

Classification: True vs. False and Positive vs. Negative

developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

bookmark_border Stay organized with collections Save and categorize content based on your preferences.

Estimated Time: 5 minutes

In this section, we'll define the primary building blocks of the metrics we'll use to evaluate classification models. But first, a fable:

An Aesop's Fable: The Boy Who Cried Wolf (*compressed*)

A shepherd boy gets bored tending the town's flock. To have some fun, he cries out, "Wolf!" even though no wolf is in sight. The villagers run to protect the flock, but then get really mad when they realize the boy was playing a joke on them.

[Iterate previous paragraph N times.]

One night, the shepherd boy sees a real wolf approaching the flock and calls out, "Wolf!" The villagers refuse to be fooled again and stay in their houses. The hungry wolf turns the flock into lamb chops. The town goes hungry. Panic ensues.

Let's make the following definitions:

- "Wolf" is a **positive class**.
- "No wolf" is a **negative class**.

We can summarize our "wolf-prediction" model using a 2x2 confusion matrix that depicts all four possible outcomes:

True Positive (TP): <ul style="list-style-type: none">• Reality: A wolf threatened.• Shepherd said: "Wolf."• Outcome: Shepherd is a hero.	False Positive (FP): <ul style="list-style-type: none">• Reality: No wolf threatened.• Shepherd said: "Wolf."• Outcome: Villagers are angry at shepherd for waking them up.
False Negative (FN): <ul style="list-style-type: none">• Reality: A wolf threatened.• Shepherd said: "No wolf."• Outcome: The wolf ate all the sheep.	True Negative (TN): <ul style="list-style-type: none">• Reality: No wolf threatened.• Shepherd said: "No wolf."• Outcome: Everyone is fine.

A **true positive** is an outcome where the model *correctly* predicts the *positive* class. Similarly, a **true negative** is an outcome where the model *correctly* predicts the *negative* class.

A **false positive** is an outcome where the model *incorrectly* predicts the *positive* class. And a **false negative** is an outcome where the model *incorrectly* predicts the *negative* class.