

Appendix 1: Precision and recall

Performance metrics for evaluating artificial intelligence (AI) tools and systems

First published 17 June 2025

2 mins read

For artificial intelligence (AI) tools involving prediction, pattern recognition, information retrieval and classification, precision and recall will usually be the performance metric to use. Recall tells you the model's ability to locate all relevant instances in a data set. Precision tells you how successful it is at identifying only the relevant data points.

For example, you are testing a tool that seeks to predict which postcodes will be targeted for domestic burglaries. When testing the tool on historic data that it has not seen before, you make a note of the number of the following.

True positives (TP)

Where the model correctly predicts a positive outcome (the actual outcome was positive). In the context of our example, it predicts that there will be burglaries in certain postcodes, and there were.

True negatives (TN)

The model correctly predicted a negative outcome (the actual outcome was negative). In our example, it predicts that there would not be burglaries in certain postcodes, and there were not.

False positives (FP)

The model incorrectly predicted a positive outcome (the actual outcome was negative). Also known as a Type I error. It predicts that a postcode will be targeted by burglars, but it was not.

False negatives (FN)

The model incorrectly predicted a negative outcome (the actual outcome was positive). Also known as a Type II error. It predicted that a postcode would not be targeted, and it was.

You put the information into a table (confusion matrix).

	Positive – burglaries in postcode	Negative – burglaries in postcode
Positive – predicts burglaries in postcode	True positives	False positives
Negative – predicts burglaries in postcode	False negatives	True negatives

Work out recall

How many burglaries did your model predict?

$$\text{Recall} = \text{TP} \div (\text{TP} + \text{FN})$$

Work out precision

Out of the number of burglary-affected postcodes predicted, how many were in fact affected by burglaries?

$$\text{Precision} = \text{TP} \div (\text{TP} + \text{FP})$$

Overall performance

Look at the relationship between precision and recall to find a measure of overall performance:

$$2 \times (\text{precision} \times \text{recall}) \div (\text{precision} + \text{recall})$$

A score of 1 is perfect (very unlikely), while 0 is imperfect.

Tags

Artificial intelligence