WIKIPEDIA
The Free Encyclopedia

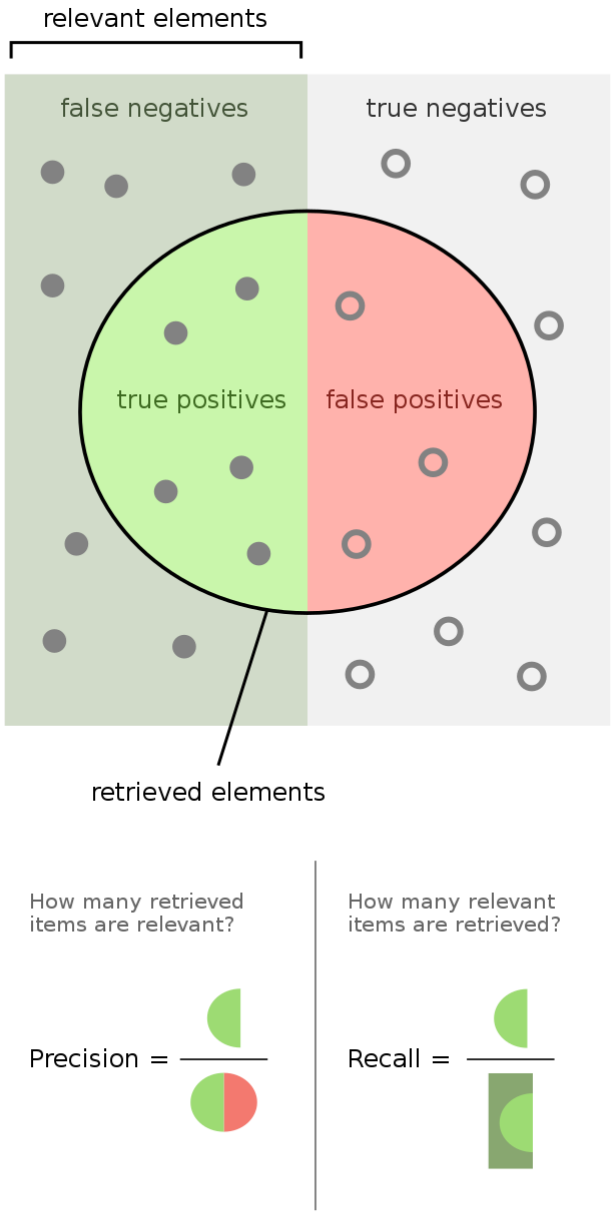
# Precision and recall

In pattern recognition, information retrieval, object detection and classification (machine learning), **precision** and **recall** are performance metrics that apply to data retrieved from a collection, corpus or sample space.

**Precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. Written as a formula: $\frac{relevant\_retrieved\_instances}{all\_\mathbf{retrieved}\_instances}$.

**Recall** (also known as sensitivity) is the fraction of relevant instances that were retrieved. Written as a formula: $\frac{relevant\_retrieved\_instances}{all\_\mathbf{relevant}\_instances}$. Both precision and recall are therefore based on relevance.

Consider a computer program for recognizing dogs (the **relevant** element) in a digital photograph. Upon processing a picture which contains ten cats and twelve dogs, the program identifies eight dogs. Of the eight elements identified as dogs, only five actually are dogs (true positives), while the other three are cats (false positives). Seven dogs were missed (false negatives), and seven cats were correctly excluded (true negatives). The program's precision is then 5/8 (true positives / selected elements) while its recall is 5/12 (true positives / relevant elements).




Precision and recall

Adopting a hypothesis-testing approach from statistics, in which, in this case, the null hypothesis is that a given item is *irrelevant* (i.e., not a dog), absence of type I and type II errors (i.e., perfect specificity and sensitivity of 100% each) corresponds respectively to perfect precision (no false positive) and perfect recall (no false negative).

More generally, recall is simply the complement of the type II error rate (i.e., one minus the type II error rate). Precision is related to the type I error rate, but in a slightly more complicated way, as it also depends upon the prior distribution of seeing a relevant vs. an irrelevant item.

