socrative

Name _____
Date _____

MIAGE AD - CC 08/02/2021

Score _____

1. Cochez les égalités (toujours) justes, supposant A, B et C matrices réelles de tailles compatibles, et toutes les opérations valides.

(A) $A(B + C) = AB + AC$

(B) $(A + B)^{-1} = A^{-1} + B^{-1}$

(C) $({}^tA)^{-1} = {}^t(A^{-1})$

(D) $AB = AC$ implique $B = C$

2.
$$M = \begin{pmatrix} 5 & -3 & 0 \\ 6 & -4 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Lequels sont des vecteurs propres de M parmi les choix suivants ?

(A) (1, -1, -1)

(B) (1, 1, 2)

(C) (1, 2, 1)

(D) (1, 2, -1)

3. En ACP (Analyse en Composantes Principales) :

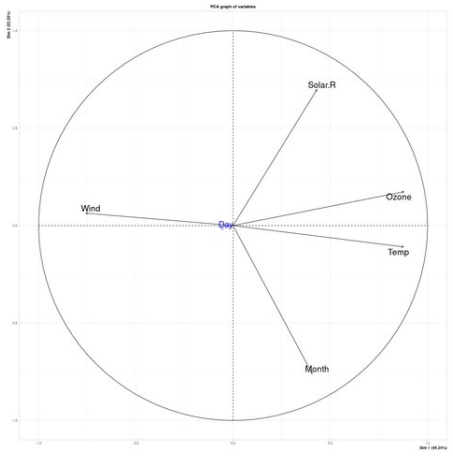
(A) L'inertie du nuage des individus est égale à celle du nuage des variables.

(B) On centre toujours les variables.

(C) La contribution d'un individu à la construction d'un axe est d'autant plus grande que sa coordonnée sur cet axe est élevée.

(D) Les variables supplémentaires ne peuvent pas être qualitatives.

4. On réalise l'ACP d'un jeu de données sur la qualité de l'air, dont le cercle des corrélation est indiqué à droite (plan 1-2).



Temp = température
Ozone = niveau d'ozone
Wind = vitesse du vent
Solar.R = rayonnement solaire
Month = mois dans l'année (1 à 12)
Day = jour dans le mois (1 à 31) → variable supplémentaire

Les inerties projetées sur les deux premiers axes valent respectivement environ 50% et 22% de l'inertie totale.

(A) Le vent souffle en général fort quand il fait chaud.

(B) Le niveau d'ozone a tendance à être bas les jours froids.

(C) Le niveau d'ozone est souvent plus bas en fin de mois.

(D) Le rayonnement solaire est nettement plus intense les jours chauds.

5. Cochez les affirmations justes concernant l'AFC - Analyse Factorielle des Correspondances.

(A) Elle a pour but de modéliser l'écart à l'indépendance entre deux variables.

(B) La "marge ligne" est la probabilité marginale calculée ligne par ligne.

(C) La "marge colonne" est la probabilité marginale calculée colonne par colonne.

(D) On peut l'appliquer à un couple de variables respectivement quantitative et qualitative (éventuellement après transformation).

6. Considérant le tableau croisé des variables en entrée de l'AFC après normalisation :

(A) si deux lignes sont proportionnelles, alors elles ont exactement les mêmes coordonnées.

(B) si une colonne et une ligne prennent les mêmes valeurs, alors elles ont exactement les mêmes coordonnées.

(C) la somme sur chaque ligne ou colonne vaut 1.

(D) une ligne qui est proche du profil moyen sera nécessairement proche du centre de gravité et donc de l'origine des axes.

7. Considérant le tableau croisé des variables en entrée de l'AFC, il y a indépendance si

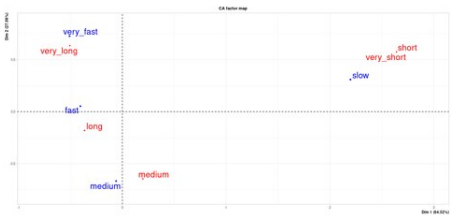
(A) le tableau est constant (même valeur dans toutes les cases).

(B) la i-eme ligne est égale à la i-eme colonne.

(C) toutes les lignes sont proportionnelles.

(D) toutes les colonnes sont proportionnelles.

8. On réalise l'AFC d'un jeu de données sur la la distance de freinage de véhicules en fonction de la vitesse. Voir le plan 1-2 sur la figure à droite. Les inerties projetées sur les deux premiers axes valent respectivement environ 50% et 22% de l'inertie totale.



