Applied Intelligence
https://doi.org/10.1007/s10489-021-03041-7

General Performance Score for classification problems

Isaac Martín De Diego[1] · Ana R. Redondo[1] · Rubén R. Fernández[1] · Jorge Navarro[1,2] · Javier M. Moguerza[1]

Accepted: 24 November 2021

Abstract
Several performance metrics are currently available to evaluate the performance of *Machine Learning (ML)* models in classification problems. *ML* models are usually assessed using a single measure because it facilitates the comparison between several models. However, there is no silver bullet since each performance metric emphasizes a different aspect of the classification. Thus, the choice depends on the particular requirements and characteristics of the problem. An additional problem arises in multi-class classification problems, since most of the well-known metrics are only directly applicable to binary classification problems. In this paper, we propose the *General Performance Score (GPS)*, a methodological approach to build performance metrics for binary and multi-class classification problems. The basic idea behind *GPS* is to combine a set of individual metrics, penalising low values in any of them. Thus, users can combine several performance metrics that are relevant in the particular problem based on their preferences obtaining a conservative combination. Different *GPS*-based performance metrics are compared with alternatives in classification problems using real and simulated datasets. The metrics built using the proposed method improve the stability and explainability of the usual performance metrics. Finally, the *GPS* brings benefits in both new research lines and practical usage, where performance metrics tailored for each particular problem are considered.

Keywords Performance metrics · Binary classification · Multi-class classification · Combination of information · Explainability

1 Introduction

Supervised Learning is the set of *Machine Learning (ML)* techniques that use labelled data. The task of these techniques is to learn a function that maps an input to a label, learning from examples of input-label pairs. When the label is categorical, the task addressed by these methods is referred to as classification. Based on the characteristics of the labels, several types of classification problems are defined: binary, multi-class, multi-labelled, and hierarchical [24].

In the literature, there are several metrics to evaluate the performance of *ML* models in classification problems [25]. Most of these metrics are defined for binary classification, of which some can be generalised for more than two classes. In practice, data analysts focus mainly on selecting the algorithm with the best predictive performance, disregarding the selection of the specific performance metric [6]. However, no general performance metric exists. Consequently, the proper definition of a performance metric, based on the problem domain and requirements, is crucial. Performance metrics are used to rank *ML* models and to evaluate if the selected one meets the classification requirements. Therefore, the choice of the right metric is crucial, especially when the cost of misclassification varies between classes.

In general, given a classification *ML* model, the information regarding its performance is summarised into a confusion matrix. This matrix is built by comparing the observed and predicted classes for a set of observations. It contains all the information needed to calculate most of the classification

✉ Isaac Martín De Diego
isaac.martin@urjc.es

Ana R. Redondo
anaisabel.rodriguez@urjc.es

Rubén R. Fernández
ruben.rodriguez@urjc.es

Jorge Navarro
j.navarro.2016@alumnos.urjc.es; jnavarro@sensowave.es

Javier M. Moguerza
javier.moguerza@urjc.es

[1] Data Science Laboratory, Rey Juan Carlos University, C/ Tulipán, s/n, 28933 Móstoles, Spain

[2] 2 Sensowave, Av. de Castilla, 1, 28830 San Fernando de Henares, Spain

Published online: 31 January 2022

performance metrics. Among them, *Accuracy (ACC)* is one of the most common. It represents the ratio of correctly predicted observations. However, in many binary classification problems, alternative measures that combine two metrics regarding the classification task in both classes are more appropriate.

In this paper, several performance metrics used in classification problems are discussed. The *General Performance Score (GPS)*, a new family of classification metrics, is presented. The *GPS* is obtained from the combination of several metrics estimated through a $K \times K$ confusion matrix, with $K \geq 2$. Therefore, this family of metrics performs for both binary and multi-class classification. Several instances of *GPS* are presented and compared with well-known alternative metrics from a theoretical and practical level.

The main contributions of the paper are listed as follows:

- A novel family of performance metrics, *GPS*, is developed for both binary and multi-class classification.
- *GPS* is configurable depending on the problem domain by combining appropriate performance metrics.
- *GPS* performance metrics allow a high explainability of the performance of the *ML* models.

The rest of the paper is structured as follows. Section 2 presents an overview of binary and multi-class classification metrics based on the confusion matrix. The proposed metrics family is described in Section 3 for both binary and multi-class classification. Experiments on simulated and real case studies with different number of classes are detailed in Section 4. Finally, Section 5 concludes and provides further research lines.

2 State of the art

2.1 Binary classification

In a binary classification problem, with classes -1 and $+1$, the performance metrics achieved by the selected *ML* classifier are obtained from the well-known 2×2 confusion matrix (see Table 1). This matrix relates the observed values to the ones predicted by the classifier. Notice that many *ML* models return probabilities. In these cases, a threshold on these probabilities can be used to obtain binary predictions. The elements of a confusion matrix are:

Table 1 Confusion matrix for binary classification

		Observed	
		-1	$+1$
Predicted	-1	TN	FN
	$+1$	FP	TP

- *True Positive* (TP): the observed $+1$ instances that are predicted as $+1$.
- *True Negative* (TN): the observed -1 instances that are predicted as -1.
- *False Positive* (FP): the observed -1 instances but predicted as $+1$.
- *False Negative* (FN): the observed $+1$ instances but predicted as -1.

FP and FN are also known as type I and type II errors, respectively. The relative importance of these errors depends on the problem under consideration [5, 21]. For instance, in anomaly detection problems, the number of observed $+1$ is usually much smaller than the number of observed -1. On the one hand, the FP are false alarms that should be treated by the system. This implies several actions with an associated cost. On the other hand, the FN are those anomalies that are not detected by the system and thus, could potentially damage it.

The performance metrics that can be obtained from a confusion matrix are summarised in Table 2. The most intuitive one is the *ACC* [9], which represents the ratio of correctly predicted instances among all instances in the dataset. The complementary metric is the *Error Rate (ERR)*, which evaluates the model by its proportion of incorrect predictions. Both metrics are commonly used by researchers to select a model. However, these two metrics turn out to be an over-optimistic estimation of the ability of the classifier over the

Table 2 Performance metrics based on a confusion matrix

Symbol	Metric	Formula
ACC	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
ERR	Error Rate	$\frac{FP+FN}{TP+TN+FP+FN}$
PPV	Precision	$\frac{TP}{TP+FP}$
TPR	Sensitivity/Recall	$\frac{TP}{TP+FN}$
TNR	Specificity	$\frac{TN}{TN+FP}$
NPV	Negative Predictive Value	$\frac{TN}{TN+FN}$
BA	Balanced Accuracy	$\frac{TPR+TNR}{2}$
GM	Geometric Mean	$(TPR \cdot TNR)^{1/2}$
FM	Fowlkes-Mallows Index	$(PPV \cdot TPR)^{1/2}$
F_1^+	F_1^+-score	$2 \cdot \frac{PPV \cdot TPR}{PPV+TPR}$
F_1^-	F_1^--score	$2 \cdot \frac{NPV \cdot TNR}{NPV+TNR}$
MK	Markedness	$PPV + NPV - 1$
BM	Bookmaker Informedness	$TPR + TNR - 1$
UPM	Unified Performance Measure	$2 \cdot \frac{F_1^+ \cdot F_1^-}{F_1^+ + F_1^-}$
MCC	Matthews Correlation Coefficient	$\frac{TP \cdot TN - FP \cdot FN}{((TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN))^{1/2}}$
KP	Cohen's Kappa	$\frac{ACC}{ACC + \frac{(TP+TN) \cdot (FP+FN)}{2 \cdot (TP \cdot TN - FP \cdot FN)}}$

majority class [4]. Consequently, they are sensitive to imbalanced data.

