Algorithmes pour l'IA – TD 3
Partitionnement et apprentissage automatique supervisé

T1: Partitionnez le jeu de données de la 1ère table ci-dessous en utilisant :

1. k-means avec k = 3 (voir 2ème tableau pour les distances euclidiennes) et en partant de trois points existants choisis au hasard.
2. le partitionnement hiérarchique ascendant avec la distance euclidienne, agrégée en utilisant la distance minimum
3. DBScan avec la distance euclidienne, une distance maximale epsilon de 3, et un nombre minimal d'objets par partition de 2

Comparez les résultats obtenus.

O1	2	10
O2	2	5
O3	8	4
O4	5	8
O5	7	5
O6	6	4
O7	1	2
O8	4	9

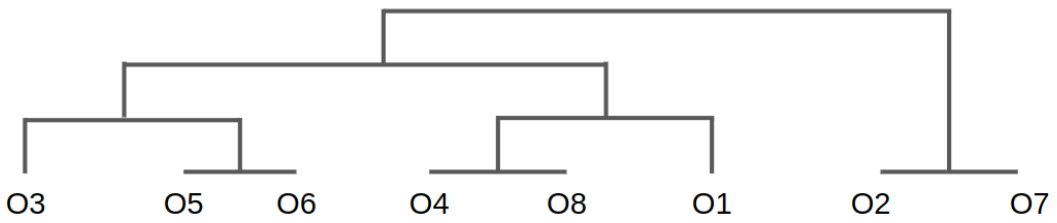
	O1	O2	O3	O4	O5	O6	O7	O8
O1	0,00	5,00	8,49	3,61	7,07	7,21	8,06	2,24
O2		0,00	6,08	4,24	5,00	4,12	3,16	4,47
O3			0,00	5,00	1,41	2,00	7,28	6,40
O4				0,00	3,61	3,61	7,21	1,41
O5					0,00	1,41	6,71	5,00
O6						0,00	5,39	5,39
O7							0,00	7,62
O8								0,00

K-Means :

1. On choisit aléatoirement trois centroïds initiaux : O2, O5, O8.
2. En utilisant le tableaux des distances, on regroupe chaque point des données avec le centroïd le plus proche :
   - O2, O7
   - O5, O3, O6
   - O8, O1, O4
3. On calcule les nouveau centroïd avec les moyennes des valeurs des variables descripteur des points regroupés dans la même partition : C1 = (1,5 ; 3,5), C2 = (7 ; 4,33), C3 = (3,67; 9)
4. On groupe de nouveau les points du jeu de données en fonctions de leur distance au nouveau centroïds (itération 2). Par exemple, la distance entre O1 et C1 est $\sqrt{(2-1,5)^2+(10-3,5)^2}$.
5. On calcule les nouveaux centroïds. Comme les partitions n'ont pas changées, les centroïds non plus et on s'arrête.

Partitionnement hiérarchique :

1. On crée 8 partitions de taille 1, chacune contenant un des points de données.
2. On trouve le couple de partitions les plus proches entre elles en utilisant la distance dans le tableau fournit. Il y a plusieurs choix possibles, et on choisit, aléatoirement, (O5) et (O6). Une nouvelle partition est créée : (O5,O8).
3. On trouve les deux partitions les plus proches : (O4) et (O8). Nouvelle partition : (O4,O8).
4. On trouve les deux partitions les plus proches : (O3) et (O5,06). Ici on a utilisé l'agrégation de la distance par le minimum. O3 et O5 sont les points les plus proches des deux partitions (ceux avec la distance minimum), donc on utilise leur distance (1,41) pour évaluer la distance entre les deux partitions. Nouvelle partition : (O3,O5,O6).
5. On trouve les deux partitions les plus proches : (O1) et (O4,O8). Nouvelle partition : (O1,O4,O8).
6. On trouve les deux partitions les plus proches : (O2) et (O7). Nouvelle partition : (O2,O7).
7. On trouve les deux partitions les plus proches : (O3,O5,O6) et (O1,O4,O8). Nouvelle partition : (O1,O3,O4,O5,O6,O8).
8. On trouve les deux partitions les plus proches : (O1,O3,O4,O5,O6,O8) et (O2,O7). Nouvelle partition : (O1,O3,O4,O5,O6,O8).
9. Il ne reste plus qu'une partition : (O1,O2,O3,O4,O5,O6,O7,O8). On s'arrête.



