Lima 05 de mayo de 2025

Señores
Miembros del Comité de Ética de la Universidad César Vallejo

Asunto
Justificación de la no necesidad de instrumento de medición en la investigación

Reciba un cordial saludo por parte de **Alcedo Javier Carlos José**, con código **7002694051** e identificado con DNI N° **71329761**; y **Pachas Luicho Freddy Amós** con código e identificado con DNI N° **73887870;** estudiantes del 9° ciclo de la carrera de Ingeniería de Sistemas. En la actualidad, estamos trabajando en el desarrollo del proyecto de investigación titulado: **Plataforma de Reclutamiento 4.0: Aplicación de Machine Learning para la Evaluación Autónoma de Perfiles Laborales, Lima, 2025.**

Mediante la presente, deseamos expresar que en el proyecto de investigación desarrollado no se requiere la validez de un instrumento tradicional de recolección de datos, pues las fichas de registro empleadas, están elaboradas en base a fórmulas previamente validadas por expertos del área, y por lo que en páginas siguientes se demuestra la existencia de estas en artículos y/o libros académicos.

Agradecemos su atención y quedamos a su disposición para cualquier aclaración que considere pertinente.

Alcedo Javier Carlos José
DNI 71329761

Pachas Luicho Freddy Amós
DNI 73887870

Instrumento de recolección de datos para la evaluación del modelo de machine learning

Ficha de registro
La presente ficha, tiene como objeto recolectar datos que permitan calcular cada uno de los indicadores del modelo elaborado, en las diferentes épocas de entrenamiento, a fin de garantizar la calidad del mismo.
Nombre de Proyecto
Plataforma de reclutamiento con Machine Learning para la Evaluación de Perfiles Laborales, Lima 2025
Investigadores
• Alcedo Javier Carlos José • Pachas Luicho Freddy Amos

Código de Ficha		Fecha aplicación	

Fórmulas aplicadas

- Precisión

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

- Recall

$$recall = \frac{tp}{tp \ + \ fn}$$

- Accuracy

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{n^+ + n^-} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Donde:
- TP: Total de positivos
- TN: Total de negativos
- FN: Falsos negativos
- FP: Falsos positivos

Tabla de registro

n° Epocas	Verdaderos Positivos	Falsos positivos	Total Positivos	Verdaderos negativos	Falsos negativo	Total negativos	Precisión	Recall	Accuracy

Comentarios

Evidencias de fórmulas

a) Precisión

The Stata Journal
Volume 20, Issue 1, March 2020, Pages 131-148

https://doi.org/10.1177/1536867X20909693

Article and Columns

When to consult precision-recall curves

Jonathan Cook[1] **and Vikram Ramadas**[2]

Abstract
Receiver operating characteristic (ROC) curves are commonly used to evaluate predictions of binary outcomes. When there is a small percentage of items of interest (as would be the case with fraud detection, for example), ROC curves can provide an inflated view of performance. This can cause challenges in determining which set of predictions is better. In this article, we discuss the conditions under which precision-recall curves may be preferable to ROC curves. As an illustrative example, we compare two commonly used fraud predictors (Beneish's [1999, *Financial Analysts Journal* 55: 24–36] *M* score and Dechow et al.'s [2011, *Contemporary Accounting Research* 28: 17–82] *F* score) using both ROC and precision-recall curves. To aid the reader with using precision-recall curves, we also introduce the command `prcurve` to plot them.

Keywords
st0591, prcurve, precision-recall curves, classifier evaluation, ROC curves

[1]Public Company Accounting Oversight Board Washington, DC, jacook@uci.edu
[2]Public Company Accounting Oversight Board Washington, DC, vnramadas@ucdavis.edu

1 Introduction

Recent developments in machine learning have increased interest in predictive modeling. An important component of building a predictive model is evaluating model efficacy. For evaluating predictions of binary outcomes, receiver operating characteristic (ROC) curves are the most common tool. In this article, we discuss when it may be advisable to consult an alternative tool—precision-recall (PR) curves—and introduce a command, `prcurve`, for doing so.

In some settings, we may be interested in predicting an outcome that is relatively rare (for example, fraud). In these settings with a rare outcome, ROC curves can be shifted outward relative to what would be found under a more balanced distribution. This outward shift can hinder comparisons of predictors. Our suggestion, and that of some recent literature (for example, Saito and Rehmsmeier [2015]), is to compare the PR plots for these predictors. There are, of course, other reasons for preferring PR curves to ROC, for example, having a loss function (or objective function) that better aligns with the output provided by the PR curve.

2 Review of PR and ROC curves

We assume that each observation belongs to one of two classes: positive or negative. In most economic applications, positive is coded as 1 and negative is coded as 0. To make predictions, we have a continuous "score". For example, the predictive probabilities from a logistic regression could be used as a score. We do not require that the score be a probability. Instead, we focus on how our score ranks the instances. Throughout this article, we refer to a set of ordinal scores as a "classifier".

