ISEN ALL IS DIGITAL! OUEST | yncréa	**Projet électif** **CCSI3 – CIR3**

Projet de 30h : Analyse Exploratoire de Données par Statistiques Descriptives, Régression Linéaire Simple et Classification Ascendante Hiérarchique

Objectifs :

- Calculer et interpréter des mesures de statistiques descriptives pour comprendre un jeu de données bidimensionnelles.
- Appliquer le principe de la régression linéaire simple pour modéliser une éventuelle relation linéaire entre les coordonnées des points.
- Découvrir et appliquer l'algorithme de Classification Ascendante Hiérarchique (CAH) pour identifier des groupes homogènes dans un ensemble de données.
- Visualiser les résultats du clustering à l'aide d'un dendrogramme et interpréter les regroupements obtenus.
- Développer des compétences en programmation Python pour l'analyse de données.
- Travailler en groupe pour mener à bien un projet d'analyse de données et rédiger un rapport concis et précis.

Instructions :

1. **Formation des groupes** : Constituez des groupes de 7 étudiants maximum.
2. **Préparation d'un rapport :** Rédigez un rapport court et précis présentant votre démarche, votre code source en Python, vos résultats et leur interprétation. N'oubliez pas d'indiquer les noms de tous les membres du groupe sur la première page du rapport.
3. **Nommage du fichier :** Le fichier du rapport devra être nommé selon le format suivant : PE_site_NomDuGroupe.
4. **Algorithme sur papier :** Avant de commencer la programmation de la CAH, élaborez un algorithme sur papier pour les étapes clés de cet algorithme.

Données :

Nous considérerons l'ensemble de points bidimensionnels suivant :

$M_1(1, 1)$, $M_2(1, 2)$, $M_3(1, 5)$, $M_4(3, 4)$, $M_5(4, 3)$, $M_6(6, 2)$, $M_7(0, 4)$.

Dans l'ensemble des parties, vous expliquerez ou redémontrerez les formules données.

Vous ferez un programme en Python pour résoudre les différentes situations.

Vous contrôlerez vos résultats en utilisant les différentes bibliothèques disponibles.

Partie 1 : Statistiques Descriptives

1. **Calculs statistiques :** Pour l'ensemble des points donnés, calculez les statistiques descriptives suivantes pour les coordonnées x et y séparément :
    - Moyenne
    - Médiane
    - Variance
    - Écart-type
    - Minimum et Maximum
    - Étendue (différence entre le maximum et le minimum)
2. **Visualisation :**
    - Créez un nuage de points pour visualiser la distribution des données.
3. **Interprétation :**

    Interprétez brièvement les statistiques descriptives et la visualisation obtenue.

    Y a-t-il une indication visuelle d'une possible relation linéaire entre les coordonnées x et y ?

Partie 2 : Régression Linéaire Simple

Introduction

Nous allons explorer s'il existe une relation linéaire entre les coordonnées x et y des points. Considérons la coordonnée x comme la variable indépendante et la coordonnée y comme la variable dépendante.

1. Calcul des Coefficients de Régression

Les coefficients de la droite de régression linéaire simple $\hat{y} = b_0 + b_1 x$ sont calculés comme suit : $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ et $b_0 = \bar{y} - b_1\bar{x}$ où $\bar{x}$ et $\bar{y}$ sont les moyennes des coordonnées *x* et *y* respectivement, et *n* est le nombre total de points.

2. Visualisation de la Droite de Régression

Pour visualiser l'ajustement linéaire, on superpose la droite de régression $\hat{y} = b_0 + b_1 x$ au nuage de points. Quel est l'intérêt d'un tel ajustement ?

3. Coefficient de Détermination R^2

Le coefficient de détermination R^2 mesure la proportion de la variance totale de *y* expliquée par la régression : $R^2 = 1 - \frac{SCE}{SCT} = \frac{SCR}{SCT}$ où :

- $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la somme des carrés des erreurs,
- $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$ est la somme des carrés totaux,
- $SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ est la somme des carrés de la régression.

