Ecole d'Ingénieur Denis Diderot
Université Paris-Cité

Rapport

Intelligence Artificielle - 2SIE

Rapport du Mini-projet – Mise au point d'un modèle d'apprentissage automatique : Cas Titanic

Auteurs

ELABJANI Nissrine, HIBAOUI Imane, ZIDANI Rababe

Encadrant

M. Christophe DENIS

Année universitaire

2024-2025

1. Introduction

Le naufrage du Titanic en 1912 est une catastrophe maritime emblématique. L'objectif de ce mini-projet est de construire un modèle d'apprentissage automatique pour prédire la survie des passagers selon leurs caractéristiques (âge, sexe, classe, etc.). La démarche comprend une exploration, un nettoyage, une transformation des données et la modélisation via des algorithmes supervisés.

2. Exploration des données

Les données exploitées dans ce projet proviennent du jeu de données **Titanic**, disponible publiquement sur la plateforme OpenML. Ce jeu de données regroupe les informations de 1309 passagers ayant embarqué à bord du célèbre paquebot Titanic.

2.1. Description des variables

Le tableau suivant présente les principales variables disponibles dans le jeu de données :

- **pclass** : Classe du billet (1 = première, 2 = deuxième, 3 = troisième).
- **sex** : Sexe du passager (`male` ou `female`).
- **age** : Âge du passager en années.
- **sibsp** : Nombre de frères/soeurs ou conjoints à bord.
- **parch** : Nombre de parents ou enfants à bord.
- **fare** : Tarif payé pour le billet.
- **embarked** : Port d'embarquement (`C` = Cherbourg, `Q` = Queenstown, `S` = Southampton).
- **survived** : Cible binaire indiquant la survie (`1` = survécu, `0` = non survécu).

2.2. Analyse des valeurs manquantes

- La variable `age` présente environ **20%** de valeurs manquantes. Cette information est critique car l'âge est un facteur potentiel de survie.
- La variable `fare` contient une unique valeur manquante.
- Enfin, la variable `embarked` comporte **deux valeurs manquantes**, ce qui reste négligeable par rapport à la taille totale du jeu.

2.3. Analyse visuelle et premiers constats

Les passagers de **première classe** présentent un taux de survie nettement plus élevé que ceux des autres classes. Les **femmes** sont majoritairement représentées parmi les survivants, ce qui reflète probablement les priorités d'évacuation adoptées à l'époque. Les passagers de **troisième classe** sont en moyenne plus jeunes et ont été plus fortement touchés par le naufrage, notamment en raison de leur position à bord et de leur accès plus restreint aux canots de sauvetage.

Les visualisations graphiques confirment que certaines variables (classe, sexe, âge) sont potentiellement discriminantes pour prédire la survie.

