkinson datasets is the only positive in comparison with the other threshold values, contrasting with negative values. This could indicate different group representations under the K-means threshold. In most datasets, the mean and median values are similar, suggesting a consistent level of disparity between datasets. In many datasets, the expert-defined threshold (ours) yields values close to other thresholds, indicating a consensus on the level of parity. In bold, the SP values that are closer to zero for each dataset.

Figure 2 shows the numeric thresholds for two datasets. The objective of the drug consumption problem is to predict the recency in time intervals of cocaine use in this case from 0 (never used) to 6 (used yesterday). The thresholds are 0

Table 4 SP_C metric computed for different discretization methods

Dataset	Discretization Methods			
	Mean	Median	K-Means	Ours
wine	−0.114	**−0.065**	−0.114	−0.114
compas	**−0.307**	**−0.307**	0.312	**−0.307**
singles	**−0.112**	**−0.112**	−0.115	−0.115
obesity	0.083	**0.028**	−0.083	**0.028**
drugs	**−0.05**	−0.181	0.05	−0.101
insurance	0.095	0.05	0.118	**0.033**
parkinson	−0.132	**−0.065**	0.091	−0.101
crime	−0.632	−0.491	−0.512	**−0.467**
older-adults	**−0.333**	−0.583	0.389	−0.611
lsac	−0.107	−0.107	**0.097**	−0.117
student	**−0.006**	0.025	**0.006**	**−0.006**

The values closest to zero are highlighted in bold for each dataset

(median), 1.2 (mean), and 3 (ours). The SP value for the mean shows no discrimination between males or females, while the SP for the median (has consumed or not) or our threshold (has consumed in the last year) shows slight differences in SP. At the same time, we can say that the problems to be solved when using the median (have used drugs or not) and the problem of predicting the recency of the use are completely different.

The objective of the crime dataset is to predict the percentage of crime in an area. Our threshold and the median coincide at 0.15 while the mean is at 0.23. Both values indicate discrimination. (The protected attribute in this case is the race of the population in the area.) The mean SP value reported when using the mean is higher.

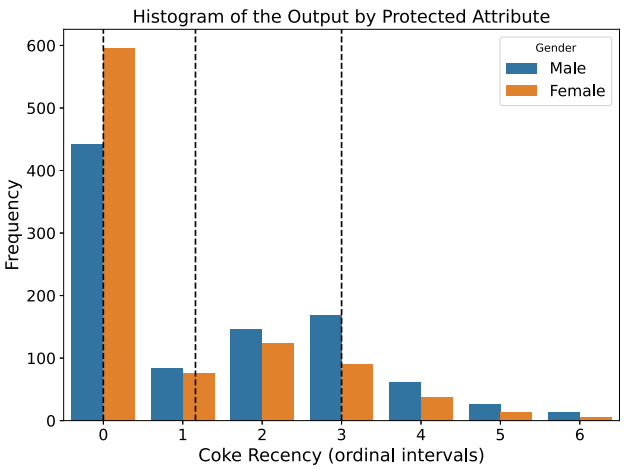
We have studied how discretization affects fairness (*H1*), the results show that the measures are different when selecting different methods and the selection of the criteria is problem-specific. Different thresholds effectively create different problems to solve. This means that fairness must be interpreted differently for each threshold choice. The case of the drug dataset is a clear example of that. Furthermore, analyzing the distribution of the data helps to understand the differences between the groups that could affect the fairness measures.

4.3 Analysis of the SP measure for classification tasks

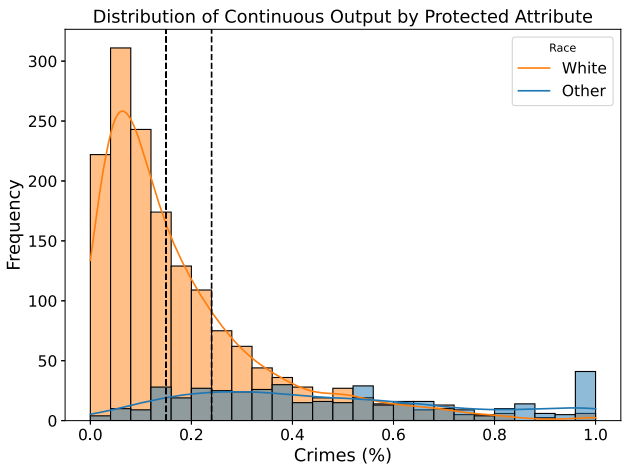
This set of experiments involves the comparison of the SP_C metric (Eq. 1). We solve each problem using a classification technique and a regression technique and then compute the metrics to compare whether there are differences in fairness associated with how the problem is solved. The comparison steps are described in Table 5. We omitted the step of *hyper-parameter tuning*, and we are running all the experiments with the default parameters in *Scikit-learn* for all the methods.

The results of the SP_C measure for all datasets are shown in Fig. 3. We implemented two sets of experiments with different techniques with the objective of comparing the results in terms of fairness (*Linear* and *Tree*). The set of experiments *Linear* appears in Fig. 3a. We used LogR for problems with binary outcomes and OrdR or LinR for ordinal or continuous output. The set of experiments using CT or RT (*Tree*) appears in Fig. 3b.

Based on the Wilcoxon test (0.5), no significant differences were observed between SP_C solving problems as a classification or regression task in the cases analyzed using the setup *Linear* or *Tree*. However, the use of different techniques to solve the same problem can affect fairness performance, with less biased results obtained when using



(a) drugs (Mean = 1.2, Median = 0, Ours = 3)



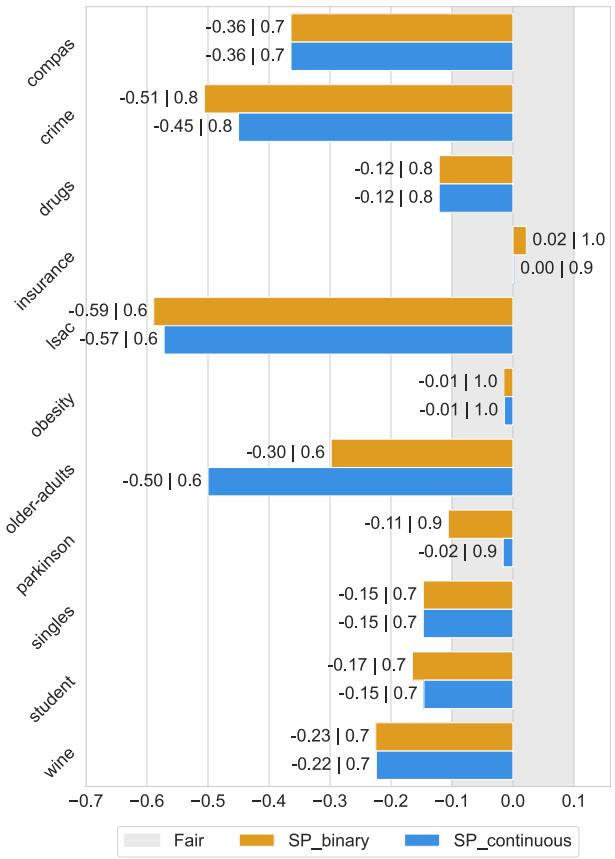
(b) crime (Mean = 0 .23, Median = 0 .15, Ours = 0 .15)