Our task is to evaluate how well our classifier predicts class. Given a threshold, we could predict that all observations with a value that is above the threshold are positive and all observations below the threshold are negative. To see how well the rating works in combination with the threshold predict class, we define precision, recall, and the false-positive rate as

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{P}}$$

$$\text{False-positive rate} = \frac{\text{FP}}{\text{N}} \tag{2}$$

where the confusion matrix in table 1 defines true positives (TP), false positives (FP), negatives (N), and positives (P). In other words, precision measures how many of the items that were predicted to be positive cases are truly positive cases. Recall is the percent of positive cases that were identified. The false-positive rate is the percent of negative cases that were incorrectly predicted to be positive.

b) Recall

Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*

Marina Sokolova[1], Nathalie Japkowicz[2], and Stan Szpakowicz[3]

[1] DIRO, Université de Montréal, Montreal, Canada
sokolovm@iro.umontreal.ca
[2] SITE, University of Ottawa, Ottawa, Canada
nat@site.uottawa.ca
[3] SITE, University of Ottawa, Ottawa, Canada
ICS, Polish Academy of Sciences, Warsaw, Poland
szpak@site.uottawa.ca

Abstract. Different evaluation measures assess different characteristics of machine learning algorithms. The empirical evaluation of algorithms and classifiers is a matter of on-going debate among researchers. Most measures in use today focus on a classifier's ability to identify classes correctly. We note other useful properties, such as failure avoidance or class discrimination, and we suggest measures to evaluate such properties. These measures – Youden's index, likelihood, Discriminant power – are used in medical diagnosis. We show that they are interrelated, and we apply them to a case study from the field of electronic negotiations. We also list other learning problems which may benefit from the application of these measures.

1 Introduction

Supervised Machine Learning (ML) has several ways of evaluating the performance of learning algorithms and the classifiers they produce. Measures of the quality of classification are built from a confusion matrix which records correctly and incorrectly recognized examples for each class. Table 1 presents a confusion matrix for binary classification, where tp are true positive, fp – false positive, fn – false negative, and tn – true negative counts.

Table 1. A confusion matrix for binary classification

Class \ Recognized	as Positive	as Negative
Positive	tp	fn
Negative	fp	tn

2 Commonly-accepted Performance Evaluation Measures

The vast majority of ML research focus on the settings where the examples are assumed to be identically and independently distributed (IID). This is the case we focus on in this study. Classification performance without focussing on a class is the most general way of comparing algorithms. It does not favour any particular application. The introduction of a new learning problem inevitably concentrates on its domain but omits a detailed analysis. Thus, the most used empirical measure, *accuracy*, does not distinguish between the number of correct labels of different classes:

$$accuracy = \frac{tp \ + \ tn}{tp \ + \ fp \ + \ fn \ + tn} \tag{1}$$

Conversely, two measures that separately estimate a classifier's performance on different classes are sensitivity and specificity (often employed in biomedical and medical applications, and in studies which involve image and visual data):

$$sensitivity = \frac{tp}{tp \ + \ fn}; specificity = \frac{tn}{fp \ + tn} \tag{2}$$

Focus on one class prevails in text classification, information extraction, natural language processing and bioinformatics, where the number of examples belonging to one class is often substantially lower than the overall number of examples. The experimental setting is as follows: within a set of classes there is a class of special interest (usually *positive*). Other classes are either left as is – multi-class classification – or combined into one – binary classification. The measures of choice calculated on the positive class[4] are:

$$precision = \frac{tp}{tp \ + fp} \ recall = \frac{tp}{tp \ + \ fn} = sensitivity \tag{3}$$

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4}$$

c) Accuracy

RESEARCH ARTICLE — Open Access

The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation

Davide Chicco[1,2*] and Giuseppe Jurman[3]

Abstract

Background: To evaluate binary classifications and their confusion matrices, scientific researchers can employ several statistical rates, accordingly to the goal of the experiment they are investigating. Despite being a crucial issue in machine learning, no widespread consensus has been reached on a unified elective chosen measure yet. Accuracy and F_1 score computed on confusion matrices have been (and still are) among the most popular adopted metrics in binary classification tasks. However, these statistical measures can dangerously show overoptimistic inflated results, especially on imbalanced datasets.

Results: The Matthews correlation coefficient (MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

Conclusions: In this article, we show how MCC produces a more informative and truthful score in evaluating binary classifications than accuracy and F_1 score, by first explaining the mathematical properties, and then the asset of MCC in six synthetic use cases and in a real genomics scenario. We believe that the Matthews correlation coefficient should be preferred to accuracy and F_1 score in evaluating binary classification tasks by all scientific communities.

Keywords: Matthews correlation coefficient, Binary classification, F_1 score, Confusion matrices, Machine learning, Biostatistics, Accuracy, Dataset imbalance, Genomics

Background

Given a clinical feature dataset of patients with cancer traits [1, 2], which patients will develop the tumor, and which will not? Considering the gene expression of neuroblastoma patients [3], can we identify which patients are going to survive, and which will not? Evaluating the metagenomic profiles of patients [4], is it possible to discriminate different phenotypes of a complex disease? Answering these questions is the aim of machine learning and computational statistics, nowadays pervasive in analysis of biological and health care datasets, and many other scientific fields. In particular, these binary classification tasks can be efficiently addressed by supervised machine learning techniques, such as artificial neural networks [5], k-nearest neighbors [6], support vector machines [7], random forest [8], gradient boosting [9], or other methods. Here the word *binary* means that the data element statuses and prediction outcomes (class labels) can be twofold: in the example of patients, it can mean healthy/sick, or low/high grade tumor. Usually scientists indicate the two classes as the negative and the positive class. The term *classification* means that the goal of the process is to attribute the correct label to each data instance (sample); the process itself is known as the classifier, or classification algorithm.

*Correspondence: davidechicco@davidechicco.it
[1] Krembil Research Institute, Toronto, Ontario, Canada
[2] Peter Munk Cardiac Centre, Toronto, Ontario, Canada
Full list of author information is available at the end of the article

This partition can be presented in a 2×2 table called *confusion matrix* $\mathbf{M} = \begin{pmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{pmatrix}$ (expanded in Table 1), which completely describes the outcome of the classification task.

