

# Evaluation metrics

Evaluation metrics are quantitative measures used to assess the performance and effectiveness of model.how well the model is performing and help in comparing different models or algorithms.

## Classification Metrics

### Confusion Matrix

#### Confusion Matrix

This is a table that visually shows how many predictions were true positives, false positives, true negatives, and false negatives.

True Positive: You predicted positive, and it's true.

True Negative: You predicted negative, and it's true.

False Positive: (Type 1 Error): You predicted positive, and it's false.

False Negative: (Type 2 Error): You predicted negative, and it's false.

#### Specificity

(True Negative Rate): The ratio of correctly predicted negative observations to all the actual negatives.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	$a/(a+b)$
	Negative	c	d	Negative Predictive Value	$d/(c+d)$
		Sensitivity	Specificity	Accuracy = $(a+d)/(a+b+c+d)$	
		$a/(a+c)$	$d/(b+d)$		

## Accuracy:

The ratio of correctly predicted instances (both true positives and true negatives) to the total instances.

Accuracy measures the proportion of correctly classified instances among the total instances. It is a useful metric when the class distribution is balanced.

Example: If a model correctly classifies 90 out of 100 instances, its accuracy is 90%.

## Precision:

The ratio of correctly predicted positive observations to the total predicted positives.

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is useful when the cost of false positives is high.

example: In spam detection, precision is important because we want to minimize the number of non-spam emails classified as spam.

## Recall

(Sensitivity or True Positive Rate): The ratio of correctly predicted positive observations to all the actual positives.

Recall measures the proportion of true positive predictions among all actual positive instances. It is useful when the cost of false negatives is high.

Example: In disease detection, recall is important because we want to identify as many positive cases as possible.

## F1 Score:

F1 score is a measure of a test's accuracy, specifically useful in binary classification problems. It is the harmonic mean of precision and recall. The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the worst performance.

The weighted average of Precision and Recall. It tries to find the balance between Precision and Recall. It is particularly useful when there is an uneven class distribution.

# ROC-AUC (Receiver Operating Characteristic - Area Under Curve):

The ROC-AUC measures the model's ability to distinguish between classes. It plots the true positive rate against the false positive rate at various threshold settings. AUC ranges from 0 to 1, with higher values indicating better performance.

Example: An AUC of 0.5 indicates no discriminatory ability, while an AUC of 1.0 indicates perfect discrimination.

## Classification Metrics

### Mean Absolute Error (MAE):

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It provides an easily interpretable measure in the same units as the target variable.

Example: If a model's predictions are off by an average of \$5, the MAE is \$5.

### Mean Squared Error (MSE):

MSE measures the average of the squares of the errors. It gives more weight to larger errors, making it useful for identifying outliers.

Example: If a model's predictions have large errors, the MSE will be high due to the squaring of errors.

### Root Mean Squared Error (RMSE):

RMSE is the square root of the mean squared error, providing error in the same units as the output variable. It is more interpretable than MSE.

Example: If the RMSE is 10, it indicates that the typical error is around 10 units.

### R-squared (Coefficient of Determination):

R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating better fit.

Example: An R-squared of 0.8 means 80% of the variance in the target variable is explained by the model.

## Adjusted R-squared:

Adjusted R-squared adjusts for the number of predictors in the model, penalizing for adding non-significant predictors. It is useful for comparing models with different numbers of predictors.

Example: If adding a predictor does not improve the model significantly, the adjusted R-squared will be lower than the R-squared.

## Clustering Metrics

### Silhouette Score:

Measures how similar an object is to its own cluster compared to other cluster. How well-separated the clusters are.

Range: -1 to 1

Higher values indicate better clustering. A value close to 1 means the data points are well-clustered, while a value close to -1 means they are poorly clustered.

### Davies-Bouldin Index:

The average similarity ratio between each cluster and its most similar one. Range: 0 to  $\infty$

Lower values indicate better clustering. A lower value means clusters are compact and well-separated from each other. A Davies-Bouldin Index of 0 indicates perfect clustering.

### Adjusted Rand Index (ARI):

The similarity between two clustering results, adjusted for chance grouping.

Range: -1 to 1

Higher values indicate better agreement between the clusterings. A value of 1 means perfect agreement, while 0 indicates random labeling and -1 indicates complete disagreement.