Fig. 2 Visualization of different thresholds for drug and crime datasets

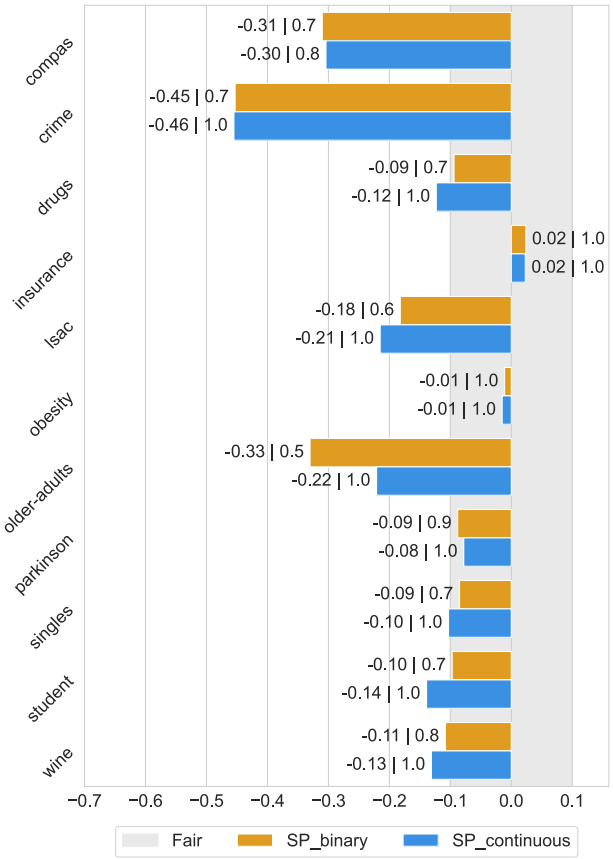
Table 5 Steps to compare fairness metrics designed for classification solving the regression problem as a classification or a regression task

Classification setting	Regression setting
• Transform into a binary classification task by applying a threshold	• Apply the regression method
• Apply the classification method	• Transform the output to binary by applying a threshold
• Compute fairness metric (SP_C)	• Compute fairness metric (SP_C)

The results of the Classification Setting can be observed in Fig. 3 as SP_binary while the Regression Setting values are show as SP_continuous



(a) SP_C values for LogR (SP _binary) or OrdR/LinR (SP _continuous)



(b) SP_C values for CT (SP _binary) or RT (SP _continuous)

Fig. 3 SP_C values using different methods. The light gray horizontal shadow shows the range in which the metric values are considered fair. SP_binary correspond to the values of SP_C when the problem is solved as a binary classification. SP_continuous represent the values of SP_C when the problem is solved as a regression task. The annotation of each bar represents the SP_C|Accuracy. The accuracy of the solution for the problems solved as regression tasks are computed after transforming the output to binary by applying the threshold specified in Table 3

Decision Trees. These findings corroborate *H2* from the introduction section.

The SP_C metric remains valid in two scenarios: when applied to classification tasks directly (SP_ binary) and when applied to regression tasks after converting their output to binary (SP_continuous). The solutions do not differ from each other in terms of fairness (except for the older-adults dataset). This study supports *H4*; classification metrics can be effectively used to find unfairness by being aware that the solved problem is different once a discretization procedure has been used to obtain a binary classification problem.

The analysis of the point (or method) to convert to a binary classification is problem-specific, as we check in Sect. 4.2, so the advantage of using this methodological choice seems to be the simplification of the problem and the use of measures that are better studied and already implemented in tools like

Table 6 SP_R values for *Linear* and *Tree* methods in comparison with SP_C for the Regression Setting values, including mean squared error (MSE)

		Linear			Tree		
	Dataset	MSE	SP_R	SP_C	MSE	SP_R	SP_C
The lowest values are the FL	compas	0.061	0.213	−0.364 (U)	0.041	**0.183**	−0.304 (U)
	crime	0.020	**0.298**	−0.450 (U)	0.000	0.307	−0.455 (U)
	drugs	0.068	0.121	−0.121 (U)	0.000	**0.077**	−0.123 (U)
	insurance	0.010	−0.025	0.002 (F)	0.000	**−0.019**	0.023 (F)
	obesity	0.002	**−0.028**	−0.014 (F)	0.000	−0.029	−0.015 (F)
	older-adults	0.108	0.241	−0.500 (U)	0.000	**0.186**	−0.221 (U)
	parkinson	0.042	0.047	−0.016 (F)	0.000	**0.044**	−0.078 (F)
The highest values are the FL	lsac	0.025	**−0.081**	−0.572 (U)	0.002	−0.082	−0.215 (U)
	singles	0.104	−0.112	−0.147 (U)	0.004	**−0.087**	−0.103 (U)
	student	0.022	**−0.040**	−0.146 (U)	0.000	−0.044	−0.139 (U)
	wine	0.025	−0.055	−0.224 (U)	0.000	**−0.048**	−0.131 (U)

In bold, the SP_R values closest to zero for each dataset

AIF360. However, this could lead to an oversimplification of the problem; there is a clear boundary, for example, when a student is considered approved or not even when the grades can be close, there is a norm to fix the boundaries; however, there is not a big difference between a recidivism score between 4 or 5 and the first is considered low risk while the second is considered high risk in binary classification.

4.4 Analysis of the SP measure for regression tasks

The objective of this experiment is the verification and analysis of fairness with the measure SP_R designed for regression tasks.

SP_R is computed as the distance between the average predicted outcome between the unprivileged and privileged groups. When this measure is negative, it means that the average predicted outcome for the privileged value is higher than the average predicted outcome for the unprivileged group. Optimal fairness is approached as SP_R tends toward zero, reflecting the minimal disparity between the predicted outcomes of the groups.

The analysis of this measure depends on what is considered a positive outcome in the original problem. For compas, obesity, drugs, insurance, parkinson, older-adults, and crime, the *FL* of the output is the smallest, so a positive value of SP_R indicates that the average predicted outcome for the unprivileged group is higher, therefore unfair for this group. (The opposite is considered when *FL* has the highest values.) The output is considered close to fair when the values are close to zero. However, the threshold at which a value is explicitly regarded as 'fair' has not been defined.

Table 6 shows the results of the SP_R and the SP_C measures calculated in the previous section using the *Linear* algorithms and the *Tree* algorithms (specifically using the Regression Setting of Table 5). The values of the SP_C measure are marked as Unfair (U) or Fair (F) according to the bounds of the measure considered fair in the interval $[-0.1, 0.1]$.

The table has a separation of the datasets according to the interpretation of SP_R in the upper part, which contains the datasets where *FL* has the lowest values. At the bottom of the table are the datasets with the highest *FL*. We normalized the output values before computing the measure for better understanding.

There are differences between the SP_R measure when using the *Linear* or *Tree* methods, the *Tree* method is lower than the *Linear* method in 7 of the 11 datasets. The less biased values (i.e., the values closer to zero) of the measure for each dataset are marked in bold in Table 6. These differences are not significant (Wilcoxon's p-value of 0.38). These results also support the analysis of the differences in fairness when using different methods (*H2*) even when the differences between the results of the settings *Tree* and *Linear* are not statistically significant.

