GRINS FOUNDATION

GRINS – Growing Resilient, INclusive and Sustainable

"9. Economic and financial sustainability of systems and territories"

Codice Identificativo: **PE00000018**

Finanziato nell'ambito del Piano Nazionale di Ripresa e Resilienza PNRR
Missione 4 – Componente 2
Investimento 1.3 – Creazione di "Partenariati estesi alle università, ai
centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base"

SPOKE 4 - D4.2.1 – Definition of risk measures of digital finance and credit ratings of SMEs and new green/young enterprises based on artificial intelligence

A Rank Graduation Box for SAFE AI

December 2024

Document data		2
Title	Spoke 4 **BAC SUST-AI: Sustainable Artificial Intelligence in Finance** Contribution to deliverable D4.2.1 Definition of risk measures of digital finance and credit ratings of SMEs and new green/young enterprises based on artificial intelligence A Rank Graduation Box for SAFE AI	
Owner	Università di Pavia	
Contributors	Arianna Agosto Paola Cerchiello Paolo Giudici	
Document version	Final	
Last version date	31/12/2024	

Executive summary

Machine Learning (ML) methods are boosting the applications of Artificial Intelligence (AI) in all human activities. Differently from ordinary computer software and applications, AI not only converts inputs into outputs, but can also change the surrounding environment, with the risk of creating harm for individuals, organizations and the environment.

This is the reason why authorities, regulators and standard bodies around the world have begun to monitor the risks arising from the adoption of AI methods. For example, the European Union has introduced the Artificial Intelligence Act (AI Act), which puts forward a number of key compliance requirements to AI in terms of sustainability, accuracy, fairness and explainability, compulsory for high risk applications. The United States Department of Commerce has introduced an AI risk management framework, based on a similar set of compliance requirements, to be voluntarily adopted by organizations that employ AI applications.

The above developments require, to be practically implemented, the availability of a set of statistical metrics that can actually measure whether AI applications are compliant, as well as the risk of not being compliant.

A consistent set of metrics for AI compliance and risk management does not exist yet. To fill the gap, Giudici and Raffinetti (2023, Finance Research Letters) introduced a model for the evaluation of the compliance of the AI applications in finance to the European AI Act.
According to the Authors' perspective, such compliance can be translated into four "S.A.F.E." key principles: ``S'' for sustainability; ``A'' for accuracy; ``F'' for fairness; ``E'' for explainability. Their approach is based on the employment of Shapley-Lorenz values, a normalized variant of the well known Shapley values which inherits from the latter its computational complexity.

In this deliverable[1], the previous approach is extended in three main ways: a) we develop a model valid for all applications of AI, and not only for those in finance; b) we provide a model that is less computational intensive, and easier to interpret; c) we develop a model that is scalable, and generalizable to all AI compliance requirements.

In more detail, with reference to a) the paper considers different cases and a simulation study; with reference to b) we introduce metrics that are interrelated by means of a common mathematical root: the Lorenz curve, which is related to the Receiver Operating Characteristic curve. This allows the integration of all measures into an agnostic score that can be employed to assess the trustworthiness of any AI application. With reference to c), we develop a model that is based on ranks, and that it appears as a ``box'', as it can be enriched by new metrics as new compliance requirements emerge, in particular for generative AI, whose risks are still under investigation. For these reasons, we call our approach a ``Rank Graduation Box'' (RGB).

The deliverable is organized as follows: a methodology section introduces the theoretical framework of our proposal, and the description of the proposed RGB metrics; an application section presents some experimental scenarios, to assess the validity of the proposal; a conclusion section concludes the paper with some final remarks and comments.

---

[1] Golnoosh Babaei, Paolo Giudici, Emanuela Raffinetti (2025), A Rank Graduation Box for SAFE AI, *Expert Systems with Applications*, Volume 259, 125239, https://doi.org/10.1016/j.eswa.2024.125239

In summary, the key ingredients of the proposal are the Lorenz curve, the dual Lorenz curve and the concordance curve, which are statistical tools widely used to summarize the distribution of income and wealth. In the paper we propose to extend these tools to provide a set of metrics that can assess the compliance of AI applications, within a common unified framework.

A Rank Graduation Box for SAFE AI

Abstract: The growth of Artificial Intelligence applications requires to develop risk management models that can balance opportunities with risks. We contribute to the development of Artificial Intelligence risk models proposing a Rank Graduation Box (RGB), a set of integrated statistical metrics that can measure the "Sustainability" , "Accuracy", "Fairness" and "Explainability" of any Artificial Intelligence application. Our metrics are consistent with each other, as they are all derived from a common underlying statistical methodology: the Lorenz curve. The validity of the metrics is assessed by means of their practical application to both simulated and real data. The results from the comparison of alternative machine learning models to simulated data are aligned with the generating models, indicating regression models as the most accurate for Gaussian data, Random Forest models as the most robust and fair; and leading to model explanations similar to the true ones from the generating model. The results from the application of Random Forest models to real data show that the proposed RGB metrics are more interpretable and more consistent with the expectations, with respect to standard metrics such as AUC, RMSE and Shapley values. The evidence also shows that the RGB metrics are very general, and can be applied to any machine learning method, regardless of the underlying data and model.

1. Background and motivation

Machine Learning (ML) methods are boosting the applications of Artificial Intelligence (AI) in all human activities. Differently from ordinary computer software and applications, AI not only converts inputs into outputs, but can also change the surrounding environment, with the risk of creating harms for individuals, organisations and the environment.

This is the reason why policy makers, regulators and standard bodies around the world are issuing regulations and recommendations that AI developers, deployers and users should follow to manage the risks arising from the adoption of AI methods.

AI risk management requires to develop a consistent set of AI risk metrics that can be employed to monitor the compliance of AI applications. Such a set of metrics is not common practice, yet.

Recently, Giudici & Raffinetti (2023) have summarised the requirements in the existing regulations and recommendations into four main measurable "S.A.F.E." key principles: "S" for sustainability; "A" for accuracy; "F" for fairness; "E" for explainability.

Sustainability refers to the robustness of an AI system to extreme events, such as cyber attacks or environmental issues. The measurement of robustness is well known in the statistics and machine learning literature and it is usually conducted in terms of the difference between the accuracy of two different predictions, obtained respectively under normal and perturbed input data. Accuracy can be measured by the AUC or by the RMSE (see, e.g. Nair et al. (2022)).

The measurement of Accuracy is well known in the statistics and machine learning: the Root Mean Square Error (RMSE) is routinely employed for a

continuous response; the Area Under the ROC curve (AUC or AUROC) is typically employed for a binary response (see e.g. Hand & Till (2001) and Gneiting (2011). More recently, Raffinetti (2023) and Giudici & Raffinetti (2024)have introduced a new accuracy measure that can be applied to both type of responses, generalising the AUC to the continuous case.

Fairness is one of the most important requirements for AI applications. Unfairness indicates that the output leads to an unequal treatment of different population groups, by gender, age and nationality, for example. There are several recent research papers that deal with the measurement of fairness. Most of them are based on parity measures, which calculate the difference in accuracy of the AI output obtained separately on different population groups. Accuracy can be measured employing AUC or RMSE, see e.g. Le Quy et al. (2022).

Explainability is a requirement that has emerged with the development of highly accurate machine learning models which have, however, so many parameters that it is very difficult, to reconduct the output to the inputs which determine it. The measurement of explainability requires to attach to each input variable an importance weight which expresses its influence on the final output. A model is explainable if there exists at least one input variable which significantly impacts the output. In white-box models, such as regression models, such weights are functions of the estimated parameters. Instead, to explain black-box models, such as random forests and deep learning, further post-processing of the output is necessary. The most employed techniques are based on Shapley values (see, e.g. (Shapley, 1953; Lundberg & Lee, 2017; Giudici & Raffinetti, 2021).

A recent paper Giudici et al. (2024) suggested to measure the S.A.F.E. principles using state of the art metrics, such as AUC, RMSE and Shapley values which are not, however, consistent with each other. A more consistent measurement model was put forward by Giudici & Raffinetti (2023), who replaced traditional metrics with new ones based on Lorenz Zonoids, the multivariate extension of the Gini coefficient. While mathematically sound, their approach suffers from computational complexity, especially when a large number of ex-

3

planatory variables are considered.

What described so far can be summarised in a concise picture of the state of the art for AI risk measurement, which is reported in Table 1.

AI principle	Main reference	Main contribution
Sustainability	Nair et al. (2022)	Change in RMSE or AUROC with respect to extreme data
	Giudici et al. (2024)	Regularisation and model selection
	Giudici & Raffinetti (2023)	Concentration of Shapley-Lorenz values
Accuray	Hand & Till (2001); Gneiting (2011)	RMSE (continuous case); AUROC (binary case)
	Raffinetti (2023); Giudici & Raffinetti (2024)	RGA (binary and continuous case)
	Giudici et al. (2024)	AUROC, RMSE
	Giudici & Raffinetti (2023)	Lorenz Zonoids
Fairness	Le Quy et al. (2022)	Parity measures (difference in RMSE or AUROC)
	Giudici et al. (2024)	Variability of Shapley values
	Giudici & Raffinetti (2023)	Variability of Shapley-Lorenz values)
Explainability	Shapley (1953)	Shapley-values in game theory
	Lundberg & Lee (2017)	Shapley-values in machine learning
	Giudici & Raffinetti (2021)	Shapley-Lorenz values: normalised Shapley
	Giudici et al. (2024)	Shapley values
	Giudici & Raffinetti (2023)	Shapley-Lorenz values

Table 1: Literature review on the key principles for AI risk management

The aim of this paper is to propose a set of metrics for AI risk measurement that are consistent with each other, as the metrics proposed by Giudici & Raffinetti (2023), but are also simple to understand and to implement, as the traditional metrics proposed in Giudici et al. (2024).

