esprit Se former autrement HONORIS UNITED UNIVERSITIES	**EXAMEN** Semestre : 1 ☐ 2 ☒ Session : Principale ☒ Rattrapage ☐

Module: Machine Learning Fundamentals **Classes:** 3A **Date:** 28/05/2025
Heure: 11h00 **Durée:** 1h30

Documents autorisés: OUI ☐ NON ☒ **Calculatrice autorisée:** OUI ☒ NON ☐
Internet autorisé: OUI ☐ NON ☒ **Nombre de pages:** 10

ETUDIANT(E) Identifiant: . . . . . . Noms & Prénom:	Classe: . . . . . . Salle: . . . . .

Questions à choix unique : (4 points)

Pour chaque question, cocher la bonne réponse.

Question 1: Laquelle des métriques suivantes est la plus utilisée pour évaluer la performance d'un modèle de régression linéaire ?

- ☐ Accuracy.
- ☑ Mean Squared Error (MSE).
- ☐ F1-score.
- ☐ Confusion Matrix.

Question 2: Quelle méthode d'imputation est adaptée pour remplacer les valeurs manquantes d'une variable catégorielle ?

- ☐ Imputation par la moyenne.
- ☑ Imputation par la modalité la plus fréquente.
- ☐ Imputation par les voisins les plus proches.
- ☐ Imputation par la médiane.

Question 3: Dans l'algorithme KNN, quel rôle joue la normalisation des données avant l'entraînement ?

- ☑ Éviter que les variables à grande échelle dominent la distance calculée.
- ☐ Améliorer la vitesse d'entraînement de l'algorithme.
- ☐ Réduire le nombre de voisins considérés.
- ☐ Transformer les variables catégorielles en numériques.

Question 4: Quelle est la première étape du processus CRISP-DM ?

- ☐ Modélisation.
- ☐ Préparation des données.
- ☐ Compréhension des données.
- ☑ Compréhension du métier.

Exercice 1 : (10 points)

Un enseignant souhaite mieux comprendre les habitudes de participation de ses étudiants afin de les segmenter en groupes de comportement similaires et d'adapter ses méthodes pédagogiques.

Il collecte, pour 10 étudiants, le nombre moyen d'interventions orales par semaine et le nombre moyen de messages postés sur le forum en ligne par semaine.

Le tableau ci-dessous dispose des données de 10 étudiants, identifiés de A à J.

Étudiants	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Interventions orales/semaine	5.0	4.5	4.7	1.0	1.5	1.2	3.0	2.8	3.2	5.2
Message en ligne/semaine	1.0	1.2	1.0	5.0	4.8	4.5	2.5	3.0	2.7	0.8

1. **(1 point)** Quel type d'apprentissage automatique convient à notre étude de cas ? Justifier votre réponse.

   Le type d'apprentissage automatique qui convient à cette étude de cas est l'apprentissage non supervisé car les variables fournies sont des caractéristiques mesurées (interventions orales/semaine et messages en ligne/semaine), sans étiquette prédéfinie.

2. **(0.5 point)** Le dataset est crée à l'aide de la bibliothèque Pandas en Python. Compléter les pointillées pour que ce script soit correctement implémenté.

[1]:
```
import pandas as pd
data = {
    "Étudiant": ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J'],
    "Interventions_orales": [5, 4.5, 4.7, 1, 1.5, 1.2, 3, 2.8, 3.2, 5.2],
    "Messages_enligne": [1, 1.2, 1, 5, 4.8, 4.5, 2.5, 3, 2.7, 0.8]
}

df = pd.DataFrame(data)
```

3. **(0.5 point)** Donner l'instruction Python qui permet d'afficher la description statistique ci-dessous du dataset.

