Métriques d'évaluation des Modèles :

Les métriques d'évaluation des modèles d'apprentissage sont des mesures utilisées pour évaluer les performances d'un modèle par rapport aux données d'entraînement et de test. Le choix des métriques appropriées dépend du type de problème à résoudre (classification, régression, clustering, etc.).

I- Métriques d'évaluation des Modèles Supervisés

I-1 Métriques pour les problèmes de Régression :

- ***Erreur absolue moyenne (Mean Absolute Error)*** :(Willmott, et al . 2005) : L'erreur absolue moyenne est la moyenne de la différence entre les valeurs réelles et les valeurs prédites. Cependant, ils ne nous donnent aucune idée de la direction de l'erreur, c'est-à-dire si nous sommes sous-prédits ou sur-prédits.

$$MAE = \frac{1}{N}\sum_{i}^{N}|y_i - \hat{y}|$$

- ***Erreur quadratique moyenne (Mean Squared Error***) : L'erreur quadratique moyenne (MSE) est assez similaire à l'erreur absolue moyenne, la seule différence étant que MSE prend la moyenne du carré de la différence entre les valeurs réelles et les valeurs prédites. L'avantage de MSE est qu'il est plus facile de calculer le gradient, tandis que l'erreur absolue moyenne nécessite des outils de programmation linéaire compliqués. Comme nous prenons le carré de l'erreur, l'effet des erreurs plus importantes devient plus prononcé que l'erreur plus petite, donc le modèle peut désormais se concentrer davantage sur les erreurs plus importantes. (Mishra, 2018)

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y})^2$$

- ***Erreur quadratique moyenne racine*** (Root Mean Squared Error) :Selon (Neill el al., 2018) L'erreur quadratique moyenne racine est la racine carrée de la moyenne du carré de toutes les erreurs. L'utilisation de RMSE est très courante et elle est considérée comme une excellente métrique d'erreur à usage général pour les prédictions numériques. C'est une bonne mesure de la précision pour la comparaison entre les erreurs de prédiction de différents modèles. Où Oi sont les observations, Si les valeurs prédites d'une variable et n le nombre d'observations disponibles pour l'analyse.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(S_i - O_i)^2}$$

- ***Le coefficient de corrélation de Pearson*** : (karl Pearson, 1895)(Wu et al.,2019) a défini le coefficient de corrélation comme suit : Le coefficient de corrélation de Pearson est généralement représenté par la lettre r, si nous avons un ensemble de données {x1, ..., xn} contenant n valeurs et la prédiction de l'ensemble de données {y1, ..., yn} contenant n valeurs , alors cette formule pour r est :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Où n est la taille de l'échantillon, xi est l'échantillon indexé avec i, yi est la prédiction du système correspondant, et $\bar{x}$, $\bar{y}$sont les moyennes de xi et yi, respectivement.

- ***Coefficient de Détermination (R^2)***: il mesure la proportion de la variance dans la variable dépendante qui est prévisible à partir de la variable indépendante.

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i^N (y_i - \bar{y})^2}$$

I-2- Métriques pour les problèmes de Classification :

- ***Matrice de confusion*** (Confusion Matrix) : (Kohavi et al., 1998) Matrice de confusion comme son nom l'indique nous donne une matrice en sortie et décrit les performances complètes du modèle (Mishra, 2018). C' est un outil qui permet de savoir à quel point le modèle est « confus », ou qu'il se trompe. Il s'agit d'un tableau avec en colonne les différents cas réels et en ligne les différents cas d'usage prédits.

Prenons l'exemple d'un test médical, la matrice sera la suivante :

		REEL *Si le patient est atteint ou non*	
		Est atteint	N'est pas atteint
PREDICTION *Ce que notre modèle prédisait*	Est atteint	Nombre de **Vrai positif**	Nombre de **Faux positif**
	N'est pas atteint	Nombre de **Faux négatif**	Nombre de **Vrai négatif**

On obtient donc les quatre valeurs suivantes :

- Vrai positif (TP), les valeurs réelles et prédites sont identiques et positives. Le patient est malade et le modèle le prédit.
- Vrai négatif (TN), les valeurs réelles et prédites sont identiques et négatives. Le patient n'est pas malade et le modèle prédit qui ne l'est pas.

- Faux positif (FP), les valeurs réelles et prédites sont différentes. Le patient n'est pas malade, mais le modèle prédit qui l'est.
- Faux négatif (FN), les valeurs réelles et prédites sont différentes. Le patient est malade, mais le modèle prédit qu'il ne l'est pas.