Partie 3 : Estimation de l'erreur en régression linéaire simple

Contexte

Dans le cadre de la régression linéaire simple, on modélise la relation entre une variable dépendante y et une variable explicative x selon le modèle : $y_i = b_0 + b_1 x_i + \varepsilon_i$ où :

- y_i est la valeur observée,
- x_i est la valeur explicative,
- b_0 et b_1 sont les coefficients estimés,
- ε_i est l'erreur aléatoire pour l'observation i.

1. Résidus et somme des carrés des erreurs (SCE)

Les résidus sont les écarts entre les valeurs observées et les valeurs prédites :

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

La somme des carrés des erreurs (SCE) est donnée par : $\text{SCE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

2. Estimation de la variance des erreurs : MSE

Une estimation non biaisée de la variance des erreurs aléatoires σ^2 est donnée par le MSE (Mean Squared Error) :

$$s^2 = \text{MSE} = \frac{\text{SCE}}{n-2}$$

Où n est le nombre total d'observations. Vous expliquerez le coefficient n-2.

3. Écart-type des erreurs

L'écart-type des erreurs est simplement la racine carrée de la variance :

$$s = \sqrt{s^2} = \sqrt{\frac{\text{SCE}}{n-2}}\,.$$

4. Interprétation des Résultats

Interpréter les coefficients b_0 et b_1 donnent respectivement l'ordonnée à l'origine et la pente de la droite de régression. La valeur de R^2 indique la qualité de l'ajustement :
- Si R^2 est proche de 1, le modèle explique bien les données.
- Si R^2 est faible, la relation linéaire est probablement insuffisante pour modéliser les données.

Le modèle linéaire est-il un bon ajustement pour ces données ? Justifiez votre réponse en justifiant les formules utilisées.

Partie 4 : Régression Linéaire Simple avec Tests Statistiques

Objectif

L'objectif de cette partie est d'examiner s'il existe une relation linéaire significative entre les coordonnées x et y de l'ensemble de nos points, et d'évaluer la force et la fiabilité de cette relation à l'aide de tests statistiques. Le modèle linéaire que nous allons ajuster est de la forme : $\hat{y}_i = b_0 + b_1 x_i$

où $\hat{y}_i$ est la valeur prédite de y, x_i est la coordonnée x.

1. Estimation de la Variance des Erreurs σ_ε^2

Pour estimer la variance des erreurs (résidus), on utilise : $s^2 = MSE = \frac{SCE}{n-2}$

où $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la somme des carrés des erreurs, et n-2 représente les degrés de liberté. L'écart-type des erreurs est : $s = \sqrt{MSE}$

2. Erreurs Standards des Coefficients

Les erreurs standards permettent de mesurer la variabilité des estimations des coefficients :

- Erreur standard de la pente : $SE_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

- Erreur standard de l'ordonnée à l'origine : $SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

3. Test d'Hypothèse pour la Pente (b_1)

Hypothèses :

$H_0 : b_1 = 0$ (pas de relation linéaire)

$H_1 : b_1 \neq 0$ (relation linéaire significative)

Statistique de test : $t = \frac{b_1 - 0}{SE_{b_1}}$

Vous expliquerez la loi suivie pat t. On la compare à une valeur critique ou à une p-valeur avec un niveau de signification $\alpha = 0,05$.

4. Test d'Hypothèse pour l'Ordonnée à l'Origine (b_0)

Hypothèses : $H_0 : b_0 = 0 \quad \text{vs} \quad H_1 : b_0 \neq 0$

Statistique de test : $t = \frac{b_0 - 0}{SE_{b_0}}$

5. Intervalles de Confiance pour les Coefficients

Un intervalle de confiance à $(1-\alpha)\times 100\%$ pour un coefficient β est :

$$\beta \pm t_{\alpha/2,n-2} \times SE_\beta$$

où $t_{\alpha/2,n-2}$ est la valeur critique de la loi t de Student avec $n-2$ degrés de liberté.

7. Interprétation des Tests Statistiques

Interprétez les résultats :
- Si la p-valeur est inférieure à α, on rejette H_0
- Si l'intervalle de confiance ne contient pas 0, cela confirme également la significativité du coefficient.

Question

Sur la base des tests statistiques effectués, pouvez-vous conclure qu'il existe une relation linéaire significative entre les coordonnées x et y ? Justifiez votre réponse à l'aide des p-valeurs et des intervalles de confiance après avoir expliqué leur construction.

