[redacted]

Exercice 2 : Arbre de Décision [5pts]

1. Donner un **pseudocode** du fonctionnement de l'algorithme des arbres de décision. **[2pts]**

***Procédure** construire-arbre(X)*
***SI** tous les individus I appartiennent à la même modalité de la variable décisionnelle*
        ***ALORS** créer un nœud feuille portant le nom de cette classe : Décision*
SINON
        choisir le meilleur attribut pour créer un nœud // l'attribut qui sépare le mieux, le test associé à ce nœud sépare X en des branches
        construire-arbre(Xd), …,construire-arbre(Xg)
FIN

2. En déduire la variable la **plus décisive** par rapport à l'appartenance d'un individu à l'origine orientale. (Donner le calcul complet et la formule utilisée) **[3pts=2pts(formule)+1pt(calcul)]**

Indice de Gini :
1-somme(fréquence de la classe décisionnelle dans le nœud)

- ✓ ***11 individus : Oriental = Oui : 6, Oriental = Non : 5 donc***
  - ***Indice de Gini avant séparation au NIVEAU DE LA RACINE :***

IG(Oriental) = 1-((5/11)²+(6/11)²) = 0.4958678

- ***Indice de Gini de la variable Yeux :***

4 Noir : 4 Oui, 0 Non IG(Y=Noir) = 1-((4/4)²+(0/4)²) = 0
3 Brun : 2 Oui, 1 Non IG(Y=Brun) = 1-((2/3)²+(1/3)²) = 0.
4 Bleu : 0 Oui, 4 Non IG(Y=Bleu) = 1-((0/4)²+(4/4)²) =0

- ***Indice de Gini de la variable Cheveux :***

4 Noir : 3 Oui, 1 Non IG(Ch=Noir) = 1-((3/4)²+(1/4)²) = 0.
4 Blanc : 3 Oui, 1 Non IG(Ch=Blanc) = 1-((3/4)²+(1/4)²) = 0.
3 Blond : 0 Oui, 3 Non IG(Ch=Blond) = 1-((0/3)²+(3/3)²) = 0

- ***Indice de Gini de la variable Taille :***

6 Petit : 3 Oui, 3 Non IG(T=Petit)=1-((3/6)²+(3/6)²) = 0.5
5 Grand : 3 Oui, 2 Non IG(T=Grand)= 1-((3/5)²+(2/5)²) = 0.

- ✓ ***La variable la plus décisive est celle qui maximise IG(avant séparation)-[IG(fils1)+……+IG(filsn)], donc la couleur des Yeux est la variable la plus décisive par rapport à l'appartenance d'un individu à l'origine orientale***

	Yeux	Cheveux	Taille	Oriental
1	Noir	Noir	Petit	Oui
2	Noir	Blanc	Grand	Oui
3	Noir	Blanc	Petit	Oui
4	Noir	Noir	Grand	Oui
5	Brun	Noir	Grand	Oui
6	Brun	Blanc	Petit	Oui
7	Bleu	Blond	Grand	Non
8	Bleu	Blond	Petit	Non
9	Bleu	Blanc	Grand	Non
10	Bleu	Noir	Petit	Non
11	Brun	Blond	Petit	Non

Exercice 3 : Régression Linéaire Multiple [5pts]

On propose de construire le meilleur modèle permettant de prédire une variable Y via la régression linéaire multiple.
Les variables explicatives sont : X_1, X_2, X_3 et X_4.

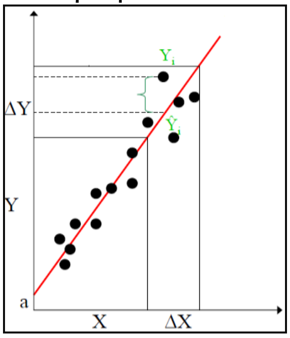
Les résultats des modèles global et réduit sont affichés ci-dessous.

