Solution DS-2

1) Indiquer la signification du paramètre minsplit.
Voir cours Arbre de décision.

2) Remplacer le ? de la ligne 2 de l'arbre par la valeur adéquate.
?=1

3) Déterminer le nombre de règles issues de ete arbre.
Il suffit de ompter le nombre de * soit 10

4) En utilisant l'arbre obtenu, classer l'individu ALPHA ayant les caractéristiques suivantes :
(`age=27 ; stab_emploi= 3 ; autre_garant=2 ; trav_etrang=2 ; montant=2000 ; etat_compte=2; duree_credit=15 ; part_mens=2 ; ressources=2 ; autres_credits=3`)

Sachant que l'individu a la modalité 2 de la variable `etat_compte` cet individu se retrouve dans le noeud 3 de l'arbre, il est par conséquent classé 1 donc bon (i.e. avec une probabilité de 0.86).

5) Indiquer la règle correspondant au noeud numéro 40 de l'arbre.

La régle est la suivante : `Si etat_compte=2 et montant < 9908.5 et ressources=1,2 et duree_credit>=20.5 et age< 25.5 alors l'individu est classé mauvais`.

6) Expliquer le principe de la sélection pas à pas *forward*.

La sélection pas à pas se fait par la minimisation du Critère AKAIKE $AIC = -2LL + 2*(j+1)$ où j est le nombre de paramètres (variables) du modèle et LL est sa déviance. La procédure FORWARD consiste à partir du modèle réduit à une constante puis de rajouter à chaque itération la variable qui minimise le critère AIC. La sélection s'arrête lorsque l'ajout d'une variable n'améliore plus le critère AIC.

7) Classer l'individu ALPHA à l'aide du modèle obtenu par la régression logistique.

On calcule $exp(a)/(1 + exp(a))$ où
$a = 4.12 * 10^{-1} + 6.924 * 10^{-1} - 8.494 * 10^{-5} * 2000 - 1.139 * 10^{-1} + 4.92 * 10^{-2} + 2.139 + 6.576 * 10^{-1} - 15 * 2.05 * 10^{-2} - 7.983 * 10^{-1} = 2.560$.
D'où $P(Y = 1|Alpha) = 0.928$ qui est supérieure à 0.5 donc ALPHA est classé bon.

8) Comparer les variables sélectionnées par l'arbre de décision `arbre.full` et celle de la sélection pas à pas *forward* de la régression logistique.

- Alor que la régression logistique n'a pas retenu les variables `part_mens` et `age`, l'arbre de décision n'a pas retenu les 3 variables suivantes : `part_mens, trav_etrang,autres_credits`.

9) Indiquer en justifiant votre réponse, quelle sont les varaibles les plus significatives

selon le modèle donné par la régression logistique.

Il s'agit en premier lieu de la variable `etat_compte` avec une significativité presque nulle (p-value = 6.19 $10^{-}11$ pour la modalité etat_compte3), puis de la variable `ressources` (p-value = 0.000306 pour la modalité ressources5), puis de la variable `autres_credits` (p-value = 0.014279 pour la modalité autres_credits3) puis la variable `duree_credits` (p-value = 0.044169 pour la modalité duree_credits3) et enfin la variable `montant` (p-value = 0.047)