Deux variables à respectivement 4 et 5 modalités :
vitesse : fast, medium, slow, very_slow (du plus rapide au plus lent).
distance : very_long, long, medium, short, very_short (de la plus longue à la plus courte).

Tableau des effectifs :

vitesse (distance)	long	medium	short	very_long	very_short
fast	8	2	0	5	0
medium	5	10	0	2	0
slow	0	2	2	0	2
very_fast	2	1	0	9	0

On donne également les valeurs propres associés aux axes :

```
> r$eig[,1]
   dim 1    dim 2    dim 3
    0.69     0.29     0.09
```

(A) La première valeur propre est assez proche de 1, indiquant une nette corrélation.

(B) La première valeur propre n'est pas assez proche de 1 pour suggérer une dépendance.

(C) Les troisième et cinquième colonnes sont identiques : c'est pour cela que les représentations de "short" et "very_short" sont confondues.

(D) "very_long" et "very_fast" sont très proches sur le graphe et assez loin de l'origine, donc les véhicules roulant très vite ont besoin d'une grande distance pour s'arrêter.

9. Cochez les affirmations justes concernant l'ACM - Analyse des Correspondances Multiples.

(A) Elle nécessite en entrée un tableau transformé depuis celui des effectifs par variable et par modalité : le tableau disjonctif complet, ou bien le tableau de Burt.

(B) Elle revient à effectuer une ACP sur le tableau disjonctif complet transformé (normalisé).

(C) L'inertie moyenne sur un axe valant 1/J, on ignore en principe dans l'analyse les dimensions d'inertie inférieure à 1/J.

(D) L'inertie totale dépend du nombre d'individus.

10.

	CSP	CSP père	CSP mère
a	Ouvrier	POuvrier	MOuvrière
b	Employé	POuvrier	MEmployée
c	Employé	POuvrier	MEmployée
d	Cadre	PEmployé	MEmployée
e	Cadre	PEmployé	MEmployée
f	Employé	PCadre	MEmployée
g	Cadre	PCadre	MCadre
h	Cadre	PCadre	MCadre

On a interrogé 8 personnes en leur demandant d'indiquer leur catégorie socio-professionnelle (CSP) ainsi que celle de leur père et de leur mère. On a obtenu les données suivantes (image à droite). On se place dans le cadre de l'ACM.

(A) L'inertie totale du nuage des individus vaut 3.

(B) L'inertie totale du nuage des individus vaut 4.

(C) Dans l'espace des modalités, Ouvrier (ou MOuvrière) présente l'inertie la plus faible.

(D) Dans l'espace des modalités, MEmployée présente l'inertie la plus faible.

11. Arbre de décision : cochez les affirmations justes.

(A) L'algorithme construit un arbre où chaque noeud divise les données selon les valeurs prises en plusieurs variables.

(B) L'algorithme construit un arbre où chaque noeud divise les données selon la valeur prise en une variable.

(C) Les feuilles de l'arbre non élagué ne contiennent chacune qu'un seul individu.

(D) Pour déterminer la classe d'un nouveau vecteur en entrée, on descend dans l'arbre selon le chemin donné par les résultats des tests successifs à chaque noeud, puis on prédit la classe majoritaire à cette feuille.

12. Arbre de décision : cochez les affirmations justes.

(A) Il faut en général élaguer l'arbre obtenu, afin de réduire l'erreur de généralisation (calculée sur un ensemble de test).

(B) Il vaut mieux garder l'arbre non élagué, puisqu'il obtient de meilleures performances sur l'ensemble d'apprentissage.

(C) Un arbre de décision est forcément binaire.

(D) Un arbre permet de traiter des variables continues et symboliques.

13.

Jour	Météo	Je viens de manger	En retard au boulot	Vais-je courir ?
1	Soleil	Oui	Non	Oui
2	Pluie	Oui	Oui	Non
3	Soleil	Non	Oui	Oui
4	Pluie	Non	Non	Non
5	Pluie	Non	Non	Oui
6	Soleil	Oui	Non	Oui
7	Pluie	Non	Oui	Non

Cochez les affirmations justes concernant le mini-jeu de données ci-contre, où quelqu'un se demande s'il doit aller courir.

On rappelle le calcul de l'indice de Gini I_G pour une partition P en P_1, P_2, ..., P_K :

$I_G(P) = 1 - \text{somme}[1 \to K] \text{ des } a_i^2$

avec a_i = taille du groupe P_i divisée par l'effectif total.