[2]:
```

df.describe()
```

	Interventions_orales	Messages_enligne
count	10.000000	10.000000
mean	3.210000	2.650000
std	1.602394	1.656133
min	1.000000	0.800000
25%	1.825000	1.050000
50%	3.100000	2.600000
75%	4.650000	4.125000
max	5.200000	5.000000

4. **(0.5 point)** déterminer les valeurs du premier quartile **(Q1)**, et de la médiane pour la colonne **Messages_enligne**.

   Q1= 1.05 **Médiane=** 2.60

Dans la suite, on s'intéresse à segmenter les étudiants en groupes de comportement similaires à l'aide de l'algorithme de Clustering **K-Means**, en se basant uniquement sur les deux variables mesurées.

5. **(0.5 point)** Extraire les données numériques des colonnes **Interventions_orales** et **Messages_enligne** sous forme d'un tableau `X`.

[3]:
```
X = df[["Interventions_orales", "Messages_enligne"]].values
```

6. **(1 point)** Remplir les pointillés pour créer un modèle de **K-Means** avec **2 clusters**.

[4]:
```
from sklearn.cluster import KMeans

model = KMeans(n_clusters=2)
```

7. **(0.5 point)** Compléter le champ vide pour entraîner le modèle de **K-Means**.

[5]:
```
model.  fit(  X )
```

8. **(0.5 point)** Compléter le champ vide par la méthode correcte permettant de prédire à quel groupe appartient chaque étudiant.

[6]:
```
model.  predict(  X )
```

9. **(0.5 point)** L'instruction ci-dessus retourne :

[6]:
```
array([1,1,1,0,0,0,1,0,1,1])
```

Interpréter le résultat, sachant que les étudiants sont nommés de A à J dans l'ordre du tableau.

Les étudiants ont été regroupés en deux clusters (0 et 1): les étudiants A, B, C, G, I, J sont dans le groupe 1, et les étudiants D, E, F, H sont dans le groupe 0.

10. **(0.5 point)** On souhaite connaître les coordonnées des centroïdes des groupes (clusters) afin de mieux comprendre leur position dans l'espace des données. Quelle est l'instruction Python à utiliser pour afficher les coordonnées des centres des clusters ? Cocher la bonne réponse.

- ☐ `centers= model.centers_`
- ☐ `centers= model.predict_centers()`
- ☑ `centers= model.cluster_centers_`