Partie 5 : Classification Ascendante Hiérarchique (CAH)

Introduction à la CAH – Classification Ascendante Hiérarchique

La CAH (classification ascendante hiérarchique) est un algorithme de machine learning de la catégorie non supervisée. Elle permet d'identifier des groupes homogènes dans une population, on parle aussi de clustering.

La Classification Ascendante Hiérarchique est donc un algorithme de clustering. C'est à dire qu'à partir des données, l'algorithme va chercher à créer des groupes d'individus homogènes :

- Des groupes dans lesquels les individus se ressemblent
- Des groupes qui se distinguent le plus possible les uns des autres

La plupart des méthodes de clustering demandent à l'utilisateur de choisir le nombre de groupes qu'il souhaite créer. Ce n'est pas le cas de la CAH qui va calculer toutes les combinaisons possibles. Elle les représente ensuite via un dendrogramme qui permettra au Data Scientist de choisir le nombre de clusters le plus adapté à ses données et à son objectif.

La construction des groupes se fait étape par étape.
La CAH construit systématiquement un dendrogramme (une sorte d'arbre) qui va résumer tous les regroupements qui ont été faits. Ce dendrogramme est vraiment très utile et facilite l'utilisation de la CAH.
Pour décider si des individus ou des groupes sont proches ou éloignés, vous pourrez choisir votre distance. En particulier j'utilise souvent la distance de Ward pour les segmentations client. Elle permet d'éviter que des individus atypiques se retrouvent seuls isolés dans un groupe.

L'algorithme de la CAH

Etape 0 : Initialisation

On considère que chaque élément est seul et isolé dans son groupe.

Etape itérative

On fusionne les 2 groupes les plus proches et on les relie dans le dendrogramme. Le trait qui nous permet de relier les 2 groupes dans le dendrogramme est d'autant plus long que la distance entre les groupes est élevée.

Arrêt :

Lorsqu'il n'y a plus qu'un seul groupe

On se propose de réaliser une classification des 8 points suivants en utilisant la méthode d'agglomération au plus proche voisin : $M_1(1, 1)$, $M_2(1, 2)$, $M_3(1, 5)$, $M_4(3, 4)$, $M_5(4, 3)$, $M_6(6, 2)$, $M_7(0, 4)$.

Dans un premier temps, la distance utilisée sera la distance euclidienne.

Question 1

a. Ecrivez une fonction dist qui calcule la distance euclidienne séparant deux points.

b. Ecrivez une fonction dist1 qui calcule la distance séparant deux points par

$$dist_1(M_i, M_j) = |x_i - x_j| + |y_i - y_j|.$$

c. Ecrivez une fonction dist_inf qui calcule la distance séparant deux points par

$$dist_\infty(M_i, M_j) = Max\{|x_i - x_j|, |y_i - y_j|\}.$$

d. Expliquer la distance de Ward

Question 2

Ecrivez une fonction dist_min qui prend pour argument un tableau de points t et retourne un couple de points *((x, y), (x_0, y_0))* situés à une distance minimale l'un de l'autre.

Question 3

Placer les points dans un repère orthonormé et remplir une matrice avec le carré de la distance euclidienne des points tracés. Regrouper sur le dessin, en les entourant d'une courbe, les deux points les plus proches pour former une classe Γ_1.

Question 4

Remplir une deuxième matrice de distances en calculant les distances au plus proche voisin de la classe Γ_1 avec les 5 points restants. Entourer la nouvelle classe Γ_2.
On aura au préalable écrit la distance d'un ensemble à un point.

Remarque : La classe obtenue n'est pas nécessairement unique.

Question 5

Poursuivre ainsi la classification jusqu'à ce que tous les points soient en une seule classe.
On détaillera étape par étape les classes obtenues.

Question6

Essayer de résoudre les questions 3 à 5 à la machine.

Un dendrogramme est un diagramme fréquemment utilisé pour illustrer l'arrangement de groupes générés par un regroupement hiérarchique ou hiérarchisant.

Les dendrogrammes sont par exemple souvent utilisés en biologie pour illustrer des regroupements de gènes, ou des filiations (arbre phylogénétique), mais aussi dans de nombreux autres domaines utilisant des notions de regroupement hiérarchique ou de coalescence, de l'arbre généalogique aux logiciels de fouille d'images.