DBScan :

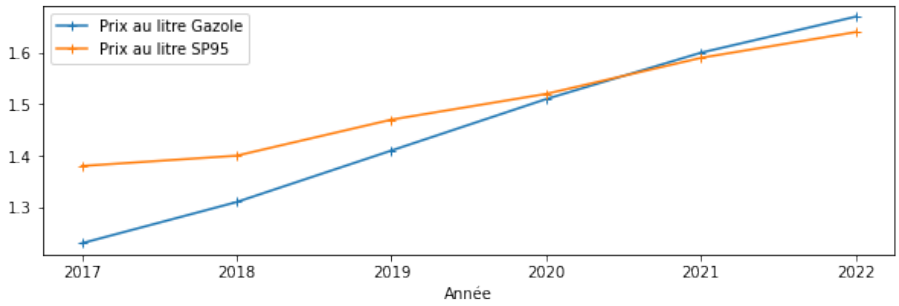
1. On choisit un point des données au hasard : O1.
2. Les epsilon-voisins de O1 sont (O8). Ce sont les points à une distance de moins de epsilon (3) du point O1. On les ajoute à la partition : (O1,O8).
3. On cherche les epsilon-voisins de O8 (le point ajouté) : O1, O4. On les ajoute à la partition : (O1,O8,O4).
4. On cherche les epsilon-voisins de O4 (le point ajouté). Il sont déjà tous dans la partition.

5. On choisit un nouveau point des données pas encore inclus dans un partition, au hasard : O2.
6. On cherche les epsilon-voisins de O2 : Il n'y en a pas. Comme la partition courante a une taille inférieur à la taille minimale (2), on considère les points qui s'y trouvent (O2) comme représentant des valeurs aberrantes (*outlier*).
7. On choisit un nouveau point des données pas encore inclus dans un partition, au hasard : O3.
8. On cherche les epsilon-voisins de O3 : O5, O6. On les ajoute à la partition : (O3,O5,O6).
9. On cherche les epsilon-voisins de O5 (un des points ajoutés). Il sont déjà tous dans la partition.
10. On cherche les epsilon-voisins de O6 (un des points ajoutés). Il sont déjà tous dans la partition.
11. On choisit un nouveau point des données pas encore inclus dans un partition, au hasard : O7.
12. On cherche les epsilon-voisins de O7 : Il n'y en a pas. Comme la partition courante a une taille inférieur à la taille minimale (2), on considère les points qui s'y trouvent (O7) comme représentant des valeurs aberrantes (*outlier*).
13. Il n'y a plus de points à partitionner, on s'arrête.

Comparaison: Si on coupe le partitionnement hiérarchique au niveau 3 on obtient les mêmes partitions que K-Means. DBScan donne aussi presque les mêmes résultats (avec la distance donnée), mais considère O2 et O7 comme représentant des valeurs aberrantes plutôt qu'une partition.

T2: Le tableau ci-dessous montre le prix moyen du gazole et de l'essence sans-plomb 95 en France sur plusieurs années. Calculez le modèle de régression linéaire pour ces deux séries de données. D'après vos modèles, quels seront les prix moyen des deux types de carburant en 2023.

Année	2017	2018	2019	2020	2021	2022
Prix au litre Gazole	1.23	1.31	1.41	1.51	1.60	1.67
Prix au litre SP95	1.38	1.40	1.47	1.52	1.59	1.64



Moyennes : $\bar{x} = 2019,5$; $\bar{y_1} = 1,455$; $\bar{y_2} = 1,5$

$a_1 = \frac{\sum_i (x_i - \bar{x}) \times (y_{1i} - \bar{y_1})}{\sum_i (x_1 - \bar{x_1})} = \frac{\sum_i (x_i - 2019,5) \times (y_{1i} - 1,455)}{\sum_i (x_1 - 2019,5)} = 0,09057142857142857$

$a_2 = \frac{\sum_i (x_i - \bar{x}) \times (y_{2i} - \bar{y_2})}{\sum_i (x_2 - \bar{x_2})} = \frac{\sum_i (x_i - 2019,5) \times (y_{1i} - 1,5)}{\sum_i (x_2 - 2019,5)} = 0,054857142857142875$