The *Precision*, also known as *Positive Predictive Value (PPV)*, can be considered as the probability of success when an instance is classified as +1. The *Sensitivity*, also known as *Recall* or *True Positive Rate (TPR)*, can be understood as the probability that an observed +1 is classified as +1 by the *ML* classifier. The *Specificity*, also known as *True Negative Rate (TNR)*, is the proportion of −1 instances that are correctly predicted. Similarly, the *Negative Predictive Value (NPV)* is the proportion of −1 instances correctly classified by the *ML* classifier. The main drawback of these metrics is that they do not consider all the confusion matrix elements. For example, the *Sensitivity* only focuses on positive examples, while *Specificity* only focuses on the negative ones. The main goal of *ML* classifiers is to improve the *Sensitivity*, without losing the *Specificity*. However, there is a trade-off between these two metrics since increasing the *Sensitivity* implies a decrease in the *Specificity* and vice versa. The same relationship appears between *Sensitivity* and *Precision*. Besides, *Precision* and *NPV* are sensitive to imbalanced data. Each of these four metrics cannot be used separately to evaluate the performance of a *ML* method because none of them takes into consideration the entire confusion matrix. This is, they do not take into account all information that the classifier provides. Hence, these metrics are adequate for capturing a partial perspective of the classifier performance, but are individually insufficient.

Regarding the four basic metrics, given three of them, the remaining fourth can be obtained. For instance, given *PPV*, *TPR*, and *TNR*, the *NPV* is defined as follows:

$$NPV = \frac{1}{1 + \frac{(1-PPV)}{PPV}\frac{TPR}{(1-TPR)}\frac{TNR}{(1-TNR)}} \tag{1}$$

The *Balanced Accuracy (BA)* is the arithmetic mean of *Sensitivity* and *Specificity*. That is, the average of two rates: positive instances correctly classified and negative instances correctly classified. The *BA*, unlike *Accuracy*, is robust for evaluating classifiers over imbalanced datasets.

Another useful metric is the geometric mean of *Sensitivity* and *Specificity*, konwn as *Geometric Mean (GM)* [25]. It can be used both with balanced and imbalanced data. Likewise, *Fowlkes-Mallows Index (FM)* [12] is defined as the geometric mean of *Sensitivity* and *Precision*. In contrast to *GM*, *FM* will approach zero with a random classification.

Notice that the harmonic mean is more intuitive than the arithmetic mean when computing a mean of ratios. Thus, the F_1^+ (usually called F_1-score [23]) is defined as the harmonic mean of *Precision* and *Recall*. Therefore, to achieve a high F_1^+ value, it is necessary to have both high values of *Precision* and *Recall*. Even though the F_1^+ is popular in statistics, it can be misleading since it does not consider the TN. Thus, this performance metric does not consider the ratio of −1 instances correctly classified by the *ML* classifier. Besides, F_1^+ is not invariant to class swapping.

Furthermore, it is possible to define the F_1^- [22] as the harmonic mean of *Specificity* and *NPV*. The F_1^- is a trade-off between the success of predicting an observation as −1 and the ratio of right predictions in the negative class. The F_1^- has the same strengths and weaknesses as the F_1^+, but focusing on the negative class. That is, it considers the TN but not the TP.

Markedness (MK) is defined as the distance of the sum of *Precision* and *NPV* to 1, while *Bookmaker Informedness (BM)* is defined as the distance between 1 and the sum of *Specificity* and *Sensitivity* [20]. Again, both measures complement each other, but do not provide an overall view of the different perspectives provided by the four metrics involved in their definitions. *MK* is sensitive to changes in data distributions and, hence, it is not appropriate for imbalanced data [25]. On the contrary, *BM* is suitable with imbalanced data. Nevertheless, it does not change concerning the differences between *Specificity* and *Sensitivity* [25].

In [22], a new metric that considers all the elements in the confusion matrix has been recently proposed. The *Unified Performance Measure (UPM)* is defined as the harmonic mean of F_1^+ and F_1^-. Thus, *UPM* assess the performance on both the positive and the negative class. This performance metric has high values only when the four fundamental metrics, *PPV*, *TPR*, *PNR*, and *NPV*, also have high values. In addition, *UPM* is suitable with imbalanced data [22].

In the same way, *Matthews Correlation Coefficient (MCC)* [16] also includes all the elements of the confusion matrix. *MCC* is defined as the geometric mean of the regression coefficients of the problem and its dual. It can be also formulated as follows:

$$MCC = \frac{1 - \frac{FP \cdot FN}{TP \cdot TN}}{(PPV \cdot TPR \cdot TNR \cdot NPV)^{1/2}} \tag{2}$$

However, *MCC* differs from the above-mentioned metrics as it takes values in the range $[-1, 1]$. On the one hand, $MCC = 1$ means that both classes are perfectly classified, as it occurs in the alternative metrics. On the other hand, $MCC = -1$ reveals a total disagreement between the observed and the predicted classes. $MCC = 0$ indicates a random prediction. It has been proven that *MCC* is not as stable as *UPM* [22].

The Cohen's Kappa coefficient measures the accordance between the *ML* classifier and the observed classes as follows:

$$KP = \frac{ACC - Pr(e)}{1 - Pr(e)} \tag{3}$$

Table 3 Confusion matrix for multi-class classification

		Observed			
		C_1	C_2	...	C_K
Predicted	C_1	C_{11}	C_{12}	...	C_{1K}
	C_2	C_{21}	C_{22}	...	C_{2K}
	...	...	...	...	...
	C_K	C_{K1}	C_{K2}	...	C_{KK}

where $Pr(e)$ is the hypothetical probability of agreement by chance, using the observed data to calculate the probabilities that each observer will randomly rank each category. The Cohen's Kappa coefficient also takes values from −1 to +1. The Cohen's kappa coefficient is more informative than *Accuracy* when working with imbalanced data. However, it is likely to give low values for imbalanced data [2].

Finally, the *Receiver Operating Characteristics (ROC)* graph is a technique for visualising, organising, and selecting classifiers based on their performance [8]. In this case, a set of confusion matrices is obtained by modifying parameters in the model. *ROC* graphs are two-dimensional representations in which two inversely related variables are plotted. For instance, *TPR* is usually plotted versus *False Positive Rate (FPR)* ($FPR = 1 - TNR$). These two metrics are calculated for each confusion matrix. Then, the *ROC* curve is plotted with *TPR* against the *FPR* where *TPR* is on the y-axis and *FPR* is on the x-axis. The *Area Under the ROC Curve (AUC)* [3] is the performance metric obtained from *ROC*. It is defined as the proportion of the unit square under the *ROC* curve. Thus, it takes values in the range [0, 1]. No realistic classifier should have an *AUC* less than 0.50. Although the *AUC* is generally used, it presents some drawbacks. For instance, the *AUC* lacks clinical interpretability because it does not reflect when diagnostic tests are presented in terms of gains and losses to individual patients [13].

2.2 Multi-class classification

Consider a multi-class classification problem with K classes to be predicted by a *ML* classifier. As in the binary classification, most performance metrics are obtained from the confusion matrix (see Table 3). In this matrix, the element C_{ij} $(i,j = 1,\dots,K)$ represents the number of the elements in class j classified as class i.

A common approach when dealing with multi-class classification problems is the One vs Rest technique [1]. It consists on facing each of the classes against the rest of them. Thus, the model is trained and evaluated on a binary setting where one of the classes is set to positive and the others to negative. This process is repeated for all classes obtaining a binary confusion matrix for each class. An instance of this approach is the generalisation of F_1^+ to multi-class classification, the $Macro - F_1^+$ [19]:

$$Macro - F_1^+ = \frac{\sum_{i=1}^{K} F_{1,i}^+}{n}, \tag{4}$$

where $F_{1,i}^+$ is the F_1^+ value obtained from the confusion matrix when the i-th class is faced against the rest of the classes. Analogously, *Macro-Precision*, *Macro-Recall*, $Macro - F_1^-$, and *Macro-Accuracy* can be defined. Notice that $Macro - F_1^+$ is an arithmetic mean of harmonic means.