The above cat and dog example contained 8 − 5 = 3 type I errors (false positives) out of 10 total cats (true negatives), for a type I error rate of 3/10, and 12 − 5 = 7 type II errors (false negatives), for a type II error rate of 7/12. Precision can be seen as a measure of quality, and recall as a measure of quantity. Higher precision means that an algorithm returns more relevant results than irrelevant ones, and high recall means that an algorithm returns most of the relevant results (whether or not irrelevant ones are also returned).

# Introduction

In a classification task, the precision for a class is the *number of true positives* (i.e. the number of items correctly labelled as belonging to the positive class) *divided by the total number of elements labelled as belonging to the positive class* (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class). Recall in this context is defined as the *number of true positives divided by the total number of elements that actually belong to the positive class* (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been).

Precision and recall are not particularly useful metrics when used in isolation. For instance, it is possible to have perfect recall by simply retrieving every single item. Likewise, it is possible to have near-perfect precision by selecting only a very small number of extremely likely items.

In a classification task, a precision score of 1.0 for a class C means that every item labelled as belonging to class C does indeed belong to class C (but says nothing about the number of items from class C that were not labelled correctly) whereas a recall of 1.0 means that every item from class C was labelled as belonging to class C (but says nothing about how many items from other classes were incorrectly also labelled as belonging to class C).

Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. Brain surgery provides an illustrative example of the tradeoff. Consider a brain surgeon removing a cancerous tumor from a patient's brain. The surgeon needs to remove all of the tumor cells since any remaining cancer cells will regenerate the tumor. Conversely, the surgeon must not remove healthy brain cells since that would leave the patient with impaired brain function. The surgeon may be more liberal in the area of the brain they remove to ensure they have extracted all the cancer cells. This decision increases recall but reduces precision. On the other hand, the surgeon may be more conservative in the brain cells they remove to ensure they extracts only cancer cells. This decision increases precision but reduces recall. That is to say, greater recall increases the chances of removing healthy cells (negative outcome) and increases the chances of removing all cancer cells (positive outcome). Greater precision decreases the chances of removing healthy cells (positive outcome) but also decreases the chances of removing all cancer cells (negative outcome).

Usually, precision and recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. *precision at a recall level of 0.75*) or both are combined into a single measure. Examples of measures that are a combination of precision and recall are the F-measure (the weighted harmonic mean of precision and recall), or the Matthews correlation coefficient, which is a geometric mean of the chance-corrected variants: the regression coefficients Informedness (DeltaP') and Markedness (DeltaP).[1][2] Accuracy is a weighted arithmetic mean of Precision and Inverse Precision (weighted by Bias) as well as a weighted arithmetic mean of Recall and Inverse Recall (weighted by Prevalence).[1] Inverse Precision and Inverse Recall are simply the Precision and Recall of the inverse problem where positive and negative labels are exchanged (for

both real classes and prediction labels). Recall and Inverse Recall, or equivalently true positive rate and false positive rate, are frequently plotted against each other as ROC curves and provide a principled mechanism to explore operating point tradeoffs. Outside of Information Retrieval, the application of Recall, Precision and F-measure are argued to be flawed as they ignore the true negative cell of the contingency table, and they are easily manipulated by biasing the predictions.[1] The first problem is 'solved' by using Accuracy and the second problem is 'solved' by discounting the chance component and renormalizing to Cohen's kappa, but this no longer affords the opportunity to explore tradeoffs graphically. However, Informedness and Markedness are Kappa-like renormalizations of Recall and Precision,[3] and their geometric mean Matthews correlation coefficient thus acts like a debiased F-measure.

# Definition

For classification tasks, the terms *true positives*, *true negatives*, *false positives*, and *false negatives* (see Type I and type II errors for definitions) compare the results of the classifier under test with trusted external judgments. The terms *positive* and *negative* refer to the classifier's prediction (sometimes known as the *expectation*), and the terms *true* and *false* refer to whether that prediction corresponds to the external judgment (sometimes known as the *observation*).