Values SP_R demonstrate unfairness to the unprivileged group for all datasets except obesity and insurance. However, the reported values are close to zero, which shows less bias. The results of measuring SP_C in these datasets by converting the output to binary are fair (F) for the methods *Linear* and *Tree*.

Both models (*Tree* and *Linear*) show a strong positive bias for the cases of compas, crime and older-adults, indicating that the unprivileged group is consistently predicted to have higher outcomes. The values for the drugs dataset also show bias even when they are not as high as the previous ones. The values for parkinson, obesity, and insurance are closer to zero, indicating less bias; the SP_C for these datasets are Fair (F).

For the datasets in which the output is favorable when it has higher values, the biases are not high, but can be appreciated in singles and lsac. The values for student and wine are closer

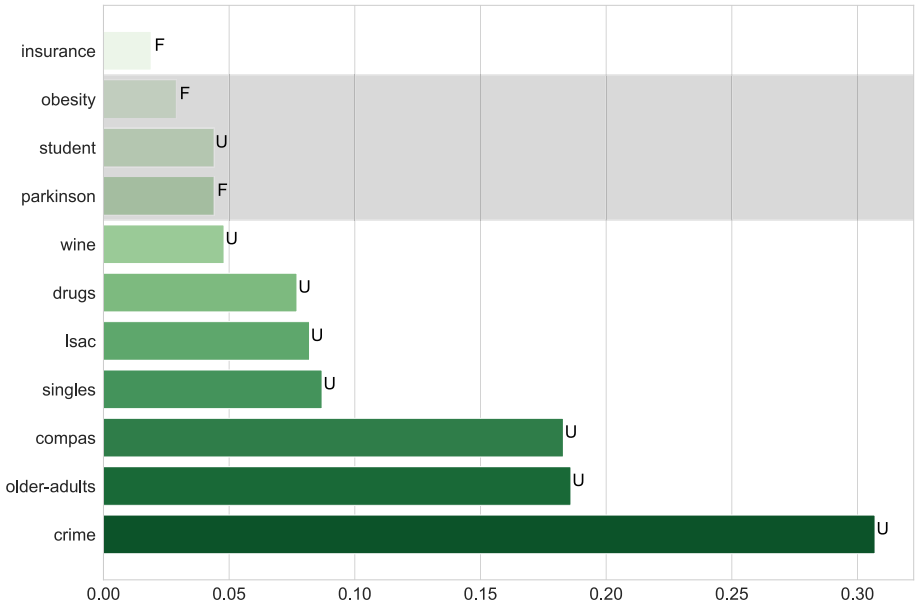


Fig. 4 Absolute value of SP_R for each dataset. The values U (unfair) and F (fair) indicates if the results of the SP_C metric. The gray band highlights inconsistencies in comparison to the SP_C metric

Table 7 Steps to compare fairness metrics designed for regression solving the regression problem as a classification or a regression task

Classification setting	Regression setting
• Transform into a binary classification task by applying a threshold	• Apply the regression method
• Apply the classification method	
• Transform the output back to continuous	
• Compute fairness metric (SP)	• Compute fairness metric (SP)

The Classification Setting results can be observed as the CT values in Figs. 5b and 5d, while the Regression Setting results are the RT values in the same figures

to zero; nevertheless, when the SP_C measure is computed, the predictions are considered unfair.

Calculating the absolute value for SP_R will give us an order of the datasets from lowest to highest bias (see Fig. 4). The main issue when using this measure is the lack of general boundaries to specify which intervals can be considered fair. Notice the cases parkinson (considered fair in SP_C) compared to student or wine (considered unfair in SP_C); these datasets have similar values in SP_R.

This analysis contributes to understanding SP when defined as Eq. 2; other definitions of this criteria require a similar analysis. However, the absence of clear fairness boundaries makes it difficult to compare the results across different problems.

4.5 Analysis of the hyper-parameter optimization

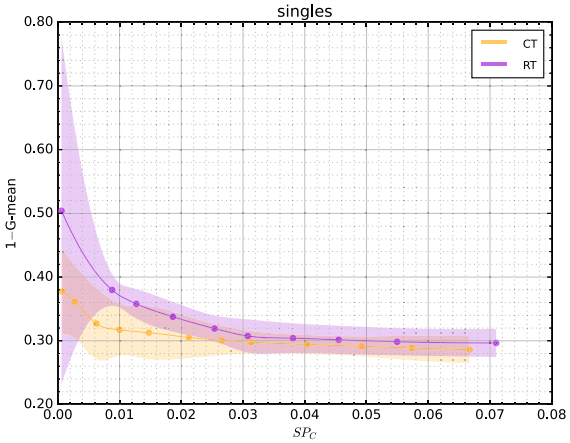
The hyper-parameter optimization implementation from [8] has been adapted to incorporate regression analysis[3]. The experimentation uses the *Tree* methods, specifically the CT and RT algorithms, to minimize error and SP. The objectives of the classification problems were to optimize the combination of Geometric Mean and SP_C. For regression problems, the objectives were the mean squared error (MSE) and SP_R were the target. The multiobjective optimization resulted in multiple models, each offering different trade-offs between fairness and error due to varying combinations of the algorithm's hyper-parameters.

To obtain the SP_C measure with the RT result, we convert the output to binary to compute the objectives for the classification setting; this is the same procedure described in 5 (Regression Setting). On the other hand, to obtain the SP_R measure from the CT results, we use the predicted probability result from the CT to transform the output back to continuous to compute the objectives for the regression setting. Table 7 shows this procedure in the column Classification setting.

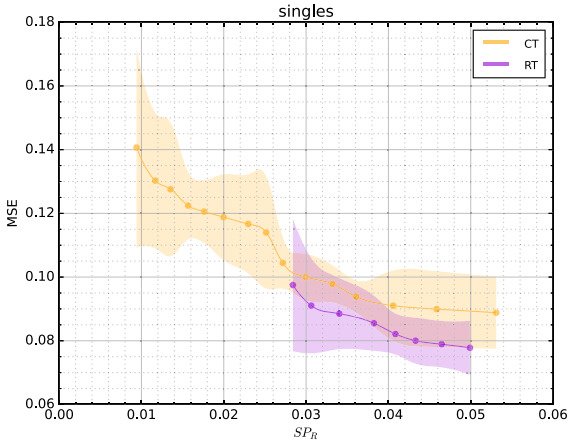
The multiobjective algorithm gives as a result several outputs with different trade-offs between fairness and error measures; a common way of presenting these results is using a Pareto diagram. Figure 5 shows the Pareto fronts of the solutions obtained in two of all datasets studied.

Figure 5a and 5c shows the average Pareto that can be obtained using the geometric mean as accuracy and the met-

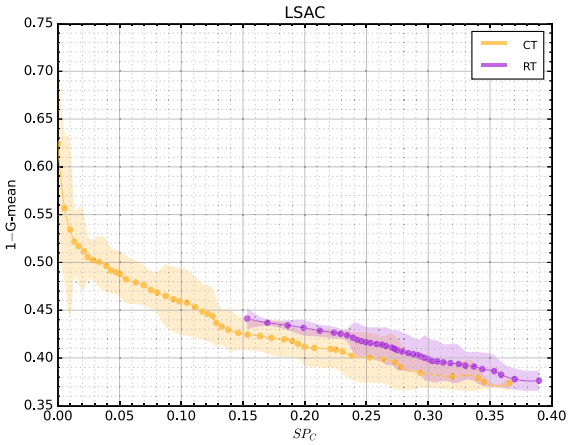
[3] The complete implementation can be found in Github (last date accessed: December 11, 2023).