Clearly $\text{TP} + \text{FN} = n^+$ and $\text{TN} + \text{FP} = n^-$. When one performs a machine learning binary classification, she/he hopes to see a high number of true positives (TP) and true negatives (TN), and less false negatives (FN) and false positives (FP). When $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ 0 & n^- \end{pmatrix}$ the classification is perfect.

Since analyzing all the four categories of the confusion matrix separately would be time-consuming, statisticians introduced some useful statistical rates able to immediately describe the quality of a prediction [22], aimed at conveying into a single figure the structure of $\mathbf{M}$. A set of these functions act classwise (either actual or predicted), that is, they involve only the two entries of $\mathbf{M}$ belonging to the same row or column (Table 2). We cannot consider such measures fully informative because they use only two categories of the confusion matrix [39].

Accuracy. Moving to global metrics having three or more entries of $\mathbf{M}$ as input, many researchers consider computing the accuracy as the standard way to go. Accuracy, in fact, represents the ratio between the correctly predicted instances and all the instances in the dataset:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{n^+ + n^-} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Table 1 The standard confusion matrix M

	Predicted positive	Predicted negative
Actual positive	True positives TP	False negatives FN
Actual negative	False positives FP	True negatives TN

True positives (TP) and true negatives (TN) are the correct predictions, while false negatives (FN) and false positives (FP) are the incorrect predictions

$\begin{pmatrix} 0 & n^- \end{pmatrix}$.

perfect misclassification $\mathbf{M} = \begin{pmatrix} 0 & n^+ \\ n^- & 0 \end{pmatrix}$.

As anticipated (Background), accuracy fails in providing a fair estimate of the classifier performance in the class-unbalanced datasets. For any dataset, the proportion of samples belonging to the largest class is called the *no-information error rate* $\text{ni} = \frac{\max(n^+, n^-)}{n^+ + n^-}$; a binary dataset is (perfectly) balanced if the two classes have the same size, that is, $\text{ni} = \frac{1}{2}$, and it is unbalanced if one class is much larger than the other, that is $\text{ni} \gg \frac{1}{2}$. Suppose now that $\text{ni} \neq \frac{1}{2}$, and apply the trivial majority classifier: this algorithm learns only which is the largest class in the training set, and attributes this label to all instances. If the largest class is the positive class, the resulting confusion matrix is $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ n^- & 0 \end{pmatrix}$, and thus accuracy $=$ ni. If the dataset is highly unbalanced, $\text{ni} \approx 1$, and thus the accuracy measure gives an unreliable estimation of the goodness of the classifier. Note that, although we achieved this result by mean of the trivial classifier, this is quite a common effect: as stated by Blagus and Lusa [99], several classifiers are biased towards the largest class in unbalanced studies.

Finally, consider another trivial algorithm, the coin tossing classifier: this classifier randomly attributes to each sample, the label positive or negative with probability $\frac{1}{2}$. Applying the coin tossing classifier to any binary dataset gives an accuracy with expected value $\frac{1}{2}$, since $\langle \mathbf{M} \rangle = \begin{pmatrix} n^+/2 & n^+/2 \\ n^-/2 & n^-/2 \end{pmatrix}$.

Matthews correlation coefficient (MCC). As an alternative measure unaffected by the unbalanced datasets issue, the Matthews correlation coefficient is a contingency matrix method of calculating the *Pearson product-moment correlation coefficient* [22] between actual and predicted values. In terms of the entries of $\mathbf{M}$, MCC reads as follows:

Instrumento de recolección de datos para los indicadores de dimensión reducción de sesgos de evaluación

Ficha de registro
La presente ficha, tiene como objeto recolectar datos que permitan calcular la reducción de segos durante el proceso de evaluación de perfiles al aplicar el modelo elaborado.
Nombre de Proyecto
Plataforma de reclutamiento con Machine Learning para la Evaluación de Perfiles Laborales, Lima 2025
Investigadores
• Alcedo Javier Carlos José • Pachas Luicho Freddy Amos

Código de Ficha		Fecha de aplicación:	

Fórmula aplicada

- Tasa de reducción

$$RRT\ (\%) = \frac{TCT - TC\ t}{TCT}\ X\ 100$$

- Paridad demográfica

$$\forall a, a' \in \mathcal{A}:\ P(\widehat{Y} = 1 \mid A = a) = P(\widehat{Y} = 1 \mid A = a').$$

Tabla de registro

# época	P(Ŷ=1 | A=a)	P(Ŷ=1 | A=a′)	**Paridad Demográfica**	n°de cvs con sesgos aceptada (sin modelo)	n°de cvs con sesgos aceptada (con modelo)	**Tasa de reducción**

Comentarios

Evidencia de fórmulas

a) Tasa de reducción

Redalyc
Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal

Barajas Pérez, Juan Saúl; Montes-Belmont, Roberto; Castrejón Ayala, Federico; Flores -Moctezuma, Hilda Elizabeth; Serrato Cruz, Miguel Angel

Propiedades antifúngicas en especies del género Tagetes
Revista Mexicana de Micología, núm. 34, diciembre, 2011, pp. 83-88
Sociedad Mexicana de Micología
Xalapa, México

Disponible en: http://www.redalyc.org/src/inicio/ArtPdfRed.jsp?iCve=88321339008

Revista Mexicana de Micología
ISSN (Versión impresa): 0187-3180
gerardo.mata@inecol.edu.mx
Sociedad Mexicana de Micología
México

¿Cómo citar? | Número completo | Más información del artículo | Página de la revista

www.redalyc.org
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

sco de 5 mm de diámetro de *S. rolfsii* o *M. fruticola* en cada
:perimento. Se incubaron a 25-27 °C por un máximo de 30
as.

