# Étude de Data Mining

Analyse Exploratoire et Prédiction de la Survie des Passagers du Titanic
à l'aide des Arbres de Décision

Travail réalisé par :

**CHERFAOUI HOUSSAM ABDERRAHMANE**

14 décembre 2025

# Table des matières

# 1 Introduction et Problématique

Le naufrage du Titanic constitue l'un des événements maritimes les plus marquants de l'histoire. Au-delà de son aspect historique, ce jeu de données représente un cas d'étude classique en data mining, permettant d'analyser comment des caractéristiques socio-démographiques et économiques influencent les chances de survie.

La problématique principale de cette étude est la suivante : *peut-on prédire la survie d'un passager du Titanic à partir de ses caractéristiques individuelles ?*

L'objectif est double :

— Réaliser une analyse exploratoire des données (EDA) afin de comprendre la structure et les relations entre les variables.

— Construire et interpréter un modèle de classification basé sur un arbre de décision pour prédire la variable `Survived`.

# 2 Présentation du Jeu de Données

Le jeu de données utilisé est le dataset **Titanic**, disponible publiquement sur Kaggle et le dépôt UCI. Il contient des informations sur 891 passagers.

## 2.1 Dimensions et Structure

Le dataset comprend 891 observations et 12 variables. Les types de données sont répartis comme suit :

— Variables quantitatives : PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare.

— Variables qualitatives : Name, Sex, Ticket, Cabin, Embarked.

## 2.2 Variable Cible

La variable cible de l'étude est :

— **Survived** : variable binaire indiquant si le passager a survécu (1) ou non (0).

# 3 Analyse Exploratoire des Données (EDA)

## 3.1 Analyse des Valeurs Manquantes

L'analyse des valeurs manquantes révèle :

— Age : 177 valeurs manquantes.

— Cabin : 687 valeurs manquantes.

— Embarked : 2 valeurs manquantes.

Ces absences nécessitent un traitement approprié avant la modélisation.

## 3.2 Analyse Univariée

### 3.2.1 Variable Cible

La distribution de la variable `Survived` montre que le nombre de passagers n'ayant pas survécu est supérieur à celui des survivants, indiquant un léger déséquilibre des classes.

### 3.2.2 Variables Quantitatives

L'âge des passagers suit une distribution approximativement normale, légèrement asymétrique à droite, avec une majorité de passagers jeunes. Les variables `SibSp` et `Parch` sont fortement asymétriques à droite, montrant que la plupart des passagers voyageaient seuls. La variable `Fare` présente une forte asymétrie à droite avec de nombreux outliers correspondant à des billets très chers.

### 3.2.3 Variables Qualitatives

L'analyse montre une majorité de passagers masculins. Le port d'embarquement `S (Southampton)` est dominant, suivi de `C (Cherbourg)` puis `Q (Queenstown)`.

## 3.3 Analyse Bivariée

### 3.3.1 Survived vs Sex

Les femmes présentent un taux de survie largement supérieur à celui des hommes, indiquant que le sexe est un facteur déterminant de survie.

### 3.3.2 Survived vs Pclass

Les passagers de première classe ont le taux de survie le plus élevé, suivis de la deuxième classe, tandis que la troisième classe présente le taux de survie le plus faible.

### 3.3.3 Survived vs Embarked

Les passagers embarqués à Cherbourg ont le taux de survie le plus élevé, suivis de Queenstown, tandis que Southampton présente le taux le plus faible.

# 4 Prétraitement des Données

## 4.1 Variables Exclues du Modèle

Certaines variables ont été exclues car elles n'apportent pas d'information prédictive exploitable :

— PassengerId : identifiant unique.

— Name : variable textuelle non structurée.
— Ticket : trop grand nombre de valeurs uniques.
— Cabin : trop grand nombre de valeurs manquantes.

### 4.2 Traitement des Valeurs Manquantes et Encodage

Les valeurs manquantes de `Age` ont été imputées par la moyenne, celles de `Embarked` par le mode. Les variables catégorielles ont été encodées numériquement afin d'être compatibles avec l'arbre de décision.

