BE2 - partie 2 - entraînement - Régression logistique
MOD - Statistique appliquée aux sciences de l'ingénieur.

Exercice 2 : Régression logistique
Nous cherchons à prédire la variable Pain en fonction de 4 autres variables. Nous disposons pour ceci d'un fichier de 60 observations nommé neuralgia.txt.

1. Importer le fichier neuralgia.txt sous R. Décrire les variables.
2. Partager le fichier en un fichier d'apprentissage (80%) et un fichier de test (20%). On utilisera la commande $u = sample(1 : n, p)$ pour choisir p éléments dans $1 : n$. L'ensemble $data[-u,]$ est alors le complémentaire de l'ensemble $data[u,]$.
3. Réaliser sur le fichier d'apprentissage une régression logistique pour prédire la variable **Pain**. Quel est l'évenement modélisé ? Rappeler la forme du modèle logit. Dans cette question, on considère un modèle sans interaction.
4. Analyser le résultat des commandes $Anova(, type = "III", test.statistic = "LR")$ et $Anova(, type = "III", test.statistic = "Wald")$. Interpréter la sortie de *summary*. Comparer à que donne la commande $anova(, test = "chisq")$. Quelles sont les variables influentes ? Simplifier le modèle.
   RQ : il faut charger et installer le package "car" pour faire fonctionner la routine *Anova*.
5. Réaliser maintenant une procédure forward pour le critère AIC. Expliquer le principe et donner le modèle obtenu. On adaptera le code vu en cours. Le modèle initial est réduit à la constante, le modèle le plus complet est le modèle contenant toutes les interactions.
6. A l'aide du fichier test, comparer les matrices de confusions pour les deux modèles (on utilisera la commande predict pour réaliser les prévisions sur le fichier test puis la commande *table* pour obtenir la matrice de confusion).
7. On se fixe un modèle. Etudier la sensibilité des qualités prédictives à l'échantillon (prendre 50 ensembles d'apprentissage constitué de 80% des individus) puis étudier la sensibilité des résultats à la proportion apprentissage/test (10%,20%,30%,40%).