To this purpose, we extend the approach by Giudici & Raffinetti (2023) in two main ways: a) we employ the Concordance curve and, specifically, the Rank Graduation Accuracy measure (RGA) to extend the AUROC metrics to the ordinal and continuous cases; b) in analogy with the construction of the RGA, we propose a "Rank Graduation Box (RGB)", a set of consistent metrics, all based on the notion of concordance between two cumulative distributions. Doing so, we obtain a set of metrics that is easy to interpret, similarly to the well known Area Under the ROC Curve; that is easy to implement and compute,

using a similar logic; that is generalisable to other AI compliance principles, that may emerge from future recommendations, as long as they require a comparison between two cumulative distributions.

We remark that both Giudici et al. (2024) and Giudici & Raffinetti (2023) consider financial applications. Finance provides very challenging machine learning problems and is highly regulated, thus justifying the development of AI risk metrics. However, the AI risk metrics that we are going to present can be similarly applied to other important fields of application of AI, such as medicine, automotive, robotics and fault tolerance.

The rest of the paper is organized as follows: Section 2 introduces the theoretical framework of our proposal and the description of the proposed RGB metrics; Section 3 presents the application of our proposal, respectively to a simulated and a real dataset, to assess the validity of the proposal; Section 4 concludes the paper with some final remarks and comments.

2. Methodology

In this section, we describe our proposal. Specifically, in Subsection 2.1, we show the functioning of the Lorenz, dual Lorenz and concordance curves, which represent the common mathematical background of all our proposed metrics; in Subsection 2.2 we define the proposed Rank Graduation Box metrics to assess Sustainability, Accuracy, Fairness and Explainability of AI applications. Finally, in Section 2.3, we introduce a set of statistical tests that can be employed to assess the statistical significance of the proposed metrics.

2.1. Theoretical background

The basic ingredients of our proposal are the Lorenz curve and the related notions of dual Lorenz curve and concordance curve, which are statistical tools widely used to summarise the distribution of income and wealth (see, e.g. Lorenz, 1905).

5

Formally, let Y^* and Y^{**} be any two statistical distributions (continuous, ordinal or binary), each defined on a set of n data points, which we are going to compare.

The Y^* values can be ordered in a non-decreasing sense to build the Lorenz curve L_{Y^*}. For $i = 1, \ldots, n$, the Lorenz curve is defined by the pairs: $(i/n, \sum_{j=1}^{i} y^*_{r^*_j}/(n\bar{y}^*))$, where r^*_j indicates the non-decreasing ranks of Y^* and $\bar{y}^*$ indicates the mean of Y^*.

The same Y^* values can be ordered in a non-increasing sense to build the dual Lorenz curve, L'_{Y^*}. For $i = 1, \ldots, n$, the dual Lorenz curve is defined by the pairs: $(i/n, \sum_{j=1}^{i} y^*_{r^*_{n+1-j}}/(n\bar{y}^*))$, where r^*_{n+1-j} indicates the non-increasing ranks of Y^*.

The concordance curve C can be derived ordering the Y^* values with respect to the ranks of the Y^{**} values. Let r^{**}_i, for $i = 1, \ldots, n$, denote the non-decreasing ranks of Y^{**}. For $i = 1, \ldots, n$, the concordance curve is defined by the pairs: $(i/n, \sum_{j=1}^{i} y^*_{r^{**}_j}/(n\bar{y}^*))$.

To illustrate the construction of the Lorenz, dual Lorenz and concordance curves, we now introduce a toy example. Let Y^* and Y^{**} be two vectors of $n = 10$ points, with $Y^* = \{12, 27, 48, 3, 34, 11, 0, 46, 75, 28\}$ and $Y^{**} = \{15, 12, 18, 24, 21, 27, 33, 30, 36, 39\}$. It follows that the mean value of Y^*, $\bar{y}^*$, is equal to 28.4. The x-axis and y-axis values of the Lorenz curve are reported in Table 2. The curve is then obtained joining $(n+1)$ points: the $(0,0)$ point with all points whose coordinates are specified in Table 2.

The x-axis values of the dual Lorenz curve are the same as those of the Lorenz curve, whereas the y-axis values are obtained cumulating the Y^* values in the reverse order. The resulting coordinate pairs of the dual Lorenz curve are reported in Table 3, and the dual Lorenz curve is then obtained joining the $(0,0)$ point with all the other points in Table 3.

For the concordance curve, the x-axis values are as before, whereas the y-axis are obtained cumulating the Y^* values according to the ranks of Y^{**}. Both are reported in Table 4. The concordance curve connects the $(0,0)$ point with all the other points in Table 4.

6

ID	y_i^*	$y_{r_i^*}^*$	$\sum_{j=1}^i y_{r_j^*}^*$	y-axis values $\left(\dfrac{\sum_{j=1}^i y_{r_j^*}^*}{n\bar{y}^*}\right)$	x-axis values $\left(\dfrac{i}{n}\right)$
1	12	0	0	0	0.1
2	27	3	3	0.01	0.2
3	48	11	14	0.05	0.3
4	3	12	26	0.09	0.4
5	34	27	53	0.19	0.5
6	11	28	81	0.29	0.6
7	0	34	115	0.40	0.7
8	46	46	161	0.57	0.8
9	75	48	209	0.74	0.9
10	28	75	284	1	1

Table 2: Construction of the Lorenz curve

ID	y_i^*	$y_{r_{n+1-j}^*}^*$	$\sum_{j=1}^i y_{r_{n+1-j}^*}^*$	y-axis values $\left(\dfrac{\sum_{j=1}^i y_{r_{n+1-j}^*}^*}{n\bar{y}^*}\right)$	x-axis values $\left(\dfrac{i}{n}\right)$
1	12	75	75	0.26	0.1
2	27	48	123	0.43	0.2
3	48	46	169	0.60	0.3
4	3	34	203	0.71	0.4
5	34	28	231	0.81	0.5
6	11	27	258	0.91	0.6
7	0	12	270	0.95	0.7
8	46	11	281	0.99	0.8
9	75	3	284	1	0.9
10	28	0	284	1	1

Table 3: Construction of the dual Lorenz curve

Having clarified the calculation of the coordinates of the Lorenz, dual Lorenz and concordance curves with a toy example, we now represent them in four distinct stylised scenarios, in Figure 1.

Reading Figure 1, from the top left diagram and clockwise, it follows that: [a] when $r_i^{**} = r_i^*$, for all $i = 1, \ldots, n$, the concordance curve coincides with the Lorenz curve (full concordance); [b] when $r_i^{**} = r_{n+1-i}^*$, for all $i = 1, \ldots, n$, the concordance curve coincides with the dual Lorenz curve (full discordance); [c] when r_i^{**} are all equal to each other, for all $i = 1, \ldots, n$, the concordance curve overlaps with the 45-degree line (no association) meaning that, as all the Y^{**} values are tied, the Y^* values have to be replaced by their mean value $\bar{y}^*$ (see
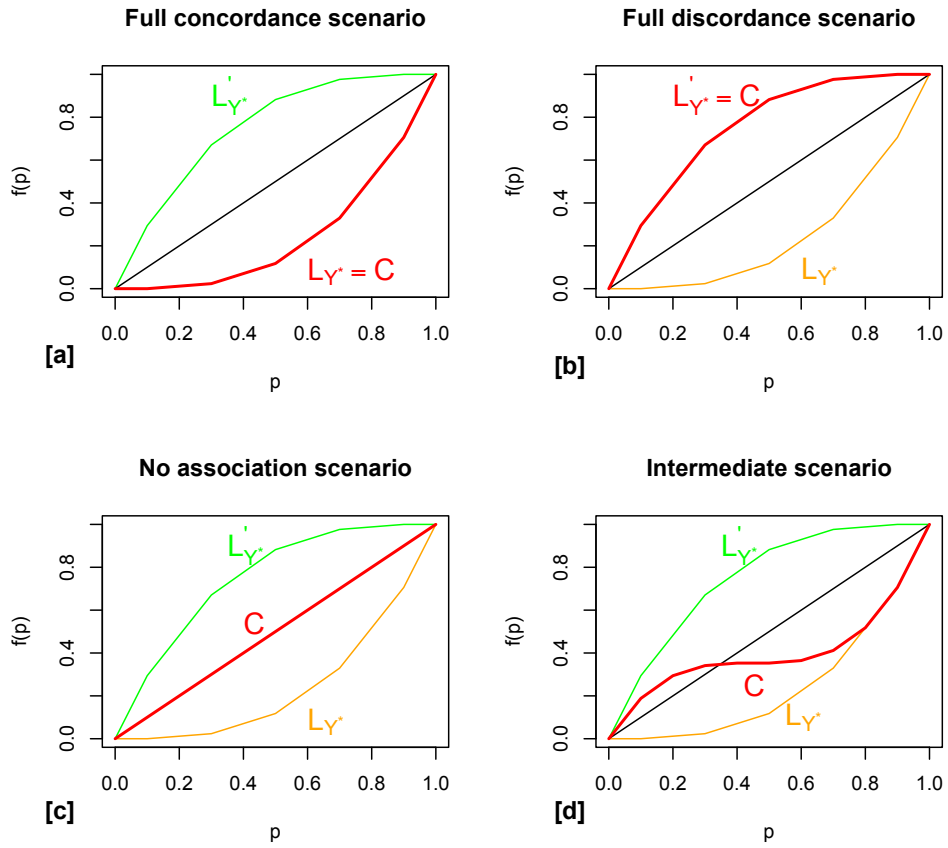
Figure 1: The L_{Y^*} and L'_{Y^*} Lorenz curves, the concordance curve C and the 45-degree line in the case of: [a] a full concordance scenario; [b] a full discordance scenario; [c] a no association scenario; [d] an intermediate scenario. Note that p (on the x-axis) and $f(p)$ (on the y-axis) are the cumulative values of the x and y coordinates of the L_{Y^*}, L'_{Y^*} and C curves.