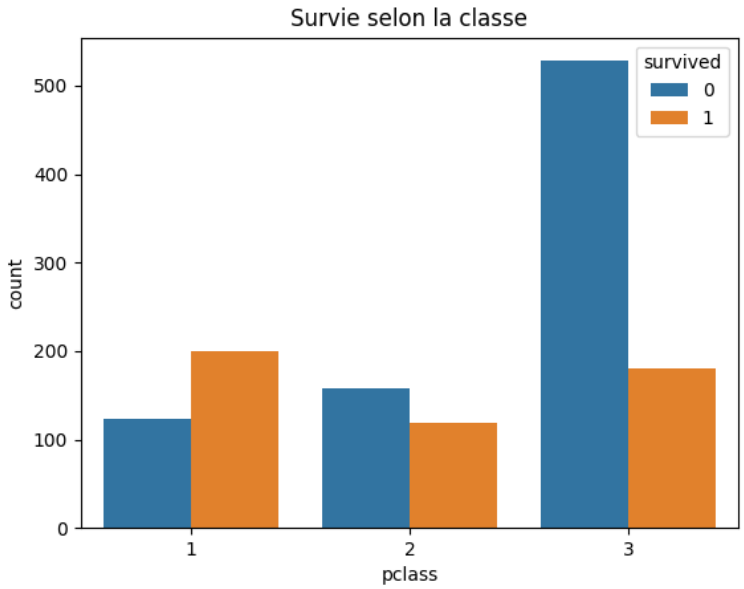

FIGURE 1 – Répartition des survivants selon la classe

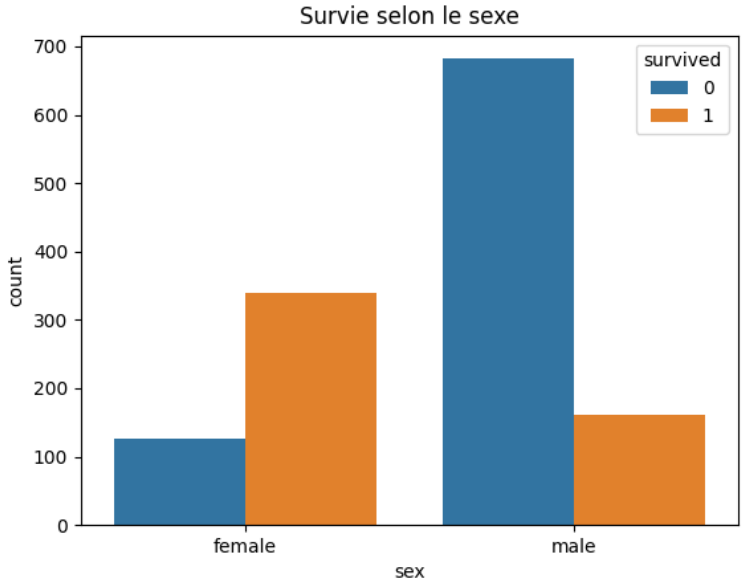

FIGURE 2 – Répartition des survivants selon le sexe

3. Prétraitement

Le prétraitement a consisté à rendre les données exploitables par les modèles de machine learning. Les étapes clés sont :

1. **Gestion des valeurs manquantes** :
   — `age` : remplacé par la médiane (stratégie robuste aux extrêmes),
   — `fare` : remplacé par la médiane,

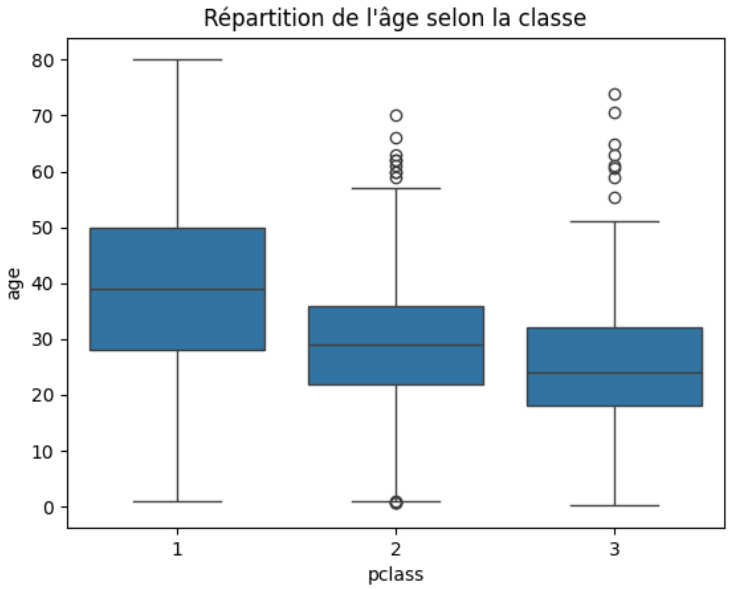

FIGURE 3 – Répartition de l'âge selon la classe

— `embarked` : remplacé par la valeur la plus fréquente (mode).