An alternative to macro averages are micro averages. Since a FP for a given class is a FN for another class, all errors are considered the same in multi-class micro averages. The same reasoning applies to *TP* and *TN*. Thus, $FP = FN$ and $TP = TN$. In this context, the *Micro-Accuracy* (or multi-class accuracy) is defined as the ratio between the correctly predicted instances and the dataset size. Furthermore, the *Micro-Accuracy* equals the *Micro-Recall*, the *Micro-Precision*, and the *Micro-*F_1. When the dataset is imbalanced, *Micro-Accuracy* provides an overoptimistic estimation of the classifier performance over the majority class. Notice that these metrics are invariant to class swapping since $TP = TN$ and $FP = FN$.

There are also specific approaches to extend binary metrics to a multi-class setting such as multi-class *MCC* [10] and multi-class Cohen's Kappa coefficient [11]. Considering the $K \times K$ confusion matrix in Table 3, MCC_K for multi-class classification is defined as:

$$MCC_K = \frac{\sum_{ijl} C_{ii}C_{jl} - C_{ij}C_{li}}{(\sum_i(\sum_j C_{ij} \sum_{j'i',i'\neq i} C_{i'j'}))^{1/2}(\sum_i(\sum_j C_{ji} \sum_{j'i',i'\neq i} C_{j'i'}))^{1/2}} \tag{5}$$

The range of multi-class *MCC* is different from the binary *MCC*. In this case, the minimum value might be between −1 and 0 depending on the labels distribution, while the maximum value is the same.

Regarding the multi-class Cohen's Kappa coefficient, it is defined as follows:

$$KP = \frac{\sum_k^K C_{kk} \cdot \sum_i^K \sum_j^K C_{ij} - \sum_k^K p_k \cdot t_k}{(\sum_i^K \sum_j^K C_{ij})^2 - \sum_k^K p_k \cdot t_k} \tag{6}$$

where $p_k = \sum_i^K C_{ki}$ and $t_k = \sum_i^K C_{ik}$.

MCC and Cohen's Kappa are close in multi-class classification. The only difference between them is that the denominator is slightly lower in Cohen's Kappa coefficient, justifying slightly higher final scores.

3 General Performance Score

Several performance metrics to evaluate *ML* classifiers have been presented in the previous section. However, in some cases it is necessary to jointly consider a set of metrics that emphasise different aspects of the classifier. Thus, it is necessary to define an approach that combines a set of metrics into a single one. In this section, *GPS*, an approach to perform this combination, is presented.

Definition 1 Let $p_1, \cdots, p_n$ be n different performance metrics that describe the output of a *ML* model for a classification task, then the General Performance Score (*GPS*) is defined as follows:

$$GPS(p_1, \dots, p_n) = \frac{n}{\sum_{i=1}^{n} \frac{1}{p_i}} \tag{7}$$

Notice that the *GPS* is the harmonic mean of the set of different performance metrics $p_1, \cdots, p_n$. The harmonic mean is a measure of central tendency, which is useful when averaging rates like those obtained from the confusion matrix.

It can be proven that the *GPS* is also equal to:

$$GPS(p_1, \dots, p_n) = \frac{n \cdot \prod_{i=1}^{n} p_i}{\sum_{j=1}^{n} \prod_{\substack{i=1 \\ i \neq j}}^{n} p_i} \tag{8}$$

The *GPS* has the following properties:

Property 1 *When the set of n performance metrics are defined in* [0, 1], *the GPS is maximum*, *i.e.*, *equal to* 1, $\iff$ *all the performance metrics are maximum*, *i.e.*, *equal to* 1.

Property 2 *GPS is equal to* 0, *if at least one performance metric is equal to* 0.

Notice that the harmonic mean minimises the impact of large values while maximizing the impact of small values. Therefore, high values of *GPS* denotes that all of the the involved metrics have high values. Furthermore, it is possible to calculate the *GPS* standard deviation based on the standard deviation of the harmonic mean [17].

Property 3 *The standard deviation of GPS is*:

$$sd(GPS) = \frac{GPS^2}{(n-1)} \sqrt{\sum_{i=1}^{n} \left(\frac{1}{p_i} - \frac{1}{GPS} \right)^2} \tag{9}$$

It is clear that the standard deviation is minimum (and takes the zero value) when all the performance metrics (p_i) are the same. To study the maximum value for *sd*(*GPS*), first consider the binary case.

Property 4 *Given two performance metrics*, *the standard deviation of GPS is maximum when one of the metrics is* 1 *and the other is* $\frac{1}{3}$. *In this case*, *GPS*$= \frac{1}{2}$, *and* $sd(GPS) = \frac{1}{2\sqrt{2}}$.

Proof Given two performance metrics p_1 and p_2, the maximum distance between them is achieved when one metric is equal to 1 and the other is equal to 0. However, in that case, the *sd*(*GPS*) is not defined. To examine the maximum of the function, let $x = 1/p_1$ and $y = 1/p_2$. Thus, $x, y \geq 1$. Without loss of generality, we assume that $x \geq y$. Then, the *GPS* is:

$$\frac{2}{x+y} \tag{10}$$

and the *sd*(*GPS*) is:

$$2 \cdot \sqrt{2} \cdot \frac{(x-y)}{(x+y)^2} \tag{11}$$

The partial derivatives of the previous expression are:

$$f_x(x,y) = 2 \cdot \sqrt{2} \cdot \frac{(3 \cdot y - x)}{(x+y)^3} \tag{12}$$

$$f_y(x,y) = 2 \cdot \sqrt{2} \cdot \frac{(y - 3 \cdot x)}{(x+y)^3} \tag{13}$$

Given that $x \geq y$, we require that $x = 3y$. Thus, when $y = 1$, the derivative is 0 at $x = 3$. That is, $p_1 = 1/3$, and $p_2 = 1$. In such a case, $GPS = \frac{2 \cdot 1/3}{1+1/3} = \frac{1}{2}$, and $sd(GPS) = \frac{1}{2\sqrt{2}}$. Figure 1 shows the value for the $sd(GPS)$ at $x \in [1, 100]$ and $y \in [1, 10]$. Figure 1 shows the value for the *sd*(*GPS*) for all the values of x in [1, 100] at several values of y. It can be shown that the maximum is achieved for $y = 1$, $x = 3$.

It is straightforward to show the following property.

Property 5 *Given a set of n performance metrics* $p_1, \dots, p_n$, *the standard deviation of GPS is maximum when* $p_i = 1$ $\forall i = 1, \dots, n-1$, *and* $p_n = \frac{1}{n+1}$. *In such a case*, *GPS*$= \frac{1}{2}$, *and* $sd(GPS) = \frac{1}{4}\sqrt{\frac{n}{n-1}}$.

Proof Let be $x_i = 1/p_i$. Since $p_i \leq 1$, then $x_i \geq 1$, $\forall i$. Let be $s = \sum_{i=1}^{n} x_i$. Then, $GPS = n/s$, and

$$sd(GPS) = \frac{n^2}{s^2(n-1)} \sqrt{\sum_{i=1}^{n} \left(x_i - s/n \right)^2}$$

In order to maximise this expression, s needs to be as small as possible. Then x_i maximise the difference to the mean