ID	y_i^*	$y_{r_j^*}^{**}$	$\sum_{j=1}^i y_{r_j^*}^{**}$	y-axis values $\left(\frac{\sum_{j=1}^i y_{r_j^*}^{**}}{n\bar{y}^*}\right)$	x-axis values $\left(\frac{i}{n}\right)$
1	12	27	27	0.10	0.1
2	27	12	39	0.14	0.2
3	48	48	87	0.31	0.3
4	3	34	121	0.43	0.4
5	34	3	124	0.44	0.5
6	11	11	135	0.48	0.6
7	0	46	181	0.64	0.7
8	46	0	181	0.64	0.8
9	75	75	256	0.90	0.9
10	28	28	284	1	1

Table 4: Construction of the concordance curve.

Ferrari & Raffinetti, 2015); [d] in general, the distance between the concordance curve C and the Lorenz curve measures how the ranks of Y^{**} differ from the ranks of Y^*, in terms of the ranked values.

Figure 1 shows that the concordance curve C is always between the L_{Y^*} Lorenz curve and the L_{Y^*}' dual Lorenz curve.

We can leverage this observation building a summary measure of the distance between the concordance curve and the two Lorenz curves, which will be the kernel of all our Rank Graduation Box risk metrics. We can divide the area between the concordance curve C and the dual Lorenz curve by its maximum value, corresponding to the area between the Lorenz and dual Lorenz curves. Such a normalised summary will be named "Rank Graduation" (RG·) measure.More formally, the Rank Graduation measure is defined as:

$$RG\cdot := \frac{\sum_{i=1}^n \left\{ \frac{1}{n\bar{y}^*} \left(\sum_{j=1}^i y_{r_{n+1-j}^*}^* - \sum_{j=1}^i y_{r_j^*}^{**} \right) \right\}}{\sum_{i=1}^n \left\{ \frac{1}{n\bar{y}^*} \left(\sum_{j=1}^i y_{r_{n+1-j}^*}^* - \sum_{j=1}^i y_{r_j^*}^* \right) \right\}}$$
$$= \frac{\sum_{i=1}^n i y_{r_i^*}^{**} - \sum_{i=1}^n i y_{r_{n+1-i}^*}^*}{\sum_{i=1}^n i y_{r_i^*}^* - \sum_{i=1}^n i y_{r_{n+1-i}^*}^*}, \tag{1}$$

where, for any $i = 1, \ldots, n$: r_i^* are the non-decreasing ranks of the y_i^* values; r_i^{**} are the non-decreasing ranks of the y_i^{**} values; r_{n+1-i}^* are the non-increasing

ranks of the y_i^* values; $\bar{y}^*$ is the mean of the y_i^* values.

In summary, in this subsection we have shown how to extend the mathematical framework underlying the Lorenz curve to obtain a Rank Graduation measure, which will be the common basis of the Rank Graduation Box that is going to be described in the next Subsection.

2.2. Rank Graduation Box metrics

In this section, we will describe how to employ the Rank Graduation measure to derive AI risk metrics for "Sustainability", "Accuracy", "Fairness" and "Explainability". We remark that the derivation is rather general, and other metrics can be developed, once specified from regulatory viewpoint, using the same logic, and put in a comprehensive "Rank Graduation Box" aimed at assessing the overall risk of AI applications.

We first consider "Accuracy". Accuracy requires that the output of an AI application is "close" to the observed (or expected) output.

Given a target variable Y and a set of K predictors, a machine learning model can be applied to obtain predictions, $\hat{Y}$. The RGA measure can then be derived from RG· in equation (1) letting $Y^* = Y$ and $Y^{**} = \hat{Y}$.

We can then interpret Figure 1 in terms of accuracy, once Y^* is replaced by Y and Y^{**} by $\hat{Y}$. The best scenario [a] occurs when the predicted ranks of the response variable Y are equal to the observed ranks, with the concordance curve C perfectly overlapping the Lorenz curve L_Y (top left graph); the worst scenario [b] occurs when the predicted ranks of the response variable are equal to the reversed observed ranks, with the concordance curve C perfectly overlapping the dual Lorenz curve $L_Y^{'}$ (top right graph); the no association [c] scenario is achieved for a model that produces random predictions, leading to a concordance curve overlapping the 45 degree-line (bottom left graph); in the general scenario [d] (bottom right graph), the concordance curve C lies in the area between the Y response variable Lorenz curve, L_Y, and its dual, $L_Y^{'}$.

It can be shown that $0 \leq \text{RGA} \leq 1$, with $\text{RGA} = 1$ for a perfectly concordant model ($C \equiv L_Y$); $\text{RGA} = 0$ for a perfectly discordant model ($C \equiv L_Y^{'}$); RGA

= 0.5 for random predictions (C coincides with the 45-degree line).

We remark that the RGA metric was originally introduced by Raffinetti (2023) to measure accuracy, independently of the nature of the response variable. Indeed, when the response is binary, the RGA measure coincides with the AUROC (see, e.g. Giudici & Raffinetti, 2024). It can however be calculated, in the same manner, also for ordinal and continuous responses. In the remaining part of this subsection we show how the reasoning behind the RGA can be extended to other metrics, leading to a unified framework to assess the compliance and risk of AI, based on the RG· metric.

We now consider "Sustainability". Sustainability is related to the robustness of AI applications. The output of AI may be affected by extreme data, either internal or generated by cyber attacks, which can distort the inputs of an AI application and, consequently, its output.

The measurement of robustness is usually conducted in terms of an appropriate distance between the model predictions and those obtained under (intentional or non intentional) data perturbations.

The distance provided by the RG· metric can be quite useful in this context, as it is based on a distance between ranks which is, by definition, more robust to outliers and extreme data than distances based on the actual values, such as the Root Mean Square Error.

We propose to measure the robustness of AI applications by means of a "Rank Graduation Robustness" measure, denoted with RGR, which can be obtained setting in equation (1) $Y^* = \hat{Y}$ and $Y^{**} = \hat{Y}^p$, with $\hat{Y}$ and $\hat{Y}^p$ the predicted values obtained using the non-perturbed data and the predicted values obtained using perturbed data, respectively.

We can then interpret Figure 1 in terms of robustness, once Y^* is replaced by $Y^* = \hat{Y}$ and Y^{**} by $Y^{**} = \hat{Y}^p$.

From Figure 1, note that the case of maximum robustness [a] occurs when the ranks of the response predicted values $\hat{Y}$ correspond to the ranks of the predicted values obtained using the perturbed data, with the concordance curve C perfectly overlapping the Lorenz curve $L_{\hat{Y}}$; the worst case [b] occurs when

11

the ranks of the $\hat{Y}$ response predicted values are in inverse correspondence with the ranks of the predicted values obtained using the perturbed data, with the concordance curve C perfectly overlapping the dual Lorenz curve $L'_{\hat{Y}}$; the [c] case occurs when the concordance curve C overlaps the 45-degree implying that perturbations on the input data lead to random predictions; the general case [d] occurs when the concordance curve C lies in the area between the $\hat{Y}$ Lorenz curve, $L_{\hat{Y}}$, and its dual, $L'_{\hat{Y}}$.

It can be shown that RGR takes values in the close range $[0, 1]$. RGR $= 1$ for a fully robust model ($C \equiv L_{\hat{Y}}$); RGR $= 0$ for a fully perturbed model ($C \equiv L'_{\hat{Y}}$); RGR $= 0.5$ if the perturbations lead to a random model.

We now consider the measurement of "Explainability". The standard metrics for explainability are Shapley Lorenz and their variants, such as Shapley Lorenz values. Both are computationally intensive and their calculation, when a large amount of input variables is involved, may become infeasible.

In this paper, we propose to overcome this drawback adapting equation (1) to the measurement of explainability, leading to a "Rank Graduation Explainability" (RGE) metric. More precisely, we let $Y^* = \hat{Y}$ and $Y^{**} = \hat{Y}^{(-X_k)}$, where $\hat{Y}$ are the predicted values from a full model (which includes all K predictors) and $\hat{Y}^{(-X_k)}$ are the predicted values from a reduced model (which excludes the k-th predictor under evaluation).