L'étude de ces valeurs prédictives permet de définir si le modèle est fiable, dans quels cas il commet des erreurs et dans quelle mesure.

La précision de la matrice peut être calculée en prenant la moyenne des valeurs situées sur la « diagonale principale », **constitue la base des autres types de mesures**.

- ***La précision*** (Precision) (Mishra, 2018) mesure la proportion d'observations positives correctement prédites parmi toutes les observations prédites comme positives.

$$\frac{TP}{TP + FP}$$

où TP (True Positives) est le nombre d' observations positives correctement prédites et FP (False Positives) est le nombre d'observations négatives incorrectement prédites comme positives.

- ***Rappel*** (Recall):(Kulhare, 2017), le rappel mesure la proportion d'observations positives correctement prédites parmi toutes les observations réellement positives.

$$\frac{TP}{TP + FN}$$

où FN (False Negatives) est le nombre d'observations positives incorrectement prédites comme négatives.

- ***F1-Score*** : C'est la moyenne harmonique de la précision et du rappel. Il combine à la fois la précision et le rappel en une seule mesure.

$$F1 = 2X\frac{PrécisionXRappel}{Précision + Rappel}$$

- ***Exactitude*** (Accuracy) : Elle mesure la proportion totale d'observations correctement classées parmi toutes les observations.

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

où TN (True Negatives) est le nombre d'observations négatives correctement prédites.

- ***Courbe ROC et Aire sous la Courbe (ROC-AUC***) : (Receiver Operating Characteristic) est un graphique de la sensibilité (taux de vrai positif TP) en fonction de la spécificité (taux de vrai négatif TN) pour différents seuils de classification. L'AUC mesure l'aire sous la courbe ROC et fournit une mesure agrégée de la performance du modèle sur tous les seuils de classification.

- ***La validation croisée (cross-validation***) est une technique couramment utilisée pour évaluer les performances des modèles supervisés de manière robuste et fiable. Elle permet d'estimer la capacité d'un modèle à généraliser à de nouvelles données en simulant le processus d'apprentissage sur plusieurs sous-ensembles des données disponibles.

Fonctionnement de la validation croisée :

1. Diviser les données en deux ensembles distincts : un ensemble d'entraînement (training set) et un ensemble de test (test set) pour évaluer les performances du modèle.

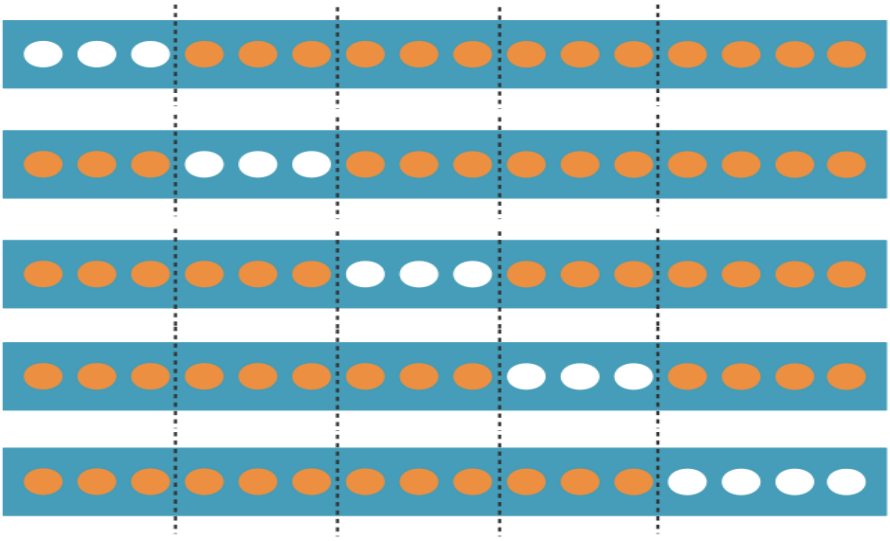
2. Entraîner et évaluer le modèle sur l'ensemble d'entraînement et ses performances sont évaluées sur l'ensemble de test en utilisant une métrique appropriée (comme la précision, le rappel, l'exactitude, etc.).

3. Répéter le processus : La validation croisée consiste à répéter cette procédure plusieurs fois en changeant la manière dont les données sont divisées en ensembles d'entraînement et de test à chaque itération. Il existe différentes techniques de validation croisée, notamment la validation croisée k-fold, la validation croisée leave-one-out (LOOCV), la validation croisée leave-p-out, etc….