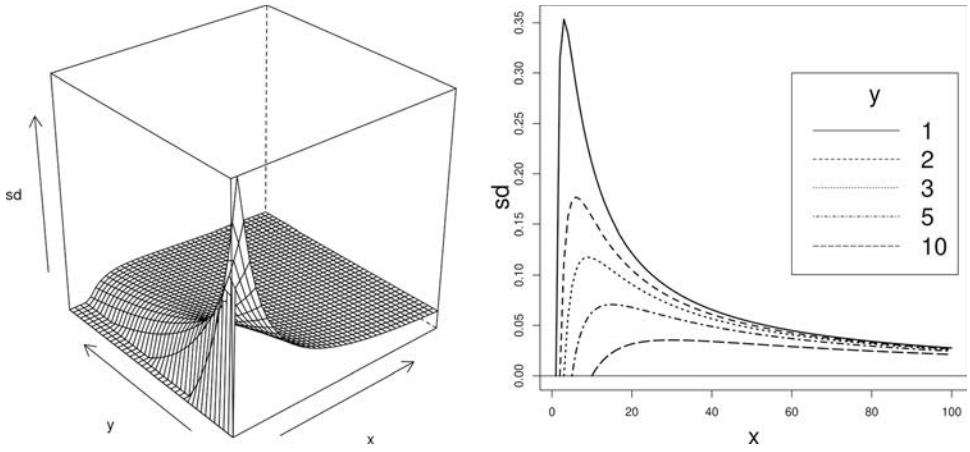

Fig. 1 The standard deviation of *GPS* when two performance metrics are considered: (a) 3D representation. (b) Standard deviation for a fixed y

value s/n for all i. To minimise s, $x_i = 1, \forall i < n$. Thus, $s = x_n + n - 1$. Now, the standard deviation is:

$$\begin{aligned} sd(GPS) &= \frac{n^2}{(x_n+n-1)^2(n-1)}\sqrt{(n-1)\left(\frac{x_n+n-1}{n}-1\right)^2+\left(\frac{x_n+n-1}{n}-x_n\right)^2} \\ &= \frac{n^2}{(x_n+n-1)^2(n-1)}\sqrt{(n-1)\left(\frac{x_n-1}{n}\right)^2+(n-1)^2\left(\frac{x_n-1}{n}\right)^2} \\ &= \frac{n^2}{(x_n+n-1)^2(n-1)}\sqrt{n(n-1)\left(\frac{x_n-1}{n}\right)^2} \\ &= \frac{(x_n-1)}{(x_n+n-1)^2}n\sqrt{\frac{n}{n-1}} \end{aligned}$$

The derivative of this expression is:

$$\frac{\partial sd(GPS)}{\partial x_n} = -\frac{(x_n-(n+1))}{(x_n+n-1)^3}n\sqrt{\frac{n}{n-1}}$$

The root of the derivative is $x_n = n+1$. Through the second derivative it can be demonstrated that it is a maximum. Thus, the standard deviation of *GPS* is achieved for $x_i = 1, \forall i < n$, and $x_n = n+1$. Therefore, $GPS = \frac{n}{n-1+n+1} = \frac{1}{2}$, and $sd(GPS) = \frac{1}{4}\sqrt{\frac{n}{n-1}}$.

3.1 Binary classification

In binary classification, a well-known particular case of *GPS* is the F_1^+-score. It corresponds to *GPS* parameterised with the *Precision* (*PPV*) and *Recall* (*TPR*):

$$GPS(PPV, TPR) = F_1^+ = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} \tag{14}$$

On the other hand, the F_1^--score is *GPS* parameterised with the *Specificity* (*TNR*) and *Negative Predictive Value* (*NPV*):

$$GPS(NPV, TNR) = F_1^- = 2 \cdot \frac{NPV \cdot TNR}{NPV + TNR} \tag{15}$$

The *UPM* [22] is another performance metric that belongs to the *GPS* family. The *UPM* is equals to *GPS* parameterised with *Precision* (*PPV*), *Recall* (*TPR*), *Specificity* (*TNR*) and *Negative Predictive Value* (*NPV*):

$$\begin{aligned} &GPS(PPV, TPR, TNR, NPV) = UPM \\ &= 4 \cdot \frac{PPV \cdot TPR \cdot TNR \cdot NPV}{PPV \cdot TPR \cdot NPV + PPV \cdot TPR \cdot TNR + NPV \cdot TNR \cdot PPV + NPV \cdot TNR \cdot TPR} \end{aligned} \tag{16}$$

Given that the combined harmonic mean of two sets of variables is equal to the harmonic mean of the harmonic means of the two sets [18], the previous expression can be easily simplified to:

$$GPS(PPV, TPR, TNR, NPV) = GPS(F_1^+, F_1^-) = 2 \cdot \frac{F_1^+ \cdot F_1^-}{F_1^+ + F_1^-} \tag{17}$$

This instance of *GPS* overcomes one of the main shortcomings of the F_1^+ and F_1^-, which is that they do not consider *TP* and *TN*, respectively. Thus, both metrics are misleading for imbalanced classes. Further, it performs properly for imbalanced classification problems, since it is built using information regarding the performance of a classifier on both classes. In addition, it improves the stability and explainability of the existing metrics [22].

Another possible instance of *GPS* is the combination of the *Specificity* (*TNR*) and *Sensitivity* (*TPR*):

$$GPS(TPR, TNR) = 2 \cdot \frac{TPR \cdot TNR}{TPR + TNR} \tag{18}$$

This same combination is performed by the *GM* and *BA* (see Section 2) that use the geometric and arithmetic mean, respectively. Since the harmonic mean is lower or equal than the geometric mean, and the geometric mean is lower or equal than the arithmetic mean, then:

$$GPS(TPR, TNR) \leq GM \leq BA \tag{19}$$

Let us consider two different *ML* models: ML_1 and ML_2. Let the performances of these models be as follows: *Specificity* = 0.4 and *Sensitivity* = 0.6, for ML_1, and *Specificity* = 0.1 and *Sensitivity* = 0.9, for ML_2. On the one hand, notice that *BA* = 0.5 for both models. On the other hand, *GM* is equal to 0.49 and 0.30 for ML_1 and ML_2, respectively, penalising the low value of *Specificity*. The proposed *GPS* results are: 0.48 and 0.18 for ML_1 and ML_2, respectively. Thus, as explained before, it can be seen that *GPS* is more sensitive to smaller values than to larger values in the involved metrics.

3.2 Multi-class classification

In this section, several instances of *GPS* in a multi-class classification problem are discussed. Lets consider a multi-class confusion matrix with *K*-classes (see Table 3). Applying a technique for switching from multi-class confusion matrices to binary matrices, it is possible to obtain *K* different binary confusion matrices. For instance, in this case the One vs Rest technique is used. Let be UPM_k (k in $1, \dots, K$) the calculated *UPM* for each of these *K* confusion matrices. Then, *GPS* can be parameterised with UPM_k in order to create a multi-class performance metric as follows:

$$GPS_{UPM} = GPS(UPM_1, UPM_2, \dots, UPM_k) = \frac{K \cdot \prod_{k=1}^{K} UPM_k}{\sum_{k'=1}^{K} \prod_{\substack{k=1 \\ k \neq k'}}^{K} UPM_k} \tag{20}$$

Consider a uniform confusion matrix such that all the elements in the matrix are equal, the following property can be defined:

Property 6 *Given a K-class classification problem. The value of* $\mathrm{GPS}_{\mathrm{UPM}}$ *for a uniform confusion matrix is*:

$$2 \cdot \frac{(K-1)}{K^2} \tag{21}$$

Proof Let consider all the elements in the uniform confusion matrix equal to x. First, notice that all *UPM*s in a uniform confusion matrix are equal. Since GPS_{UPM} is an harmonic mean of the *UPM*s, its value is equal to the value of the *UPM*s. Thus, it is enough to calculate one *UPM*. The UPM_k in a uniform confusion matrix is equal to:

Table 4 Binary confusion sub-matrices from a 3×3 confusion matrix

a)

		Observed	
		C_1	$C_2 \vee C_3$
Predicted	C_1	C_{11}	$C_{12} + C_{13}$
	$C_2 \vee C_3$	$C_{21} + C_{31}$	$C_{22} + C_{23} + C_{32} + C_{33}$

b)

		Observed	
		C_2	$C_1 \vee C_3$
Predicted	C_2	C_{22}	$C_{21} + C_{23}$
	$C_1 \vee C_3$	$C_{12} + C_{32}$	$C_{11} + C_{13} + C_{31} + C_{33}$

c)