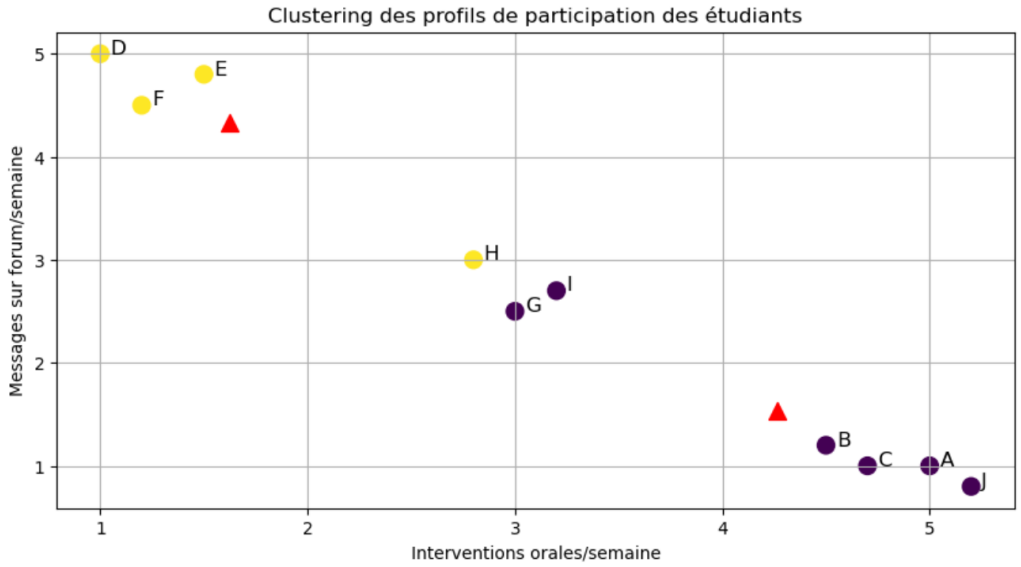
[7]:
```
centers
```

[7]:
```
array([[1.625     , 4.325     ],
       [4.26666667, 1.53333333]])
```

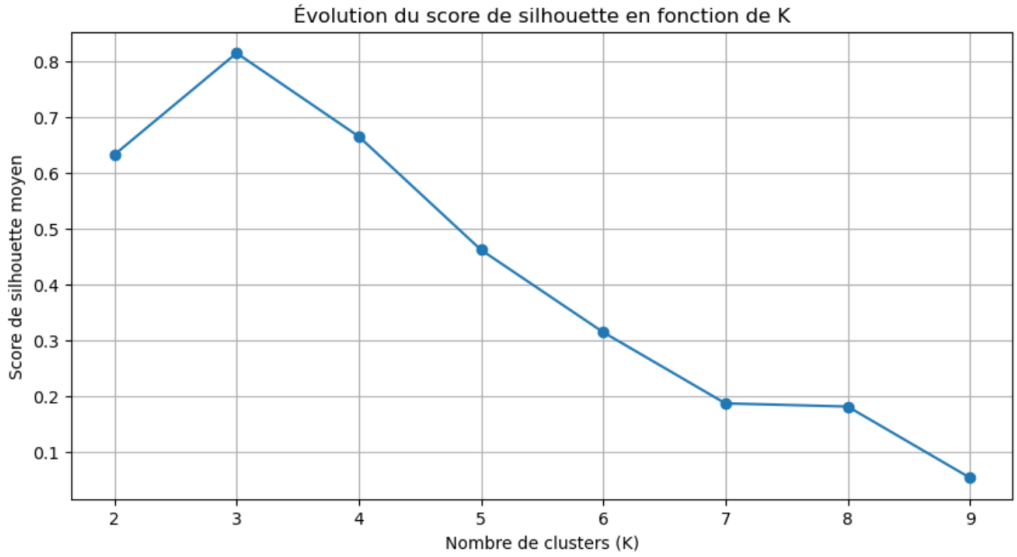
11. **(1 point)** Le code suivant est utilisé pour visualiser les résultats du modèle de **K-Means**. Compléter les champs manquants pour tracer les centroïdes des clusters, et l'ensemble des données, présentés dans la figure ci-dessous.

[8]:
```
import matplotlib.pyplot as plt

plt.figure(figsize=(10,5))

plt.scatter( X[:,0] , X[:,1] ,s=100, c=model.predict(X))

plt.scatter( centers[:,0] , centers[:,1]  , c='r',marker='^',s=100)

plt.xlabel("Interventions orales/semaine")
plt.ylabel("Messages sur forum/semaine")
plt.title("Clustering des profils de participation des étudiants")
plt.grid(True)
```

ETUDIANT(E) Identifiant: . . . . . . Noms & Prénom:	 Classe: . . . . . . Salle: . . . . .



12. **(1 point)** À partir du graphique ci-dessous, montrant l'évolution du score de silhouette en fonction du nombre de clusters, que peut-on conclure sur le choix optimal du nombre de clusters ? justifier votre réponse.