4. Calculer la performance moyenne : Une fois que toutes les itérations sont terminées, la performance moyenne du modèle sur l'ensemble de test est calculée en agrégeant les performances obtenues lors de chaque itération.

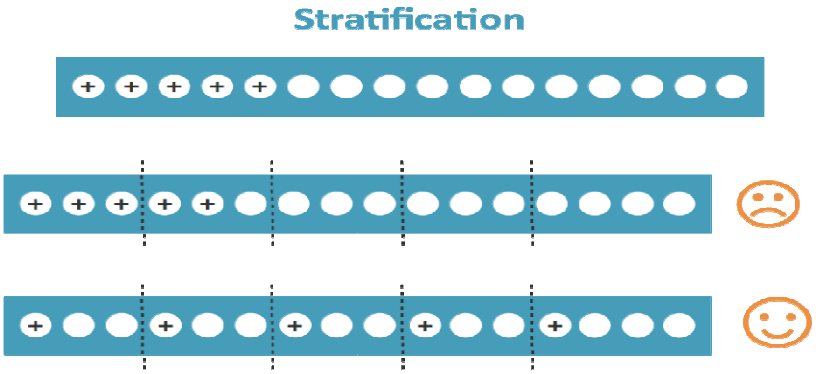
Important : La validation croisée permet de maximiser l'utilisation des données disponibles pour l'entraînement et l'évaluation du modèle, ce qui est particulièrement important lorsque les ***ensembles de données sont limités en taille*** ou bien les ***classes sont déséquilibrées***.
Exemple :
Validation croisée à 5 folds Chaque point appartient à un jeu de test (en blanc) et pour les autres validations aux jeux d'entraînements (en orange)
La validation croisée permet donc d'évaluer un modèle en ayant la moyenne des performances et l'erreur type sur chacun des folds ou en évaluant les prédictions faites sur l'ensemble des données.

Pour des raisons de temps de calcul, on utilise généralement 5 ou 10 folds.



Pour cette méthode, il est important d'appliquer la stratification.

La stratification : est un processus qui consiste à diviser les données connues en folds homogènes avant l'échantillonnage, c'est-à-dire répartir les étiquettes pour que chaque fold ressemble au maximum à un petit jeu de données connues.



II- Métriques pour modèles non Supervisés « Clustering »

- ***Indice de Silhouette : C'*** est une mesure de la cohésion intra-cluster et de la séparation inter-cluster dans un ensemble de données . Il mesure à quel point les objets d'un même cluster sont similaires entre eux et à quel point ils sont différents des objets des autres clusters. Il varie de -1 à 1, où une valeur élevée indique que les objets sont bien groupés dans leur propre cluster et séparés des autres clusters.

La formule de l'indice de silhouette S pour un point i est la suivante :

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Où :- a_i est la distance moyenne entre le point i et tous les autres points dans le même cluster (cohésion intra-cluster).
- b_i est la distance moyenne entre le point i et tous les points dans le cluster voisin le plus proche, où le cluster voisin est celui auquel le point i n'appartient pas (séparation inter-cluster).
=>L'indice de silhouette global pour un ensemble de données est la moyenne des indices de silhouette de tous les points dans l'ensemble de données.
- Un indice de silhouette proche de 1 indique que le point est bien classé par rapport aux autres points de son cluster et mal classé par rapport aux points des autres clusters.
- Un indice de silhouette proche de -1 indique que le point est mal classé par rapport aux points de son cluster et bien classé par rapport aux points des autres clusters.
- Un indice de silhouette proche de 0 indique que le point est proche du seuil de décision entre les clusters et pourrait être affecté à un cluster ou à un autre.

- ***Inertie (Inertia):*** Elle mesure la cohésion (lien) intra-cluster d'un ensemble de données. c'est-à-dire à quel point les points à l'intérieur d'un cluster sont similaires les uns aux autres.

La formule de l'inertie dépend de la mesure de distance utilisée. Pour les algorithmes de clustering tels que K-means, la distance euclidienne est couramment utilisée. La formule de l'inertie dans ce cas est : la somme des carrés des distances des échantillons au centre de leur cluster le plus proche:

C un ensemble de k clusters, C_i le ième cluster et μ_i le centre du cluster C_i.

$$I = \sum_{i}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Une inertie plus faible indique une meilleure cohésion intra-cluster et donc un meilleur partitionnement des données.

FIN