We remark that the stronger is the effect of a variable X_k on explaining Y, the larger is the divergence between the ranks of the full model predicted values and those associated with the reduced model. Thus, for interpretational purpose, once let $Y^* = \hat{Y}$ and $Y^{**} = \hat{Y}^{(-X_k)}$, we will define RGE as RGE $= 1-$ RG$\cdot$.

Interpreting Figure 1 in terms of explainability, note that the case of mimimum explainability [a] occurs when the ranks of the response predicted values $\hat{Y}$ correspond to the ranks of the $\hat{Y}^{-X_k}$ predicted values (obtained by fitting the model without the k-th predictor), with the concordance curve C perfectly overlapping the Lorenz curve $L_{\hat{Y}}$ (left graph); the case of maximum explainability [b] occurs when the ranks of the $\hat{Y}$ response predicted values is in inverse cor-

12

respondence with the ranks of the $\hat{Y}^{(-X_k)}$ predicted values (obtained by fitting the model on the data without the k-th predictor), with the concordance curve C perfectly overlapping the dual Lorenz curve $L'_{\hat{Y}}$; the case [c] arises when the concordance curve C is equivalent to the 45-degree line meaning that the $\hat{Y}^{-X_k}$ predicted values are random predictions; the general case [d] occurs when the concordance curve C lies in the area between the $\hat{Y}$ Lorenz curve, $L_{\hat{Y}}$, and its dual, $L'_{\hat{Y}}$. Based on the previous considerations, it holds that $0 \leq \text{RGE} \leq 1$. Specifically: $RGE = 1$ if the k-th predictor provides the maximum explanation $(C \equiv L'_{\hat{Y}})$; $RGE = 0$ if the k-th predictor does not contribute to the explanation of the response $(C \equiv L_{\hat{Y}})$; $RGE = 0.5$ if the model without the k-th predictor corresponds to the random model (C coincides with the 45-degree line).

We finally consider the measurement of "Fairness". Fairness is typically measured by parity measures, which compare the AI outputs obtained separately on different population groups. This group based comparison is unconditional on the values of the other input variables, and it is not useful to assess, besides unfairness, its reasons. Furthermore, it may lead to paradoxes: the output may be fair marginally, but not conditionally on a control variable. For example, a credit score may be fair with respect to the gender of the borrower overall, but not conditionally on the level of income of the borrower.

We propose to overcome the difficulties of parity measures adapting formula (1) to the fairness context, leading to the "Rank Graduation Fairness" (RGF) metric.

More precisely, given a set G of group variables, converted into G binary variables, to calculate RGF we let $Y^* = \hat{Y}$ and $Y^{**} = \hat{Y}^{(-X_g)}$, where $\hat{Y}$ are the predicted values obtained with a full model (including all the G group variables as predictors); and $\hat{Y}^{(-X_g)}$ are the predicted values obtained with a reduced model (that excludes the g-th group binary variable under evaluation).

We can now interpret Figure 1 in terms of the RGF measure. The maximum fairness scenario [a] arises when the ranks of the response predicted values $\hat{Y}$ correspond to the ranks of the $\hat{Y}^{(-X_g)}$ predicted values (obtained by fitting the model without the g-th group variable), with the concordance curve C per-

13

fectly overlapping the Lorenz curve $L_{\hat{Y}}$ (left graph); the maximum unfairness condition [b] arises when the ranks of the $\hat{Y}$ response predicted values are in inverse correspondence with the ranks of the $\hat{Y}^{(-X_g)}$ predicted values (obtained by fitting the model on data without the g-th group variable), with the concordance curve C perfectly overlapping the dual Lorenz curve $L_{\hat{Y}}^{'}$; in the case [c], the concordance curve C coincides with the 45-degree line, indicating that the $\hat{Y}^{(-X_g)}$ predicted values without the g-th group variable correspond to a random model; in the intermediate scenario [d], the concordance curve C lies in the area between the $\hat{Y}$ Lorenz curve, $L_{\hat{Y}}$ and its dual, $L_{\hat{Y}}^{'}$. It can be shown that $0 \leq \text{RGF} \leq 1$. More precisely: $RGF = 1$ $(C \equiv L_{\hat{Y}})$ in the case of maximum fairness; $RGF = 0$ $(C \equiv L_{\hat{Y}}^{'})$ in the case of maximum unfairness. $RGF = 0.5$ if the model without the g-th group variable corresponds to the random model.

For the sake of clarity, the toy example introduced in Section 2.1 is further exploited in the Appendix Section to show how the RGA, RGR, RGE and RGF can be practically computed.

In summary, in this subsection we have shown how to adapt the Rank Graduation measure, illustrated in the previous subsection, to measure different compliance aspects of AI: Accuracy, by means of RGA; Sustainability, by means of RGR; Explainability, by means of RGE; and fairness, by means of RGF. The application of the metrics to an AI application can assess its compliance and risks.

2.3. Statistical tests

The introduced $RG\cdot$ metric can be complemented with statistical tests that can indicate whether the found degree of sustainability, accuracy, fairness or explainability is significantly high or not, taking into account the sampling variability.

To this aim, we now derive two statistical tests based on the $RG\cdot$ metric. The first test can be applied to RGA and RGR. For both cases, $RG\cdot$ can be expressed in terms of the covariance operator as follows:

$$RG\cdot = \frac{1}{2} \frac{cov(Y^*_{r(Y^{**})}, F(Y^*))}{cov(Y^*, F(Y^*))} + \frac{1}{2}, \tag{2}$$

where $Y^*_{r(Y^{**})}$ represents the Y^* variable re-ordered according to the ranks of Y^{**} and F is the cumulative continuous distribution function of Y^*.

It follows that the RG· is a linear function of the ratio:

$$\psi(Y^*, Y^{**}) = cov(Y^*_{r(Y^{**})}, F(Y^*))/cov(Y^*, F(Y^*)). \tag{3}$$

We recall that Y^* and Y^{**} have two different roles: for RGA, they represent the target variable Y to be predicted (i.e., $Y^* = Y$) and the predicted values $\hat{Y}$ (i.e., $\hat{Y} = Y^{**}$); for RGR, they represent the predicted values $\hat{Y}$, provided by the model fitted on non-perturbed data, and the predicted values $\hat{Y}^p$, derived from the same model fitted on perturbed data (i.e., $Y^{**} = \hat{Y}^p$).

Given two alternative models (Mod_1 and Mod_2), the statistics in (3) can be used to test the following hypotheses:

$$H_0 : \psi(Y^*_{Mod_1}, Y^{**}_{Mod_1}) = \psi(Y^*_{Mod_2}, Y^{**}_{Mod_2}) \quad \text{vs} \quad H_1 : \psi(Y^*_{Mod_1}, Y^{**}_{Mod_1}) \neq \psi(Y^*_{Mod_2}, Y^{**}_{Mod_2}), \tag{4}$$

where

$$\psi(Y^*_{Mod_1}, Y^{**}_{Mod_1}) = cov(Y^*_{r(Y^{**}_{Mod_1})}, F(Y^*_{Mod_1}))/cov(Y^*_{Mod_1}, F(Y^*_{Mod_1}))$$

and

$$\psi(Y^*_{Mod_2}, Y^{**}_{Mod_2}) = cov(Y^*_{r(Y^{**}_{Mod_2})}, F(Y^*_{Mod_2}))/cov(Y^*_{Mod_2}, F(Y^*_{Mod_2}))$$

are functions that derive from the application of (3), respectively to $RG\cdot_{Mod_1}$ and $RG\cdot_{Mod_2}$.

Note that the estimator of $\psi(Y^*_{Mod_1}, Y^{**}_{Mod_1})$ can be expressed as a function of two dependent U-statistics, denoted with U_1 and U_2:

$$\hat{\psi}(Y^*_{Mod_1}, Y^{**}_{Mod_1}) = \frac{U_1}{U_2} = \frac{\frac{1}{4\binom{n}{2}} \sum_{i=1}^{n}(2i-1-n)Y^*_{r(Y^{**}_{i\,Mod_1})}}{\frac{1}{4\binom{n}{2}} \sum_{i=1}^{n}(2i-1-n)Y^*_{r(Y^*_{i\,Mod_1})}}. \tag{5}$$

15

Similarly, the estimator of $\psi(Y^*_{Mod_2}, Y^{**}_{Mod_2})$ can be defined as a function of two dependent U-statistics, U_3 and U_4:

$$\hat{\psi}(Y^*_{Mod_2}, Y^{**}_{Mod_2}) = \frac{U_3}{U_4} = \frac{\frac{1}{4\binom{n}{2}}\sum_{i=1}^{n}(2i-1-n)Y^*_{r(Y^{**}_{i Mod_2})}}{\frac{1}{4\binom{n}{2}}\sum_{i=1}^{n}(2i-1-n)Y^*_{r(Y^*_{i Mod_2})}} \qquad (6)$$

It follows that $\delta = \psi(Y^*_{Mod_1}, Y^{**}_{Mod_1}) - \psi(Y^*_{Mod_2}, Y^{**}_{Mod_2})$ can be estimated by $\hat{\delta}$, a function of four dependent U-statistics:

$$\hat{\delta} = \hat{\psi}(Y^*_{Mod_1}, Y^{**}_{Mod_1}) - \hat{\psi}(Y^*_{Mod_2}, Y^{**}_{Mod_2}) = \frac{U_1}{U_2} - \frac{U_3}{U_4}. \qquad (7)$$

According to Hoeffding (1948), a function of several dependent U-statistics has a normal distribution, provided that the sample size is large enough. Thus, the estimator in equation (7) has a limiting normal distribution, whose variance $\widehat{Var(\hat{\delta})}$ can be estimated by means of the Jackknife method (see, e.g. Efron & Stein, 1981) each time omitting the pairs $(Y^*_{Mod_1}, Y^{**}_{Mod_1})$ and $(Y^*_{Mod_2}, Y^{**}_{Mod_2})$.