$b_1 = \bar{y_1} - a_1 \bar{x} = 1,455 - 0,09057142857142857 \times 2019,5 = -181,45399999999998$

$b_2 = \bar{y_2} - a_2 \bar{x} = 1,5 - 0,054857142857142875 \times 2019,5 = -109.32900000000004$

Gazole en 2023 : $0,09057142857142857 \times 2023 - 181,45399999999998 = 1,7720000000000198$

SP95 en 2023 : $0,054857142857142875 \times 2023 - 109.32900000000004 = 1,6920000000000073$

T3: Le tableau suivant montre les information sur des voitures arrivées dans un garage : si le conducteur était fatigué, si il y avait de la neige, si il y a eu un accident, si il y avait des passager dans la voirture, en plus du conducteur, et si la voiture est réparable. En utilisant comme variables “descripteurs” Fatigué, Neige, Accident et Passager et comme variable à prédire Réparable, construire l'arbre de décision correspondant.

Nom	Fatigué	Neige	Accident	Passager	Réparable
Arsène	V	V	V	F	F
Barnabé	F	V	F	V	V
Charles	F	V	V	F	V
Davis	F	F	F	F	V
Edgar	F	F	V	F	F
Ferdinand	V	F	V	V	F
Gaston	F	F	F	F	V
Henri	F	V	F	V	V
Igor	F	V	V	V	V
Jason	F	F	V	V	F
Kevin	F	V	F	V	V
Léonard	F	V	V	F	F

Entropie du noeud racine : 0.9072088417741441

Entropie de Fatigué=V : 0
Entropie de Fatigué=F : 0.7204024419616616
Gain d'information pour Fatigué : 0.9072088417741441-((1/2)*0+(1/10)*0.7204024419616616)
Gain d'information pour Fatigué : 0.8351685975779779

Entropie de Neige=V : 0.6934668959574882
Entropie de Neige=F : 1.0575424759098897
Gain d'information pour Neige : 0.9072088417741441-((1/7)*0.6934668959574882+(1/5)*1.057542475909
Gain d'information pour Neige : 0.5966336471696678

Entropie de Accident=V : 1.0327742411757737
Entropie de Accident=F : 0

Gain d'information pour Accident : 0.9072088417741441-((1/7)*1.0327742411757737+(1/5)*0)
Gain d'information pour Accident : 0.7596696644633193

Entropie de Passager=V : 0.7799500009615417
Entropie de Passager=F : 1.0
Gain d'information pour Passager : 0.9072088417741441-((1/6)*0.7799500009615417+(1/6)*1.0)
Gain d'information pour Passager : 0.6105505082805538

On coupe sur la variable Fatigué

Entropie de Fatigué=V : 0
On ne peut pas descendre plus.

Entropie de Fatigué=F : 0.7204024419616616
Entropie de Neige=V : 0.43839067638965634
Entropie de Neige=F : 1.0
Gain d'information pour Neige : 0.7204024419616616-((1/6)*0.43839067638965634+(1/4)*1.0)
Gain d'information pour Neige : 0.3973373292300522

Entropie de Accident=V : 1.0575424759098897
Entropie de Accident=F : 0
Gain d'information pour Accident : 0.7204024419616616-((1/5)*1.0575424759098897+(1/5)*0)
Gain d'information pour Accident : 0.5088939467796836

Entropie de Passager=V : 0.5150849518197796
Entropie de Passager=F : 0.8843587129994475
Gain d'information pour Passager : 0.7204024419616616-((1/5)*0.5150849518197796+(1/5)*0.884358712
Gain d'information pour Passager : 0.44051370899781617

On coupe sur la variable Accident

Entropie de Accident=V : 1.0575424759098897
Entropie de Neige=V : 0.7799500009615417
Entropie de Neige=F : 0
Gain d'information pour Neige : 1.0575424759098897-((1/3)*0.7799500009615417+(1/2)*0)
Gain d'information pour Neige : 0.7975591422560425

Entropie de Passager=V : 1.0
Entropie de Passager=F : 1.0566416671474375
Gain d'information pour Passager : 1.0575424759098897-((1/2)*1.0+(1/3)*1.0566416671474375)
Gain d'information pour Passager : 0.20532858686074396