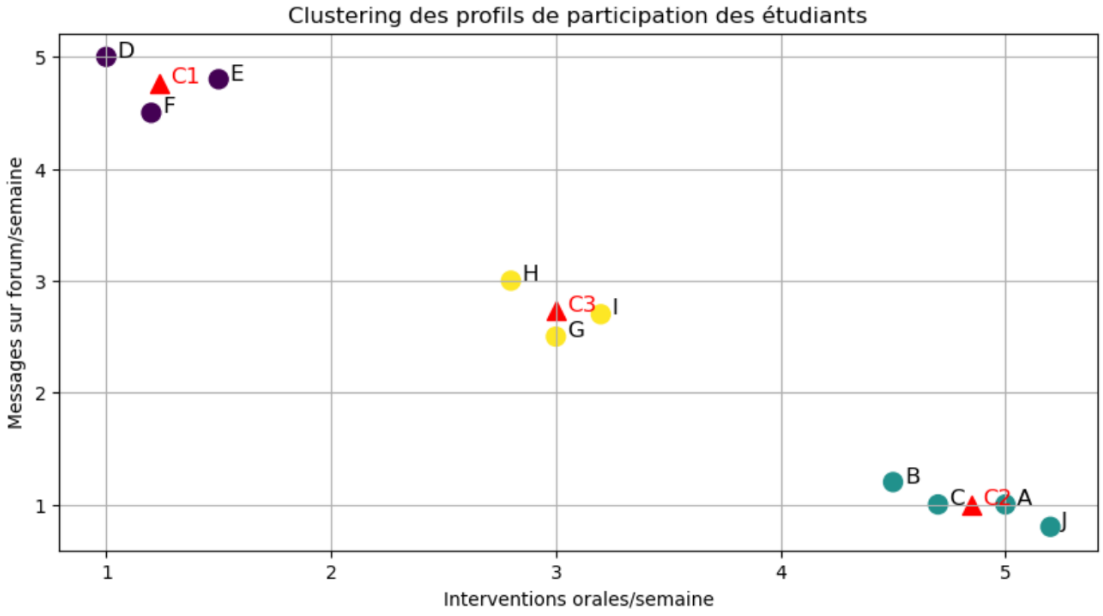
Le graphique montre que le score de silhouette augmente jusqu'à atteindre un maximum pour $K = 3$, puis diminue légèrement au-delà. Il dépasse 0.8, ce qui indique une excellente qualité de regroupement. Ainsi, le choix optimal est $K = 3$ clusters.

En considérant dans la suite le choix optimal du nombre de Clusters, les centroïdes issus du mod-

èle de **K-Means** sont respectivement de coordonnées :

$$C1(1.23, 4.77), \quad C2(4.85, 1.0), \quad C3(3.0, 2.73).$$

La figure ci-dessous présente le nuage de points ainsi que les centroïdes optimaux du modèle.



13. (a) **(0.5 point)** Relier chaque centroïde à l'une des descriptions suivantes:

    i. **(D1)** Un groupe d'étudiants modérés (équilibrés).

    ii. **(D2)** Un groupe d'étudiants très actifs en classe mais peu en ligne.

    iii. **(D3)** Un groupe d'étudiants très actifs sur le forum mais réservés en présentiel.

    C1 $\longrightarrow$ (D3) C2 $\longrightarrow$ (D2) C3 $\longrightarrow$ (D1)

    (b) **(1 point)** Disons qu'on a un **nouvel étudiant "Z"** avec les valeurs suivantes :

    Interventions orales = 4.0 et **Messages sur le forum = 2.0**.

    Calculer la distance euclidienne d(C1,Z) entre le centroïde $C1$ et le nouvel étudiant "Z".

    d(C1,Z) = $\sqrt{(1.23-4)^2 + (4.77-2)^2} = 3.91737$

    (c) **(0.5 point)** Prédire à quel Cluster cet étudiant "Z" appartient, sachant que d(C2,Z)=3.91 et d(C3,Z)=1.24.

    L'étudiant Z sera donc affecté au cluster 3, car il est proche du centroide C3.

Exercice 2 : (6 points)

Dans cet exercice, on va utiliser un algorithme d'arbre de décision afin de prédire le statut d'activité d'un client, c'est-à-dire déterminer si un client est actif ou inactif. Pour cela, on va se baser sur plusieurs critères qui peuvent influencer le comportement des clients :

- Âge : L'âge du client, exprimé en années.
- Temps passé sur le site (en minutes) : La durée moyenne d'une visite du client sur le site web.
- Satisfaction client : Une variable indiquant si le client est satisfait, insatisfait ou neutre par rapport aux services ou produits proposés.
- Nombre de visites : Le nombre total de visites du client sur le site durant une période donnée.