		Observed	
		$C_1 \vee C_2$	C_3
Predicted	$C_1 \vee C_2$	$C_{11} + C_{12} + C_{21} + C_{22}$	$C_{13} + C_{23}$
	C_3	$C_{31} + C_{32}$	C_{33}

		Observed	
		$\bar{k}$	k
Predicted	$\bar{k}$	$(K-1)^2 x$	$(K-1)x$
	k	$(K-1)x$	x

The *Precision* and *Recall* are $\frac{1}{K}$, and the *NPV* and *Specificity* are $\frac{(K-1)^2}{(K-1)^2+(K-1)} = \frac{K-1}{K}$. Then, *UPM* is equal to:

$$\frac{4}{\frac{1}{1/K} + \frac{1}{1/K} + \frac{1}{(K-1)/K} + \frac{1}{(K-1)/K}} = \frac{4}{2 \cdot K + 2 \cdot \frac{K}{K-1}} = 2 \cdot \frac{(K-1)}{K^2}$$

As an example, let us consider a 3-classes classification problem. The 3×3 multi-class confusion matrix can be divided into 3 binary confusion sub-matrices (see Table 4). Then, $GPS(UPM_1, UPM_2, UPM_3)$ is defined as follows:

$$GPS_{UPM} = GPS(UPM_1, UPM_2, UPM_3) = \frac{3 \cdot \prod_{k=1}^{3} UPM_k}{\sum_{k'=1}^{3} \prod_{\substack{k=1 \\ k \neq k'}}^{3} UPM_k} \tag{22}$$

Notice that in the particular case of ordered classes, the confusion matrix in Table 4b could be omitted. When the order is relevant, merging the first and last classes could be meaningless for the application domain perspective. Then, the *GPS* implementation parameterised with *UPM* for ordered classes is defined as follows:

$$GPS(UPM_1, UPM_3) = 2 \cdot \frac{UPM_1 \cdot UPM_3}{UPM_1 + UPM_3} \tag{23}$$

Furthermore, alternative context-aware definitions of performance metrics could be useful. For instance, consider a multi-class classification problem where only the *Recall* of each class is relevant. Thus, the base metrics are:

$$Recall_k = \frac{C_{kk}}{\sum_{k'=1}^{K} C_{k',k}}, k = 1, \ldots, K.$$

In such a case, *GPS* is defined as follows:

$$GPS_{Recall} = GPS(Recall_1, \ldots, Recall_k) = \frac{K \cdot \prod_{k=1}^{K} Recall_k}{\sum_{k'=1}^{K} \prod_{\substack{k=1 \\ k \neq k'}}^{K} Recall_k} \quad (24)$$

Notice that when $K = 2$, then GPS_{Recall} is equal to the harmonic mean of *Specificity* and *Sensitivity*, presented in (18).

4 Experiments

In this section, several experiments on real and artificial datasets are considered. The properties and performance of *GPS*-based metrics are discussed and compared with alternative performance metrics. The first and second experiments consider a binary classification problem with simulated confusion matrices and real datasets, respectively. In the third experiment, a battery of simulated confusion matrices obtained from a multi-class classification problem is considered. Finally, in the fourth experiment, several definitions of *GPS* for two real dataset in multi-class classification problem are explored.

4.1 Simulated confusion matrices in binary classification

In this experiment, five confusion matrices are generated to compare *GPS*-based metrics against several alternatives. These confusion matrices are reported in Table 5. The confusion matrix a) presents a good classifier with adequate results in both classes. The confusion matrix b) is a random confusion matrix with the same values in all its cells. In the confusion matrices c) and d) only one class is correctly classified, negative class in c) and positive class in d). Finally, the confusion matrix e) presents a conservative classifier (most of the model predictions are negative) in an imbalanced dataset (most of the instances are positive).

Table 6 shows the results of the metrics for these confusion matrices. In this experiment, the *GPS*(*PPV*, *TPR*, *TNR*, *NPV*) has been considered. First, when the classification model works properly, as in a), all metrics achieve high values. The *GPS* instance presents low values in the confusion matrices c), d) and e) since at least one of its performance metrics presents low values. Regarding the random confusion matrix b), the *GPS* value is 0.5. It is interesting to remark that in this case, all the performance metrics used in its definition have the same value. Thus, the standard deviation of *GPS* is 0.0.

In confusion matrix e), the *Precision* and *Specificity* are very high, but the *Recall* and *NPV* are very low. In addition, it can be observed in the confusion matrices c) and d) that these metrics are sensitive to swapping the classes and to imbalanced data. The *Balanced Accuracy* obtains very similar values for the last four confusion matrices, although they represent totally different scenarios. It can be observed that the F_1^+ and the F_1^- metrics are sensible to imbalanced data. In the confusion matrix c), F^- achieves a high value while the positive class is almost entirely misclassified. On the other hand, in confusion matrix d), F_1^+ achieves a high value while the negative class is almost entirely misclassified. Moreover, they are sensitive to swapping the classes. The *Geometric Mean* value in the confusion matrices c) and d) is similar to the random confusion matrix b). The *Fowlkes-Mallows* Index obtains very similar values to F_1^+.

Table 5 Simulated 2×2 confusion matrices

	TN	FN	FP	TP
a)	40	10	10	40
b)	25	25	25	25
c)	90	4	5	1
d)	1	5	4	90
e)	5	94	0	1

Table 6 Performance metrics in the simulated binary confusion matrices

	a)	b)	c)	d)	e)
Accuracy	0.80	0.50	0.91	0.91	0.06
Precision	0.80	0.50	0.17	0.96	1.00
Sensitivity/Recall	0.80	0.50	0.20	0.95	0.01
Specificity	0.80	0.50	0.95	0.20	1.00
NPV	0.80	0.50	0.96	0.17	0.05
Balanced Accuracy	0.80	0.50	0.57	0.57	0.50
F_1^+-score	0.80	0.50	0.18	0.95	0.02
F_1^--score	0.80	0.50	0.95	0.18	0.09
Geometric Mean	0.80	0.50	0.43	0.43	0.10
Fowlkes-Mallows Index	0.80	0.50	0.18	0.95	0.10
Markedness	0.60	0.00	0.12	0.12	0.05
Bookmaker Informedness	0.60	0.00	0.15	0.15	0.01
Cohen's Kappa	0.60	0.00	0.13	0.13	0.00
MCC	0.60	0.00	0.13	0.13	0.02
GPS	0.80	0.50	0.30	0.30	0.03

Both *Markedness*, *Bookmaker Informedness* and *Cohen's Kappa* get low values for the last three confusion matrices, and 0.00 for the random confusion matrix b). Given the low performance on the non-predominant class, *GPS* achieves values lower than 0.50 for the confusion matrices c) and d). However, *MCC* achieves higher values for these confusion matrices (0.13 in both cases) than for the random confusion matrix b) (0.00). Moreover, *MCC* returns similar values for the confusion matrices b) (random) and a) (high *Precision* and low *Recall*).

4.2 Binary classification with real datasets

The performance of *GPS* for binary classification is also evaluated on several real datasets from the UCI Machine Learning Repository [7]. In this experiment, the following datasets are considered:

- Pima Indians and Vote datasets: two imbalanced datasets for the positive class.
- Ionosphere: an imbalanced dataset for the negative class.
- Sonar: a balanced dataset.
- Adult and Credit datasets: two very imbalanced datasets for the positive class
- Hepatitis: a very imbalanced dataset for the negative class.

Each dataset has been randomly split into two sets: training (80%) and testing (20%) sets. A *Random Forest (RF)* model with the following parameters has been trained on the training set: number of trees equals to 500, each tree grows to the maximum number of terminal nodes as possible, and the square root of the number of variables in the dataset is used as the number of variables randomly sampled as candidates at each split. Then, the metrics *MCC* and *GPS* are estimated over the testing sets. This process is repeated 100 times. Finally, the global performance metric values are obtained as the mean of the 100 performance score in the testing sets. The *Mean*, *Standard Deivation* (*SD*) and *Coefficient of Variation* (*CV*) for both *GPS* and *MCC* are shown in Table 7.