Let us define an experiment from $P$ positive instances and $N$ negative instances for some condition. The four outcomes can be formulated in a 2×2 contingency table or confusion matrix, as follows:

| | | Predicted condition | | Sources: [4][5][6][7][8][9][10][11][12] | |
|---|---|---|---|---|---|
| | Total population = P + N | **Positive (PP)** | **Negative (PN)** | Informedness, bookmaker informedness (BM) = TPR + TNR − 1 | Prevalence threshold (PT) = $\frac{\sqrt{\text{TPR}\times\text{FPR}}-\text{FPR}}{\text{TPR}-\text{FPR}}$ |
| **Actual condition** | **Positive (P)** | **True positive (TP), hit** | **False negative (FN), type II error, miss, underestimation** | True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{\text{TP}}{\text{P}}$ = 1 − FNR | False negative rate (FNR), miss rate = $\frac{\text{FN}}{\text{P}}$ = 1 − TPR |
| | **Negative (N)** | **False positive (FP), type I error, false alarm, overestimation** | **True negative (TN), correct rejection** | False positive rate (FPR), probability of false alarm, fall-out = $\frac{\text{FP}}{\text{N}}$ = 1 − TNR | True negative rate (TNR), specificity (SPC), selectivity = $\frac{\text{TN}}{\text{N}}$ = 1 − FPR |
| | Prevalence = $\frac{\text{P}}{\text{P + N}}$ | Positive predictive value (PPV), precision = $\frac{\text{TP}}{\text{PP}}$ = 1 − FDR | False omission rate (FOR) = $\frac{\text{FN}}{\text{PN}}$ = 1 − NPV | Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$ | Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$ |
| | Accuracy (ACC) = $\frac{\text{TP + TN}}{\text{P + N}}$ | False discovery rate (FDR) = $\frac{\text{FP}}{\text{PP}}$ = 1 − PPV | Negative predictive value (NPV) = $\frac{\text{TN}}{\text{PN}}$ = 1 − FOR | Markedness (MK), deltaP (Δp) = PPV + NPV − 1 | Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$ |
| | Balanced accuracy (BA) = $\frac{\text{TPR + TNR}}{2}$ | $F_1$ score = $\frac{2\text{PPV}\times\text{TPR}}{\text{PPV + TPR}} = \frac{2\text{TP}}{2\text{TP + FP + FN}}$ | Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV}\times\text{TPR}}$ | Matthews correlation coefficient (MCC) = $\sqrt{\text{TPR}\times\text{TNR}\times\text{PPV}\times\text{NPV}} - \sqrt{\text{FNR}\times\text{FPR}\times\text{FOR}\times\text{FDR}}$ | Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP + FN + FP}}$ |

Precision and recall are then defined as:[23]

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Recall in this context is also referred to as the true positive rate or sensitivity, and precision is also referred to as positive predictive value (PPV); other related measures used in classification include true negative rate and accuracy.[23] True negative rate is also called specificity.

$$\text{True negative rate} = \frac{tn}{tn + fp}$$

## When to use precision versus recall

They are both useful in cases where there is imbalanced data. Use precision when the cost of false positives is high: In situations where misclassifying an instance as positive has serious consequences, precision is a better choice. For example, in medical diagnosis, mistakenly diagnosing a healthy person with a disease can lead to unnecessary treatment and expenses. Use recall when the cost of false negatives is high: When the consequence of misclassifying an instance as negative is severe, recall is a better choice. For example, in fraud detection, failing to detect a fraudulent transaction can result in significant financial loss.

## Probabilistic Definition

Precision and recall can be interpreted as (estimated) conditional probabilities:[24] Precision is given by $\mathbb{P}(C = P|\hat{C} = P)$ while recall is given by $\mathbb{P}(\hat{C} = P|C = P)$,[25] where $\hat{C}$ is the predicted class and $C$ is the actual class (i.e. $C = P$ means the actual class is positive). Both quantities are, therefore, connected by Bayes' theorem.