[redacted]

Modèle Global	*Coefficients*
Constante	10
X_1	1
X_2	1.5
X_3	2.5
X_4	5

Modèle Réduit	*Coefficients*
Constante	15
X_3	3
X_4	1

1. Expliquer la méthode des **moindres carrées** (possibilité d'utiliser un schéma). **[1pt]**



$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Cette méthode consiste à chercher des coefficients prédicteurs qui représentent la droite qui minimise la somme au carré des distances entre $Y_{réelle}$ et $Y_{prédite}$

2. Décrire l'**utilité** de la régression linéaire. **[2pts]**

La régression linéaire permet :
- de prédire une variable décisionnelle quantitative via un modèle global ou un modèle avec sélection de variables pertinentes (modèle réduit via le critère AIC)
- d'avoir une vision sur l'ordre de pertinence des variables prédictives à l'aide des valeurs des coefficients prédicteurs
- de détecter les anomalies, les individus atypiques ou aberrants

3. Donner la **formule** du modèle global et celle du modèle réduit. **[1pt]**

$$Y_{global} = 10 + X_1 + 1.5\, X_2 + 2.5\, X_3 + 5\, X_4$$
$$Y_{réduit} = 15 + 3\, X_3 + X_4$$

4. Calculer les **prédictions** d'un individu I(X_1=1, X_2=1, X_3=1, X_4=2). **[0.5pt]**

$$Y_{global} = 10 + 1 + 1.5 *1 + 2.5 *1 + 5 *2 = 25$$
$$Y_{réduit} = 15 + 3 *1 + 2 = 20$$

5. Sachant que la valeur réelle de la variable décisionnelle de l'individu I est $\mathbf{Y_{réelle} = 20}$, Comparer les deux prédictions obtenues et **Conclure**. **[0.5pt]**

$$Y_{réduit} = Y_{réelle}$$

Donc on peut constater que prédire Y à partir des valeurs de X_3 et X_4 est mieux que prédire à l'aide de toutes les variables prédictives // le modèle réduit mieux que le modèle global

[redacted]

Exercice 4 : Segmentation avec K-Means [5pts]

Un opérateur téléphonique souhaite analyser les données de ces clients afin d'identifier ceux qui sont susceptibles de changer d'opérateur (les clients susceptibles de *churner*). Les données disponibles sont composées des variables quantitatives *Age_C* : les âges des clients, *Durees_A* : les durées d'appels par jour, *Nbre_A* : les nombres d'appels par jour, *Nbre_SMS* : les nombres des SMS envoyés par jour et une variable qualitative *Churn_C* qui prend la valeur 1 si le client a déjà churné et la valeur 0 si le client est encore fidèle à son opérateur téléphonique. L'échantillon étudié est composé de 87 clients de la classe 0 et 56 clients de la classe 1.

On propose tout d'abords d'appliquer une méthode descriptive de groupage (Clustering) des clients via la méthode K-means en utilisant les variables **quantitatives** disponibles.

1. Décrire les étapes de l'algorithme de la méthode **K-means**. **[2pts]**

Algorithme K-moyennes

Entrée : k le nombre de groupes cherché

Début

- **Choisir aléatoirement les centres des groupes**

Répéter

- **Affecter chaque cas au groupe dont il est le plus proche au son centre (utiliser une distance adéquate)**
- **Recalculer le centre de chaque groupe**
- **jusqu'à ce que (stabilisation des centres) ou (nombre d'itérations =t) ou (stabilisation de l'inertie totale de la population)**

Fin

2. Citer deux **inconvénients** de la méthode K-means. **[1pt]**

- **Le choix de k est subjectif dans le cas où le nombre de classes est inconnu au sein de l'échantillon.**
- **Les résultats de l'algorithme du k-means sont sensibles à l'initialisation aléatoires des centres.**

3. L'application de la méthode K-means, en fixant le nombre de groupe **K=2**, génère **la matrice de confusion** suivante en croisant la variable *Churn_C* avec les résultats de la classification de K-means:

	Groupe 1	Groupe 2
Classe 0	7	80
Classe 1	52	4

- Calculer les taux de **bonne classification** de chaque classe de la variable *Churn_C* et le taux de bonne classification **total**. **[1pt]**

TBC_G1=((52)/(56))*100= 92.85%; TBC_G2=((80)/(87))*100= 91,85%; TBC_total=((52+80)/(56+87))*100=92,3%

- Est-ce que la méthode K-means génère une bonne classification des clients ? Justifier votre réponse. **[1pt]**

Le groupe 1 est bien représenté par la classe 1 et le groupe 2 est bien représenté par la classe 2. De plus, les taux de bonnes classifications de chaque groupe et le taux de bonne classification totale dépassent 90%. Ainsi, la méthode K-means génère une bonne classification des clients.