The correlation between both metrics is very high (Pearson correlation coefficient equals 0.98). However, *GPS* presents a lower standard deviation, which indicates that *GPS* is more stable. Furthermore, *MCC* obtains higher *CV* values, meaning that it is more dispersed than *GPS*. In addition, the *GPS* is easier to interpret since it is defined in the range [0, 1] as most performance metrics. Thus, it can be concluded that the proposed *ML* model performs properly for Vote and Ionosphere datasets. Better classifiers could probably be found for Sonar, Adult, and Pima Indians datasets. Finally, given the low values for *GPS*, the proposed classification technique shows a poor performance for Credit and Hepatitis datasets.

4.3 Simulated confusion matrices in multi-class classification

In this experiment, different simulated 3×3 confusion matrices are generated and presented in Table 8. The confusion matrices a) and b), show good classifiers on balanced datasets. The confusion matrices c) and d) correspond to very high imbalanced data. The confusion matrices e) and f) correspond to classifiers on imbalanced data. In the confusion matrix g) results from a bad classifier are presented. Finally, the confusion matrices h) and i) show very bad classifiers, completely wrong in their predictions. The following metrics have been calculated: *Accuracy*, *Macro-Accuracy*, *Macro-Precision*, *Macro-Recall*, *Macro-*F_1^+, *Macro-*F_1^-, *Micro-*F_1^+, *Micro-*F_1^-, *MCC* and GPS_{UPM}.

Table 9 shows the results of the metrics for these multi-class confusion matrices. First, when the classes are balanced and the classification error is not high, as in a) and b), all performance metrics achieve higher values. Notice that the metrics *Accuracy*, *Micro-*F_1^+ and *Micro-*F_1^- have the same results for all the proposed confusion matrices. In the confusion matrices c) and d), corresponding to imbalanced data, *ACC* and *Macro-Accuracy* are unreliable measures for model performance. The good performance of the model for the majority class implies high *ACC* and *Macro-Accuracy*, even when the performance of the model is low for

Table 7 Mean, *Standard Deviation (SD)* and *Coefficient of Variation (CV)* of the performance metrics *GPS* and *MCC* for real datasets

	GPS			*MCC*		
	Mean	*SD*	*CV*	Mean	*SD*	*CV*
Pima Indians	0.71	0.06	0.08	0.46	0.07	0.15
Sonar	0.82	0.06	0.07	0.67	0.11	0.16
Ionosphere	0.92	0.03	0.03	0.85	0.06	0.07
Hepatitis	0.56	0.18	0.32	0.37	0.19	0.51
Vote	0.95	0.02	0.02	0.90	0.05	0.06
Adult	0.79	0.00	0.00	0.60	0.00	0.00
Credit	0.61	0.01	0.02	0.39	0.01	0.03

Table 8 Simulated 3×3 confusion matrices

	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{31}	C_{32}	C_{33}
a)	90	10	10	10	90	10	10	10	90
b)	90	30	30	30	90	30	30	30	90
c)	30	0	30	0	9000	0	30	0	9000
d)	30	0	30	0	30	0	30	0	9000
e)	90	60	0	60	90	0	30	30	90
f)	90	60	0	60	90	60	0	60	90
g)	50	100	0	0	50	100	100	0	50
h)	0	150	0	0	0	150	150	0	0
i)	0	150	150	0	0	0	150	0	0

Table 9 Performance metric values in the simulated 3×3 confusion matrices

	a)	b)	c)	d)	e)	f)	g)	h)	i)
ACC	0.82	0.60	1.00	0.99	0.60	0.53	0.33	0.00	0.00
Macro-Accuracy	0.88	0.73	1.00	0.99	0.73	0.69	0.55	0.33	0.33
Macro-Precision	0.82	0.60	0.83	0.83	0.60	0.54	0.33	0.00	0.00
Macro-Recall	0.82	0.60	0.83	0.83	0.67	0.54	0.33	0.00	0.00
Macro-F_1^+	0.82	0.60	0.83	0.83	0.61	0.54	0.33	0.00	0.00
Macro-F_1^-	0.91	0.80	1.00	0.89	0.79	0.75	0.67	0.50	0.43
Micro-F_1^+	0.82	0.60	1.00	0.99	0.60	0.53	0.33	0.00	0.00
Micro-F_1^-	0.82	0.60	1.00	0.99	0.60	0.53	0.33	0.00	0.00
Cohen's Kappa	0.73	0.40	0.99	0.66	0.40	0.28	0.00	−0.50	−0.50
MCC	0.73	0.40	0.99	0.66	0.41	0.28	0.00	−0.50	−0.61
GPS_{UPM}	0.86	0.68	0.86	0.80	0.68	0.62	0.44	0.00	0.00

the other classes. By contrast, GPS_{UPM} penalises the poor performance of the model in any of the classes.

The GPS_{UPM} obtains the lowest possible value when all observations are wrongly classified. The GPS_{UPM} is similar in the confusion matrices a) and c). Nevertheless, its standard deviation is minimum in a) 0.0, but 0.15 in c). This evinces a non-homogeneous performance along the different classes in the problem. The same occurs in cases b) (standard deviation 0.0) and e) (standard deviation 0.07). Note that following Property 5, the maximum standard deviation is 0.31. The GPS_{UPM} value for confusion matrix g) implies a near-random performance. In fact, notice that the expected random value in each element of the diagonal is equal to the observed value 50 (450 observations to be distributed in 9 cells). Following Property 6, the GPS_{UPM} for a uniform 3×3 confusion matrix is 4/9.

The confusion matrices h) and i) show non-zero values for *Macro*-F_1^-, even though all the observations are misclassified. In these two confusion matrices, *MCC* obtains different values. Moreover, negative *MCC* values are difficult to interpret. This difficulty arises from the fact that the minimum *MCC* value depends on the distribution of the observed label. Finally, Cohen's Kappa coefficient achieves similar results to *MCC* in all the cases except for the example i). In that case, Cohen's Kappa coefficient performs similar to *GPS* providing the same values for h) and i).

4.4 Multi-class classification with real datasets

In the last experiment, *GPS*-based metrics are evaluated on multi-class datasets. Firstly, the three classes Connect-4 dataset [7] is used. Secondly, the four classes Vehicle dataset [7] is considered. Both datasets have been divided in training set (80%) to fit a *ML* model and testing set (20%).

In the Connect-4 dataset, a *RF* model with the following parameters has been trained on the training set: number of trees equals to 500, each tree grows to the maximum number of terminal nodes as possible, and the square root of the number of variables in the dataset is used as the number of variables randomly sampled as candidates at each split. For each observation in the testing set, the *ML* model returns the probability of belonging to each class. Given these probabilities, different thresholds are used to classify the elements. Thus, a set of confusion matrices is obtained.

Three *GPS*-based instances are considered to show that it can be built up depending on the particular problem specifications. First, the GPS_{UPM} as a summary metric is calculated. Next, the GPS_{Recall} is considered as a metric that focuses on the relevant instances retrieved from all the

relevant instances of all the classes in the problem. Finally, the $GPS_{Recall,Precision_3}$ is considered. In this case, it is calculated from the three *Recalls* and the *Precision* of class 3.

In Table 10, the confusion matrices that maximise the GPS_{UPM}, the GPS_{Recall} and the $GPS_{Recall,Precision_3}$ values respectively in the test dataset are presented. Table 11 shows the value of the metrics for each of confusion matrix. Notice that, in this case:

$$Precision_k = \frac{C_{kk}}{\sum_{k'=1}^{3} C_{k,k'}}, k = 1, \ldots, 3.$$

The GPS_{UPM} achieves its maximum value, 0.69 ± 0.08, in the confusion matrix a). The standard deviation of *GPS* has been calculated using Property 3 in Section 3. Notice that the range of the six basic metrics (three *Precisions* and three *Recalls*) is minimal for this case: a) 0.56, b) 0.64, c) 0.75. When only the *Recalls* are relevant, the maximum of GPS_{Recall} is 0.67 ± 0.07, corresponding to confusion matrix b). Since the *Precisions* are not considered, they can have more extreme values (range equals 0.64), while less extreme values are allowed for the *Recalls* (range equals 0.21). Finally, when the $GPS_{Recall,Precision_3}$ is used, a higher value of $Precision_3$ is obtained. In this case, the maximum value is achieved in confusion matrix *c*), (0.72 ± 0.08).

