3. Analyse des résultats de l'ACP

3.1. Sélectionner les individus (les années) qui sont bien représentés sur le plan factoriel en analysant les qualités de leurs représentations (cos2) dans le tableau 2. [1pt]

Cos2 >= 0.8 : 1969, 1970, 1971, 1972, 1976, 1977, 1982

3.2. Sélectionner les variables corrélées avec les premières composantes principales à partir du tableau 3. [1pt]

IMM et DCT, EXP et NET, INT et SUB

3.3. Commenter les positions des années bien représentées sur le plan factoriel par rapport aux variables corrélées avec les deux premières composantes principales. [1pt]

Les années (individus) 1969, 1970, 1971 et 1972 sont corrélées positivement avec NET, IMM et INT. 1982 est corrélée positivement avec EXP et DCT. 1976 et 1977 n'ont pas de corrélations.

Exercice 2 : Clustering – Kmeans et CAH [3pts]

On souhaite découper un échantillon de patientes atteintes de la maladie de l'ostéoporose en groupes homogènes.

1. Un premier médecin propose de construire deux groupes de patientes via la méthode K-means : patientes atteintes de fractures de la hanche et patientes atteintes de tassements vertébraux. Afin de quantifier la qualité du découpage, le médecin propose de croiser le résultat de l'algorithme K-means sur un échantillon de patientes caractérisé par une variable qui identifie les patientes réellement fracturées à la hanche par les caractères FH et les patientes atteintes de tassements vertébraux par le caractère TV.
Le résultat du croisement génère la table de confusion suivante :

	1	2
FH	4	54
TV	72	7

En se basant sur les résultats de la table de confusion, quantifier la qualité du découpage obtenu. [1pt]

Individus bien classés : 72/76+54/61 = 83.3%

2. Un deuxième médecin propose d'appliquer la classification hiérarchique ascendante afin de classifier les patientes. Après la construction du dendrogramme, le médecin propose de faire le découpage de son dendrogramme au niveau de la plus forte perte d'inertie interclasses.

2.1. Citer l'avantage de la classification hiérarchique ascendante par rapport à la méthode K-means dans le choix du nombre de classes à construire. [1pt]

L'algorithme de la CAH n'est pas basé sur un choix aléatoire comme dans le cas de l'algorithme du K-means et il permet d'obtenir toutes les classifications possibles à partir de la notion du dendrogramme.

2.2. Le découpage du dendrogramme au niveau de la plus forte perte d'inertie interclasses génère deux groupes de patientes.

En croisant le résultat de classification avec la variable de l'échantillon d'apprentissage qui identifie les patientes réellement fracturées à la hanche et les patientes atteintes de tassements vertébraux, on obtient la table de confusion suivante :

	1	2
FH	56	2
TV	9	70

Quantifier la qualité du découpage obtenu et comparer les résultats de la classification hiérarchique ascendante

aux résultats de la méthode K-means. Conclure. [1pt]

Individus mal classés : 56/65+70/72 = 83.3 %

On obtient le même pourcentage de classification pour les deux méthodes.

Exercice 3 : Réseau de Neurones – Perceptron Simple [5pts]

1. Donner un pseudocode du fonctionnement de l'algorithme du réseau de neurones modèle Perceptron Simple. [2pts]

i. **On note S la base d 'apprentissage.**
ii. **S est composée de couples (x, c) où :**

x est le vecteur associé à l 'entrée (x0, x1, … , xn)

c la sortie correspondante souhaitée

// On cherche à déterminer les coefficients (w0, w1, … , wn).

iii. **Initialiser aléatoirement les coefficients wi.**

Répéter

Prendre un exemple (x, c) dans S

Calculer la sortie o du réseau pour l 'entrée x

Mettre à jour les poids :

Pour i de 0 à n

wi = wi + ε * (c – o) * xi

Fin Pour

Fin Répéter

2. On considère les données suivantes et on désire déterminer les coefficients synaptiques obtenus après exactement deux itérations dans le but d'effectuer l'apprentissage de la fonction ET logique. [3pts]

X1	X2	Y
0	0	0
0	1	0
1	0	0
1	1	1

On donne les valeurs initiales des poids :

w0 = 0,1 ; w1 = 0,2 ; w2 = 0,05.