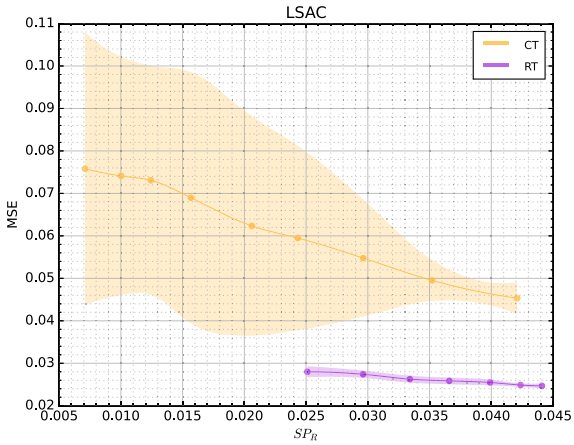
(a) The Pareto front of the CT dominates the Pareto front of the RT, indicating better performance in terms of error-fairness trade-offs. The plot highlights how lower error values are achieved at the expense of fairness in the RT, with the error increasing past a threshold of 0.03. Conversely, the CT achieves lower error values with a trade-off in fairness, with the error beginning to increase around a fairness value of 0.01.



(b) The RT solutions consistently outperform those of the CT in terms of error minimization. However, the CT finds more fair solutions by compromising on error. This diagram illustrates the trade-off between fairness and error, where the CT sacrifices error for better fairness, and the RT optimizes error with a decrease in fairness.



(c) The CT Pareto front outperforms the RT by offering a broader range of solutions that balance fairness and error. However, the trade-off becomes apparent as the CT reaches higher fairness solutions at the expense of increased error, demonstrating the challenge of finding an optimal balance between these two metrics.



(d) The RT dominates the CT in terms of performance, offering solutions with lower error values. While the CT achieves better fairness, it does so at the expense of error, making its solutions less competitive. This diagram clearly shows the trade-off where the RT prioritizes error minimization, and the CT sacrifices error to improve fairness.

Fig. 5 Average Pareto set for CT (yellow) and RT (purple), depicting the trade-offs between error and fairness across multiple scenarios. Each plot shows how different algorithms achieve varying levels of fairness and error, illustrating the challenge of balancing these two objectives in practical applications

ric SP_C as objectives when solving the problem with a CT (yellow) or a RT (purple). Using the RT algorithm while optimizing SP_C shows poorer performance, CT solutions dominate the RT, and in the case of the lsac datasets, the solutions are less fair than CT solutions. Calculating SP_C in these examples is performed using the procedure in Table 5.

Figure 5b and 5d shows the Paretos that optimize MSE and SP_R. The results clearly show that the solutions using CT can reach the fairest results with the cost of error; nevertheless, those Paretos (purple) are dominated by the solutions of RT. The steps described in 7 are used in these examples to calculate SP_R.

The comparison between which solution of the RT is more equitable (using SP_C or SP_R) is challenging due to the difference between the definitions. When a continuous output is needed, calculating fairness using the measure SP_C has the same limitations to solving the problem as a classification task; even when the problem is solved as a continuous one, the measure is computed using a threshold that strictly divides the solutions into favorable or unfavorable sets.

The goal of these experiments is not to give the best solutions to the problems but to explore the consequences of different methodological choices. These experiments show the wide range of solutions that can be obtained using the same method (*H3*). It also seems plausible that the solutions using regression techniques are better when continuous output is needed (*H5*), not only because it is not always possible to cast the solutions back to continuous using probabilities, because not all algorithms bring this possibility, but also because the error when doing this is higher, as can be checked in the Paretos shown in Fig. 5.

5 Discussion

In this section, we critically examine the principal findings derived from the investigation concerning the hypothesis presented in the introduction of this study. The discussion further clarifies both the theoretical and practical implications of these findings.

Our study shows that the choice of discretization procedure affects the fairness results (*H1*). When transforming regression problems into binary classification tasks, different discretization techniques lead to varying fairness outcomes, as demonstrated by the SP_C metric across datasets.

We compared fairness metrics in two scenarios: when problems are solved as classification tasks (using thresholds) and when they are solved as regression tasks. While we found no significant differences in fairness between these approaches, the specific technique chosen may influence fairness results (*H2*). Our study of statistical parity in both classification and regression tasks shows that the choice of solution method affects the fairness outcomes.

In this work, the importance of the selection of hyperparameters is also studied for regression tasks (*H3*), highlighting the different trade-offs between fairness and error that can be obtained using a multiobjective approach. Multiobjective optimization produces diverse solutions, demonstrating its effectiveness in improving fairness.

Fairness metrics designed for classification applied to regression problems after converting the output to classes are valid (*H4*). Nevertheless, the analysis of the point (or method) to make this conversion is problem-specific. It is necessary to use specific error measures for the regression task to have an overall view of the solution, as is checked in Sect. 4.5. The analysis shows that the solutions using regression-specific techniques seem more suitable when continuous output is required (*H5*).

The analysis we performed also reveals the complexities involved in interpreting and defining fairness also for regression tasks, especially considering different datasets. We found that similar statistical parity values in the regression metrics were interpreted differently when using classification metrics for the same problem. Regression problems lack clear thresholds to determine what constitutes a fair value. This makes it difficult to assess solution fairness and can lead to varying interpretation. The use of a multiobjective approach can help clarify how fair the solution is.

5.1 Theoretical and practical implications

Our study contributes to the theoretical understanding of fairness in machine learning, particularly in regression tasks. The key theoretical implications include:

- Fairness metrics in regression: We demonstrate the adaptability of classification-based fairness measures to regression tasks, highlighting the need for careful interpretation of the implication of the results.
- Discretization impact: We establish the sensitivity of fairness measures to discretization methods, emphasizing how problem formulation affects fairness outcomes.
- Method selection: We show that the choice of machine learning method significantly impacts fairness in regression tasks, extending beyond accuracy considerations.
- Hyper-parameter optimization: We extend multiobjective optimization to regression tasks, providing a framework for balancing fairness and accuracy.
- Continuous vs. discretized assessment: We contribute to understanding how fairness assessments differ between continuous and discretized outputs, crucial for evaluation frameworks.

The framework and findings presented in this study have several practical implications for machine learning practitioners.

Table 8 Summary of Hypotheses and Key Findings

Hypothesis	Description	Key result
H1	Fairness sensitivity to discretization methods	Different discretization techniques can significantly impact fairness outcomes in regression problems
H2	Fairness differences using different methods	The choice of machine learning method can influence fairness results, even when fairness metrics show no significant differences
H3	Impact of hyper-parameter optimization on fairness	Multiobjective optimization reveals diverse solutions with different trade-offs between fairness and error
H4	Effectiveness of classification measures for regression problems	Classification fairness metrics can be validly applied to regression problems after output conversion, but interpretation requires a problem-specific analysis
H5	Superiority of regression techniques for continuous output	Regression-specific techniques seems more suitable when continuous output is required, offering better performance

- Our general procedure provides a structured approach for practitioners to assess and improve fairness in regression problems, offering clear decision points throughout the machine learning pipeline. We enable them to make more informed decisions when building and deploying models in real-world applications by highlighting the impact of various methodological choices on fairness and model performance.
- The proposed framework can serve as a practical tool for auditing existing regression models for fairness, allowing organizations to identify and mitigate potential biases in their systems.
- Our comparison of different fairness metrics helps practitioners choose appropriate measures for their specific regression tasks, considering the strengths and limitations of each approach.

