EVALUATION METRICS

Dr. E.Milgo

1. Evaluation Metrics for Classification

To assess the performance of a classification model and ensure it generalizes well to unseen data, the model needs to be evaluated statistically. The goal is to measure the accuracy and reliability of the model, identify overfitting or underfitting and compare the performance of different models.

Confusion Matrix

A **confusion matrix** is a table that summarizes the performance of a classification model by displaying the counts of true positives, false positives, true negatives, and false negatives.

	PREDICTED		
ACTUAL		**TRUE**	**FALSE**
	TRUE	True Positive (TP)	False Negative(FN)
	FALSE	False Positive(FP)	True Negative (TN)

100 patients

TRUE FIGURE FROM THE HOSPITAL

70 actually had Covid

30 did not have Covid

THE FIGURES FROM THE MODEL

60 patients have COVID

40 Patients do not have COVID

		PREDICTED		
ACTUAL		**TRUE**	**FALSE**	**Total**
	TRUE	TP = 50	FN =20	70
	FALSE	FP = 10	TN =20	30
	Total	60	40	**100**

To summarize,

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- **True Positives (TP):** Correctly predicted positive classes.
- **True Negatives (TN):** Correctly predicted negative classes.
- **False Positives (FP):** Incorrectly predicted positive classes (**Type I error)**.
- **False Negatives (FN):** Incorrectly predicted negative classes (**Type II error**).

Accuracy: Measures the proportion of correctly predicted instances. Can be misleading in imbalanced datasets.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision: Measures the proportion of correctly predicted positive instances out of all predicted positive instances. Useful when the cost of FP is high.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall / Sensitivity / True Positive Rate (TPR): Measures the proportion of correctly predicted positive instances out of all actual positive instances. Useful when the cost of FN is high.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score: Harmonic means of precision and recall. Useful when seeking a balance between precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

False Positive Rate (FPR): Measures the proportion of actual negative instances that are incorrectly classified as positive. It quantifies how many false alarms a classification model generates.

$$\text{FPR} = \frac{FP}{FP+TN}$$

ROC Curve: Plots the **True Positive Rate (TPR)** against the False Positive Rate (FPR) at various threshold settings. The **AUC (Area Under the Curve):** Measures the entire area under the ROC curve where; AUC = 1: Perfect classifier and AUC = 0.5: Random classifier

2. Model Evaluation for Regression Models

Regression models are evaluated using metrics that assess how well their predictions match the actual values. Key metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. These metrics help determine the accuracy and reliability of the model's predictions.

- **Mean Absolute Error (MAE):** The average absolute difference between predicted and actual values.
- **Mean Squared Error (MSE):** The average squared difference between predicted and actual values.
- **R-squared (R^2):** The proportion of variance in the dependent variable that is predictable from the independent variables.

3. Cross-Validation

Cross-validation is a crucial technique in machine learning used to evaluate a model's performance on unseen data while reducing overfitting. It ensures the model generalizes well by testing it on different subsets of the data.

Common Cross-Validation Methods

K-Fold Cross-Validation

- Splits the dataset into K equal subsets (folds).

- The model is trained on K-1 folds and tested on the remaining fold.
- The process repeats K times, with each fold serving as a test set once.
- The final performance is averaged across all folds.

Stratified K-Fold Cross-Validation

- Similar to K-Fold but ensures that each fold maintains the same class distribution as the original dataset.
- Particularly useful for imbalanced datasets to prevent biased evaluation.