Examen de rattrapage
« Algorithmes Data Mining II »

Exercice 1 (QCM) :

1- Après la construction du modèle, prédire la classe d'une nouvelle observation avec un arbre de décision prend plus de temps qu'un KNN.
   a. Vrai
   b. Faux

2- Lesquelles des affirmations suivantes sont vraies à propos du « Forward Feature Selection » ?
   a. C'est un algorithme avide (greedy), il ajoute les features qui améliore le plus l'accuracy de la validation croisée
   b. Il cherche un sous ensemble de features qui donne l'erreur de test la plus faible
   c. "Forward selection" est plus rapide que la « backward selection » si le nombre de features pertinents pour la prédiction est très petit.

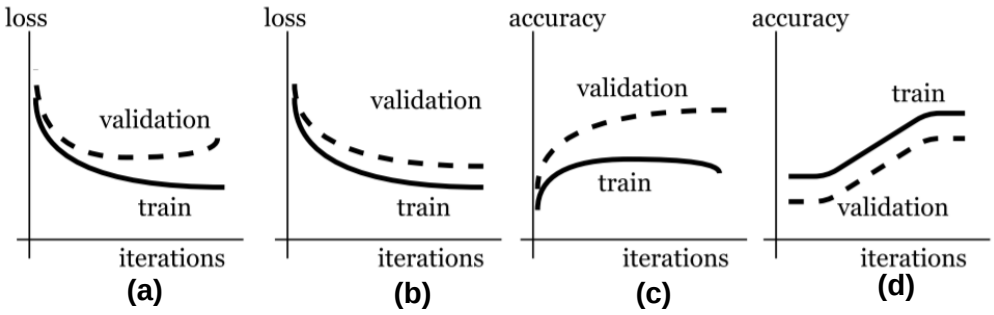
3- Les valeurs aberrantes (outliers) sont toujours dues au bruit.
   a. Vrai
   b. Faux

4- Considérant un arbre de décision entrainé sur la matrice $\mathrm{X} = \begin{bmatrix} 6 & 3 \\ 2 & 7 \\ 9 & 6 \\ 4 & 2 \end{bmatrix}$ avec les labels $y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$. Soit f1 et f2, les features correspondant au premier et deuxième colonne respectivement. Parmi les conditions ci-dessous considéré au niveau du nœud racine donne le plus grand gain d'information ?
   a. f1 > 2
   b. f1 > 4
   c. f2 > 3
   d. f2 > 6

5- Une valeur AUC (Area under the ROC curve) de 0 correspond à un prédicteur aléatoire.
   a. Vrai
   b. Faux

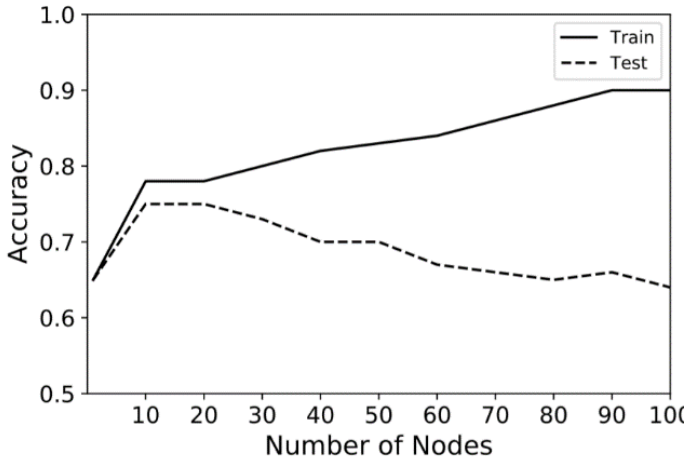
6- En termes bias-variance, un 1-NN (KNN avec k=1) comparé à un 3-NN (KNN avec k=3) a
   a. Une variance plus élevée
   b. Une variance plus faible
   c. Un bias plus élevée
   d. Un bias plus faible

7- Parmi les courbes présentés ci-dessous, préciser le ou les courbe illustrant le problème du surapprentissage (overfitting).



a. Figure (a)
b. Figure (b)
c. Figure (c)
d. Figure (d)

8- Le graphique ci-dessous est une représentation graphique de l'accuracy du train et de test pour des arbres de décision de différentes tailles (l'accuracy en fonction du nombre des nœuds). Le même ensemble de train (un ensemble fini) est utilisé pour la construction des arbres. Que se passe-t-il des courbes de train et de test si la quantité des données utilisée pour l'apprentissage s'approche de l'infini ?



Exercice 3 (Arbre de décision et Bagging) :

On dispose de la base de données, présentée dans Tableau 1, représentant les résultats d'analyse de 15 patients concernant l'infection par COVID19 :

1. Détailler les étapes de construction de l'arbre de décision (en utilisant l'ensemble du Train,) permettant de prédire l'infection par le virus COVID-19 en utilisant le Gain d'information pour le choix des attributs.
2. Donner la matrice de confusion du modèle entrainé puis calculer la précision et le recall sur l'ensemble du train puis sur l'ensemble de test.
3. Commenter le résultat de l'évaluation du modèle.
4. Tester la performance sur un ensemble de test est une des stratégies pour estimer la performance de généralisation d'un modèle d'apprentissage. Décrire d'autres stratégies d'évaluation des modèles.

5. On veut construire un modèle ensembliste en utilisant la méthode du Bagging avec l'arbre de décision comme modèle de base.
   5.1. Rappeler le principe du « Bagging ». Comment les datasets utilisées pour l'entrainement des modèles de l'ensemble sont générées ?
   5.2. Construire un ensemble avec 3 arbres de décision
6. Bagging, Boosting et Stacking sont trois approches pour la combinaison des modèles. Quelles sont les différences entre ces 3 approches ?

Tableau 1: Base de données COVID19

	N°	Age	Difficulté Respiratoire (DR)	Fièvre (FV)	Toux (TX)	Fièvre et Toux (FAT)	COVID19
Train	1	Vieux	Oui	Non	Oui	Oui	Positif
	2	Jeune	Oui	Oui	Non	Non	Positif
	3	Vieux	Non	Oui	Non	Non	Négatif
	4	Enfant	Non	Non	Oui	Oui	Positif
	5	Jeune	Oui	Non	Oui	Non	Négatif
	6	Enfant	Oui	Oui	Non	Oui	Négatif
	7	Vieux	Non	Non	Oui	Oui	Positif
	8	Vieux	Oui	Oui	Non	Non	Négatif
	9	Jeune	Non	Oui	Oui	Non	Négatif
	10	Jeune	Non	Non	Oui	Oui	Positif
Test	11	Vieux	Oui	Oui	Non	Non	Positif
	12	Jeune	Oui	Non	Oui	Oui	Négatif
	13	Vieux	Non	Oui	Oui	Non	Négatif
	14	Enfant	Non	Non	Oui	Oui	Positif
	15	Jeune	Oui	Oui	Non	Non	Négatif