By implementing these methodological choices and considering their impacts, practitioners can significantly improve the fairness of their regression models in real-world applications.

6 Conclusions and future work

Our investigation makes different contributions to the understanding of fairness in ML pipelines, particularly focusing on the impact of methodological choices on regression tasks. We presented a general procedure for measuring fairness in regression tasks, and we revised different definitions of statistical parity in the regression setting. Our method is instantiated using two distinct definitions of statistical parity for both classification and regression tasks. These definitions were used to carry out extensive experimentation of all the steps of the procedure, providing insights into each alternative option at each step and highlighting how every decision affects fairness.

This work advances the field by introducing a structured framework for evaluating fairness in regression tasks, facilitates informed decision-making by researchers and practitioners within the ML pipeline and lays the groundwork for creating fairer AI systems in regression contexts.

We offer a novel comparison between fairness measures designed for regression problems and those designed for classification tasks, specifically for statistical parity. Table 8 presents a summary of the conclusions drawn from the experimental analysis, disaggregated by each hypothesis explored.

We presented a review of several measures of statistical parity for regression problems. This review highlights the diversity of approaches to quantifying fairness in regression tasks and highlights the need for rigorous assessment when selecting fairness metrics. This variability emphasizes the importance of understanding each metric's assumptions and limitations. As the field of fairness in regression evolves, further research is needed to develop more robust and widely applicable fairness metrics for continuous and ordinal output spaces. One limitation of our study that future research could address is the exploration of continuous protected attributes.

Our investigation highlights a significant limitation in the current state of fairness metrics for regression tasks: the absence of clear intervals or boundaries within which a measure can be considered fair. This limitation leads to ambiguity in interpreting results and assessing the fairness of regression models. Future work should focus on establishing statistically grounded thresholds for fairness in continuous output spaces, potentially through extensive empirical studies across diverse datasets or theoretical frameworks that consider the nature of regression problems. Furthermore, the formulation of standardized guidelines for the interpretation of fairness metrics within regression contexts would significantly advantage both practitioners and researchers.

Future research directions arising from this work are diverse. Firstly, the proposed general procedure for measuring fairness in regression problems can be instantiated with other fairness criteria to study the differences between

dissimilar definitions of fairness in the regression setting. This could include exploring group fairness notions beyond statistical parity, such as equal opportunity or predictive parity in regression contexts, and developing more robust and widely applicable fairness metrics for continuous and ordinal output spaces. Secondly, given the insufficient research on continuous protected attributes, an analysis of metrics is necessary to quantify unfairness when such attributes are present. Third, defining clearer fairness bounds for regression tasks is crucial; this could involve developing statistical methods to establish meaningful thresholds for fairness metrics in continuous output spaces, addressing the current lack of clear intervals for fair results interpretation. In addition, there is a need for new tools and visualization techniques specifically designed to measure and interpret fairness in continuous data, which could greatly aid practitioners in assessing and mitigating bias in regression models. Finally, given the diversity of approaches to quantifying fairness in regression tasks highlighted in our review, it is necessary to develop guidelines to select appropriate fairness metrics based on specific characteristics and assumptions of the problem. These directions aim to address the current limitations in fairness assessment for regression tasks and contribute to the evolving field of fair machine learning in continuous output spaces.

A Detailed comparison of statistical parity measures for classification and regression approaches across different methods

The experimental results comparing *Linear* and *Tree*-based methods across both binary classification and continuous/ordinal regression tasks are presented in Table 9, and this table is a complementary material for the results presented in Sect. 4.3 and Sect. 4.4. For each dataset, we report four metrics: accuracy (Acc) and statistical parity for classification (SP_C), alongside mean squared error (MSE) and statistical parity for regression (SP_C). The analysis encompasses the 11 datasets studied. For both Linear and Tree-based approaches, we evaluate their performance when solving the problem as a binary classification task versus treating it as a continuous/ordinal prediction problem. This comparison allows us to assess not only the predictive performance through traditional metrics (Acc, MSE) but also the fairness implications through statistical parity measures (SP_C, SP_R). Notably, the Tree-based methods generally achieve higher accuracy and lower MSE compared to Linear approaches, while exhibiting varying patterns in fairness metrics across different datasets.

It is important to note that these results are presented to illustrate methodological choices in fairness assessment rather than to establish performance benchmarks. All algorithms were executed using their default parameters without any hyper-parameter tuning or optimization leaving the anal-

Table 9 Performance and fairness metrics for Linear and Tree-based methods across binary and continuous/ordinal approaches

	Linear							Tree						
	Binary				Continuous/Ordinal			Binary				Continuous/Ordinal		
dataset	Acc	SP_C	MSE	SP_R	Acc	SP_C	MSE	SP_R	Acc	SP_C	MSE	SP_R	Acc	SP_C	MSE	SP_R
compas	0.739	−0.364	0.155	0.001	0.739	−0.364	0.061	0.213	0.720	−0.310	0.003	0.034	0.791	−0.304	0.041	0.183
crime	0.821	−0.506	0.223	0.667	0.805	−0.450	0.020	0.298	0.739	−0.453	0.261	0.453	1.000	−0.455	0.000	0.307
drugs	0.795	−0.121	0.062	0.000	0.794	−0.121	0.068	0.121	0.718	−0.094	0.008	0.016	1.000	−0.123	0.000	0.077
insurance	0.978	0.022	0.044	−0.025	0.944	0.002	0.010	−0.025	0.982	0.024	0.000	−0.000	1.000	0.023	0.000	−0.019
lsac	0.638	−0.589	0.070	−0.364	0.626	−0.572	0.025	−0.081	0.560	−0.182	0.076	−0.076	0.958	−0.215	0.002	−0.082
obesity	0.993	−0.015	0.221	0.002	0.994	−0.014	0.002	−0.028	0.989	−0.011	0.000	0.002	1.000	−0.015	0.000	−0.029
older-adults	0.589	−0.298	0.170	0.254	0.572	−0.500	0.108	0.241	0.467	−0.330	0.001	0.013	1.000	−0.221	0.000	0.186
parkinson	0.875	−0.106	0.174	0.111	0.862	−0.016	0.042	0.047	0.916	−0.088	0.000	0.002	1.000	−0.078	0.000	0.044
singles	0.711	−0.147	0.294	0.000	0.711	−0.147	0.104	−0.112	0.684	−0.085	0.005	−0.011	0.981	−0.103	0.004	−0.087
student	0.728	−0.165	0.087	−0.122	0.729	−0.146	0.022	−0.040	0.657	−0.097	0.001	−0.005	1.000	−0.139	0.000	−0.044
wine	0.744	−0.225	0.086	0.000	0.743	−0.224	0.025	−0.055	0.754	−0.108	0.010	−0.022	1.000	−0.131	0.000	−0.048

