TP : Supervisée

Vous pouvez faire les exercices dans le langage de votre choix.

1 Exercice 1 : Position des joueurs NFL

Dans cet exercice, on se propose de développer un modèle permettant de classifier les joueurs NFL grâce à l'analyse discrimnante.

1. Télécharger le jeu de données suivant : lien.
2. Faire une rapide analyse descriptive des données.
3. Faire une analyse discriminante pour classer les joueurs des positions `QB` (quart-arrière) et `OL` (ligne offensive) en utilisant leur taille (`Ht`) et leur poids (`Wt`).
4. Quelle serait la coordonnée d'un joueur ayant une taille est de 76.5 et un poids de 335.5 sur le nouvel axe ? Quelle serait la position d'un tel joueur ?
5. Faire la même chose avec les joueurs des lignes offensives (`OL`) et défensives (`DL`), puis entre les joueurs de la ligne défensive (`DL`) et les quarts-arrières (`QB`). En utilisant les résultats de ce modèle, quelles positions sont les plus faciles à séparer avec l'analyse discriminante ?
6. Faire une analyse discriminante pour classer les joueurs des positions `QB`, `OL` et `DL`. La classification des joueurs est-elle facilitée avec ce modèle ou bien est-il préférable d'utiliser les trois modèles précédents ?

2 Exercice 2 : Prédire les réclamations

Dans cet exercice, on se propose de construire un arbre de classification permettant de prédire si une réclamation pourrait avoir lieu sur un colis envoyé par Rakuten. Le jeu de données fourni par PriceMinister Rakuten contient une variable cible (`CLAIM_TYPE`), ainsi que 12 variables explicatives.

1. Télécharger le jeu de données suivant : lien.

2. Faire une rapide analyse descriptive des données.
3. Partionner les données en échantillon d'entraînement (70%), de validation (30%) de façon aléatoire.
4. Ajuster un arbre de classification sur le jeu d'entraînement en utilisant les paramètres par défaut.
5. Faire une prédiction sur les jeux d'entraînement et de validation.
6. Estimer le taux d'erreur dans le noeud 3.
7. Calculer le taux d'observations bien classifiés global sur les échantillons d'entraînement et de validation.
8. Changer les hyperparamètres tel que l'effectif minimal pour q'un noeud puisse être séparé soit de 2 et l'effectif minimal d'un noeud terminal soit de 1 et reconstruire l'arbre.
9. Faire varier le paramètre de complexité du modèle et analyser les arbres résultant.
10. Calculer le taux d'erreur des différents modèles sur les échantillons d'entraînement et de validation. Tracer le graphique du taux d'erreur en fonction de la complexité du modèle.