Veuillez prendre connaissance des consignes ci-dessous :

- Durée de l'examen: **2h30**
- Une feuille de note recto verso 8.5" X 11" ou A4 et une calculatrice programmable sont permises.
- Le barème est donné à titre indicatif et peut être sujet à modification.
- Vous devez encadrer les résultats et donner les valeurs avec deux chiffres après la virgule en arrondissant au plus proche. Par exemple, $1,947$ doit être noté $\boxed{1{,}95}$.
- Vous pouvez répondre aux questions en français ou en anglais.

Please read carefully the following instructions:

- Exam duration: **2h30**
- A double-sided sheet of size 8.5" X 11" or A4 and a programmable calculator are allowed.
- The marking of the questions is given for references and may be subject to change.
- You must draw a box around the results and give values with two decimal places by rounding to the nearest value. For instance, 1.947 must be given as $\boxed{1.95}$.
- You are allowed to answer in French or in English.

1 Fouille des données (2 points)

Discutez si chacune des activités suivantes est une tâche de la fouille des données ou non.

(0.4 point) (a) Diviser les clients d'une entreprise selon leur sexe.
Non. Il s'agit d'une simple requête de base de données.

(0.4 point) (b) Calculer les ventes totales d'une entreprise.
Non. C'est un calcul comptable.

(0.4 point) (c) Prédire les résultats du lancer d'une paire de dés (équitables).
Non. Puisque les dés sont équitables, c'est un calcul de probabilité.

(0.4 point) (d) Surveiller le rythme cardiaque d'un patient pour détecter des anomalies.
Oui. Nous tenterions de créer un modèle capable de prédire la valeur continue du prix des actions (régression).

(0.4 point) (e) Extraire les fréquences d'une onde sonore.
Non. C'est du traitement du signal.

2 Préparation de données (1 point)

Les valeurs nulles dans les enregistrements de données peuvent faire référence à des valeurs manquantes ou non applicables.

Considérez le tableau suivant des membres d'un campus :

Nom	cours donnés	rôle
Jean	5	Professeur
Marie	1	Professeure
Bob	null	Étudiant
Lisa	null	Étudiante

Les valeurs nulles dans le tableau font référence à des valeurs non applicables, car les étudiants ne donnent pas de cours.

Supposons que nous souhaitions calculer la similarité entre les membres du campus en fonction de leur nombre de cours donnés.

(0.5 point) (a) Expliquez quelle est la limitation de l'approche pour calculer la similarité si nous remplaçons les valeurs nulles pour des 0.

Marie (professeure) sera plus similaire à Bob et Lisa qu'à Jean, l'autre professeur.

(0.5 point) (b) Expliquez quelle est la limitation de l'approche pour calculer la similarité si nous remplaçons les valeurs nulles par la valeur moyenne de cours donnés par les membres du campus (c'est-à-dire 3 cours).

À la fois Marie et Jean sont moins similaires l'un à l'autre qu'à Bob et Lisa.

3 Transformation de données (2 points)

Démontrez mathématiquement que la corrélation de Pearson entre deux attributs X et Y est un chiffre compris entre -1 et +1.

Piste: Utilisation de l'inégalité de Cauchy-Schwarz

$$(Cov(X,Y))^2 \leq Var(X) \cdot Var(Y)$$

où $Var(X) = \sigma(X)^2$ et $Var(Y) = \sigma(Y)^2$

Le coefficient de corrélation de Pearson r entre deux variables X et Y est défini comme suit :

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Notre objectif est de montrer que ce rapport est toujours compris entre -1 et $+1$.

Nous commençons par utiliser l'inégalité de Cauchy-Schwarz. Cette inégalité nous dit que le carré de la covariance entre X et Y est inférieur ou égal au produit de leurs variances.

Puisque $\text{Cov}(X,Y)$ fait partie de la formule de corrélation de Pearson, nous pouvons utiliser cette inégalité pour contraindre la valeur de r.

Nous élevons les deux membres de l'équation (1) au carré :

$$r^2 = \left(\frac{Cov(X,Y)}{\sigma_X \sigma_Y}\right)^2 = \frac{(Cov(X,Y))^2}{\sigma_X^2 \sigma_Y^2}$$

En utilisant l'inégalité de Cauchy-Schwarz :

$$(\text{Cov}(X,Y))^2 \leq \sigma_X^2 \cdot \sigma_Y^2$$

Ainsi :

$$r^2 \leq \frac{\sigma_X^2 \cdot \sigma_Y^2}{\sigma_X^2 \cdot \sigma_Y^2} = 1$$

Cela nous donne le résultat :

$$r^2 \leq 1$$

En prenant la racine carrée des deux côtés :

$$-1 \leq r \leq 1$$

4 Classification - Partie I (2 points)

Considérons le problème de prédire à quel point un joueur de hockey tirera bien contre un gardien de but particulier. L'ensemble d'entraînement contient dix exemples positifs et dix exemples négatifs. Supposons qu'il y ait deux attributs candidats pour diviser les données : ID (qui est unique pour chaque joueur) et *côté dominant* (gauche ou droit). Parmi les joueurs gauchers, neuf d'entre eux appartiennent à la classe positive et un à la classe négative. En revanche, parmi les joueurs droitiers, un seul appartient à la classe positive, tandis que les neuf autres appartiennent à la classe négative.