On dispose du jeu de données suivant, qui inclut les informations de six clients (on se concentrera uniquement sur ces six clients).

Client	Age	Temps sur site	Satisfaction client	Nombre visites	Statut client
Client1	25	15	Neutre	6	Actif
Client2	22	**NaN**	Insatisfait	1	Inactif
Client3	28	**NaN**	Neutre	4	Actif
Client4	40	12	**None**	2	Inactif
Client5	27	**NaN**	Satisfait	**NaN**	Actif
Client6	45	**NaN**	Neutre	1	Inactif

1. **(0.5 point)** Est-ce que le modèle d'arbre de décision qu'on va construire représente un arbre de décision de classification ou de régression ? Justifier votre réponse.

   Le modèle d'arbre de décision qu'on va construire est un arbre de décision de classification, car la variable cible que l'on cherche à prédire est une variable catégorielle.

2. **(0.5 point)** Indiquer le type de chaque caractéristique présente dans notre jeu de données.

   - Âge : quantitative discrète.
   - Temps sur site : quantitative continue.
   - Satisfaction client : qualitative ordinale.
   - Nombre de visites : quantitative discrète.

Avant de commencer la modélisation, il est nécessaire de gérer les valeurs manquantes dans les variables "Temps sur site", "Satisfaction client" et "Nombre visites".

3. (a) **(0.5 point)** Imputer les valeurs manquantes en utilisant **la médiane** pour la variable "Nombre visites" et **le mode** pour la variable "Satisfaction Client".
   La médiane pour la variable "Nombre visites" est égale à 2.
   Le mode pour la variable "Satisfaction client" est la modalité **Neutre**.

   (b) **(0.5 point)** Comment peut-on gérer les valeurs manquantes dans la variable "Temps sur site", en tenant compte de leur forte proportion ? Justifier votre réponse.
   On remarque que la variable "Temps sur site" contient 4 valeurs manquantes sur 6, soit environ 67% des données manquantes. Ainsi, la suppression de cette colonne est une solution raisonnable pour éviter d'introduire un biais dans le modèle, car avec seulement 2 valeurs disponibles, il est difficile d'imputer les valeurs manquantes de façon fiable.

La partie suivante est consacrée à la construction de l'arbre de décision pour prédire le "Statut client" en utilisant uniquement les variables "Age", "Satisfaction client" et "Nombre visites".

4. **(1 point)** Citer deux critères de division pour un arbre de décision et expliquer brièvement ce que mesure chacun de ces critères.

   - L'indice de Gini : mesure l'impureté d'un ensemble de données.
   - L'entropie : mesure l'homogénéité ou la pureté d'un ensemble par rapport à la variable cible.

5. **(0.5 point)** Pour effectuer la première division de l'arbre de décision, on a calculé le gain de Gini :