## 5 Modélisation par Arbre de Décision

### 5.1 Préparation des Données

Après prétraitement, le jeu de données final contient 891 observations et 8 variables explicatives. Les données ont été divisées en un ensemble d'apprentissage (70

### 5.2 Construction du Modèle

Un classifieur **Decision Tree** a été entraîné avec les paramètres suivants :
— Critère : Entropy
— Profondeur maximale : 5
— Nombre minimum d'observations par nœud : 10

## 6 Évaluation du Modèle

Le modèle atteint une accuracy d'environ **80.97%** sur l'ensemble de test.
L'analyse du rapport de classification montre :
— Une très bonne capacité à identifier les passagers non survivants (recall élevé pour la classe 0).
— Une performance plus modérée pour la classe des survivants, avec un recall plus faible.

La matrice de confusion confirme ces observations et met en évidence un déséquilibre dans la capacité de prédiction entre les deux classes.

## 7 Interprétation de l'Arbre de Décision

L'analyse de l'arbre révèle que :
— Le **Sex** est la variable la plus importante (nœud racine).

— Les femmes, en particulier en première et deuxième classe, ont une probabilité très élevée de survie.
— Les hommes suivent majoritairement un chemin menant à la non-survie, avec des ajustements liés à l'âge, au tarif et à la composition familiale.

L'analyse de l'arbre de décision met en évidence plusieurs enseignements majeurs concernant les facteurs influençant la survie des passagers du Titanic.

Tout d'abord, la variable Sex apparaît comme le critère de décision le plus important, puisqu'elle constitue le nœud racine de l'arbre. Cette position indique que le sexe est le facteur le plus déterminant dans la prédiction de la survie. Les passagers masculins suivent majoritairement une trajectoire menant à la non-survie, tandis que les passagers féminins présentent, dès le départ, une probabilité de survie nettement plus élevée.

Pour les passagères, la variable Pclass joue un rôle clé après le sexe. Les femmes voyageant en première et en deuxième classe affichent des taux de survie très élevés. Même les femmes de troisième classe présentent des chances de survie supérieures à celles de la majorité des hommes. Dans certains cas, l'âge et le prix du billet (*Fare*) interviennent pour affiner la prédiction, notamment pour distinguer des situations de survie plus ou moins favorables.

En revanche, pour les passagers masculins, le chemin décisionnel conduit le plus souvent vers la classe « non survivant ». Bien que des variables telles que l'âge, le tarif du billet et parfois le nombre de proches à bord (*Parch*) permettent d'affiner la prédiction, la probabilité de survie demeure globalement faible. Les hommes plus jeunes peuvent bénéficier d'une légère amélioration de leurs chances, mais cela reste insuffisant pour inverser la tendance générale.

Enfin, l'arbre de décision montre que, après le sexe, les variables Pclass, Age et Fare sont les plus influentes. Elles permettent de préciser les conditions spécifiques dans lesquelles certains passagers avaient plus ou moins de chances de survivre.

En résumé, le modèle confirme que le genre est le facteur prédictif le plus puissant de la survie, suivi par la classe sociale. L'âge et le prix du billet viennent ensuite affiner ces prédictions, rendant le modèle à la fois performant et facilement interprétable, en cohérence avec les faits historiques du naufrage du Titanic.

# 8 Discussion et Limites

Malgré de bonnes performances, le modèle présente certaines limites :

— Taille relativement réduite du dataset.
— Déséquilibre des classes.
— Variables limitées à des caractéristiques individuelles.

Des améliorations futures pourraient inclure l'utilisation de forêts aléatoires, la validation croisée ou l'ingénierie de nouvelles variables.

# 9 Conclusion

Cette étude a démontré l'intérêt des techniques de data mining et des arbres de décision pour l'analyse de données réelles. Les résultats montrent clairement que le sexe et la classe sociale sont les facteurs les plus déterminants dans la survie des passagers du Titanic. L'arbre de décision offre un compromis efficace entre performance prédictive et interprétabilité.