Metrics include accuracy (Acc), classification statistical parity (SP_C), mean squared error (MSE), and regression statistical parity (SP_R). All methods use default parameters without optimization

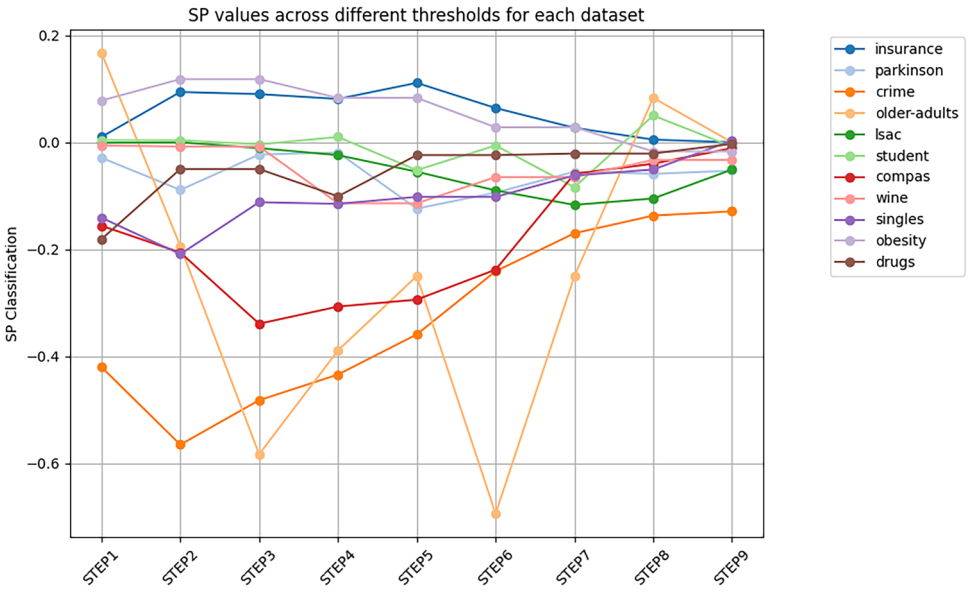


Fig. 6 Statistical parity values across nine equally spaced thresholds for each dataset. The x-axis represents different threshold steps (STEP1–STEP9), while the y-axis shows the corresponding SP values. Each line represents a different dataset

ysis of the impact of hyper-parameter tuning for Sect. 4.5. This deliberate choice allows us to focus on comparing the fundamental differences between classification and regression approaches to fairness measurement, rather than on achieving optimal performance. The reported values should therefore be interpreted as indicative of general patterns and relationships between different fairness metrics and problem formulations, rather than as definitive performance benchmarks. This approach aligns with our primary objective of understanding how different methodological choices in solving regression problems affect fairness measurements, rather than determining which algorithm performs best for each specific dataset.

B Additional analysis of threshold selection impact on statistical parity

To complement the threshold selection analysis presented in Sect. 4.2, we conducted an extended examination of how different numerical thresholds affect statistical parity (SP) across all datasets. Figure 6 shows the variation in SP values across nine equally spaced thresholds (STEP1–STEP9) derived from the range of each dataset's output variable. The results reveal distinct patterns in how threshold selection impacts fairness measurements:

Datasets such as insurance, parkinson, and student demonstrate relatively stable SP values across different thresholds, suggesting that fairness assessments for these problems are somewhat robust to threshold selection. In contrast, crime and older-adults show high threshold sensitivity, with SP values varying substantially (e.g., older-adults ranges from +0.2 to −0.7).

The visible variations in some datasets highlight how threshold selection can fundamentally alter fairness assessments. These findings provide additional support for our approach of prioritizing theoretically grounded, domain-specific thresholds over purely numerical thresholds in fairness assessments.

C Description of the datasets

Table 10 presents a summary of the characteristics of the datasets used in this study, including the output type (Output), the dataset name (Dataset), the protected attribute (PA), and the value to predict (VtP). The source of the original data is also cited (Ref). Additionally, the table provides the number of instances (# Instances) and attributes (# Attributes) post-preprocessing. The subsequent sections elaborate on the preprocessing steps for each dataset (if any), as well as previous studies that have employed these datasets.

C.1 Wine

The objective of this problem is to predict the quality of the wine considering the color as *PA*, white wines being the privileged value. The quality of the wines is between 3 and 9, we removed the lines with the highest quality (9) since the color of the wines within this category was only white. The final dataset contains a total of 6492 rows described by 14 attributes; white wines represent 75% of the dataset. It was included in the analysis due to the nature of the ordinal regression problem. Moreover, some authors also use this dataset for fairness analysis, as appears in [20, 25].

Table 10 Summary of datasets characteristics

Output type	Dataset	PA	VtP	# Instances	# Attributes	Refs
Ordinal	wine	Color	Quality	6,492	13	[39]
	compas	Race	Recidivism	5,278	9	[6]
	singles	Gender	Income	2,813	13	[40]
	obesity	Gender	Obesity	2,111	23	[41]
	drugs	Gender	Coke Recency	1,885	13	[42]
Continuous	insurance	Gender	Charges	1,338	9	[43]
	parkinson	Gender	UPDRS score	5,875	19	[44]
	older-adults	Gender	Mistakes	70	16	[45]
	crime	Race	% Crimes	1,993	98	[46]
	lsac	Race	ugpa	20,715	7	[7]
	student	Gender	Final Grade	649	39	[47]

C.2 Compas

Compas software has been used by judges to decide whether to reease an offender or not. Analysis in [6] shows that the software results are biased against black defendants. This study uses the same dataset that the authors published in their study [50]. [51] argue that similar precision can be obtained with a reduced set of attributes compared to the 137 attributes used by compas applying a more interpretable solution, such as a linear classifier.

We have considered the same attributes as taken into account by Angwin et al [6]. Compas scores for each defendant ranged from 1 to 10, with 10 being the highest risk. Analyzing the risk of criminal recidivism has been reduced to a classification task with two or three classes instead of solving a problem as an ordinal regression task, which is the nature of the problem with scores between 1 and 10.

We have solved two different problems: binary classification and ordinal regression. The target variable for the regression problem was the original numeric scale, while in binary classification, the target variable will take two values. The original analysis was performed taking into account three values representing the target variable $\{Low, Medium, High\}$, and in this paper we are combining $\{Medium, High\}$ to obtain the $High$ class.

The *PA* considered in this study is race and the privileged value 1 represents Caucasians, the group with a systematic advantage compared to African-Americans represented by the value 0. The objective is to predict the score; this value is between 1 and 10 and is considered high for values in the interval $[5, 10]$. In this way, the favorable label in the binary classification problem will be 0 representing the values between $[1, 4]$, both included. This is a well-studied problem that appears in several studies [19, 24, 52–54] that always solve the classification problem.

C.3 Singles

This dataset was extracted from the marketing dataset [40]. We have used the same preprocessing steps as the authors in [55]. First, the data are filtered, keeping only the data from single individuals. The *PA* is gender, and the objective is to predict the annual income of the individuals taking into account a total of 12 attributes. The total number of individuals in the analysis is 2813, 49% represented by females. The income to be predicted is on a scale from 1 to 9.