Therefore, a test statistic for testing the null hypothesis $H_0 : \psi(Y^*_{Mod_1}, Y^{**}_{Mod_1}) = \psi(Y^*_{Mod_2}, Y^{**}_{Mod_2})$ is:

$$Z = \frac{\hat{\delta}}{\sqrt{\widehat{Var(\hat{\delta})}}} \rightarrow N(0,1), \qquad (8)$$

where, for $i = 1, \ldots, n$: $\widehat{Var(\hat{\delta})} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\delta}_{(-i)} - \bar{\delta})^2$; $\hat{\delta}_{(-i)}$ are the values of $\hat{\delta}$ by omitting the pairs $(Y^*_{Mod_1}, Y^{**}_{Mod_1})$ and $(Y^*_{Mod_2}, Y^{**}_{Mod_2})$ at a time; $\bar{\delta}$ is the average of the values $\hat{\delta}_{(-i)}$.

For a fixed significance level α, a rejection region for the test corresponds to the region $|Z| \geq z_{\alpha/2}$. If the test statistic falls in this region, Mod_1 and Mod_2 are significantly different from each other.

Note that: in the case of RGA, $Y^*_{Mod_1} = Y^*_{Mod_2}$, being $Y^* = Y$, i.e. the observed target variable to be predicted based on two alternative models Mod_1 and Mod_2; in the case of RGR, $Y^*_{Mod_1} \neq Y^*_{Mod_2}$, being $Y^*_{Mod_1} = \hat{Y}_{Mod_1}$ and $Y^*_{Mod_2} = \hat{Y}_{Mod_2}$, i.e. the predicted values computed on non-perturbed data and provided by Mod_1 and Mod_2, respectively.

We now derive a second test, aimed to assess the significance of RGE and RGF. In this case Y^* corresponds to the predicted values obtained with a full model while Y^{**} corresponds to the predicted values generated by the reduced model, that is the model without the k-th variable, for RGE; and without the g-th group variable for RGF. It follows that a suitable test statistic can be defined as the difference between the denominator and the numerator of (3):

$$\gamma(Y^*, Y^{**}) = cov(Y^*, F(Y^*)) - cov(Y^*_{r(Y^{**})}, F(Y^*)). \tag{9}$$

By resorting to $\gamma(Y^*, Y^{**})$, the hypotheses

$$H_0 : \gamma(Y^*, Y^{**}) = 0 \quad vs \quad H_1 : \gamma(Y^*, Y^{**}) \neq 0, \tag{10}$$

can be tested.

Note that the estimator of $\gamma(Y^*, Y^{**})$ can be expressed as a function of two dependent U-statistics, denoted with $\tilde{U}_1$ and $\tilde{U}_2$:

$$\hat{\gamma}(Y^*, Y^{**}) = \tilde{U}_1 - \tilde{U}_2 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i-1-n) Y^*_{r(Y^*_i)} - \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i-1-n) Y^*_{r(Y^{**}_i)}. \tag{11}$$

By exploiting Hoeffding's Theorem (see, e.g. Hoeffding, 1948), (11) has a limiting normal distribution whose variance $Var(\hat{\gamma})$ can be estimated through the Jackknife method giving rise to the test statistic:

$$Z = \frac{\hat{\gamma}}{\sqrt{\widehat{Var(\hat{\gamma})}}} \rightarrow N(0, 1), \tag{12}$$

where, for $i = 1, \ldots, n$: $\widehat{Var(\hat{\gamma})} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\gamma}_{(-i)} - \bar{\gamma})^2$; $\hat{\gamma}_{(-i)}$ are the values of $\hat{\gamma}$ by omitting one pair (Y^*, Y^{**}) at a time; $\bar{\gamma}$ is the average of the values $\hat{\gamma}_{(-i)}$.

For a fixed significance level α, a rejection region for the test corresponds to the region $|Z| \geq z_{\alpha/2}$.

In summary, in this subsection we have shown how to improve the robustness of the results from the application of the proposed RGB metrics. Essentially, by

means of statistical tests that can verify whether the obtained assessments for Sustainability, Accuracy, Fairness and Explainability are statistically significant.

3. Empirical analysis

In this Section we apply the described Rank Graduation Box metrics. We consider both a simulated and a real data set.

Specifically, the simulation settings and the related results are presented and discussed in Section 3.1, whereas the application of our proposal to real surveys is described in Section 3.2.

3.1. Simulated data

For simulated data, we illustrate the experimental design in Subsection 3.1.1 and the obtained empirical findings in Subsection 3.1.2.

3.1.1. Experimental designs

We simulate $1,000$ times 100 five-dimensional samples (Y, X_1, X_2, X_3, X_4), from a five-dimensional Normal distribution, with mean $\boldsymbol{\mu}$ and correlation matrix $\boldsymbol{\Sigma}$. We assume, without loss of generality, that $\boldsymbol{\mu} = \{150, 20, 72, 5, 200\}$. For the correlation matrix, we consider two alternative specifications: $\Sigma = S_1$, as in Table 5; and $\Sigma = S_2$, as in Table 6, with different correlations between the target variable and the predictors. We also simulate the 1,000 samples from a non-Normal distribution, with mean and correlation matrices as before, plus a skewness parameter at $\nu = 3$ and a kurtosis parameter at $\kappa = 61$.

	Y	X_1	X_2	X_3	X_4
Y	1	0.6	0.4	0.3	0.1
X_1		1	0.2	0.7	0.3
X_2			1	0.05	0.1
X_3				1	0.5
X_4					1

Table 5: Correlation matrix S_1

	Y	X_1	X_2	X_3	X_4
Y	1	0.1	0.5	0.2	0.8
X_1		1	0.3	0.6	0.3
X_2			1	0.05	0.1
X_3				1	0.5
X_4					1

Table 6: Correlation matrix S_2

We thus obtain four experimental settings: (A) 1,000 samples of size 100 from a Normal distribution with correlation matrix S_1, (B) 1,000 samples of size 100 from a Normal distribution with correlation matrix S_2; (C) 1,000 samples of size 100 from a non-Normal distribution with correlation matrix S_1; (D) 1,000 samples of size 100 from a non-Normal distribution with correlation matrix S_2.

To measure robustness, in all four scenarios the predictors are then perturbed in the train set by replacing the left and right tails of the related distributions with outliers. Specifically, the values lower than the 15% percentile are replaced by observations sampled from a $U(-6, -4)$ distribution, whereas the values greater than the 85% percentile are replaced by observations sampled from a $U(15, 22)$ distribution.

We have then run three alternative machine learning models on the simulated data: a linear regression model, a regression tree model and a Random Forest model. While the linear regression model is explainable by design and the regression tree can be explained by visualizing its tree, the Random Forest is a full black-box model.

The proposed metrics (RGA, RGR, RGE and RGF) have then been applied to compare the three machine learning models in terms of accuracy, robustness, explainability and fairness. All models have been fit on the same training set (80% of the observations) and evaluated on the same test set (20% of the observations).

3.1.2. results

The results of the RGA, RGR, RGF and RGE metrics can be summarised by their mean and standard deviations along the 1,000 considered samples.

Table 7 and Table 8 show the results for experiment (A).

RGA	Mean	Standard deviation (sd)
Linear regression	0.8297	0.0722
Regression tree	0.7569	0.0882
Random Forest	0.7911	0.0835
RGR	Mean	Standard deviation (sd)
Linear regression	0.6543	0.1943
Regression tree	0.8222	0.1111
Random Forest	0.9445	0.0465
RGF	Mean	Standard deviation (sd)
Linear regression	0.6627	0.2917
Regression tree	0.9758	0.0801
Random Forest	0.9823	0.0181

Table 7: Summary statistics for RGA, RGR and RGF metrics, for experiment (A)

From Table 7 note that the linear regression model is the most accurate model, followed by the Random Forest and by the regression tree model. This is not surprising, as the generating model is a multivariate Gaussian distribution.

Moving to explainability, Table 8 shows that, in general, the most explainable predictor is X_1, which is indeed the variable with the highest correlation with Y. Removing variable X_1 from any of the three models leads to a change in the ranked values of about 3%-12%. The importance ranking of the predictors is almost the same for the tree models: X_1 is followed by X_2 and, then, by X_3 and X_4. This is as expected, being the true correlations with Y equal to 0.4, 0.3 and 0.1, respectively. A notable exception is the explanation for X_3 with the linear regression model, which has a very high mean, but also a very high standard deviation. Taking into account all four metrics, we can conclude that the Random Forest model is the best model, as the high performance of the linear regression on RGA may be due to its low robustness.