(0.5 point) (a) Calculez le gain d'information si nous utilisons l'ID comme attribut de séparation.

L'entropie initiale est:

$E_{initial} = -(0.5 \log 0.5 + 0.5 \log 0.5) = 1$

Si on sépare avec l'attribut ID, nous obtenons:

$$Gain = E_{initial} - (\sum_{i=1}^{20} Entropie(i)) = 1 - 20 \times \frac{1}{20}(-0 \log 0 - 1 \log 1) = 1$$

(0.5 point) (b) Calculez le gain d'information si nous utilisons le *côté dominant* comme attribut de séparation.

$$Gain = E_{initial} - \frac{10}{20}(-0.9 \log 0.9 - 0.1 \log 0.1) \times 2 = 1 - 0.469 = 0.531$$

(0.5 point) (c) En vous basant sur vos réponses aux parties (a) et (b), quel attribut sera choisi selon le gain d'information ? Justifiez votre réponse.

ID

(0.5 point) (d) Trouvez-vous que l'arbre qui sépare ayant comme seul attribut séparateur l'attribut *ID* possède une bonne capacité de généralisation ? Justifiez votre réponse.

Non. L'arbre souffre du surapprentissage.

5 Évaluation de modèles (1 point)

Supposons deux enregistrements X_i et X_j tels que $X_i^s = X_j^s$ pour $s = 1, \ldots, d$ et tels que les classes $y_i \neq y_j$. Est-il possible de créer un modèle de classification tel que le F1-score soit égal à 1 ? Justifiez votre réponse.

Non, il n'est pas possible de créer un modèle de classification avec un F1-score égal à 1 dans ce cas.

Si les enregistrements X_i et X_j sont identiques sur tous les attributs $s = 1, \ldots, d$, mais appartiennent à des classes différentes ($y_i \neq y_j$), cela signifie qu'un modèle de classification ne pourra pas les distinguer correctement. Le modèle fera inévitablement des erreurs, car il ne pourra pas différencier ces deux instances identiques qui appartiennent à des classes opposées.

Par conséquent, ceci entraînera une diminution de la précision (à cause d'un faux positif) ou du rappel (à cause des faux négatif).

Puisque pour avoir un F1-score égal à 1, il faut que la précision et le rappel soient toutes deux égales à 1, ce qui est impossible dans ce cas, le F1-score ne pourra pas être égal à 1.

6 Clustering (2 points)

(1 point) (a) Exécutez l'algorithme k-moyennes sur les points $X = \{(1,4),(1,3),(0,4),(5,1),(6,2),(4,0)\}$ pour $k = 2$ clusters. Utilisez les centroïdes initiaux $\mu_1 = (0,4)$ et $\mu_2 = (1,2)$. Détaillez tous les calculs jusqu'à la convergence de l'algorithme.

- Première itération:

  $C_1 = \{(1,4),(0,4)\}$ $\mu_1 = \{(0.5,4)\}$

  $C_2 = \{(1,3),(4,0),(5,1),(6,2)\}$ $\mu_2 = \{(4,1.5)\}$

- Deuxième itération:

  $C_1 = \{(1,4),(0,4),(1,3)\}$ $\mu_1 = \{(2/3,11/3)\}$

  $C_2 = \{(4,0),(5,1),(6,2)\}$ $\mu_2 = \{(5,1)\}$

(0.5 point) (b) Répétez l'exécution de l'algorithme k-moyennes mais à partir des centroïdes initiaux $\mu_1 = (0,4)$ et $\mu_2 = (15,0)$.

- Première itération:

  $C_1 = \{(1,4),(1,3),(0,4),(5,1),(6,2),(4,0)\}$ $\mu_1 = \{(17/6,8/3)\}$

  $C_2 = \emptyset$

(0.5 point) (c) Pourquoi les solutions obtenues en (a) et (b) ne sont-elles pas les mêmes ? L'algorithme des k-moyennes ne devrait-il pas toujours trouver la même solution après convergence ? Justifiez votre réponse.

Non. L'algorithme de k-moyennes est un algorithme de recherche local, sensible à initialisation.