C.4 Obesity

This dataset contains a total of 2111 records with 22 attributes of different people and their level of obesity, which is the target variable to predict. The gender is considered *PA*, the dataset contains a 49% of females. We follow the same preprocessing presented by Do et al [56] by combining obesity types two and three into one category to finally have an obesity scale from 1 to 5.

C.5 Drugs

The drugs dataset is the result of a poll in which individuals answered the last time they consumed drugs. For each drug, they selected one answer: Never used the drug, used it more than a decade ago, or in the last decade, year, month, week, or day. We have performed the same preprocessing presented by Do et al [56] but considering the recency of Coke consumption as the value to be predicted (from 0 to 6 according to the answer, the value 0 corresponds to individuals who had never used the drug). We selected the gender as *PA*, the 50% of the individuals are male. The final dataset contains 1885 records with 12 attributes.

C.6 Insurance

This dataset contains records of patients that indicate several characteristics, such as sex, BMI, whether the patient is a smoker or not, and the total annual medical expenses charged to these individuals. The objective is to predict medical expenses, which is a continuous variable, we use the gender as *PA*. We followed the same preprocessing as the authors [53] without subsampling the data. To simulate a real-world unfair scenario, the authors took imbalanced samples, but we keep all the data. The final dataset is made up of 1,338 records with 8 explanatory variables. There are 50% males in the data.

C.7 Parkinson

The parkinson telemonitoring dataset contains records from 42 people with early-stage parkinson's disease. These records describe, among other attributes, 16 biomedical voice measures, and each row corresponds to one recording of one of these individuals. The objective is to predict the unified parkinson's disease rating scale (UPDR) score. As [56] we have removed unnecessary columns such as the subject number and 'motor UPDRS' which is another indicator that can be considered to be predicted and has a close relationship with 'total UPDR.' The gender was considered *PA* and in the final data we have 68% of females from 5875 records with 18 attributes.

C.8 Older-adults

This dataset was collected to measure the relationship between physical fitness and cognitive performance in older adults. The data contains some physical characteristics of the individuals such as height and weight and also some physical activity tests such as a 6-minute walk test and bicep curls. The objective is to predict the number of mistakes made on the StroopTask, taking those attributes into account. As appears in [23], gender was considered as the *PA*. There are a total of 70 individuals (and rows) in the dataset, 42 of them are female who represent 60% of the data. The number of explanatory variables is 15.

C.9 Crime

This dataset includes a total of 122 variables related to crime in communities in the USA. The attribute to be predicted is the violent crimes per population. The variables 'racepctblack,' 'racePctWhite,' 'racePctAsian' and 'racePctHisp' contain the percentage of the population belonging to each racial group, and we assign the racial group with the highest percentage of the population in each community. We have grouped the rest of the racial groups into a single group considered to be of unprivileged value.

At the preprocessing stage, we also removed all the missing values and some variables not useful for the analysis as the country, state, and the fold; so, the final number of exploratory variables is 97 and the number of cases is 1993. The white group represents 78.8% of the population. This dataset is one of the most used among the papers studied [19, 21, 22, 25, 26, 29, 52–54, 56].

C.10 Lsac

The law school admission council (lsac) dataset is a well-studied problem that appears in several of the articles reviewed [19, 21, 22, 26, 29, 52, 56]. The outcome to predict in this study is the Undergraduate Grade Point Average (ugpa) of students during law school, which is a continuous variable. We consider the race of students to be a sensitive factor, being white is the privileged attribute, and for the unprivileged attribute, we grouped the values Asian, black, hisp, and other. The final dataset contains 20715 records with 6 attributes. The white race is represented by 83% of the cases.

C.11 Student

The student performance dataset contains the student achievement in secondary education of two Portuguese schools. Data attributes include student grades, demographic, social, and school-related characteristics. We have removed the variables G1 and G2 strongly related to the continuous variable to predict G3, as well as the school attribute. The gender was selected as *PA*. The final data have 649 records with 38 explanatory variables. Females are the 60% of these data. This dataset is analyzed in different studies [22, 56].

Funding Funding for open access charge: Universidad de Granada/CBUA. Professor Jorge Casillas has received financial support from Grant no. PI20/01435—funded by the National Institute of Health Carlos III (ISCIII) of Spain and co-funded by the European Union—and grant no. C-ING-206-UGR23—Applied Research Projects of the University of Granada Research and Transfer Plan 2023, funded by the Andalusia ERDF Operational Program 2021-2027.

Data availability Appendix C describes all the datasets used for the experimentation. All of them have been taken from previous studies referenced in the manuscript. The code for all the experiments can be found in the repository Fairness in regression on Github. The algorithms used are not intended to constitute a contribution of this paper, nor should the results obtained be regarded as benchmarks for the problems studied.

Declarations

Ethical approval This declaration is not applicable.

