**Exercice de Classification sur le Dataset Titanic**

**Introduction**

Dans cet examen, nous simulons un scénario réel en utilisant le dataset Titanic.

Imaginez que nous disposons d'une base de données contenant des informations détaillées sur chaque passager, telles que l'âge, le sexe, la classe et le prix du billet. Cependant, cette base de données manque une colonne essentielle indiquant si chaque passager a survécu, ce qui rend impossible la création d'un modèle prédictif de survie en utilisant une approche d'apprentissage supervisé.

Pour résoudre ce problème, nous avons consulté un historien du Titanic (notre expert), qui a analysé les données et fourni des étiquettes de survie pour chaque passager. En utilisant ces étiquettes fournies par l'expert, votre tâche sera divisée en trois parties :

**1.** **Prétraitement des données :** Nettoyer et préparer le dataset Titanic en gérant les valeurs manquantes, en encodant les variables catégorielles et en mettant à l'échelle les caractéristiques si nécessaire.

**2. Construction d'un modèle supervisé :** Utiliser les étiquettes de survie de l'expert pour entraîner un modèle supervisé capable de prédire la survie des passagers à partir des données traitées.

**3. Évaluation avec l'apprentissage non supervisé :** Enfin, tester si un modèle non supervisé, comme le clustering, peut approximer les étiquettes de l'expert, en regroupant les passagers de manière similaire aux étiquettes de survie fournies.

Ce tp vous permettra de parcourir les étapes de préparation des données, d'entraînement du modèle et d'évaluation non supervisée pour simuler un projet de classification en situation réelle.

Questions pour chaque partie

**Partie 1 : Prétraitement des données**

- Question 1.1 : Comment géreriez-vous les données manquantes dans le dataset Titanic ? Quelles sont les options disponibles, et quelle méthode choisiriez-vous pour chaque colonne ?

- Question 1.2 : La colonne "Sex" est catégorielle. Comment encoderiez-vous cette caractéristique pour qu'elle puisse être utilisée dans un modèle d'apprentissage automatique ?

- Question 1.3 : Pensez-vous qu'une mise à l'échelle des caractéristiques est nécessaire pour ce dataset ? Si oui, expliquez quelles caractéristiques devraient être mises à l'échelle et pourquoi.

- Question 1.4 : Quelles étapes suivriez-vous pour gérer les éventuels outliers dans le dataset, en particulier dans la colonne "Fare" ?

**Partie 2 : Construction d'un modèle supervisé**

- Question 2.1 : Quel algorithme d'apprentissage supervisé choisiriez-vous pour prédire la survie (par exemple, régression logistique, arbres de décision, etc.) ? Justifiez votre choix.

- Question 2.2: Comment diviseriez-vous les données en ensembles d'entraînement et de test ? Quelle proportion utiliseriez-vous, et pourquoi ?

- Question 2.3 : Après l'entraînement du modèle, comment évalueriez-vous ses performances ? Quelles métriques utiliseriez-vous, et pourquoi ?

- Question 2.4 : Si votre modèle sous-performe, quelles étapes suivriez-vous pour l'améliorer (par exemple, réglage des hyperparamètres, ingénierie des caractéristiques, etc.) ?

**Partie 3 : Évaluation avec l'apprentissage non supervisé**

- Question 3.1 : Expliquez comment vous appliqueriez une technique d'apprentissage non supervisé, comme le clustering, pour regrouper les passagers dans le dataset. Quel algorithme utiliseriez-vous ?

- Question 3.2 : Après le clustering, comment évalueriez-vous l'efficacité du modèle non supervisé ? Comment compareriez-vous les clusters résultants aux étiquettes de survie de l'expert ?

- Question 3.3 : Si le modèle non supervisé ne performe pas bien, quels changements pourriez-vous apporter pour l'améliorer (par exemple, sélection de caractéristiques, techniques de clustering différentes) ?

- Question 3.4 : Pensez-vous que l'apprentissage non supervisé peut atteindre des performances similaires à un modèle supervisé pour prédire la survie ? Pourquoi ou pourquoi pas ?

**Description des Colonnes**

Vous trouverez ci-dessous une description des colonnes disponibles dans le jeu de données du Titanic :

* **PassengerId** : Un identifiant unique pour chaque passager.
* **Pclass** : La classe du passager (1 = première classe, 2 = deuxième classe, 3 = troisième classe).
* **Name** : Le nom du passager.
* **Sex** : Le genre du passager (homme ou femme).
* **Age** : L'âge du passager en années.
* **SibSp** : Le nombre de frères, sœurs ou conjoints que le passager avait à bord du Titanic.
* **Parch** : Le nombre de parents ou d'enfants que le passager avait à bord du Titanic.
* **Ticket** : Le numéro du billet.
* **Fare** : Le prix du billet payé par le passager.
* **Cabin** : Le numéro de la cabine où le passager a séjourné.
* **Embarked** : Le port où le passager est monté à bord du Titanic (C = Cherbourg, Q = Queenstown, S = Southampton).
* **Survived** : L'indicateur précisant si le passager a survécu (1 = a survécu, 0 = n'a pas survécu). Cette colonne a été fournie par l’expert.