Secondly, *GPS*-based metrics are evaluated on the Vehicle dataset. In this case, the *ML* model selected is a *Support Vector Machines (SVM)* with linear kernel and cost equals to 1. For each observation in the testing set, the *ML* model returns the probability of belonging to each class. Given these probabilities, different thresholds are used to classify the elements. Thus, a set of confusion matrices is obtained.

In this case, six different *GPS*-based instances are considered to show that the classifier predictions that maximise the chosen performance metric will differ, depending on the *GPS* definition, leading to different confusion matrices.

Table 11 Performance metrics in the confusion matrices from the Connect-4 dataset. In bold, the maximum in each metric

	a)	b)	c)
$Precision_1$	**0.33**	0.27	0.21
$Precision_2$	**0.78**	0.72	0.62
$Precision_3$	0.88	0.91	**0.96**
$Recall_1$	0.47	0.58	**0.71**
$Recall_2$	0.62	0.66	**0.70**
$Recall_3$	**0.89**	0.79	0.60
GPS_{UPM}	**0.69**	0.68	0.62
GPS_{Recall}	0.62	**0.67**	0.66
$GPS_{Recall,Precision_3}$	0.67	0.71	**0.72**

First, the GPS_{UPM} as a summary metric is calculated. Next, the GPS_{NPV} is considered as a metric that measures the proportion of negative samples that were correctly classified respect to the total number of negative predicted samples. Later, the $GPS_{Precision}$ is considered as the inverse *NPV*, which represents the proportion of positive samples that were correctly classified with respect the total number of positive predicted samples. After, the $GPS_{NPV,Precision_1}$ is considered. In this case, it is calculated from the four *NPV*s and the *Precision* of class 1. Then, the GPS_{Recall} is considered as a metric that focuses on the relevant instances retrieved from all relevant instances of all classes of the problem. Finally, the $GPS_{Recall,Precision_4}$ is presented to show the changes related to the increase in the *Precision* of class 4.

In Table 12, the confusion matrices in the test dataset, obtained from the maximization of the different *GPS*-based instances are presented. GPS_{Recall}, and the $GPS_{Recall,Precision_4}$ values respectively in the test dataset are presented. Table 13 shows the values of the metrics for each confusion matrix.

The confusion matrix a) maximises GPS_{UPM}, being the maximum value 0.12 ± 0.04. The standard deviation of *GPS*

Table 10 Confusion matrices obtained from the maximization of different *GPS*-based in-stances in the Connect-4 dataset

a) GPS_{UPM}

		Observed		
		C_1	C_2	C_3
Predicted	C_1	606	611	636
	C_2	215	2061	353
	C_3	461	662	7905

b) GPS_{Recall}

		Observed		
		C_1	C_2	C_3
Predicted	C_1	747	768	1226
	C_2	237	2205	601
	C_3	298	361	7067

c) $GPS_{Recall,Precision_3}$

		Observed		
		C_1	C_2	C_3
Predicted	C_1	907	905	2411
	C_2	271	2324	1133
	C_3	104	105	5350

Table 12 Confusion matrices obtained from the maximization of different *GPS*-based instances in the Vehicle dataset

a) GPS_{UPM}

		Observed			
		C_1	C_2	C_3	C_4
Predicted	C_1	1	0	0	29
	C_2	0	13	25	0
	C_3	42	0	4	6
	C_4	0	29	14	4

b) GPS_{NPV}

		Observed			
		C_1	C_2	C_3	C_4
Predicted	C_1	43	0	5	32
	C_2	0	42	38	7
	C_3	0	0	0	0
	C_4	0	0	0	0

c) $GPS_{Precision}$

		Observed			
		C_1	C_2	C_3	C_4
Predicted	C_1	1	0	0	27
	C_2	0	4	15	0
	C_3	42	0	4	6
	C_4	0	38	24	6

d) $GPS_{NPV,Precision_1}$

		Observed			
		C_1	C_2	C_3	C_4
Predicted	C_1	35	0	0	8
	C_2	8	42	43	31
	C_3	0	0	0	0
	C_4	0	0	0	0

e) GPS_{Recall}

		Observed			
		C_1	C_2	C_3	C_4
Predicted	C_1	4	1	3	36
	C_2	0	9	21	0
	C_3	39	0	4	2
	C_4	0	32	15	1

f) $GPS_{Recall,Precision_4}$

		Observed			
		C_1	C_2	C_3	C_4
Predicted	C_1	1	0	0	29
	C_2	0	9	24	0
	C_3	41	0	3	4
	C_4	1	33	16	6

has been calculated using Property 3 in Section 3. When only the *NPV* is relevant, the maximum of GPS_{NPV} is 0.80, corresponding to the confusion matrix b). Notice the significant differences between the confusion matrices, depending on the chosen performance metric. In this case, since the *Recalls* are not considered, they can have more extreme values (range equals 1.00), whereas less extreme values are allowed for the *Specificity* (range equals 0.36). When only the *Precisions* are relevant, the maximum of $GPS_{Precision}$ is 0.07 ± 0.09, corresponding to confusion matrix c). The confusion matrix d) is the result of maximising $GPS_{NPV,Precision_1}$. The solution is similar to the obtained when GPS_{NPV} is chosen as performance metric (confusion matrix b)). However, in d) a high value of $Precision_1$ is required (0.81 vs 0.54). The *ML* classifier chooses the thresholds to maximise GPS_{Recall} resulting in confusion matrix e), where the maximum value is 0.06 ± 0.11. The confusion matrix f) is the achieved solution when the *Precision* in class 4 is added to the above definition of GPS_{Recall}. As expected, the main differences between confusion matrices e) and f) are presented in class 4, increasing the corresponding *Precision* from 0.02 to 0.11, and the corresponding *Recall* from 0.02 to 0.15.

5 Conclusions

In this paper, the *GPS*, a novel family of performance metrics for binary and multi-class classification problems, has been presented. It is defined as the combination of a set of performance metrics using the harmonic mean. The harmonic mean is a natural choice to combine values representing ratios, such as those from the confusion matrix. Besides, it generates conservative combinations since it penalises low values. Thus, data analysts can develop different metrics tailored for the problem domain and the domain-expert goals based on *GPS*.

Several instances of *GPS* have been presented and compared with various state-of-the-art performance metrics in both binary and multi-class classification problems. It has been shown that it is possible to use different instances of *GPS* depending on the particular problem specifications. These definitions lead to different class predictions from the classifier. Therefore, to different confusion matrices. The *GPS* has proven to be more stable and explainable than the alternatives. Further, it has been shown that previous definitions of performance metrics such as F_1^+, F_1^- and *UPM* are instances of *GPS*.