En ambas especies se evaluó el área de crecimiento
icelial hasta que el testigo llenó la caja Petri. Se tomaron
tografías digitales diariamente, y con ayuda del programa
nageJ®(versión 1.4), para análisis de imágenes (Instituto
acional de la Salud Mental de Estados Unidos), se calculó el
ea de crecimiento en milímetros cuadrados. Con el
ograma Sigma Stat V. 3.5 se realizaron regresiones lineales
ra calcular la tasa de crecimiento. Con los datos de la tasa de
ecimiento se calculó el porcentaje de reducción del
ecimiento con respecto al tratamiento testigo (RRT)
ediante la fórmula:

$$RRT\ (\%) = \frac{TCT - TCt}{TCT} \times 100$$

Donde: TCT= tasa de crecimiento en el testigo; TCt=
sa de crecimiento en el tratamiento.

Se cuantificó el total de esclerocios producidos por
rolfsii después de 30 días de incubación.

Para evaluar la esporulación de *M. fructicola*, una
:z que en el testigo se obtuvo el máximo crecimiento, se le
licionaron 2 mL de agua destilada estéril y se raspó el
icelio, se obtuvo una suspensión de la que se tomaron 10 µL
ue se depositaron en una cámara de Neubauer (Marienfeld
Co., Alemania) y con el objetivo 40x del microscopio óptico,
se realizó el conteo de esporas. Se efectuaron conteos en 4
campos microscópicos por unidad experimental en cada
tratamiento.

Con los datos obtenidos (promedio de dos
repeticiones) de todos los experimentos se realizaron análisis
de varianza con el programa SAS versión 8.0 y se aplicó el
procedimiento PROC GLM. La comparación de medias se
llevó a cabo por la diferencia mínima significativa DMS.

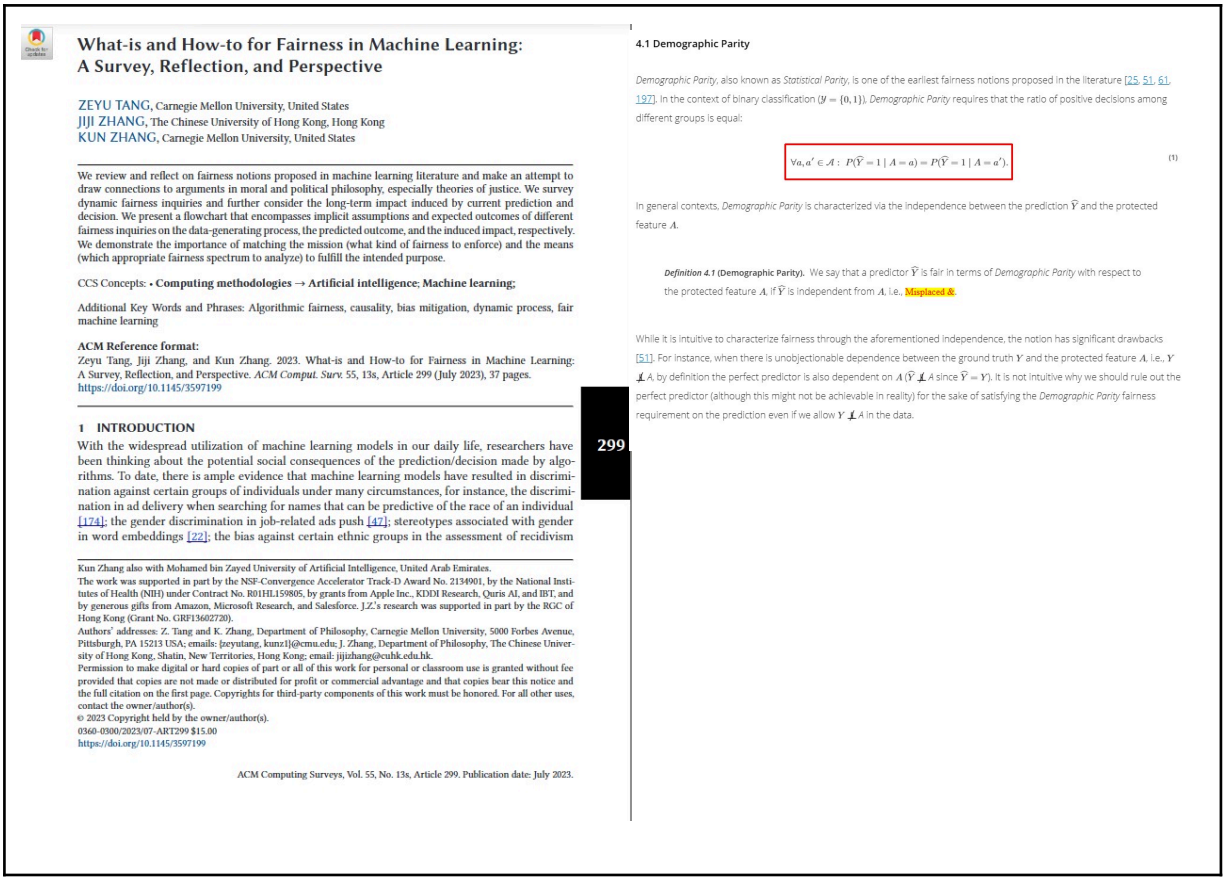
Para aceites esenciales con *S. rolfsii* se encontró que
el tratamiento con *T. filifolia* al 0.1 % fue el más notable
porque tuvo un efecto fungicida sobre todos los aislamientos
y en los demás tratamientos con aceites esenciales se encontró
que la tasa de crecimiento se redujo respecto al testigo, pero
varió entre aislamientos de *S. rolfsii* y entre tratamientos con
aceites esenciales de *Tagetes* (Tabla 1). El porcentaje de
reducción en la tasa de crecimiento con respecto al testigo
también varió, *T. coronopifolia* presentó una inhibición
mayor al 50 % en ocho aislamientos, en tanto que *T. lucida*
tuvo ese mismo comportamiento en siete aislamientos y se
redujo a cinco y a cuatro aislamientos en *T. erecta* y *T.
foetidissima*, respectivamente. El emulsificante Tween no
interfirió en el crecimiento micelial y estimuló ligeramente el
crecimiento en algunos aislamientos como el 1 y el 27. Los
aislamientos más sensibles fueron el 4, 5 y el 27 que tuvieron