20

$\mathbf{RGE}_{X_1}$	Mean	Standard deviation (sd)
Linear regression	0.0307	0.0733
Regression tree	0.2379	0.1623
Random Forest	0.1116	0.0709
$\mathbf{RGE}_{X_2}$	Mean	Standard deviation (sd)
Linear regression	0.0242	0.0483
Regression tree	0.1388	0.1491
Random Forest	0.0703	0.0545
$\mathbf{RGE}_{X_3}$	Mean	Standard deviation (sd)
Linear regression	0.2594	0.2253
Regression tree	0.1371	0.1340
Random Forest	0.0354	0.0290
$\mathbf{RGE}_{X_4}$	Mean	Standard deviation (sd)
Linear regression	0.0031	0.0130
Regression tree	0.0466	0.0885
Random Forest	0.0436	0.0387

Table 8: Summary statistics for the RGE metric, for variables $(X_1, X_2, X_3, X4)$), for experiment (A)

The results for experiment (B) are displayed in Table 9 and in Table 10.

$\mathbf{RGA}$	Mean	Standard deviation (sd)
Linear regression	0.9736	0.0149
Regression tree	0.8828	0.0483
Random Forest	0.9323	0.0353
$\mathbf{RGR}$	Mean	Standard deviation (sd)
Linear regression	0.2335	0.1431
Regression tree	0.9010	0.0700
Random Forest	0.9591	0.0312
$\mathbf{RGF}$	Mean	Standard deviation (sd)
Linear regression	0.5784	0.2469
Regression tree	1,000	0.0012
Random Forest	0.9927	0.0062

Table 9: Summary statistics for the RGA, RGR and RGF metrics for experiment (B)

Table 9 highlights that the linear regression model is again the most accurate

$\mathbf{RGE}_{X_1}$	Mean	Standard deviation (sd)
Linear regression	0.0634	0.0951
Regression tree	0.0077	0.0247
Random Forest	0.0285	0.0243
$\mathbf{RGE}_{X_2}$	Mean	Standard deviation (sd)
Linear regression	0.1312	0.1270
Regression tree	0.0461	0.0742
Random Forest	0.0770	0.0487
$\mathbf{RGE}_{X_3}$	Mean	Standard deviation (sd)
Linear regression	0.1176	0.1370
Regression tree	0.0254	0.0489
Random Forest	0.0183	0.0147
$\mathbf{RGE}_{X_4}$	Mean	Standard deviation (sd)
Linear regression	0.2579	0.2264
Regression tree	0.3926	0.1365
Random Forest	0.1750	0.0883

Table 10: Summary statistics for the RGE metric for experiment (B)

model, followed by the Random Forest and the regression tree models.

As in Scenario (A), the linear regression has a worse performance in robustness and fairness. The Random Forest and the regression tree perform better.

In terms of explainability, Table 10 shows that, for all three models, the predictor X_4 contributes the most to the explanation of the target variable Y. Removing X_4 from any of the three models leads to a change in the ranked values of about 17.5%-26%. On the other hand, X_1 in general provides the smallest contribution in explaining the target variable Y. These findings are coherent with our expectations as X_1 and X_4 have, respectively, a low and a strong correlation with Y in Experiment (B). Overall, we can conclude that, for this experiment, both linear regression and Random Forest are good models, the choice depends on which metric is retained more important.

Consider now Experiment (C), whose results are displayed in Tables 11 and in Table 12.

From Table 11 it turns out that the linear regression model has again the highest accuracy. The regression model has also a good performance in robust-

RGA	Mean	Standard deviation (sd)
Linear regression	0.8263	0.0819
Regression tree	0.7707	0.0914
Random Forest	0.7994	0.0876
RGR	Mean	Standard deviation (sd)
Linear regression	0.9743	0.0268
Regression tree	0.9378	0.0859
Random Forest	0.9613	0.0364
RGF	Mean	Standard deviation (sd)
Linear regression	0.7842	0.1078
Regression tree	0.8865	0.1375
Random Forest	0.8968	0.0705

Table 11: Summary statistics for the RGA, RGR and RGF metrics for experiment (C)

$\mathbf{RGE}_{X_1}$	Mean	Standard deviation (sd)
Linear regression	0.0955	0.0698
Regression tree	0.2502	0.1234
Random Forest	0.1069	0.0657
$\mathbf{RGE}_{X_2}$	Mean	Standard deviation (sd)
Linear regression	0.0641	0.0578
Regression tree	0.1212	0.1101
Random Forest	0.0752	0.0543
$\mathbf{RGE}_{X_3}$	Mean	Standard deviation (sd)
Linear regression	0.0068	0.0133
Regression tree	0.0204	0.0458
Random Forest	0.0217	0.0185
$\mathbf{RGE}_{X_4}$	Mean	Standard deviation (sd)
Linear regression	0.0086	0.0155
Regression tree	0.0250	0.0542
Random Forest	0.0231	0.0233

Table 12: Summary statistics for the RGE metric in experiment (C)

ness, but it is still low performing in fairness.

Table 12 shows that, also in case of non-normally distributed data, the most important predictor is X_1, followed by X_2, X_4 and X_3. The highest explanation for X_1 is reached with the regression tree, which, however, is also the model where the RGE metric achieves the largest variability. In this scenario, removing predictor X_1 from any of the three models leads to a change in the ranked values of about 9.5%-25%. Overall, no one of the three models dominate the others in terms of all metrics. The choice therefore depends on which metric is considered most important.

We finally consider experiment (D), whose results are displayed in Table 3.1.2 and in Table 14.

RGA	Mean	Standard deviation (sd)
Linear regression	0.9666	0.0239
Regression tree	0.8666	0.0598
Random Forest	0.9134	0.0523
RGR	**Mean**	**Standard deviation (sd)**
Linear regression	0.9820	0.0200
Regression tree	0.9398	0.0742
Random Forest	0.9673	0.0295
RGF	**Mean**	**Standard deviation (sd)**
Linear regression	0.5727	0.2279
Regression tree	0.9996	0.0045
Random Forest	0.9804	0.0216

Table 13: Summary statistics for the RGA, RGR and RGF metrics in experiment (D)

Table 3.1.2, shows that the linear regression is more accurate than the Random Forest and regression tree. And it also appears as the most robust model in presence of extreme observations, followed by Random Forest and regression

24

$\mathbf{RGE}_{X_1}$	Mean	Standard deviation (sd)
Linear regression	0.0073	0.0112
Regression tree	0.0070	0.0212
Random Forest	0.0126	0.0105
$\mathbf{RGE}_{X_2}$	Mean	Standard deviation (sd)
Linear regression	0.0697	0.0541
Regression tree	0.1016	0.0815
Random Forest	0.0828	0.0498
$\mathbf{RGE}_{X_3}$	Mean	Standard deviation (sd)
Linear regression	0.0036	0.0076
Regression tree	0.0082	0.0238
Random Forest	0.0110	0.0091
$\mathbf{RGE}_{X_4}$	Mean	Standard deviation (sd)
Linear regression	0.1823	0.0938
Regression tree	0.3163	0.1349
Random Forest	0.1750	0.0862

Table 14: Summary statistics for the RGE metric in experiment (D)

tree. Differently, in terms of fairness, the linear regression model has the worst performance. In terms of explainability, Table 14 shows that the predictor which mostly affect the target variable Y is X_4, followed by predictor X_2. Removing predictor X_4 from any of the three models would imply a change in the ranked values of about 17%-32%.

To robustify the presented results, helping the final choice between the three models, we now carry out statistical tests to check whether the difference in the values of the metrics among the three models is statistically significant. To this aim, we apply the tests presented in Section 2.3 to compare $s = 3$ metrics. Under the null hypotheses, there is no difference among the expected value of the s metrics. Whereas, under the alternative hypotheses the expected values differ.

The results of the test, in terms of p-values, are reported in Table 15, for the RGA, RGR and RGF metrics, and in Table 16 for RGE.

The results of the statistical tests, reported in Tables 15 and 16, highlight that the metrics computed on the different models are all significantly different

Experiment (A)	Alternative hypothesis	p-values
RGA	$H_1 : RGA_{\text{lin reg}} > RGA_{\text{Rand For}} > RGA_{\text{reg tree}}$	0.001
RGR	$H_1 : RGR_{\text{Rand For}} > RGR_{\text{reg tree}} > RGR_{\text{lin reg}}$	0.001
RGF	$H_1 : RGF_{\text{Rand For}} > RGF_{\text{reg tree}} > RGF_{\text{lin reg}}$	0.001
Experiment (B)	**Alternative hypothesis**	**p-values**
RGA	$H_1 : RGA_{\text{lin reg}} > RGA_{\text{Rand For}} > RGA_{\text{reg tree}}$	0.001
RGR	$H_1 : RGR_{\text{Rand For}} > RGR_{\text{reg tree}} > RGR_{\text{lin reg}}$	0.001
RGF	$H_1 : RGF_{\text{reg tree}} > RGF_{\text{Rand For}} > RGF_{\text{lin reg}}$	0.001
Experiment (C)	**Alternative hypothesis**	**p-values**
RGA	$H_1 : RGA_{\text{lin reg}} > RGA_{\text{Rand For}} > RGA_{\text{reg tree}}$	0.001
RGR	$H_1 : RGR_{\text{lin reg}} > RGR_{\text{Rand For}} > RGR_{\text{reg tree}}$	0.001
RGF	$RGF_{\text{Rand For}} > H_1 : RGF_{\text{reg tree}} > RGF_{\text{lin reg}}$	0.001
Experiment (D)	**Alternative hypothesis**	**p-values**
RGA	$H_1 : RGA_{\text{lin reg}} > RGA_{\text{Rand For}} > RGA_{\text{reg tree}}$	0.001
RGR	$H_1 : RGR_{\text{lin reg}} > RGR_{\text{Rand For}} > RGR_{\text{reg tree}}$	0.001
RGF	$H_1 : RGF_{\text{reg tree}} > RGF_{\text{Rand For}} > RGF_{\text{lin reg}}$	0.001

Table 15: p-values for RGA, RGR, RGF

(p-value < 0.01) in all the considered experiments.

