Apprentissage Automatique 1

TP N°6– Arbres de décisions

Exercice N°1 : Prédiction de survie ou non du passager (Titanic)

Le principe est le suivant : vous disposez un jeu de données sur des passagers du titanic :

- Un jeu de données sur les passagers pour lequel le champ 'Survived' (qui indique la survie du passager)

Le but du jeu est de construire une fonction qui prenne en entrée les informations sur un passager, et qui donne une prédiction de survie ou non du passager. Cette fonction sera construite grâce aux données dites d'entrainement, et sera évaluée sur les données de test.

1. Charger, à l'aide de pandas le fichier titanic.csv, et procéder le traitement sur le jeu de données.
2. Afficher les DataFrames obtenus.
3. Y-a-t-il a priori des données qui ne sont pas à prendre en compte pour prédire la survie du passager ? Supprimer ces données du DataFrame.
4. Y-a-t-il des données manquantes ? Sur quelles propriétés ?
5. Ajouter un champ au DataFrame pour indiquer l'existence ou non d'une donnée manquante sur cette propriété.
6. Remplacer les données manquantes par des valeurs spéciales. (Utiliser la fonction mean)
7. Ajuster les variables catégorielles en des variables numérique.
8. Construire le Modèle d'arbre de décision
9. Diviser le jeu de données en deux partie (Training, Testing)
10. Appliquer l'algorithme DecisionTreeClassifier().
11. Calculer l'apprentissage automatique.
12. Calculer la prédiction
13. Calculer le score de Training et de Testing
14. Appliquer l'algorithme de la régression Logistique sur le même jeu de données.
15. Appliquer l'apprentissage
16. Calculer le score de Training et de Testing
17. Comparer les résultats de (13,16) commenter les résultats.
18. Tester par un exemple la prédiction de deux algorithmes, commenter les résultats.