Tabla 1. Efecto de los aceites esenciales de *Tagetes* al 0.1%, sobre la tasa de crecimiento micelial (mm^2/día) de diferentes aislamientos de *S. rolfsii* y porcentaje de reducción (RRT) del crecimiento de cada tratamiento respecto al testigo

A	T1	T2	RRT	T5	RRT	T4	RRT	T5	RRT	T6	RRT	T7	RRT	DMS
1	16.4 d	15.1 a	7.9	0**	100	11.1 b	32.3	6.2 c	62.2	9.2 bc	43.9	10.3 b	37.2	3.9
4	11.7 a	12.1 a	-3.4*	0	100	2.9 bc	75.2	3.3 b	84.7	1.9 c	83.8	3.7 bc	68.4	2.5
5	15.2 a	11.1 b	27.0	0	100	4.3 d	71.7	4.2 d	72.4	5.3 de	65.1	6.1 c	59.9	1.7
12	14.3 a	15.1 a	-5.6	0	100	13.5 a	5.6	8.5 b	40.6	13.2 a	7.7	9.7 b	32.2	3.2
16	12.5 a	8.5 b	32.0	0	100	1.6 d	87.2	0.1 e	99.2	0.1 e	99.2	3.0 c	76.0	1.4
17	14.0 ab	15.9 a	-13.6	0	100	13.7 ab	2.1	4.7 c	66.4	6.2 c	55.7	11.2 b	20.0	3.6
18	14.9 a	16.0 a	-7.4	0	100	9.5 b	36.2	7.4 b	50.3	6.4 c	57.0	5.7 c	61.7	2.5
20	15.2 a	14.0 ab	7.9	0	100	8.0 de	47.4	11.5 bc	24.3	4.9 c	67.8	10.0 ab	34.2	3.4
25	15.1 a	15.7 a	-4.0	0	100	8.7 b	42.4	0.1 d	99.3	2.8 c	81.5	10.3 b	31.8	1.7
27	15.4 a	11.9 b	22.7	0	100	3.2 c	79.2	3.5 c	77.3	0.8 c	94.8	3.5 c	77.3	2.7
$\bar{X}$	14.47	13.54	6.35	0	100	7.65	47.93	5.15	64.07	5.08	65.45	7.35	49.87	

Valores seguidos de la misma letra en cada línea no difieren estadísticamente P<0.0001. A: Aislamiento de *S. rolfsii*. T1: Testigo sin AE. T2: Testigo Tween 20 al 0.02 %. T3: AE de *T. filifolia*. T4: AE de *T. foetidissima*. T5: AE de *T. lucida*. T6: AE de *T. coronopifolia*. T7: AE de *T. erecta*. DMS: diferencia mínima significativa. *Resultado con signo negativo, significa efecto estimulatorio en la tasa de crecimiento. **Valores no incluidos en el análisis estadístico, por ser concluyentes.

b) Paridad Demográfica



What-is and How-to for Fairness in Machine Learning:
A Survey, Reflection, and Perspective

ZEYU TANG, Carnegie Mellon University, United States
JIJI ZHANG, The Chinese University of Hong Kong, Hong Kong
KUN ZHANG, Carnegie Mellon University, United States

We review and reflect on fairness notions proposed in machine learning literature and make an attempt to draw connections to arguments in moral and political philosophy, especially theories of justice. We survey dynamic fairness inquiries and further consider the long-term impact induced by current prediction and decision. We present a flowchart that encompasses implicit assumptions and expected outcomes of different fairness inquiries on the data-generating process, the predicted outcome, and the induced impact, respectively. We demonstrate the importance of matching the mission (what kind of fairness to enforce) and the means (which appropriate fairness spectrum to analyze) to fulfill the intended purpose.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**;

Additional Key Words and Phrases: Algorithmic fairness, causality, bias mitigation, dynamic process, fair machine learning

ACM Reference format:
Zeyu Tang, Jiji Zhang, and Kun Zhang. 2023. What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective. *ACM Comput. Surv.* 55, 13s, Article 299 (July 2023), 37 pages.
https://doi.org/10.1145/3597199

1 INTRODUCTION

With the widespread utilization of machine learning models in our daily life, researchers have been thinking about the potential social consequences of the prediction/decision made by algorithms. To date, there is ample evidence that machine learning models have resulted in discrimination against certain groups of individuals under many circumstances, for instance, the discrimination in ad delivery when searching for names that can be predictive of the race of an individual [174]; the gender discrimination in job-related ads push [47]; stereotypes associated with gender in word embeddings [22]; the bias against certain ethnic groups in the assessment of recidivism

Kun Zhang also with Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates.
The work was supported in part by the NSF-Convergence Accelerator Track-D Award No. 2134901, by the National Institutes of Health (NIH) under Contract No. R01HL159805, by grants from Apple Inc., KDDI Research, Quris AI, and IBT, and by generous gifts from Amazon, Microsoft Research, and Salesforce. J.Z.'s research was supported in part by the RGC of Hong Kong (Grant No. GRF13602720).
Authors' addresses: Z. Tang and K. Zhang, Department of Philosophy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA; emails: {zeyutang, kunz1}@cmu.edu; J. Zhang, Department of Philosophy, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; email: jijizhang@cuhk.edu.hk.

