ACT6100 H2019	Analyse de données Série 6

Arbres

1. On considère un noeud avec 6 observations

$$R = \{A, A, B, B, B, C\}$$

et pour lequel on utilise le mode pour réaliser une prédiction.

(a) Si on considère les sous-groupes

$$R_1 = \{A, A, B, C\} \text{ et } R_2 = \{B, B\},$$

calculer l'impact de la division sur l'index Gini et sur l'entropie.

(b) Pour un noeud homogène, c'est-à-dire un noeud pour lequel $\widehat{p}$ est très près de 0 ou de 1, vérifier que l'index Gini et l'entropie sont numériquement semblables. MAT8594

2. Pour les arbres représentés aux Figures 1, 2 et 3, faire la représentation de l'espace des variables explicatives (si vous faites l'exercice à la main, n'indiquez pas les variables explicatives sur votre graphique), indiquer les différentes régions obtenues et indiquer la prédiction faite par le modèle pour chacune de ces régions.

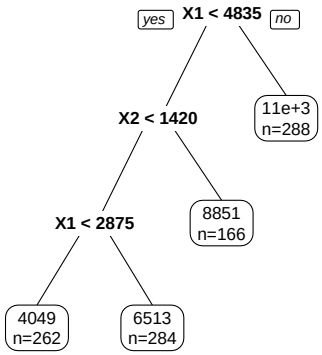

FIGURE 1 – Arbre de décision 1.

(a)

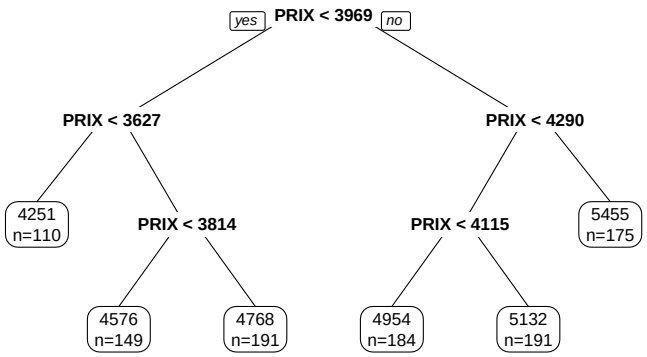

FIGURE 2 – Arbre de décision 2.

(b)

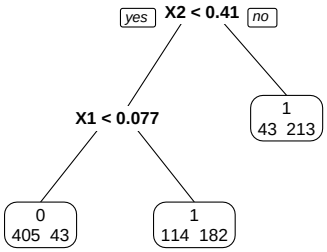

FIGURE 3 – Arbre de décision 3.

(c)

3. On considère la base de données *freMPL2* disponible dans la libraire *CASdatasets*.
Partie I

(a) Pour la première question, on s'intéresse uniquement aux variables suivantes :
— **ClaimInd** : une variable indicatrice d'un sinistre (1) ou non (0) ;
— **Exposure** : l'exposition (en année), à considérer cette fois comme une variable explicatrice et non comme un terme d'*offset* ;
— **DrivAge** : l'âge du conducteur et
— **BonusMalus** : un indicateur du niveau de risque de l'assuré.
Pour l'analyse demandée, considérer que toutes les variables sont numériques. Construire un arbre de régression complet en posant le paramètre de complexité comme étant nul (`cp = 0`). Utiliser ensuite la fonction `plotcp` pour déterminer la valeur optimale du paramètre de complexité. Enfin, utiliser la fonction `prune` pour élaguer l'arbre complet avec la valeur optimale du paramètre de complexité.

(b) On s'intéresse maintenant à la variable **ClaimAmount** qui indique le cout total des réclamations. Construire une base de données contenant uniquement les observations pour lesquelles un sinistre a été observé. Par la suite, diviser cette base de données en une base d'entrainement (80%) et une base de validation (20%).

(c) En utilisant les variables explicatives **DrivAge** et **BonusMalus**, sélectionner la valeur du paramètre de complexité qui minimuser l'erreur quadratique moyenne de validation. Par la suite, élaguer l'arbre complet avec la valeur optimale du paramètre de complexité.

Partie II

(a) La première partie a permis d'obtenir un modèle pour la survenance d'un sinistre. Écrire l'équation de ce modèle.

(b) La première partie a également permis d'obtenir un modèle pour la sévérité d'un sinistre. Écrire l'équation de ce modèle.

(c) On suppose que l'assureur a un portefeuille contenant 1 000 contrats et dont la composition est similaire à la base de données *freMPL2*, utiliser la simulation afin d'obtenir la distribution du cout total du portefeuille, c'est-à-dire la fonction de répartition de

$$S = X_1 + \ldots + X_{1\,000},$$

où X_i correspond au cout total du contrat i avec $X_i = I_i Y_i$ (approche individuelle dans le cours ACT3400).