1ère itération :

Observation à traiter

$$\begin{cases} x_0 = 1 \\ x_1 = 0 \\ x_2 = 0 \\ y = 0 \end{cases}$$

⇨

Appliquer le modèle

$$0.1\times1+0.2\times0+0.05\times0=0.1$$
$$\Rightarrow \hat{y}=1$$

⇨

Màj des poids

$$\begin{cases} \Delta a_0 = 0.1\times(-1)\times1 = -0.1 \\ \Delta a_1 = 0.1\times(-1)\times0 = 0 \\ \Delta a_2 = 0.1\times(-1)\times0 = 0 \end{cases}$$

2ème itération :

Observation à traiter

$$\begin{cases} x_0 = 1 \\ x_1 = 1 \\ x_2 = 0 \\ y = 0 \end{cases}$$

$\Rightarrow$

Appliquer le modèle

$$0.0\times1+0.2\times1+0.05\times0=0.2$$
$$\Rightarrow \hat{y}=1$$

$\Rightarrow$

Màj des poids

$$\begin{cases} \Delta a_0 = 0.1\times(-1)\times1 = -0.1 \\ \Delta a_1 = 0.1\times(-1)\times1 = -0.1 \\ \Delta a_2 = 0.1\times(-1)\times0 = 0 \end{cases}$$

➔ **Les poids seront : w0=-0,1, w1=-0,1 et w2=0 après deux itérations**

Exercice 4 : Régression Linéaire Multiple [6pts]

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	t	Sig.
1	(Constant)	3129.231	641.355	4.879	.000
	MT	4.423	1.588	2.785	.009
	RG	1.676	3.291	.509	.614
	PRIX	-13.526	8.305	-1.629	.114
	BR	-3.410	6.569	-.519	.608
	INV	1.924	.778	2.474	.019
	PUB	8.547	1.826	4.679	.000
	FV	1.497	2.771	.540	.593
	TPUB	-2.15E-02	.401	-.054	.958

a. Dependent Variable: VENTES

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.898[a]	.806	.752	256.29

On propose de construire un modèle permettant de prédire les ventes semestrielles d'un produit agroalimentaire via la régression linéaire multiple.
Les variables explicatives sont : MT : Besoin du marché, RG : Remises aux grossistes, PRIX : Prix du produit, BR : Budget de Recherche, INV : Investissement, PUB : Publicité, FV = Frais de ventes, TPUB = Total budget publicité de la branche.

Les résultats du modèle sont affichés ci-dessus.

1. En étudiant individuellement les coefficients liés à chaque variable explicative, déterminer les variables significativement pertinentes dans l'augmentation des ventes semestrielles du produit. [1pt]
PUB, MT ont les plus faibles valeurs significatives donc elles sont les plus pertinentes.

2. En se basant sur les résultats des coefficients de détermination, conclure sur la qualité du modèle construit. [1pt]
R^2 = 0.806 donc il s'agit d'un bon modèle.

3. On propose de sélectionner les variables pertinentes du modèle en utilisant la méthode de sélection pas à pas descendante (Backward). Expliquer brièvement les étapes de cette méthode. [2pts]
AIC, le critère de choix du modèle le plus performant : itération par itération on élimine la variable la moins pertinente.

4. Donner l'équation du modèle permettant de prédire les ventes semestrielles en fonction de toutes les variables explicatives. [2pts]
VENTES = 3129,231 + 4,423 MT + 1,676 RG - 13,526 PRIX – 3,410 BR +1,924 INV + 8,328 PUB + 1,497 FV – 0,00215 TPUB