0360-0300/2023/07-ART299 $15.00
https://doi.org/10.1145/3597199

4.1 Demographic Parity

Demographic Parity, also known as *Statistical Parity*, is one of the earliest fairness notions proposed in the literature [25, 51, 61, 197]. In the context of binary classification ($\mathcal{Y} = \{0, 1\}$), *Demographic Parity* requires that the ratio of positive decisions among different groups is equal:

$$\forall a, a' \in \mathcal{A}: \ P(\widehat{Y} = 1 \mid A = a) = P(\widehat{Y} = 1 \mid A = a'). \tag{1}$$

In general contexts, *Demographic Parity* is characterized via the independence between the prediction $\widehat{Y}$ and the protected feature A.

Definition 4.1 **(Demographic Parity).** We say that a predictor $\widehat{Y}$ is fair in terms of *Demographic Parity* with respect to the protected feature A, if $\widehat{Y}$ is independent from A, i.e., Misplaced &.

While it is intuitive to characterize fairness through the aforementioned independence, the notion has significant drawbacks [51]. For instance, when there is unobjectionable dependence between the ground truth Y and the protected feature A, i.e., $Y \not\perp\!\!\!\perp A$, by definition the perfect predictor is also dependent on A ($\widehat{Y} \not\perp\!\!\!\perp A$ since $\widehat{Y} = Y$). It is not intuitive why we should rule out the perfect predictor (although this might not be achievable in reality) for the sake of satisfying the *Demographic Parity* fairness requirement on the prediction even if we allow $Y \not\perp\!\!\!\perp A$ in the data.

Instrumento de recolección de datos para los indicadores de dimensión reducción de tiempo de evaluación

Ficha de registro			
La presente ficha, tiene como objeto recolectar datos que permitan calcular la reducción de tiempo efectuada en el proceso de evaluación de perfiles al aplicar el modelo elaborado.			
Nombre de Proyecto			
Plataforma de reclutamiento con Machine Learning para la Evaluación de Perfiles Laborales, Lima 2025			
Investigadores			
• Alcedo Javier Carlos José • Pachas Luicho Freddy Amos			
Código de Ficha		Fecha de aplicación:	
Fórmula aplicada			
• Tiempo promedio de evaluación			

$$\mu = \frac{1}{N} \sum_{i=1}^{N} s_i$$

- Variación

$$\text{Variacion} = \frac{\text{VF - VI}}{\text{VI}} \text{ x } 100$$

Donde:
- VF: Tiempo con modelo
- VI: Tiempo tradicional

Tabla de registro

N° de CVs	Tiempo modelo	Tiempo tradicional	N° de CVs seleccionados	**Tiempo promedio de evaluación**	**% de Variación**

Comentarios

Evidencia de fórmulas

a) Tiempo promedio de evaluación

HAL open science

By Means of the Means: Arithmetic, Harmonic, Geometric ...

Luciano da Fontoura Costa

To cite this version:

Luciano da Fontoura Costa. By Means of the Means: Arithmetic, Harmonic, Geometric ... 2023. hal-04152919

HAL Id: hal-04152919
https://hal.science/hal-04152919v1

Preprint submitted on 5 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Given some numeric data (discrete or continuous), such as a set S of numbers, it is possible to assign a respective scalar value to that set, which will henceforth be considered to be real. The mapping that implements this association is often called a *functional*.

A generic functional F can be mathematically expressed as:

$$F: \ f(x) \longrightarrow F(f(x)) \in \mathbb{R} \tag{7}$$

As already observed, the means of a set S of values are functionals on S. However, not every functional is a mean, including the mode, median, extremes and the range of S, to name but a few cases among infinite possibilities.

When considered from the perspective of statistics and probability (e.g. [2, 13, 3, 4, 5, 6, 7, 8]), the concept of mean relates directly to the *expectance* of a random variable X, which constitutes a functional that can be expressed in discrete form as:

$$E[X] = \sum_{i=1}^{N} x_i p_i(x) \tag{8}$$

where x_i are N the possible values of the discrete random variable X and p_i corresponds to the respective relative frequencies.

where k is often taken as corresponding to a positive integer value.