### No-Skill Classifiers

The probabilistic interpretation allows to easily derive how a no-skill classifier would

Terminology and derivations from a confusion matrix

**condition positive (P)**
the number of real positive cases in the data
**condition negative (N)**
the number of real negative cases in the data

---

**true positive (TP)**
A test result that correctly indicates the presence of a condition or characteristic
**true negative (TN)**
A test result that correctly indicates the absence of a condition or characteristic
**false positive (FP), Type I error**
A test result which wrongly indicates that a particular condition or attribute is present
**false negative (FN), Type II error**
A test result which wrongly indicates that a particular condition or attribute is absent

---

**sensitivity, recall, hit rate, or true positive rate (TPR)**

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

**specificity, selectivity or true negative rate (TNR)**

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

**precision or positive predictive value (PPV)**

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

**negative predictive value (NPV)**

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$

**miss rate or false negative rate (FNR)**

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

**fall-out or false positive rate (FPR)**

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

**false discovery rate (FDR)**

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

**false omission rate (FOR)**

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

**Positive likelihood ratio (LR+)**

$$\mathrm{LR+} = \frac{\mathrm{TPR}}{\mathrm{FPR}}$$

**Negative likelihood ratio (LR-)**

perform. A no-skill classifiers is defined by the property that the conditional probability

$$\mathrm{LR-} = \frac{\mathrm{FNR}}{\mathrm{TNR}}$$

**prevalence threshold (PT)**

$$\mathrm{PT} = \frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}} + \sqrt{\mathrm{FPR}}}$$

**threat score (TS) or critical success index (CSI)**

$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

---

**Prevalence**

$$\frac{\mathrm{P}}{\mathrm{P} + \mathrm{N}}$$

**accuracy (ACC)**

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

**balanced accuracy (BA)**

$$\mathrm{BA} = \frac{TPR + TNR}{2}$$

**F1 score**

is the harmonic mean of precision and sensitivity:

$$\mathrm{F_1} = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

**phi coefficient (φ or $r_\varphi$) or Matthews correlation coefficient (MCC)**

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

**Fowlkes–Mallows index (FM)**

$$\mathrm{FM} = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} = \sqrt{PPV \times TPR}$$

**informedness or bookmaker informedness (BM)**

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$

**markedness (MK) or deltaP (Δp)**

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

**Diagnostic odds ratio (DOR)**

$$\mathrm{DOR} = \frac{\mathrm{LR+}}{\mathrm{LR-}}$$

*Sources:* Fawcett (2006),[13] Piryonesi and El-Diraby (2020),[14] Powers (2011),[15] Ting (2011),[16] CAWCR,[17] D. Chicco & G. Jurman (2020, 2021, 2023),[18][19][20] Tharwat (2018).[21] Balayla (2020)[22]

$\mathbb{P}(C = P, \hat{C} = P) = \mathbb{P}(C = P)\mathbb{P}(\hat{C} = P)$ is just the product of the unconditional probabilites since the classification and the presence of the class are independent.

For example the precision of a no-skill classifier is simply a constant $\mathbb{P}(C = P|\hat{C} = P) = \frac{\mathbb{P}(C = P, \hat{C} = P)}{\mathbb{P}(\hat{C} = P)} = \mathbb{P}(C = P)$, i.e. determined by the probability/frequency with which the class P occurs.

A similar argument can be made for the recall: $\mathbb{P}(\hat{C} = P|C = P) = \frac{\mathbb{P}(C = P, \hat{C} = P)}{\mathbb{P}(C = P)} = \mathbb{P}(\hat{C} = P)$ which is just (the typically threshold dependent) probability for a positive classification.