References

1. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for AI. Commun. ACM **64**(7), 58–65 (2021). https://doi.org/10.1145/3448250
2. Castelluccia, C., Métayer, D.L.: Understanding algorithmic decision-making: Opportunities and challenges. Tech. rep, European Parliament Study (2019)
3. Makhlouf, K., Zhioua, S., Palamidessi, C.: Machine learning fairness notions: bridging the gap with real-world applications. Inform. Process Manag. (2021). https://doi.org/10.1016/j.ipm.2021.102642
4. Kearns, M.: The Science of Socially Aware Algorithm Design. Oxford University Press, Cambridge (2019)
5. Casillas, J.: Bias and Discrimination in Machine Decision-Making Systems, pp. 13–38. Springer, Cham (2023)
6. Angwin, J., Larson, J., Mattu, S., et al.: Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed: 2022-05-28 (2016)
7. Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series. https://api.semanticscholar.org/CorpusID:151073942 (1998)
8. Valdivia, A., Sánchez-Monedero, J., Casillas, J.: How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. Int. J. Intell. Syst. **36**, 1619–1643 (2021). https://doi.org/10.1002/int.22354
9. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. MIT Press (2023)
10. Feuerriegel, S., Dolata, M., Schwabe, G.: Fair AI: Challenges and Opportunities. Bus. Inf. Syst. Eng. **62**(4), 379–384 (2020). https://doi.org/10.1007/s12599-020-00650-3
11. Dwork, C., Hardt, M., Pitassi, T., et al.: Fairness through Awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. Association for Computing Machinery, New York, NY, USA, p 214-226, https://doi.org/10.1145/2090236.2090255 (2012)
12. Mehrabi, N., Morstatter, F., Saxena, N., et al.: A survey on bias and fairness in machine learning. ACM Comput Surv 54(6). https://doi.org/10.1145/3457607 (2021)
13. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. Association for Computing Machinery, New York, NY, USA, FairWare 18, p 1-7, https://doi.org/10.1145/3194770.3194776 (2018)
14. Kusner, M.J., Loftus, J.R., Russell, C., et al.: Counterfactual fairness. In: Proceedings of Neural Information Processing Systems (2017)
15. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops, pp 13–18, https://doi.org/10.1109/ICDMW.2009.83 (2009)
16. de Langhe, B., Fernbach, P.: The dangers of categorical thinking. https://hbr.org/2019/09/the-dangers-of-categorical-thinking (2019)
17. Corbett-Davies, S., Gaebler, J.D., Nilforoshan, H., et al.: The measure and mismeasure of fairness. J Mach Learn Res 24(1) (2024)
18. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data (2016). https://doi.org/10.1089/big.2016.0047
19. Komiyama, J., Takeda, A., Honda, J., et al.: Nonconvex Optimization for Regression with Fairness Constraints. In: Proceedings of the 35th International Conference on Machine Learning, vol 80. PMLR, Stockholm, Sweden, pp 2737–2746 (2018)
20. Zhao, C., Chen, F.: Rank-based multi-task learning for fair regression. In: Proceedings of the IEEE International Conference on Data Mining, (ICDM 2021). Institute of Electrical and Electronics Engineers Inc., Beijing, China, pp 916–925, (2019) https://doi.org/10.1109/ICDM.2019.00102
21. Agarwal, A., Dudík, M., Wu, Z.: Fair regression: Quantitative definitions and reduction-based algorithms. In: Proceedings of the 36th International Conference on Machine Learning, (ICML 2019). International Machine Learning Society (IMLS), Long Beach, CA, USA, pp 166–183 (2019)
22. Chzhen, E., Denis, C., Hebiri, M., et al.: Fair regression via plug-in estimator and recalibration. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Neural information processing systems foundation, Vancouver, Canada (2020)
23. Yan, S., Kao, H.T., Lerman, K., et al.: Mitigating the Bias of Heterogeneous Human Behavior in Affective Computing. In: Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII 2021). Institute of Electrical and Electronics Engineers Inc., Nara, Japan, https://doi.org/10.1109/ACII52823.2021.9597439 (2021)
24. Fitzsimons, J., Al Ali, A., Osborne, M., et al.: A general framework for fair regression. Entropy (2019). https://doi.org/10.3390/e21080741
25. Calders, T., Karim, A., Kamiran, F., et al.: Controlling Attribute Effect in Linear Regression. In: Proceedings of the 2013 IEEE International Conference on Data Mining (ICDM 2013). IEEE, Dallas, USA, pp 71–80, https://doi.org/10.1109/ICDM.2013.114 (2013)
26. Steinberg, D., Reid, A., O'Callaghan, S., et al.: Fast fair regression via efficient approximations of mutual information. Preprint at (2020a) https://doi.org/10.48550/ARXIV.2002.06200
27. Steinberg, D., Reid, A., O'Callaghan, S.: Fairness measures for regression via probabilistic classification. Preprint at https://doi.org/10.48550/ARXIV2001.06089 (2020b)
28. Yan, S., Huang, D., Soleymani, M.: Mitigating Biases in Multimodal Personality Assessment. In: Proceedings of the International Conference on Multimodal Interaction (ICMI '20). Association for Computing Machinery, Inc, pp 361–369, https://doi.org/10.1145/3382507.3418889 (2020)
29. Narasimhan, H., Cotter, A., Gupta, M., et al.: Pairwise Fairness for Ranking and Regression. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), New York, USA, pp 5248–5255 (2020)
30. Kleindessner, M., Samadi, S., Zafar, M.B., et al.: Pairwise fairness for ordinal regression. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2022)
31. Mann, H., Whitney, D.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**, 50–60 (1947)
32. Cumbo, L.A., Ampry-Samuel, A., Rosenthal, H.K., et al.: A local law to amend the administrative code of the city

of new york, in relation to automated employment decision tools 2021/144. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page (2021)
33. Filippi, G., Zannone, S., Hilliard, A., et al.: Local law 144: A critical analysis of regression metrics. Preprint at https://doi.org/10.48550/ARXIV.2302.04119 (2023)
34. Bellamy, R.K.E., Dey, K., Hind, M., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. https://doi.org/10.1147/JRD.2019.2942287 (2019)
35. Bird, S., Dudik, M., Edgar, R., et al.: Fairlearn: A toolkit for assessing and improving fairness in ai. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/, accessed: 2019-11-16 (2020)
36. Saleiro, P., Kuester, B., Hinkson, L., et al.: (2018) Aequitas: A bias and fairness audit toolkit. https://doi.org/10.48550/ARXIV.1811.05577
37. Adebayo, J.: FairML : ToolBox for diagnosing bias in predictive modeling. S.M. diss., Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology., Massachusetts, USA (2016)
38. Fabris, A., Messina, S., Silvello, G., et al.: Algorithmic fairness datasets: the story so far. Data Mining Know. Discovery (2022). https://doi.org/10.1007/s10618-022-00854-z
39. Cortez, P., Cerdeira, A., Almeida, F., et al.: Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. **47**(4), 547–553 (2009). https://doi.org/10.1016/j.dss.2009.05.016
40. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. 0172-7397, Springer New York, NY (2009)
41. Palechor, F.M., de la Hoz, Manotas A.: Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. Data Brief **25**, 104344 (2019). https://doi.org/10.1016/j.dib.2019.104344
42. Fehrman, E., Muhammad, A., Mirkes, E., et al.: The five factor model of personality and evaluation of drug consumption risk. Studies in Classification, Data Analysis and Knowledge Organization book series pp 231–242 (2017)
43. Lantz, B.: Machine Learning with R. Packt Publishing, Community experience distilled. (2013)
44. Tsanas, A., Little, M., McSharry, P., et al.: Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. IEEE Trans. Biomed. Eng. **57**(4), 884–893 (2010). https://doi.org/10.1109/TBME.2009.2036000
45. Ramnath, U., Rauch, L., Lambert, E.V., et al.: The relationship between functional status, physical fitness and cognitive performance in physically active older adults: A pilot study. PLoS ONE **13**, 4 (2018). https://doi.org/10.1371/journal.pone.0194918
46. Redmond, M., Baveja, A.: A data-driven software tool for enabling cooperative information sharing among police departments. Eur. J. Oper. Res. **141**(3), 660–678 (2002). https://doi.org/10.1016/S0377-2217(01)00264-8
47. Cortez, P., Silva, A.: Using data mining to predict secondary school student performance. In: Proceedings of the 15th European Concurrent Engineering Conference 2008, (ECEC 2008), 5th Future Business Technology Conference, (FUBUTEC 2008). EUROSIS, p 5-12 (2008)
48. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12:2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html (2011)
49. Frank, E., Hall, M.: A simple approach to ordinal classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **2167**, 145–156 (2001)
50. Larson, J., Roswell, M., Atlidakis, V.: Propublica compas analysis. https://github.com/propublica/compas-analysis (2016)
51. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Sci. Adv. **4**, 1–5 (2018). https://doi.org/10.1126/sciadv.aao5580
52. Berk, R., Heidari, H., Jabbari, S., et al.: A convex framework for fair regression. Preprint at https://doi.org/10.48550/ARXIV.1706.02409 (2017)
53. Chi, J., Tian, Y., Gordon, G.J., et al.: Understanding and mitigating accuracy disparity in regression. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021). PMLR, pp 1866–1876 (2021)
54. Aghaei, S., Azizi, M.J., Vayanos, P.: Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making. https://dblp.org/rec/journals/corr/abs-1903-10598.bib, arXiv:1903.10598 (2019)
55. Fish, B., Kun, J., Lelkes, A.: A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics Publications, Miami, USA, pp 144–152 (2016)
56. Do, H., Putzel, P., Martin, A.S., et al.: Fair generalized linear models with a convex penalty. In: Chaudhuri K, Jegelka S, Song L, et al (eds) Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 162. PMLR, pp 5286–5308 (2022)