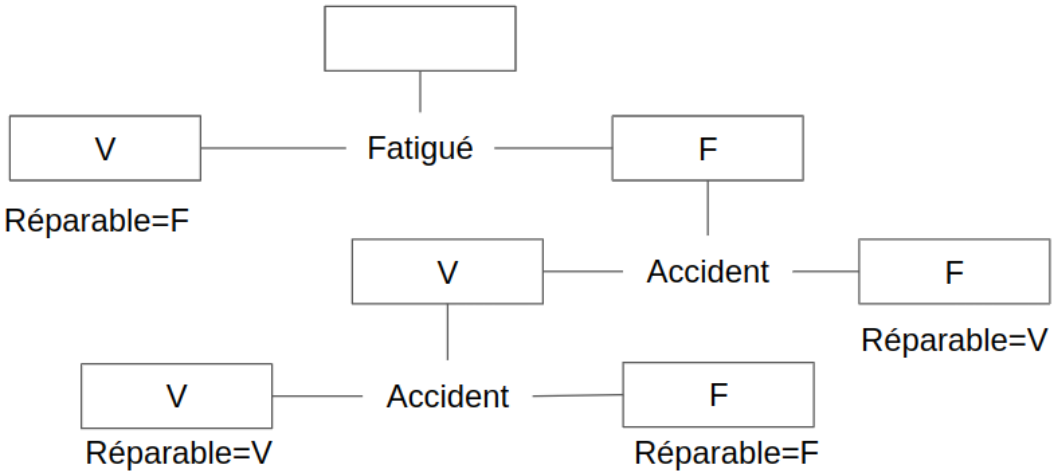
On coupe sur la variable Neige

Entropie de Accident=F : 0
On ne peut pas descendre plus.

Entropie de Neige=V : 0

On ne peut pas descendre plus.

Entropie de Neige=F : 0
On ne peut pas descendre plus.



T4: Le tableau suivant montre les mêmes variables, mais avec le coût de réparation ou de remplacement plutôt que si la voiture est réparable. Construire l'arbre de décision pour la régression.

Nom	Fatigué	Neige	Accident	Passager	Coût
Arsène	V	V	V	F	3000
Barnabé	F	V	F	V	600
Charles	F	V	V	F	2500
Davis	F	F	F	F	50
Edgar	F	F	V	F	2000
Ferdinand	V	F	V	V	2600
Gaston	F	F	F	F	50
Henri	F	V	F	V	600
Igor	F	V	V	V	2600
Jason	F	F	V	V	2100
Kevin	F	V	F	V	600
Léonard	F	V	V	F	2600

Coefficient de variation (CV) du noeud n = écart type du coût pour les éléments du noeud / moyenne des coût pour l'ensemble du jeu de données.

CV du noeud racine : 0.7010449902483555

CV de Fatigué=V : 0.17586075387022945
CV de Fatigué=F : 0.6714514149244878
Gain d'information pour Fatigué : 0.7010449902483555-((1/2)*0.17586075387022945+(1/10)*0.67145141
Gain d'information pour Fatigué : 0.545969471820792

CV de Neige=V : 0.6964735044400097
CV de Neige=F : 0.7568527704496734
Gain d'information pour Neige : 0.7010449902483555-((1/7)*0.6964735044400097+(1/5)*0.756852770449
Gain d'information pour Neige : 0.45017822123841944

CV de Accident=V : 0.21063125284183462
CV de Accident=F : 0.18730408702249204
Gain d'information pour Accident : 0.7010449902483555-((1/7)*0.21063125284183462+(1/5)*0.1873040
Gain d'information pour Accident : 0.6334939938664522

CV de Passager=V : 0.6345828349179219
CV de Passager=F : 0.8189807174499165
Gain d'information pour Passager : 0.7010449902483555-((1/6)*0.6345828349179219+(1/6)*0.818980717
Gain d'information pour Passager : 0.4587843981870492

On coupe sur la variable Accident

== CV de Accident=V : 0.21063125284183462

CV de Fatigué=V : 0.17586075387022945
CV de Fatigué=F : 0.17912779636337328
Gain d'information pour Fatigué : 0.21063125284183462-((1/2)*0.17586075387022945+(1/5)*0.1791277
Gain d'information pour Fatigué : 0.08687531663404524