Of particular interest are the *central moments*, which take into account a translation by the respective expected value $E[X]$, being defined as:

$$\mathcal{M}_k[X] = E\left[(X - E[X])^k\right] = \\ = \int_{-\infty}^{\infty} (x - \mu_x)^k \, p(x) \, dx \tag{13}$$

The important quantity know as *variance* of a random variable X is a particular case of a central moment, obtained for $k = 2$:

$$\sigma_X^2 = \mathcal{M}_2[X] = \int_{-\infty}^{\infty} (x - \mu_x)^2 \, p(x) \, dx \tag{14}$$

The *standard deviation* of X can then be defined as:

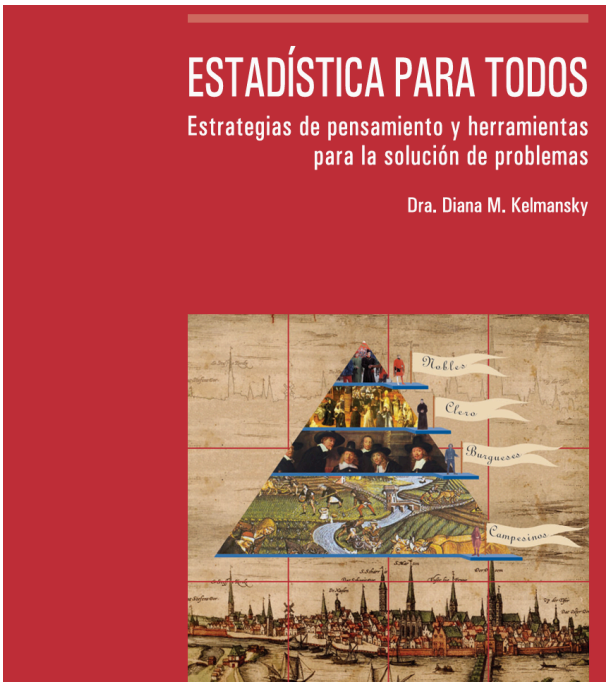
$$\sigma_X = +\sqrt{\sigma_X^2} \tag{15}$$

6 Arithmetic Mean

The arithmetic mean of a set of sampled values s_i, $i = 1, 2, \ldots, N$, can be defined as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{16}$$

b) Variación porcentual



Una tasa (o velocidad) es un **cociente que refleja una cierta cantidad por unidad.** Por ejemplo, un automóvil se desplaza a 45 km por hora (la unidad es una hora), o la tasa de robos en un barrio, 3 robos por cada 1.000 hogares (la unidad es 1.000 hogares).

Un porcentaje es un **número entre 0 y 100** que mide la **proporción** de un total. Por ejemplo, cuando decimos que una camisa tiene un 10% de descuento, si el precio original (el total) es $ 90, el descuento es de $ 9. Si decimos que el 35% de la población está a favor de un período de cuatro días de trabajo a la semana, y la población tiene 50.000 habitantes, entonces son 17.500 (50.000 x 0,35 = 17.500) los que están a favor. La proporción de los que están a favor es 0,35.

- Un porcentaje del 35% es lo mismo que una proporción de 0,35
- Para convertir **un porcentaje** en **una proporción**, se **divide** al porcentaje por 100.
- Para convertir una proporción en un porcentaje, se multiplica la proporción por 100.

3.3.2 Variaciones relativas

Cuando un porcentaje se utiliza para determinar un aumento o reducción relativa (relativa al valor inicial), se denomina **variación porcentual.**

Supongamos que la cantidad de accidentes por año en una ciudad pasó de 50 a 60, mientras que la cantidad de accidentes en otra ciudad pasó de 500 a 510. Ambas ciudades tuvieron un **aumento** de 10 accidentes por año, pero para la primera ciudad, esta diferencia como porcentaje del número inicial de accidentes, es mucho mayor.

Variación porcentual: se toma el valor "después de" y se le resta el "antes de", luego se divide ese resultado por el "antes de". Así, se obtiene una proporción. Para transformarla en un porcentaje se multiplica el resultado por 100.

Para la primera ciudad, esto significa que la cantidad de accidentes aumentó en un

Instrumento de recolección de datos para los indicadores de la dimensión eficiencia de evaluación

Ficha de registro			
La presente ficha, tiene como objeto recolectar datos que permitan calcular el nivel de eficiencia en la evaluación que conlleva la aplicación del modelo de machine learning.			
Nombre de Proyecto			
Plataforma de reclutamiento con Machine Learning para la Evaluación de Perfiles Laborales, Lima 2025			
Investigadores			
• Alcedo Javier Carlos José • Pachas Luicho Freddy Amos			
Código de Ficha		Fecha aplicación	

Fórmulas aplicadas

- Rendimiento de modelo (F1 Score)

$$F1 = \frac{2tp}{2tp + fp + fn}.$$

Donde:
- TP: Total de positivos
- FN: Falsos negativos
- FP: Falsos positivos

- Área bajo la curva

$$\mathrm{AUC} = \sum_i \left((1 - \beta_i \cdot \Delta\alpha) + \frac{1}{2}(\Delta(1-\beta) \cdot \Delta\alpha) \right)$$

Tabla de registro

n° Epocas	Falsos positivos	Total Positivos	Falsos negativo	F1 Score	Área Bajo la Curva

Comentarios

Evidencias de fórmulas

a) Rendimiento

Thresholding Classifiers to Maximize F1 Score

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy

University of California, San Diego,
La Jolla, California, 92093-0404, USA
{zlipton,celkan,muralib}@cs.ucsd.edu

Abstract. This paper provides new insight into maximizing F1 scores in the context of binary classification and also in the context of multilabel classification. The harmonic mean of precision and recall, F1 score is widely used to measure the success of a binary classifier when one class is rare. Micro average, macro average, and per instance average F1 scores are used in multilabel classification. For any classifier that produces a real-valued output, we derive the relationship between the best achievable F1 score and the decision-making threshold that achieves this optimum. As a special case, if the classifier outputs are well-calibrated conditional probabilities, then the optimal threshold is half the optimal F1 score. As another special case, if the classifier is completely uninformative, then the optimal behavior is to classify all examples as positive. Since the actual prevalence of positive examples typically is low, this behavior can be considered undesirable. As a case study, we discuss the results, which can be surprising, of applying this procedure when predicting 26,853 labels for Medline documents.