Some very specific no-skill classifiers are implemented in sklearn and are named dummy classifiers there.[26]

# Imbalanced data

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy can be a misleading metric for imbalanced data sets. Consider a sample with 95 negative and 5 positive values. Classifying all values as negative in this case gives 0.95 accuracy score. There are many metrics that don't suffer from this problem. For example, balanced accuracy[27] (bACC) normalizes true positive and true negative predictions by the number of positive and negative samples, respectively, and divides their sum by two:

$$\text{Balanced accuracy} = \frac{TPR + TNR}{2}$$

For the previous example (95 negative and 5 positive samples), classifying all as negative gives 0.5 balanced accuracy score (the maximum bACC score is one), which is equivalent to the expected value of a random guess in a balanced data set. Balanced accuracy can serve as an overall performance metric for a model, whether or not the true labels are imbalanced in the data, assuming the cost of FN is the same as FP.

Another metric is the predicted positive condition rate (PPCR), which identifies the percentage of the total population that is flagged. For example, for a search engine that returns 30 results (retrieved documents) out of 1,000,000 documents, the PPCR is 0.003%.

$$\text{Predicted positive condition rate} = \frac{TP + FP}{TP + FP + TN + FN}$$

According to Saito and Rehmsmeier, precision-recall plots are more informative than ROC plots when evaluating binary classifiers on imbalanced data. In such scenarios, ROC plots may be visually deceptive with respect to conclusions about the reliability of classification performance.[28]

Different from the above approaches, if an imbalance scaling is applied directly by weighting the confusion matrix elements, the standard metrics definitions still apply even in the case of imbalanced datasets.[29] The weighting procedure relates the confusion matrix elements to the support set of each considered class.

# F-measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

This measure is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean. There are several reasons that the F-score can be criticized, in particular circumstances, due to its bias as an evaluation metric.[1] This is also known as the $F_1$ measure, because recall and precision are evenly weighted.

It is a special case of the general $F_\beta$ measure (for non-negative real values of $\beta$):

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\beta^2 \cdot \mathrm{precision} + \mathrm{recall}}$$

Two other commonly used $F$ measures are the $F_2$ measure, which weights recall higher than precision, and the $F_{0.5}$ measure, which puts more emphasis on precision than recall.

The F-measure was derived by van Rijsbergen (1979) so that $F_\beta$ "measures the effectiveness of retrieval with respect to a user who attaches $\beta$ times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure $E_\alpha = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$, the second term being the weighted harmonic mean of precision and recall with weights $(\alpha, 1-\alpha)$. Their relationship is $F_\beta = 1 - E_\alpha$ where $\alpha = \frac{1}{1+\beta^2}$.

# Limitations as goals

There are other parameters and strategies for performance metric of information retrieval system, such as the area under the ROC curve (AUC)[30] or pseudo-R-squared.

## See also

- Uncertainty coefficient, also called *proficiency*
- Sensitivity and specificity
- Confusion matrix
- Scoring rule
- Base rate fallacy

## References


1. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://web.archive.org/web/20191114213255/https://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf) (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63. Archived from the original (http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf) (PDF) on 2019-11-14.
2. Perruchet, P.; Peereman, R. (2004). "The exploitation of distributional information in syllable processing". *J. Neurolinguistics*. **17** (2–3): 97–119. doi:10.1016/s0911-6044(03)00059-9 (https://doi.org/10.1016%2Fs0911-6044%2803%2900059-9). S2CID 17104364 (https://api.semanticscholar.org/CorpusID:17104364).
3. Powers, David M. W. (2012). "The Problem with Kappa" (https://www.aclweb.org/anthology/E12-1035). *Conference of the European Chapter of the Association for Computational Linguistics (EACL2012) Joint ROBUS-UNSUP Workshop*.
4. Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves" (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240215). *PLoS One*. **15** (10). doi:10.1371/journal.pone.0240215 (https://doi.org/10.1371%2Fjournal.pone.0240215).
5. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).
6. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512).
7. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
8. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
9. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.

10. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).
11. Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).
12. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
13. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).
14. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512).
15. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
16. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
17. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
18. Chicco D.; Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).
19. Chicco D.; Toetsch N.; Jurman G. (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).
20. Chicco D.; Jurman G. (2023). "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification" (https://doi.org/10.1186/s13040-023-00322-4). *BioData Mining*. **16** (1). doi:10.1186/s13040-023-00322-4 (https://doi.org/10.1186%2Fs13040-023-00322-4). PMC 9938573 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9938573).

21. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
22. Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves" (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240215). *PLoS One*. **15** (10). doi:10.1371/journal.pone.0240215 (https://doi.org/10.1371%2Fjournal.pone.0240215).
23. Olson, David L.; and Delen, Dursun (2008); *Advanced Data Mining Techniques*, Springer, 1st edition (February 1, 2008), page 138, ISBN 3-540-76916-1
24. Fatih Cakir, Kun He, Xide Xia, Brian Kulis, Stan Sclaroff, *Deep Metric Learning to Rank* (http://cs-people.bu.edu/fcakir/papers/fastap_cvpr2019.pdf), In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
25. Information Retrieval Models, Thomas Roelleke, ISBN 9783031023286, page 76, https://books.google.com/books?id=YX9yEAAAQBAJ&pg=PA76
26. "Sklearn.dummy.DummyClassifier" (https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html).
27. Mower, Jeffrey P. (2005-04-12). "PREP-Mt: predictive RNA editor for plant mitochondrial genes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1087475). *BMC Bioinformatics*. **6**: 96. doi:10.1186/1471-2105-6-96 (https://doi.org/10.1186%2F1471-2105-6-96). ISSN 1471-2105 (https://www.worldcat.org/issn/1471-2105). PMC 1087475 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1087475). PMID 15826309 (https://pubmed.ncbi.nlm.nih.gov/15826309).
28. Saito, Takaya; Rehmsmeier, Marc (2015-03-04). Brock, Guy (ed.). "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800). *PLOS ONE*. **10** (3): e0118432. Bibcode:2015PLoSO..1018432S (https://ui.adsabs.harvard.edu/abs/2015PLoSO..1018432S). doi:10.1371/journal.pone.0118432 (https://doi.org/10.1371%2Fjournal.pone.0118432). ISSN 1932-6203 (https://www.worldcat.org/issn/1932-6203). PMC 4349800 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800). PMID 25738806 (https://pubmed.ncbi.nlm.nih.gov/25738806).
    - Suzanne Ekelund (March 2017). "Precision-recall curves – what are they and how are they used?" (https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used). *Acute Care Testing*.
29. Tripicchio, Paolo; Camacho-Gonzalez, Gerardo; D'Avella, Salvatore (2020). "Welding defect detection: coping with artifacts in the production line" (https://link.springer.com/article/10.1007/s00170-020-06146-4). *The International Journal of Advanced Manufacturing Technology*. **111** (5): 1659–1669. doi:10.1007/s00170-020-06146-4 (https://doi.org/10.1007%2Fs00170-020-06146-4). S2CID 225136860 (https://api.semanticscholar.org/CorpusID:225136860).
30. Zygmunt Zając. What you wanted to know about AUC. http://fastml.com/what-you-wanted-to-know-about-auc/

- Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. New York, NY: ACM Press, Addison-Wesley, Seiten 75 ff. ISBN 0-201-39829-X
- Hjørland, Birger (2010); *The foundation of the concept of relevance*, Journal of the American Society for Information Science and Technology, 61(2), 217-237
- Makhoul, John; Kubala, Francis; Schwartz, Richard; and Weischedel, Ralph (1999); *Performance measures for information extraction* (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.4637), in *Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999*

- van Rijsbergen, Cornelis Joost "Keith" (1979); *Information Retrieval*, London, GB; Boston, MA: Butterworth, 2nd Edition, ISBN 0-408-70929-4


# External links

- Information Retrieval – C. J. van Rijsbergen 1979 (http://www.dcs.gla.ac.uk/Keith/Preface.html)
- Computing Precision and Recall for a Multi-class Classification Problem (http://www.text-analytics101.com/2014/10/computing-precision-and-recall-for.html)