Table 13 Performance metrics in the confusion matrices from the Vehicle dataset. In bold, the maximum in each metric

	a)	b)	c)	d)	e)	f)
$Precision_1$	0.03	0.54	0.03	0.81	0.09	0.03
$Precision_2$	0.34	0.48	0.21	0.34	0.30	0.27
$Precision_3$	0.08	NAN	0.08	0.00	0.09	0.06
$Precision_4$	0.08	NAN	0.09	NAN	0.02	0.11
$Recall_1$	0.02	1.00	0.02	0.81	0.09	0.02
$Recall_2$	0.31	1.00	0.09	1.00	0.21	0.21
$Recall_3$	0.09	0.00	0.09	0.00	0.09	0.07
$Recall_4$	0.10	0.00	0.15	0.00	0.02	0.15
$Specificity_1$	0.76	0.70	0.78	0.93	0.68	0.77
$Specificity_2$	0.80	0.64	0.88	0.34	0.83	0.81
$Specificity_3$	0.61	1.00	0.61	1.00	0.67	0.64
$Specificity_4$	0.66	1.00	0.51	1.00	0.63	0.61
NPV_1	0.69	1.00	0.70	0.93	0.68	0.69
NPV_2	0.77	1.00	0.74	1.00	0.76	0.75
NPV_3	0.66	0.74	0.66	0.74	0.68	0.66
NPV_4	0.71	0.76	0.67	0.77	0.67	0.70
GPS_{UPM}	**0.12**	NAN	0.11	NAN	0.11	0.11
GPS_{NPV}	0.70	**0.80**	0.67	0.67	0.69	0.70
$GPS_{Precision}$	0.07	NAN	**0.07**	NAN	0.05	0.07
$GPS_{NPV,Precision_1}$	0.14	0.77	0.15	**0.84**	0.30	0.14
GPS_{Recall}	0.06	NAN	0.06	NAN	**0.06**	0.06
$GPS_{Recall,Precision_4}$	0.06	NAN	0.06	NAN	0.04	**0.06**

Future work will focus on performing model selection using *GPS*. Given a set of *ML* classifiers, different performance metrics might lead to a different selection of best model. In this context, the effect of *GPS*-based metrics in the selection process could be evaluated. In addition, a sensitivity analysis to study the effect of different misclassification costs and different techniques to build binary matrices in multi-class problems will be carried out in the future. Further analysis will be carried out on the classification of datasets with a large number of categories. Notice that as the number of categories grows, the number of possible definitions of performance metrics that can be derived from the one proposed in this paper increases. Thus, a future research line would be to carry out a comparative study of the different solutions achieved through the chosen metrics within a specific problem. Furthermore, instances of *GPS* for multi-labelled, hierarchical, and non-square confusion matrices classification will be developed. The latter corresponds to binary classification problems where an output with more than two options is more informative. For instance, in a system that predicts if a patient will die in a given surgery, an output such as high-risk, medium-risk, and low-risk is more informative than a binary output. Finally, future work will focus on the use of the method when the data are in tensor form [14, 15].

Acknowledgements This research has been supported by grants from Madrid Autonomous Community (Ref: IND2018/TIC-9665) and the Spanish Science and Innovation, under the Retos-Colaboración program: SABERMED (Ref: RTC-2017-6253-1); and the Retos-Investigación program: MODAS-IN (reference: RTI-2018-094269-B-I00). Special thanks to MISC International S.L.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

1. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

2. Bland M (2008) Cohen's kappa. University of York Department of Health Sciences https://www.usersyorkacuk/~ mb55/msc/clinimet/week4/kappash2.pdf. Accessed 13 Feb 2014
3. Bradley AP (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7):1145–1159
4. Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics 21(1):6
5. Cohen P (1982) To be or not to be: Control and balancing of type i and type ii errors. Evaluation and Program Planning 5(3):247–253
6. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30
7. Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml
8. Fawcett T (2006) An introduction to roc analysis. Pattern Recognition Letters 27(8):861–874
9. Goodall DW (1967) The distribution of the matching coefficient. Biometrics, 647–656
10. Gorodkin J (2004) Comparing two k-category assignments by a k-category correlation coefficient. Computational biology and chemistry 28(5–6):367–374
11. Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. arXiv:200805756
12. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. Journal of Intelligent Information Systems 17(2–3):107–145
13. Halligan S, Altman DG, Mallett S (2015) Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. European Radiology 25(4):932–939
14. Hu C, Wang Y, Gu J (2020) Cross-domain intelligent fault classification of bearings based on tensor-aligned invariant subspace learning and two-dimensional convolutional neural networks. Knowledge-Based Systems 209:106214
15. Hu C, He S, Wang Y (2021) A classification method to detect faults in a rotating machinery based on kernelled support tensor machine and multilinear principal component analysis. Applied Intelligence 51(4):2609–2621
16. Matthews BW (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405(2):442–451
17. Norris N (1940) The standard errors of the geometric and harmonic means and their application to index numbers. The Annals of Mathematical Statistics 11(4):445–448
18. Ogbi MSZ (2012) A mathematical property of the harmonic mean. In: The 6th international days of statistics and economics. Prague University of Economics and Business, pp 873–877
19. Opitz J, Burst S (2019) Macro f1 and macro f1. arXiv:191103347
20. Powers DM (2020) Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv:201016061
21. Puthiya Parambath S, Usunier N, Grandvalet Y (2014) Optimizing f-measures by cost-sensitive classification. Advances in Neural Information Processing Systems 27:2123–2131
22. Redondo AR, Navarro J, Fernández RR, de Diego IM, Moguerza JM, Fernández-Muñoz JJ (2020) Unified performance measure for binary classification problems. In: International conference on intelligent data engineering and automated learning. Springer, pp 104–112
23. Sasaki Y, Fellow R (2007) The truth of the f-measure, manchester: Mib-school of computer science. University of Manchester p 25
24. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4):427–437
25. Tharwat A (2020) Classification assessment methods. New England Journal of Entrepreneurship 17(1):168–192

Isaac Martín De Diego Tenured Professor at the Universidad Rey Juan Carlos (URJC). Coordinator of the high-performance research group, DSLAB: Foundations and Applications of Data Science at URJC. His CV includes numerous publications related to research projects, technology transfer to the productive sector through patents and collaborative projects, extensive undergraduate and postgraduate teaching experience and experience in university management. His scientific production includes more than 50 articles indexed in JCR, more than 70 papers presented at national and international conferences, and two books for the dissemination of knowledge. Contributions in methods and techniques for data cleaning and cleaning, information representation, feature fusion from different information sources, new machine learning methods, novel metrics for evaluating artificial intelligence models, explainability techniques and visualisation of learning model results. Intense activity of technology transfer to the private sector: healthcare, video surveillance, computer vision, cybersecurity, livestock, telecommunications, and energy.

Ana R. Redondo Bachelor of Mathematical Engineering at University Complutense of Madrid (UCM). Master in Data Science at Rey Juan Carlos University (URJC). Master in Decision Systems Engineering at Rey Juan Carlos University (URJC). Researcher at Data Science Laboratory (DSLAB) of the URJC. His research interests are combination of information methods, Machine Learning algorithms and Explainable Machine Learning.

Rubén R. Fernández Bachelor of Computer Science at University of Leon (ULE). Master in Data Science at Rey Juan Carlos University (URJC). Master in Artificial Intelligence at Polytechnic University of Madrid (UPM). Researcher at Data Science Laboratory (DSLAB) in the URJC. Research interests: Explainable Machine Learning, active learning, and natural language processing. Application domains: tabular data, time series and text information retrieval.

Jorge Navarro received the B.Sc. and the M.Sc. degrees from Polytechnical University of Valencia and from King Juan Carlos University, respectively. He is currently a PhD candidate at King Juan Carlos University under a grant from Madrid Autonomous Community. He also works as a data scientist for Sensowave analyzing IoT data and his research interests include data mining and anomaly detection in time series and spatio-temporal data.

Javier M. Moguerza PhD in Mathematical Engineering at University Carlos III of Madrid (UC3M). Full Professor at Rey Juan Carlos University. Previously he has worked at Carlos III University of Madrid and at Pontificia Comillas University of Madrid (ICAI-ICADE). His research interests are focused on Operations Research (Six Sigma Quality, Optimization of Resources), the design of Machine Learning methods and Data Science. He has been responsible for the Ericsson Institutional Chair on Data Science applied to 5G at Rey Juan Carlos University from September 2016 to September 2019. He has been an academician of the Global Young Academy (GYA) since December 2010 until May 2016, and currently he belongs to the Alumni of the Global Young Academy. He is founder academician of the Young Academy of Spain, created by the Spanish Government in 2019.