# Glossary for Data Analytics

## confusion matrix

A confusion matrix is a table used to evaluate the performance of a classification model, especially in supervised machine learning. It provides a detailed breakdown of the actual versus predicted classifications, helping to identify the types of errors the model is making. Here's how it's structured:

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Actual Positive** | True Positive (TP) | False Negative (FN) |
| **Actual Negative** | False Positive (FP) | True Negative (TN) |

#### Components of the Confusion Matrix

1. **True Positives (TP)**: The number of instances where the model correctly predicted the positive class.
2. **True Negatives (TN)**: The number of instances where the model correctly predicted the negative class.
3. **False Positives (FP)**: The number of instances where the model incorrectly predicted the positive class (also called Type I error).
4. **False Negatives (FN)**: The number of instances where the model incorrectly predicted the negative class (also called Type II error).

#### Metrics Derived from the Confusion Matrix

Using the values in the confusion matrix, several important metrics can be calculated:

1. **Accuracy**: [(TP + TN) / (TP + TN + FP + FN)]
2. **Precision**: [TP / (TP + FP)]
3. **Recall (Sensitivity)**: [TP / (TP + FN)]
4. **Specificity**: [TN / (TN + FP)]
5. **F1 Score**: [2 * (Precision * Recall) / (Precision + Recall)]

These metrics help assess different aspects of a model's performance, such as its ability to detect true positives (sensitivity), avoid false positives (specificity), and balance precision and recall (F1 score).

## nominal column

A nominal column in a dataset contains categorical data without any intrinsic ordering or ranking. Examples of nominal columns include:

- **Customer ID**: Unique identifier for each customer.
- **Product Category**: Labels for different product types, such as "Electronics," "Clothing," or "Groceries."

- **Region**: Geographic areas like "North," "South," "East," and "West."
- **Payment Method**: Types of payment used, like "Credit Card," "Cash," or "Online Payment".

These columns are useful for grouping and segmenting data but do not imply any order or scale among the categories.

When we say a nominal column has "no intrinsic ordering or ranking," it means the categories in that column do not follow any natural sequence or hierarchy. The values are simply labels or names and are not meant to be compared in terms of order or value.

For example, if a column called **"Fruit Type"** contains categories like "Apple," "Banana," and "Cherry," there's no inherent reason to rank them as better or worse, or in any particular sequence. These values are just distinct types of fruit, with no relationship that makes one fruit "higher" or "lower" than another.

In contrast, columns with an order, like "Low," "Medium," "High," or numbers that represent quantities, have an inherent progression or ranking.

---

## F1 score

The F1 score is a metric used to evaluate the performance of a classification model, particularly when there is an uneven class distribution or when both precision and recall are important. It is the harmonic mean of precision and recall, giving a single score that balances the two metrics.

### Formula for F1 Score

The F1 score is calculated as:

[ \text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} ]

where:

- **Precision** is the ratio of true positives (TP) to the sum of true positives and false positives (FP). [ \text{Precision} = \frac{TP}{TP + FP} ]
- **Recall** is the ratio of true positives (TP) to the sum of true positives and false negatives (FN). [ \text{Recall} = \frac{TP}{TP + FN} ]

### Why the F1 Score Is Useful

- **Balances Precision and Recall**: The F1 score is useful when you want to find a balance between precision and recall. It is particularly helpful when high precision and high recall are both essential, or when one metric alone (e.g., accuracy) might give a misleading view of performance.

## harmonic mean

The harmonic mean is used instead of the arithmetic mean because it punishes extreme values. For instance, if precision is high but recall is low, the F1 score will reflect this imbalance, giving a lower score.

### When to Use the F1 Score

The F1 score is ideal in situations where:

- The class distribution is imbalanced (e.g., in fraud detection where fraudulent transactions are much less common than legitimate ones).
- Both false positives and false negatives carry a significant cost, and neither can be ignored.

For example, in a medical diagnosis model, you might use the F1 score to balance between precision (how many diagnosed cases are actually positive) and recall (how many true cases the model catches).

The harmonic mean is a type of average used to find the central tendency of a set of numbers. Unlike the arithmetic mean (regular average), which sums values and divides by the count, the harmonic mean emphasizes smaller values in the dataset. It is particularly useful when dealing with rates or ratios and is often applied in fields such as machine learning, physics, and finance.

**Formula for the Harmonic Mean**

For ( n ) values ( x_1, x_2, x_3, \ldots, x_n ), the harmonic mean (HM) is calculated as:

[ \text{HM} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}} ]

For two values, ( a ) and ( b ), the harmonic mean simplifies to:

[ \text{HM} = \frac{2 \cdot a \cdot b}{a + b} ]

**Why Use the Harmonic Mean?**

The harmonic mean is especially useful in situations where:

- You need to average rates, such as speeds, where time or distance is constant.
- You want to penalize extreme values, as it gives more weight to smaller values in the dataset.

For example, in the F1 score (used in evaluating model performance in classification), the harmonic mean of precision and recall gives a balanced measure that considers both values equally and punishes cases where one value is much lower than the other.

**Example of Using the Harmonic Mean**

Suppose you drive a distance at two different speeds:

- 60 km/h for half the distance
- 120 km/h for the other half

The harmonic mean speed, rather than the arithmetic mean, better reflects the overall speed because it considers the time taken at each rate. Calculating it:

[ \text{HM} = \frac{2 \times 60 \times 120}{60 + 120} = \frac{14400}{180} = 80 \text{ km/h} ]

Thus, the harmonic mean speed here is 80 km/h, which accurately captures the effective speed over the whole journey.