Keywords: machine learning, evaluation methodology, F1-score, multilabel classification, binary classification

1 Introduction

Performance metrics are useful for comparing the quality of predictions across systems. Some commonly used metrics for binary classification are accuracy, precision, recall, F1 score, and Jaccard index [15]. Multilabel classification is an extension of binary classification that is currently an area of active research in supervised machine learning [18]. Micro averaging, macro averaging, and per instance averaging are three commonly used variants of F1 score used in the multilabel setting. In general, macro averaging increases the impact on final score of performance on rare labels, while per instance averaging increases the importance of performing well on each example [17]. In this paper, we present theoretical and experimental results on the properties of the F1 metric.[1]

[1] For concreteness, the results of this paper are given specifically for the F1 metric and its multilabel variants. However, the results can be generalized to $F\beta$ metrics for $\beta \neq 1$.

Fig. 1: Confusion Matrix

of each label applying to each instance given the feature vector. For a batch of data of dimension $n \times d$, the model outputs an $n \times m$ matrix C of probabilities. In the single-label setting, $m = 1$ and C is an $n \times 1$ matrix, i.e. a column vector.

A decision rule $D(C) : \mathbb{R}^{n \times m} \to \{0,1\}^{n \times m}$ converts a matrix of probabilities C to binary predictions P. The gold standard $G \in \mathbb{R}^{n \times m}$ represents the true values of all labels for all instances in a given batch. A performance metric M assigns a score to a prediction given a gold standard:

$$M(P|G) : \{0,1\}^{n \times m} \times \{0,1\}^{n \times m} \to \mathbb{R} \in [0,1].$$

The counts of true positives tp, false positives fp, false negatives fn, and true negatives tn are represented via a confusion matrix (Figure 1).

Precision $p = tp/(tp + fp)$ is the fraction of all positive predictions that are true positives, while recall $r = tp/(tp + fn)$ is the fraction of all actual positives that are predicted positive. By definition the F1 score is the harmonic mean of precision and recall: $F1 = 2/(1/r + 1/p)$. By substitution, F1 can be expressed as a function of counts of true positives, false positives and false negatives:

$$F1 = \frac{2tp}{2tp + fp + fn}. \quad (1)$$

The harmonic mean expression for F1 is undefined when $tp = 0$, but the translated expression is defined. This difference does not impact the results below.

2.1 Basic Properties of F1

Before explaining optimal thresholding to maximize F1, we first discuss some properties of F1. For any fixed number of actual positives in the gold standard, only two of the four entries in the confusion matrix (Figure 1) vary independently. This is because the number of actual positives is equal to the sum $tp + fn$ while the number of actual negatives is equal to the sum $tn + fp$. A second basic property of F1 is that it is non-linear in its inputs. Specifically, fixing the number fp, F1 is concave as a function of tp (Figure 2). By contrast, accuracy is a linear function of tp and tn (Figure 3).

As mentioned in the introduction, F1 is asymmetric. By this, we mean that the score assigned to a prediction P given gold standard G can be arbitrarily different from the score assigned to a complementary prediction P^c given complementary gold standard G^c. This can be seen by comparing Figure 2 with Figure 5. This asymmetry is problematic when both false positives and false negatives are costly. For example, F1 has been used to evaluate the classification of tumors as benign or malignant [1], a domain where both false positives and false negatives have considerable costs.

b) Área bajo la curva

The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms

Andrew P. Bradley*

*The author is with the Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP) at the Dept. of Electrical and Computer Engineering, The University of Queensland, QLD 4072, Australia. E-mail bradley@elec.uq.edu.au.

method is to plot $P(T_p)$ against $P(F_p)$ as the *decision threshold* is varied. Selecting the operating point (decision threshold) that most closely meets the requirements for $P(F_n)$ and $P(F_p)$. The plotted values of $P(T_p)$ and $P(F_p)$ as the decision threshold is varied is called a Receiver Operating Characteristic (ROC) curve.

There is still, however, a problem with specifying performance in terms of a single operating point (usually a $P(T_p)$, $P(T_n)$ pair), in that there is no indication as to how these two measures vary as the decision threshold is varied. They may represent an operating point where sensitivity ($P(T_p)$) can be increased with little loss in specificity ($P(T_n)$), or they may not. This means that the comparison of two systems can become ambiguous. Therefore, there is a need for a *single* measure of classifier performance (often termed accuracy, but not to be confused with $P(C)$) that is invariant to the decision criterion selected, prior probabilities, and is easily extended to include cost/benefit analysis. This paper describes the results of an experimental study to investigate the use of the area under the ROC curve (AUC) as such as a measure of classifier performance.

When the decision threshold is varied and a number of points on the ROC curve ($P(F_p) = \alpha, P(T_p) = 1 \quad \beta$) have been obtained the simplest way to calculate the area under the ROC curve is to use trapezoidal integration,

$$\mathrm{AUC} = \sum_i \left((1 - \beta_i \cdot \Delta\alpha) + \frac{1}{2}(\Delta(1 - \beta) \cdot \Delta\alpha) \right). \tag{7}$$

Where,

$$\Delta(1 \quad \beta) = (1 \quad \beta_i) \quad (1 \quad \beta_i \ _1), \tag{8}$$

$$\Delta\alpha = \alpha_i - \alpha_{i-1}. \tag{9}$$

It is also possible to calculate the AUC by assuming that the underlying probabilities of