Le gain d'information en divisant selon une variable qualitative (en K groupes P_1, ..., P_K) s'obtient alors comme suit :

$G = I_G(P) - \text{somme}[1 \to K] \text{ des } a_i \, I_G(P_i)$

(A) Le gain d'information maximal vaut environ 0.54, obtenu pour la variable "Je viens de manger"

(B) Le gain d'information maximal vaut environ 0.67, obtenu pour la variable "En retard au boulot"

(C) Le gain d'information maximal vaut environ 0.28, obtenu pour la variable "Météo"

(D) Le gain d'information maximal vaut environ 0.48, obtenu pour la variable "En retard au boulot"

14.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica

Cochez les propositions menant à un gain d'information maximal concernant l'extrait du jeu de données Iris ci-contre.

On rappelle le calcul de l'indice basé sur l'entropie I_E pour une partition P en P_1, P_2, ..., P_K :

I_E = - somme[1 → K] des $a_i \log(a_i)$

avec a_i = taille du groupe P_i divisée par l'effectif total.

Le gain d'information en divisant selon une variable quantitative (en 2 groupes P_1 et P_2) s'obtient comme suit :

$G = I_E(P) - a_1 I_E(P_1) - a_2 I_E(P_2)$

(A) Découpage selon la variable Petal.Length = 4

(B) Découpage selon la variable Petal.Length = 5

(C) Découpage selon la variable Petal.Width = 0.7

(D) Découpage selon la variable Petal.Width = 1.7

15. L'objectif du clustering appliqué à un jeu de données est de

(A) Diviser les données en groupes d'individus similaires, et idéalement très différents d'un groupe à l'autre.

(B) Trouver K individus représentatifs de la répartition des données.

(C) Déterminer les plus proches voisins de chacun des individus.

(D) Prédire la classe d'un nouvel individu.

16. L'algorithme des k-means

(A) converge toujours vers une configuration maximisant globalement l'inertie inter-classes (il n'en existe pas de meilleure).

(B) converge vers une configuration maximisant localement l'inertie inter-classes (il en existe peut-être une meilleure, mais elle serait très différente).

(C) est sensible aux données aberrantes ("outliers").

(D) utilise des centres de clusters n'appartenant pas nécessairement au jeu de données.

17. Un algorithme des k-médoïdes

(A) utilise des centres de clusters appartenant au jeu de données.

(B) prend en entrée une matrice de dissimilarités (qui peuvent être des distances, mais pas forcément).

(C) est plus rapide que l'algorithme des k-means car les distances sont toutes précalculées.

(D) cherche tout comme l'algorithme des k-means à minimiser l'inertie intra-classes.

18. Qu'est-ce qu'un dendrogramme ?

(A) Un diagramme arborescent utilisé afin d'illustrer l'agencement des clusters dans un clustering de type k-means.

(B) Un diagramme arborescent utilisé afin d'illustrer l'agencement des clusters dans un clustering hiérarchique.

(C) Un diagramme en barres permettant de visualiser les clusters obtenus par l'algorithme des k-means.

(D) Un algorithme de clustering hiérarchique.

19. Dans l'algorithme de classification ascendante hiérarchique (CAH), comment choisit-on les deux groupes à fusionner (passant de k à k-1 clusters) ?

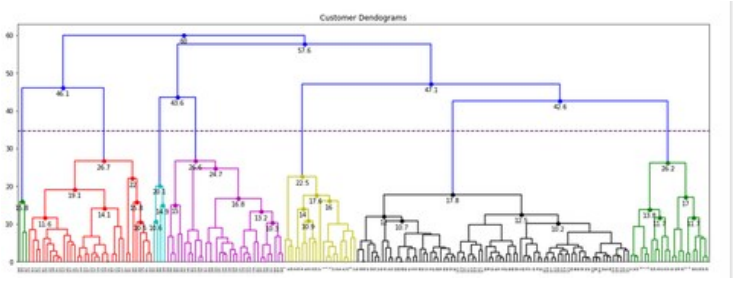
(A) Au hasard.

(B) En regardant toutes les distances inter-groupes, retenant la plus grande.

(C) En regardant toutes les distances inter-groupes, retenant la plus petite.

(D) On prend les deux plus petits groupes.

20.



Quel est le nombre de clusters retenu si l'on coupe au niveau de la ligne horizontale pointillée ? (cf. image à droite).

(A) 2

(B) 3

(C) 6

(D) 7