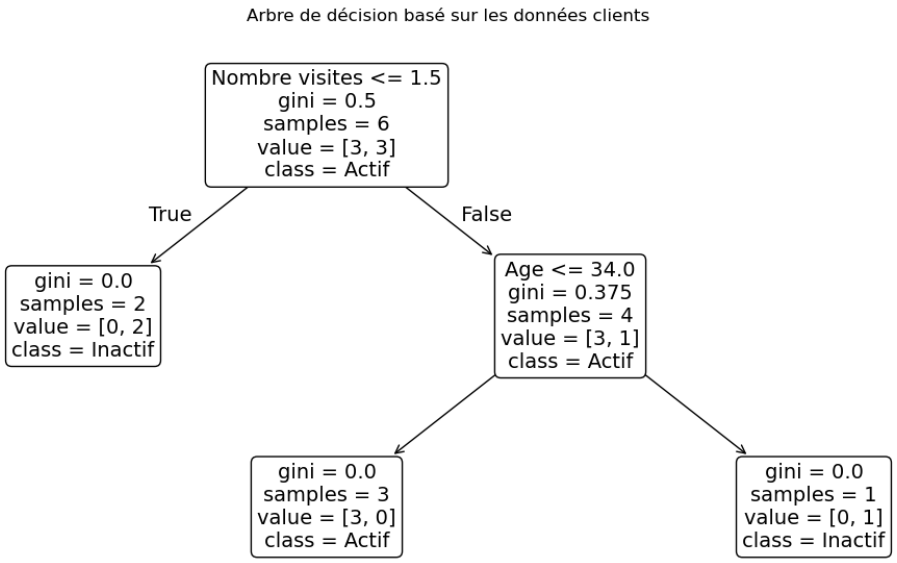
   - Le gain de Gini pour "Âge" est 0.056.
   - Le gain de Gini pour "Satisfaction client" est 0.
   - Le gain de Gini pour "Nombre visites" est 0.25.

   Identifier quelle variable doit être choisie comme racine de l'arbre de décision. Expliquer votre choix.

   Le gain de Gini le plus élevé est pour la variable "Nombre visites" = 0.25. Cela signifie que la division sur cette variable réduit le plus l'impureté des sous-ensembles résultants. Donc, cette variable doit être choisie comme racine de l'arbre de décision, car elle permet une meilleure séparation des classes dès le premier niveau.

ETUDIANT(E) Identifiant: . . . . . . Noms & Prénom:	 Classe: . . . . . . Salle: . . . . .

À partir des données clients (6 observations), l'algorithme a généré l'arbre suivant :



6. **(0.5 point)** Prédire le statut d'activité d'un nouveau client âgé de 30 ans et ayant effectué 3 visites.
   Comme le client a 3 visites ($\geq$ 1.5), on teste ensuite si son âge est inférieur à 34 ; comme $30 < 34$, on prédit que son statut est Actif.

Après avoir construit le modèle d'arbre de décision, il est maintenant nécessaire d'évaluer la qualité de nos prédictions. Pour cela, on va utiliser un nouvel ensemble de données, représentant un échantillon test. Ces données comprennent les prédictions générées par le modèle d'arbre de décision ainsi que les résultats réels, afin de pouvoir évaluer la performance du modèle.

Age	**Satisfaction client**	**Nombre visites**	**Statut client (Réel)**	**Statut client (Prédit)**
30	Neutre	3	Inactif	Actif
24	Satisfait	1	Actif	Inactif
35	Insatisfait	2	Inactif	Inactif
22	Neutre	4	Inactif	Actif

7. (a) **(0.5 point)** Compléter la matrice de confusion ci-dessous à partir des résultats réels et les prédictions du modèle.

Réel \ Prédit	**Actif**	**Inactif**
Actif	0	1
Inactif	2	1

(b) **(0.5 point)** Calculer l'Accuracy (Exactitude) globale du modèle. Interpréter le résultat obtenu.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} = \frac{0+1}{4} = 0.25$$

Le modèle prédit correctement seulement 25% des statuts (actif ou inactif). Cela signifie que son exactitude est faible sur cette base de test, ce qui indique que le modèle doit être amélioré pour mieux distinguer les clients actifs des inactifs.

(c) **(1 point)** Calculer les métriques de précision (Precision) et rappel (Recall). Interpréter les résultats obtenus.

$$Precision = \frac{VP}{VP + FP} = \frac{0}{0+2} = 0,$$

ce qui signifie que parmi toutes les fois où le modèle a prédit "Actif", aucune n'était correcte.

$$Rappel = \frac{VP}{VP + FN} = \frac{0}{0+1} = 0,$$

ce qui signifie que le modèle n'a jamais réussi à identifier un client réellement actif. Ces deux résultats montrent que le modèle est très mauvais pour prédire le statut des clients actifs.

⋆⋆ **BON COURAGE** ⋆⋆