Les dendrogrammes sont utilisés dans le domaine de l'ergonomie web avec la méthode de tri par cartes pour optimiser l'architecture de l'information d'un site web.

Question 7

Tracer un dendrogramme résumant cette classification.
Expliquer le trait en pointillé permettant de couper ce dendrogramme.
Où faut-il le couper ? Quelles sont les classes ainsi obtenues ?

Question 8

Ce que vous avez réalisé est une classification ascendante hiérarchique utilisée en big data. C'est un algorithme de machine learning permettant d'identifier des groupes homogènes.

On se propose d'utiliser un dendrogramme pour représenter le tableau suivant constitué de 29 individus et en dimension 9.

L'objectif est d'identifier des groupes d'individus homogènes, partageant des caractéristiques similaires.

Voici une rédaction complète pour étoffer votre sujet, intégrant la partie sur l'évaluation et la validation du clustering, ainsi que l'interprétation et l'analyse critique des résultats.

Partie 6- Évaluation et validation du clustering

1. Indices d'évaluation

Pour juger de la qualité des regroupements obtenus par la Classification Ascendante Hiérarchique (CAH), il est essentiel de recourir à des mesures quantitatives. Parmi les plus couramment utilisées, on trouve :

L'indice de silhouette (Silhouette Score) : Cet indice mesure la cohésion interne (la proximité entre les points d'un même cluster) et la séparation entre différents clusters. Pour chaque point, il calcule la différence entre la distance moyenne avec les autres points du même cluster et la distance moyenne avec le cluster le plus proche. La valeur du score varie entre -1 et 1, une valeur proche de 1 indiquant une bonne séparation, et une valeur proche de -1 indiquant un mauvais regroupement.

L'indice de Dunn : Il évalue la densité et la séparation des clusters. Plus cet indice est élevé, meilleurs sont la compacité et la séparation des groupes.

Cohésion et séparation : Ces mesures internes permettent de quantifier la densité des clusters (cohésion) et leur éloignement mutuel (séparation). Une bonne partition doit allier une forte cohésion et une bonne séparation.

2. Validation croisée

Pour vérifier la stabilité et la robustesse des résultats obtenus par la CAH, il est recommandé de :

Comparer avec d'autres méthodes de clustering : Par exemple, appliquer k-means, DBSCAN ou d'autres algorithmes et comparer les regroupements. Si les résultats convergent, cela renforce leur crédibilité.

Validation externe : Si des étiquettes ou classifications de référence existent, utiliser des indices comme le Rand index ou l'Adjusted Rand Index pour mesurer la concordance entre la classification hiérarchique et la vérité terrain.

3. Interprétation des résultats

Une fois la classification réalisée, il est crucial d'analyser chaque groupe pour en dégager les caractéristiques principales :

Calculer les statistiques descriptives (moyennes, médianes, écart-types) pour chaque variable dans chaque classe.

Identifier des « profils types » ou « segments de population » à partir des variables clés.

Discuter de l'intérêt pratique de ces groupes : par exemple, dans un contexte marketing, cibler des segments homogènes ; en médecine, détecter des profils de patients avec des caractéristiques communes.

4. Analyse critique de la méthode

Dans cette partie vous allez expliquer les limites du critère de distance.

Que peut-il se passer lorsque l'on utilise la méthode de linkage et surtout vous expliquerez la s notions de robustesse et de sensibilité. Est-il préférable de modifier les données ?

Si cette méthode n'est pas satisfaisante, pouvez-vous en proposer une autre ?

5.Visualisation avancée

Pour mieux comprendre la structure des données et des clusters, il est utile d'utiliser des techniques de projection en 2D ou 3D :

Analyse en composantes principales (ACP) : Réduit la dimension tout en conservant la majorité de la variance.

t-SNE : Permet une visualisation en deux ou trois dimensions en conservant la structure locale.

Heatmaps des distances : Visualisent la matrice de distances, indiquant la proximité relative des points.

A réaliser sur l'ensemble des données.