We can therefore conclude that the values of the metrics, previously discussed, are all statistically significant.

3.2. Real data

In this section, we show how the proposed metrics can be actually employed to check the compliance of AI applications on real data.

We consider the publicly available "Employee" dataset, which can be directly uploaded from the **stima** package in **R**. The data derives from a study carried out on the 473 employees of a bank, and includes information on their gender, age, educational degree (in terms of years of education), employment category (custodial, clerical, or manager), job time in months since hire, total work experience (total job time in months, since hire and from previous experiences), minority classification (that is, whether of an ethnic minority), starting salary (in dollars), and current salary (in dollars). For a better description see e.g. Ferrari & Raffinetti (2015).

Experiment (A)	Alternative hypothesis	p-values
RGE_{X_1}	$H_1 : RGE_{X_1\text{reg tree}} > RGE_{X_1\text{Rand For}} > RGE_{X_1\text{lin reg}}$	0.001
RGE_{X_2}	$H_1 : RGE_{X_2\text{reg tree}} > RGE_{X_2\text{Rand For}} > RGE_{X_2\text{lin reg}}$	0.001
RGE_{X_3}	$H_1 : RGE_{X_3\text{lin reg}} > RGE_{X_3\text{reg tree}} > RGE_{X_3\text{Rand For}}$	0.001
RGE_{X_4}	$H_1 : RGE_{X_4\text{reg tree}} > RGE_{X_4\text{Rand For}} > RGE_{X_4\text{lin reg}}$	0.001
Experiment (B)	**Alternative hypothesis**	p-**values**
RGE_{X_1}	$H_1 : RGE_{X_1\text{lin reg}} > RGE_{X_1\text{Rand For}} > RGE_{X_1\text{reg tree}}$	0.001
RGE_{X_2}	$H_1 : RGE_{X_2\text{lin reg}} > RGE_{X_2\text{Rand For}} > RGE_{X_2\text{reg tree}}$	0.001
RGE_{X_3}	$H_1 : RGE_{X_3\text{lin reg}} > RGE_{X_3\text{reg tree}} > RGE_{X_3\text{Rand For}}$	0.001
RGE_{X_4}	$H_1 : RGE_{X_4\text{reg tree}} > RGE_{X_4\text{lin reg}} > RGE_{X_4\text{Rand For}}$	0.001
Experiment (C)	**Alternative hypothesis**	p-**values**
RGE_{X_1}	$H_1 : RGE_{X_1\text{reg tree}} > RGE_{X_1\text{Rand For}} > RGE_{X_1\text{lin reg}}$	0.001
RGE_{X_2}	$H_1 : RGE_{X_2\text{reg tree}} > RGE_{X_2\text{Rand For}} > RGE_{X_2\text{lin reg}}$	0.001
RGE_{X_3}	$H_1 : RGE_{X_3\text{Rand For}} > RGE_{X_3\text{reg tree}} > RGE_{X_3\text{lin reg}}$	0.001
RGE_{X_4}	$RGE_{X_4\text{reg tree}} > H_1 : RGE_{X_4\text{Rand For}} > RGE_{X_4\text{lin reg}}$	0.005
Experiment (D)	**Alternative hypothesis**	p-**values**
RGE_{X_1}	$H_1 : RGE_{X_1\text{Rand For}} > RGE_{X_1\text{lin reg}} > RGE_{X_1\text{reg tree}}$	0.001
RGE_{X_2}	$H_1 : RGE_{X_2\text{reg tree}} > RGE_{X_2\text{Rand For}} > RGE_{X_2\text{lin reg}}$	0.001
RGE_{X_3}	$H_1 : RGE_{X_3\text{Rand For}} > RGE_{X_3\text{reg tree}} > RGE_{X_3\text{lin reg}}$	0.001
RGE_{X_4}	$H_1 : RGE_{X_4\text{reg tree}} > RGE_{X_4\text{lin reg}} > RGE_{X_4\text{Rand For}}$	0.001

Table 16: p-values for RGE, for variables (X_1, X_2, X_3, X_4)

We consider both classification and prediction machine learning problems for the employee data. In the classification problem, the response variable is binary. It is defined as a "doubling salary", obtained assigning a level of 1 to the employees that achieve a salary growth rate (ratio of the current salary to the starting salary) greater or equal to two, and a level of 0 otherwise. In the prediction problem, the response variable is directly the "salary growth", defined as the difference between two variables, the current salary and the starting salary.

We consider, for both the prediction and the classification problem, a Random Forest model. More specifically, the predicted values for the response variable are estimated for the test data (corresponding to the 30% of the whole dataset), employing the model fitted on the train dataset, using all explanatory variables, and including the 70% of the observations. We remark that the choice of a Random Forest model is taken without loss of generality, as it can ease the interpretation of the results, as we use the same type of models for both classification and prediction.

We apply the proposed RGA, RGR, RGE and RGF metrics to the predictions from the Random Forest model. In the case of RGR we consider perturbing all input variables, with a perturbation scheme different from what seen for the simulated data: we replace the values smaller than the 10th percentile with the values greater than the 90th percentile.

To improve the robustness of our results, we compare the proposed metrics with existing standard methods. This will provide a benchmark and a clearer picture of the advantages and limitations of the proposed metrics. Specifically, the RGA will be compared to the Root Mean Squared Error (RMSE) in the continuos case and to the AUC in the binary case; the RGR will be compared with the loss in accuracy when all input data are perturbed; the RGF will be compared with the difference in accuracy between the model applied separately on the different population groups; the explanations from RGE will be compared with those obtained applying Shapley values.

Table 17 report the results from the application of RGA, RGR and RGF,

along with the benchmark metrics, in the continuous case.

Accuracy	RGA	RMSE
	0.9149	8107.1686
Robustness	RGR	Accuracy Loss
	0.4697	399.6947
Fairness	RGF	Accuracy Difference
	0.9984	5244.0689

Table 17: RGA, RGR, RGF and benchmark measures, continuous response

From Table 17, it results that the RGA is equal to about 91%. We recall that the higher the RGA value, the better the concordance between the trained model and the actual values. Here, the RGA value is close to one, meaning that the predictive accuracy of the model is high. The result is statistically significant, as it turns out that, if we test the model against a random model, the p-value is very small, approximately equal to 10^{-253}, thus rejecting the null hypotheses of equality.

The benchmark measure for RGA is the RMSE which, in this case, is equal to about 8107. We cannot easily interpret this value, in absolute terms; we need to make relative comparisons. For example, as the mean salary growth is about 17409 \$, we can conclude that the RMSE is about half of the mean. The RGA , instead, is normalised and can be interpreted in both absolute and relative terms. A possible disadvantage of RGA is that, with respect to RMSE, it requires a further computational step of ordering the response values with respect to the ranks of the predictions. This, however, allows to compare models for a continuous response with models with a binary response.

The result of the application of RGR leads to a value of about 47%, which indicates that the Random Forest model is not so robust in terms of the considered perturbation of all input variables. However, tha model is significantly more robust then a random model, as the p-value of the RGR test is approximately equal to 10^{-27}. The RGR can be compared with the loss in accuracy (in RMSE) obtained moving from the model with non perturbed data to a model

with perturbed data. Such difference is equal to 399, about 5% of the original RMSE, indicating high robustness. This result seems, however, not very reliable, considering that all input variables have been perturbed. The RGR seems more suited to capture lack of robustness. A further advantage of RGR, with respect to the difference in accuracy, is that it is normalised. This allows an easier interpretation with respect to the accuracy loss, albeit at the cost of an extra computational cost.

We now discuss the obtained value for the RGF metrics, taking gender as the selected protected variable. The result is an RGF close to one, which seems to indicate that "gender" does not affect the ranks of the target variable "salary growth". The result is, however, not statistically significant, as the p-value for the statistical test on RGA is equal to 0.003: we reject the assumption of fairness. On the other hand, the standard "parity" measure which compares the RMSE of the model in the female and male group is equal to 5244: about 64% of the value of the overall RMSE, indicating a similar behaviour of the model in the two groups. The result is consistent with RGF, the difference is that RGF, although more expensive computationally, is normalised and, thus, easier to interpret.

We now consider the case of a binary response. Table 18 report the results from the application of RGA, RGR and RGF, along with the benchmark metrics, in the binary case, respectively.

