Rapport sur l'analyse et la modélisation des données Titanic

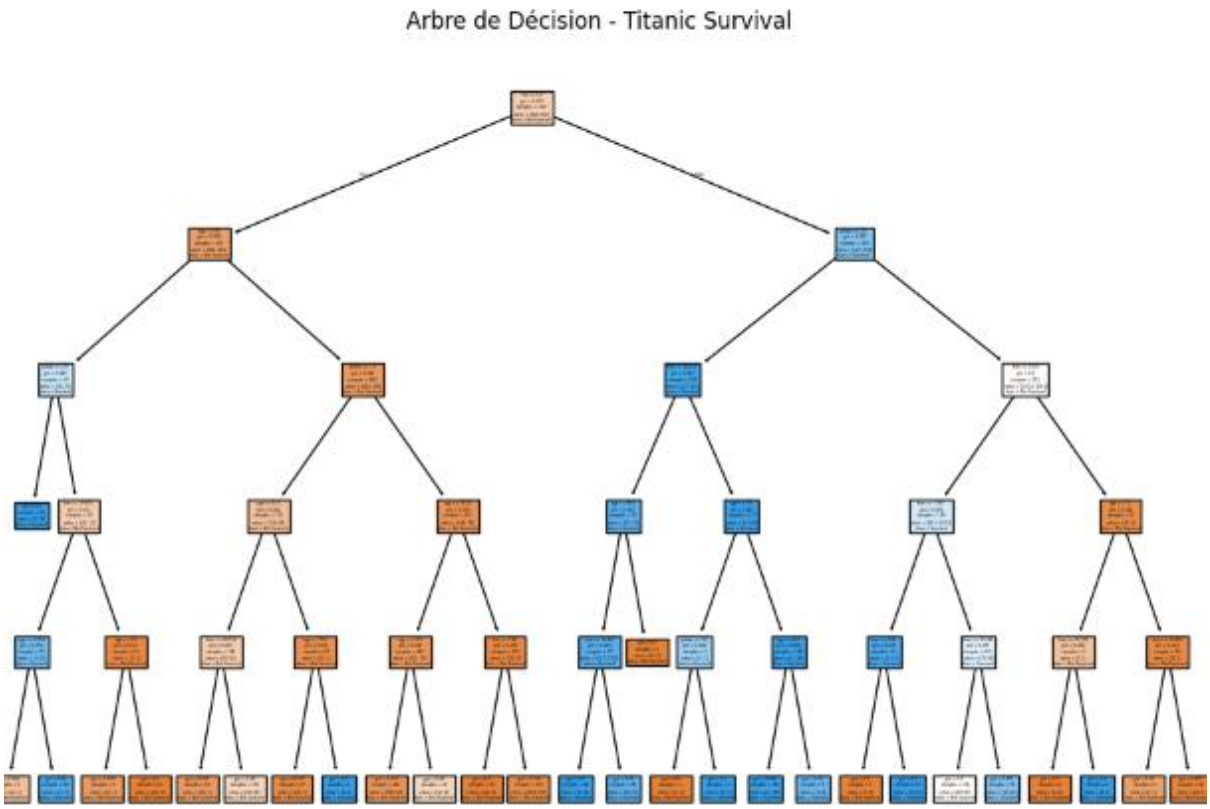
Objectif du projet

L'objectif de ce projet était de prédire si un passager aurait survécu ou non à l'accident du Titanic en utilisant un modèle basé sur un **arbre de décision**. Nous avons utilisé les données disponibles pour extraire des caractéristiques pertinentes, entraîner un modèle et visualiser les décisions prises par le modèle.

Étapes réalisées

1. **Chargement des données** :
    a. Le fichier **titanic3.xls** contenant les informations sur les passagers a été chargé.
    b. Les colonnes suivantes ont été utilisées :
        i. `pclass` : Classe sociale du passager (1ère, 2ème, 3ème classe).
        ii. `sex` : Sexe du passager (homme ou femme).
        iii. `age` : Âge du passager.
        iv. `fare` : Montant payé pour le billet.
    c. La colonne cible était `survived`, indiquant si le passager a survécu (`1`) ou non (`0`).
2. **Prétraitement des données** :
    a. **Encodage des variables catégoriques** :
        i. La colonne `sex` a été convertie en valeurs numériques (`0` pour homme, `1` pour femme).
    b. **Gestion des valeurs manquantes** :
        i. Les valeurs manquantes dans `age` et `fare` ont été remplacées par la médiane de leur colonne respective pour éviter tout biais.
    c. Suppression des lignes restantes contenant des valeurs manquantes.
3. **Création du modèle** :
    a. Un arbre de décision a été entraîné à l'aide des données d'entraînement.
    b. La profondeur maximale de l'arbre a été fixée à `5` pour éviter le surapprentissage (overfitting).
4. **Visualisation de l'arbre** :
    a. L'arbre de décision a été visualisé pour interpréter les règles utilisées par le modèle.

b. Les nœuds de l'arbre montrent comment les variables influencent les prédictions de survie.



Résultats obtenus

1. **Arbre de décision** :
    a. L'arbre montre que les principales variables influençant la survie sont :
        i. **sex** : Les femmes avaient une probabilité plus élevée de survie.
        ii. **pclass** : Les passagers de la 1ère classe avaient une meilleure chance de survie par rapport aux classes inférieures.
        iii. **fare** : Un tarif plus élevé (souvent lié à une classe sociale plus élevée) est associé à une probabilité de survie accrue.
        iv. **age** : Les enfants ont eu plus de chances de survivre.
    b. Les nœuds colorés indiquent la proportion de passagers ayant survécu ou non à chaque étape de la décision.
2. **Interprétation des résultats** :

a. Les résultats confirment les observations historiques selon lesquelles les femmes et les enfants, ainsi que les passagers de la 1ère classe, ont eu une priorité lors de l'évacuation.

Conclusion

Ce projet a démontré l'utilité des arbres de décision pour analyser et interpréter des données réelles. L'arbre de décision obtenu fournit une compréhension claire des facteurs ayant influencé les chances de survie des passagers du Titanic. Cette approche peut être étendue à d'autres jeux de données similaires pour des analyses comparatives ou prédictives.