

Alert thresholds, precision and recall

Why a single accuracy percentage can hide failure on rare events.

In short

- If only ten of a thousand observations are faults, always predicting normal reaches 99% accuracy while missing every fault.
- Precision measures how many alerts are correct, while recall measures how many actual events were detected.
- Thirty consecutive alerts around one failure are not thirty independent successes, so warnings should be grouped into event windows.

01 / Expose the denominator

If only ten of a thousand observations are faults, always predicting normal produces 99% accuracy while missing every fault. This is an illustrative calculation, not a field result. Include true positives, false positives, false negatives and true negatives in the acceptance table.

02 / Separate two questions

Precision asks how many alerts are correct. Recall asks how many actual positives were detected. A zero denominator makes a metric undefined; display that state instead of an arbitrary percentage. F1 combines precision and recall but does not automatically encode operational cost.

03 / Select thresholds on validation data

A lower threshold typically produces more warnings. Evaluate the benefit and cost together. For a pilot, write down maintenance capacity and the impact of a missed event. Freeze the threshold before the final test. An uncalibrated model score must not be described as a failure probability.

04 / Report event-level performance

Thirty consecutive alerts around one failure are not thirty independent successes. Group warnings into event windows and document repeat-notification rules. False alarms per shift and first-warning lead time are useful operational measures. Retain point-level metrics too, since they answer a different question.

Try it

Move the threshold slider in Kutup Lab and compare the number of warnings with missed labeled samples.

References and online guide

[Open the online guide](#)

scikit-learn: classification metrics

Proposed pilot steps do not claim field accuracy. Synthetic examples are not measurements from a production site.