Résumé

L'évaluation et l'interprétation des résultats issus d'une CAH sont essentielles pour assurer leur pertinence dans un contexte pratique. En combinant des indices quantitatifs, une analyse qualitative, et une critique des limites, on obtient une vision équilibrée de la qualité du clustering. Cela permet d'éviter les conclusions hâtives et d'orienter la prise de décision vers des solutions robustes et compréhensibles.

Ressources suggérées :

- Documentation des librairies Python numpy, pandas, matplotlib, et scipy.
- Supports de cours sur les statistiques descriptives, la régression linéaire simple et le clustering.

- Recherches en ligne sur la Classification Ascendante Hiérarchique et les différentes métriques de distance.

Évaluation :

L'évaluation de ce projet portera sur :

- La qualité du code Python et son adéquation aux questions posées.
- La pertinence et l'exactitude des calculs et des interprétations statistiques et de régression.
- La clarté et la précision des explications et des interprétations concernant la CAH.
- La compréhension des concepts de statistiques descriptives, de régression linéaire simple et de clustering hiérarchique.
- La rigueur de la démarche suivie et la présentation des résultats.
- La capacité à travailler en groupe et à produire un rapport cohérent.

Bon travail !

	Données 1	Données 2	Données 3	Données 4	Données 5	Données 6	Données 7	Données 8	Données 9
Individu 01	70,00	91,00	215,70	3,40	42,90	2,90	4,10	13,00	14,00
Individu 02	321,00	140,00	218,00	29,30	49,20	3,70	17,60	80,00	30,00
Individu 03	321,00	252,00	125,50	27,30	62,30	6,20	21,80	80,00	20,00
Individu 04	298,00	205,00	261,00	23,30	60,40	6,70	23,30	70,00	26,00
Individu 05	370,00	432,00	162,00	31,20	83,50	13,30	18,70	100,00	25,00
Individu 06	309,00	272,00	202,30	24,60	73,10	8,10	19,70	80,00	30,00
Individu 07	355,00	232,00	178,90	28,00	51,50	6,80	22,40	90,00	25,00
Individu 08	300,00	223,00	156,70	23,40	53,00	4,00	21,10	70,00	22,00
Individu 09	142,00	22,00	78,20	10,40	63,40	20,40	9,40	20,00	10,00
Individu 10	381,00	240,00	334,60	27,50	90,00	5,20	35,70	80,00	46,00
Individu 11	347,00	285,00	219,00	29,50	57,60	5,80	23,60	80,00	30,00
Individu 12	338,00	311,00	236,70	29,10	46,70	3,60	20,40	90,00	40,00
Individu 13	115,00	25,00	94,80	7,80	64,30	22,60	7,00	30,00	10,00
Individu 14	80,00	41,00	146,30	3,50	50,00	20,00	8,30	10,00	11,00
Individu 15	292,00	390,00	168,50	24,00	77,40	5,50	16,80	70,00	20,00
Individu 16	206,00	160,00	72,80	18,50	150,50	31,00	11,10	50,00	16,00
Individu 17	378,00	60,00	308,20	29,40	56,30	2,40	29,40	110,00	45,00
Individu 18	327,00	148,00	272,20	24,70	65,70	5,50	24,70	80,00	44,00
Individu 19	308,00	222,00	79,20	25,60	63,60	21,10	20,50	80,00	13,00
Individu 20	399,00	92,00	220,50	32,40	55,90	1,30	29,20	120,00	51,00
Individu 21	406,00	172,00	182,30	32,50	76,40	4,90	26,00	110,00	28,00
Individu 22	292,00	276,00	132,90	25,40	116,40	32,50	17,80	70,00	25,00
Individu 23	344,00	192,00	87,20	27,90	90,10	36,30	19,50	80,00	36,00
Individu 24	367,00	256,00	264,00	28,80	48,80	5,70	23,00	90,00	30,00
Individu 25	264,00	314,00	215,90	19,50	103,00	36,40	23,40	60,00	20,00
Individu 26	342,00	336,00	211,10	28,90	37,10	27,50	20,20	90,00	27,00
Individu 27	401,00	112,00	259,40	33,30	54,90	1,20	26,60	120,00	41,00
Individu 28	314,00	238,00	209,80	25,10	63,70	6,40	22,60	70,00	27,00
Individu 29	314,00	353,50	72,60	26,30	51,60	30,30	21,00	70,00	20,00