2. **Encodage des variables catégorielles** :
   — `sex` a été encodé en 0 (homme) et 1 (femme),
   — `embarked` a été transformé en variables indicatrices via `get_dummies`.
3. **Sélection des variables pertinentes** :
   — Ont été conservées les variables : `pclass`, `sex`, `age`, `sibsp`, `parch`, `fare`, `embarked_Q`, `embarked_S`.

Remarque : Une normalisation n'a pas été appliquée car les modèles utilisés (régression logistique, random forest) sont peu sensibles aux échelles de valeurs.

4. Modélisation et évaluation

Deux modèles ont été testés :
— **Régression logistique** : simple et interprétable, utilisée pour les variables binaires,
— **Random Forest** : méthode d'ensemble plus robuste aux variations des données.
Les modèles ont été entraînés sur 80% des données et testés sur 20%.
Résultats :
— Régression logistique : 77% de précision
— Random Forest : 77% également

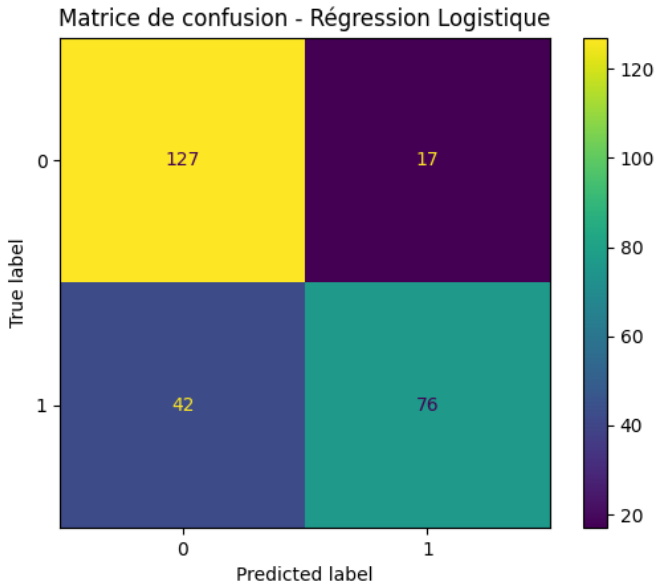

FIGURE 4 – Matrice de confusion - Régression Logistique : quelques erreurs de classification subsistent.

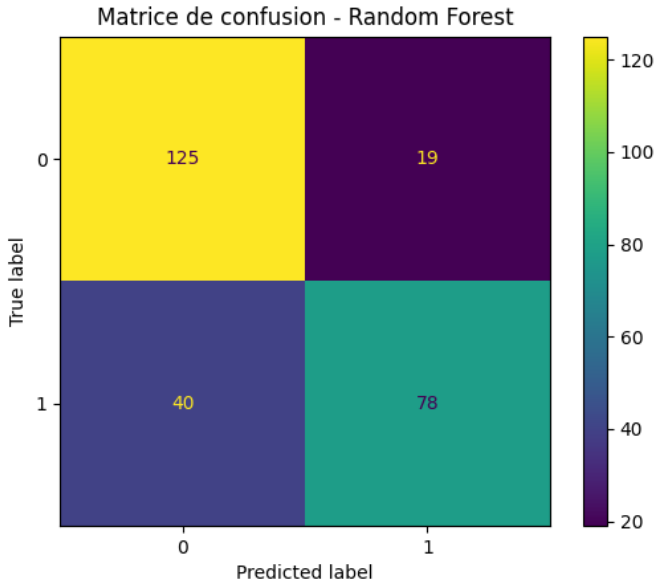

FIGURE 5 – Matrice de confusion - Random Forest : performances similaires mais plus stable.

5. Conclusion et pistes d'amélioration

Ce mini-projet a permis d'appliquer les étapes de construction d'un modèle d'apprentissage supervisé. Les résultats montrent que les femmes, les enfants et les passagers de première classe ont eu plus de chances de survivre. Pour aller plus loin, on pourrait ajouter d'autres variables (nom...) ou explorer des modèles avancés (Gradient Boosting, SVM).