CV de Neige=V : 0.13786668078393854
CV de Neige=F : 0.19986840955425814
Gain d'information pour Neige : 0.21063125284183462-((1/4)*0.13786668078393854+(1/3)*0.199868409
Gain d'information pour Neige : 0.10954177946109728

CV de Passager=V : 0.17948713031801836
CV de Passager=F : 0.2557297964612035
Gain d'information pour Passager : 0.21063125284183462-((1/3)*0.17948713031801836+(1/4)*0.255729
Gain d'information pour Passager : 0.0868697602871943

On coupe sur la variable Neige

== CV de Accident=F : 0.18730408702249204

CV de Fatigué=V : 0
Ne peut pas couper : un des sous-noeud est vide

CV de Neige=V : 0.0
CV de Neige=F : 0.0
Gain d'information pour Neige : 0.18730408702249204-((1/3)*0.0+(1/2)*0.0)
Gain d'information pour Neige : 0.18730408702249204

CV de Passager=V : 0.0
CV de Passager=F : 0.0

Gain d'information pour Passager : 0.18730408702249204-((1/3)*0.0+(1/2)*0.0)
Gain d'information pour Passager : 0.18730408702249204

On coupe sur la variable Passager

On ira pas plus loin pour la branche Accident=F, les CV sont déjà à 0.

==== CV de Neige=V : 0.13786668078393854

CV de Fatigué=V : 0
CV de Fatigué=F : 0.035897426063603674
Gain d'information pour Fatigué : 0.13786668078393854-((1/1)*0+(1/3)*0.035897426063603674)
Gain d'information pour Fatigué : 0.12590087209607065

CV de Passager=V : 0
CV de Passager=F : 0.1645026721905445
Gain d'information pour Passager : 0.13786668078393854-((1/1)*0+(1/3)*0.1645026721905445)
Gain d'information pour Passager : 0.08303245672042371

On coupe sur la variable Fatigué

Les CV étant déjà très bas, on ira pas plus loin sur cette branche.

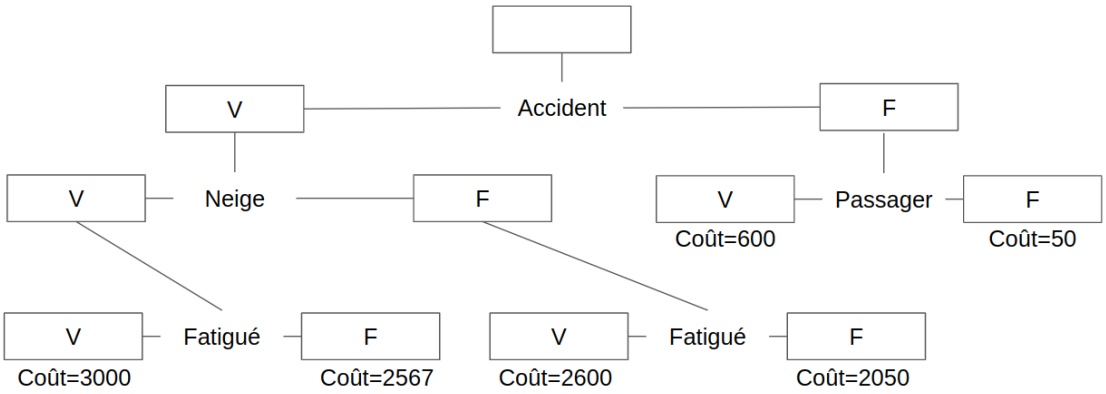
==== CV de Neige=F : 0.19986840955425814

CV de Fatigué=V : 0
CV de Fatigué=F : 0.04396518846755736
Gain d'information pour Fatigué : 0.19986840955425814-((1/1)*0+(1/2)*0.04396518846755736)
Gain d'information pour Fatigué : 0.17788581532047945

CV de Passager=V : 0.21982594233778682
CV de Passager=F : 0
Gain d'information pour Passager : 0.19986840955425814-((1/2)*0.21982594233778682+(1/1)*0)
Gain d'information pour Passager : 0.08995543838536472

On coupe sur la variable Fatigué

Les CV étant déjà très bas, on ira pas plus loin sur cette branche.


Accident
V
F
Neige
V
F
Passager
V
F
Coût=600
Coût=50
Fatigué
V
F
Fatigué
V
F
Coût=3000
Coût=2567
Coût=2600
Coût=2050