Accuracy	RGA	AUC
	0.6848	0.6848
Robustness	RGR	Accuracy Loss
	0.8591	0.0333
Fairness	RGF	Accuracy Difference
	0.9366	0.1116

Table 18: RGA, RGR, RGF and benchmark measures, binary response

Table 18 highlights the clear advantage of our proposed metrics: we can compare the performance of the random forest model under two different con-

figurations of the response. While it is not possible to directly compare RMSE with AUC; or accuracy loss/difference expressed as difference of RMSE rather than difference in AUC, the RGB metrics are the same for the binary and the continuous case. Specifically, the comparison indicates that a continuous response is better predicted than a binary response (RGA = 91% vs 68%); that the model for the continuous response is less robust than the model for the binary response (47% vs 85%); that the model which predict salary growth (continuous) is more fair than the model which predicts salary doubling (binary) (99% vs 93%). The application of the statistical tests show that the model is significantly more accurate and robust than a random model; and that, similarly to what occurs for a continuous response, the model is not significantly fair, being the p-value equal to 0.001.

The results for standard metrics are consistent with those from our metrics; however, the latter have the advantage of a better interpretability and of a wider comparability, against the disadvantage of an increased computational cost.

We now compare the application of RGE with the standard explainability metrics: Shapley values. Table 19 report the results from the application of RGE, along with Shapley values, in both the continuous and the binary case.

Variable	RGE	Shapley values	Variable	RGE	Shapley values
jobtime	0.2112	1939.8556	jobtime	0.0355	0.0830
age	0.1901	2050.5184	age	0.0151	0.1018
previous experience	0.1197	557.7384	previous experience	0.0130	0.0888
education	0.1056	1143.2048	education	0.0180	0.0193
manager	0.0986	526.5470	manager	0.0129	0.0221
clerical	0.0985	221.5846	clerical	0.0004	0.0263
minority	0.0915	284.7139	minority	0.0016	0.0180
gender	0.0633	208.1586	gender	0.0015	0.0120
custodial	0.0563	208.2397	custodial	0.0004	0.0085

Table 19: RGE and Shapley values for a random forest model to predict a continuous reponse (left) and a binary response (right)

Table 19 shows that, consistently with the results obtained for RGA, the continuous response is better predicted by the random forest model. The most

important explanation of the salary growth is age, followed by jobtime, previous experience and also education. For the binary response, the ranking is similar, but the values are all smaller. Indeed, the standard Shapley values do not allow an easy comparison between the values of the two problems, binary and continuous. However, in both cases, the ranking are consistent with those of the RGE. Once more, the comparison shows that our proposed metric is, while consistent, easier to interpret than the benchmark. From a computational complexity viewpoint, Shapley values are computationally expensive, more than the RGE, as they require calculating the difference in predictions for all possible models.

A further comparison between our proposed RGB metrics and standard ones can be conducted by means of resampling. For example, if we focus on our core metrics, without loss of generality, and for the sake of space, we can calculate the value of RGA and of RMSE, for a continuous response, in 100 different train/sample tests. We can then calculate the mean and standard deviation of the values. they are reported in Table 20.

RGA	mean	standard deviation
	0.8962	0.0167
RMSE	mean	standard deviation
	7148.0864	693.8385

Table 20: Robustess comparison between RGA and RMSE

From Table 20 note that the RGA has a small standard deviation, about 50 times smaller than the mean. An approximate 99.7% empirical confidence interval ranges from 0.84 to 0.94, indicating a high value of accuracy. On the other hand, the RMSE has a higher standard deviation, about 10 times smaller than the mean. An approximate 99.7% empirical confidence interval ranges from 5069 to 9226, indicating a variable degree of accuracy.

4. Concluding Remarks

The paper has presented a set of consistent statistical metrics, all rooted in the framework of the rank graduation approach introduced by Lorenz (1905).

The implementation instructions of the proposed metrics, along with their corresponding Python code, and all data analysed in the paper are available at: https://github.com/GolnooshBabaei/safeaipackage. The results of the paper are therefore fully transparent and reproducible.

The paper has illustrated how these metrics can practically assess the compliance of any AI application to regulatory requirements such as Sustainability (Robustness), Accuracy, Fairness and Explainability.

Our experimental results show that the proposed metrics work quite effectively and can thus be a useful tool for the different stakeholders involved in the assessment of compliance of AI applications, and their risks: developers, deployers, consumers, regulators.

Further research is needed, from a methodological viewpoint, to enhance the robustness of the proposed methodology by comparing it with alternative modelisations. Further research is also needed, from an applied viewpoint, to test the functioning of the methodology on different field domains and applications, such as insurance, health and robotics, to name a few.

We would like to mention that what presented can be enriched by further research developments, without altering the reference framework. For example, different types of adversarial attacks can be considered, as different types of perturbations; explainability of group of variables could be considered; other ethical requirements can be included, such as human oversight and environmental sustainability.

Overall, our proposed approach to assess the compliance of AI applications is scientifically sound and relatively simple to implement and interpret. It has two possible limitations, which may be overcome in future research work. First, it assumes that the response variable to be predicted is unidimensional. A multidimensional response would require a multidimensional generalisation of

33

Lorenz curves. Second, it has been applied to numerical data. Further research should assess how the approach works for non-numerical data, such as text and images.

We finally remark that our devised approach is based only on the comparison of the outputs of machine learning models. It is therefore rather secure, as it does not involve neither new data inputs nor new model elaborations.

Appendix

With this Appendix we aim at showing how the RG· metric is calculated, based on data reported in Tables 2, 3 and 4.

The single terms involved in formula (1) are displayed in Table 21.

By summing all the terms appearing in the third, fifth and seventh column of Table 21 across the 10 observations, from formula (1) it results that RG· = 0.6031.

The proposed metrics can be computed based on the previous example by replacing the terms i) $y^*_{r^*_i}$, ii) $y^*_{r^*_{n+1-i}}$ and iii) $y^*_{r^{**}_i}$ with:

- the response variable values ordered in a non-decreasing sense; ii) the response variable values ordered in a non-increasing sense; iii) the response

ID	$y^*_{r_i*}$	$iy^*_{r_i*}$	$y^*_{r_{n+1-i}*}$	$iy^*_{r_{n+1-i}*}$	$y^*_{r_i**}$	$iy^*_{r_i**}$
1	0	0	75	75	27	27
2	3	6	48	96	12	24
3	11	33	46	138	48	144
4	12	48	34	136	34	136
5	27	135	28	140	3	15
6	28	168	27	162	11	66
7	34	238	12	84	46	322
8	46	368	11	88	0	0
9	48	432	3	27	75	675
10	75	750	0	0	28	280

Table 21: The terms of the RG· formula

variable values ordered according to the non-decreasing ranks of the predicted values (in the case of RGA);

- the predicted values, computed on non-perturbed data, ordered in a non-decreasing sense; ii) the predicted values, computed on non-perturbed data, ordered in a non-increasing sense; iii) the predicted values, computed on non-perturbed data, ordered according to the non-decreasing ranks of the predicted values computed on perturbed data (in the case of RGR);

- the predicted values, provided by the full model, ordered in a non-decreasing sense; ii) the predicted values, provided by the full model, ordered in a non-increasing sense; iii) the predicted values, provided by the full model, ordered according to the non-decreasing ranks of the predicted values provided by the reduced model, i.e. without the predictor under evaluation (in the case of RGE);

- the predicted values, provided by the full model, ordered in a non-decreasing sense; ii) the predicted values, provided by the full model, ordered in a non-increasing sense; iii) the predicted values, provided by the full model, ordered according to the non-decreasing ranks of the predicted values provided by the reduced model, i.e. without the group variable under evaluation (in the case of RGF).

References

Efron, B., & Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics*, *9(3)*, 586–596. DOI: 10.1214/aos/1176345462.

Ferrari, P., & Raffinetti, E. (2015). A Different Approach to Dependence Analysis. *Multivariate Behavioral Research*, *50(2)*, 248–264. DOI: 10.1080/00273171.2014.973099.

Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial intelligence risk measurement. *Expert Systems with applications*, *235*, 121220. DOI: 10.1016/j.eswa.2023.121220.

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems With Applications*, *167*, 114104. DOI: 10.1016/j.eswa.2020.114104.

Giudici, P., & Raffinetti, E. (2023). SAFE Artificial Intelligence in Finance. *Finance Research Letters*, *13*, 104088. DOI: 10.1016/j.eswa.2020.114104.

Giudici, P., & Raffinetti, E. (2024). RGA: a unified approach of predictive accuracy. *Advances in Data Analysis and Applications*, *Volume in press*. DOI: 10.1007/s11634-023-00574-2.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106(494)*, 746–762. DOI: 10.1198/jasa.2011.r10138.

Hand, D., & Till, R. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problem. *Machine Learning*, *45*, 171–186. DOI: 10.1023/A:101092081983.

Hoeffding, W. (1948). A Class of Statistcs with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, *19(3)*, 293–325. DOI: 10.1214/aoms/1177730196.

Le Quy et al. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12(3)*, e1452. DOI: 10.1002/widm.1452.

Lorenz, M. (1905). Methods of Measuring the Concentration Wealth. *Publications of the American Statistical Association*, *9(70)*, 209–219. DOI: 10.2307/2276207.

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777).

Nair et al. (2022). Maximum Likelihood Uncertainty Estimation: Robustness to Outliers. See URL `https://arxiv.org/abs/2202.03870`.

Raffinetti, E. (2023). A Rank Graduation Accuracy Measure to mitigate Artificial Intellgence risks. *Quality and Quantity*, *57(2)*, 131–150. DOI: 10.1007/s11135-023-01613-y.

Shapley, L. (1953). A Value for n-Person Games. *In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II*, *Princeton University Press*.