Université de Sherbrooke

DEPARTEMENT D'INFORMATIQUE

Projet de forage de données

Titanic - Forage de données à partir d'une catastrophe

Partie 1 : Données

Auteurs:

Membres de l'équipe:
AIT LHAJ, Walid
BOUHAMIDI EL ALAOUI, Kaoutar
RAZAFINDRAMISA, Andrianihary
THOMAS, Eliott

CIP:
aitw2501
bouk1001
raza3902
thoe2303

Superviseur:

Nadia Tahiri

Hiver 2022

Table des Matiéres

1 Introduction

1.1 Contexte général : Domaine d'étude

Le naufrage du paquebot Titanic dans l'océan Atlantique Nord est l'un des naufrages les plus tristement célèbres de l'histoire. Le 15 avril 1912, lors de son premier voyage, le Titanic a coulé après être entré en collision avec un iceberg. Malheureusement, il n'y avait pas assez de canots de sauvetage pour tout le monde à bord, ce qui a entraîné la mort de 1502 des 2224 passagers et membres d'équipage. Cet évènement est l'une des plus grandes catastrophes maritimes de l'époque.

Cette tragédie a choqué la communauté internationnale et a mis en évidence l'insuffisance des règles de sécurité de l'époque. Elle a mené les différentes parties prenantes dans le domaine de transport touristique et maritime à instaurer de meilleures mesures de régulations pour leurs navires, notamment dans les procédures d'évacuation d'urgence.

1.2 Description du projet et problématique

La tragédie du Titanic a causé beaucoup de morts et a entrainé des modifications des régles de voyages maritimes. Bien qu'il y ait eu un élément de chance dans la survie, il semble que certains groupes de personnes étaient plus susceptibles de survivre que d'autres. C'est dans ce cadre que se présente notre projet de session de forage de données.

Le but du projet est de faire la classification des personnes qui étaient à bord du paquebot pour savoir si elles ont survécu ou non en fonction de leurs données socio-économiques, ceci en utilisant des modèles de classification. Pour ce faire, notre projet est divisé en trois phases principales: une phase de manipulation des données (transformation, nettoyage), une phase d'implémentation d'algorithmes et une phase d'interprétations des résultats.

Notre problématique s'inscrit donc comme suit: "Quelles sont les catégories de personnes les plus susceptibles de survivre à la catastrophe du Titanic?"

1.3 Importance du sujet et motivations

Nos motivations pour choisir ce sujet ont été les suivantes :

- Walid : Je n'ai pas encore regardé Titanic donc c'est ma chance de savoir ce qui s'est passé de manière scientifique.
- Kaoutar: L'histoire du film 'Titanic' a éveillé ma curiosité pour savoir comment le processus du sauvetage s'etait passé et sur quels critères ils se sont basés pour choisir les personnes à mettre sur les canaux de sauvetage .
- Andrianihary : Cette compétition est légendaire et il s'agit du premier défi à effectuer avant de commencer les compétitions d'apprentissage automatique sur Kaggle. C'est un excellent point de départ dans la science des données et, je trouve, qu'elle doit être faite par tous ceux qui veulent devenir scientifique des données.
- Eliott : Je trouve que ce sujet est l'équivalent du 'Hello World' dans le site de Kaggle et dans le domaine du forage de données en particulier. Il n'en demeure pas moins intéressant et challengeant de par sa grande popularité.

2 Description des données

Les données contiennent des informations sur les passagers comme leur nom, leur classe socio-économique, leur genre, leur âge ou leur port d'embarquement. Ces données sont divisées en deux groupes :

- L'ensemble "train.csv" pour créer les modèles ;
- Le jeu de test "test.csv" pour voir dans quelle mesure le modèle fonctionne sur la prédicition de survie de passagers non classés.

Feature	***Valeur***	***Type***
PassengerId	1-891	Entier
Survived	0,1	Entier
Pclass	1-3	Entier
Name	Nom des passagers	Objet
Sex	Male, female	Objet
Age	0-80	Réel
SibSp	0-8	Entier
Parch	0-9	Entier
Ticket	numéro du ticket	Objet
Fare	0-512	Réel
Cabin	numéro de cabine	Objet
Embarkment	S, C, Q	Objet

Table 1: Description des Données du Dataset

2.1 Description des attributs

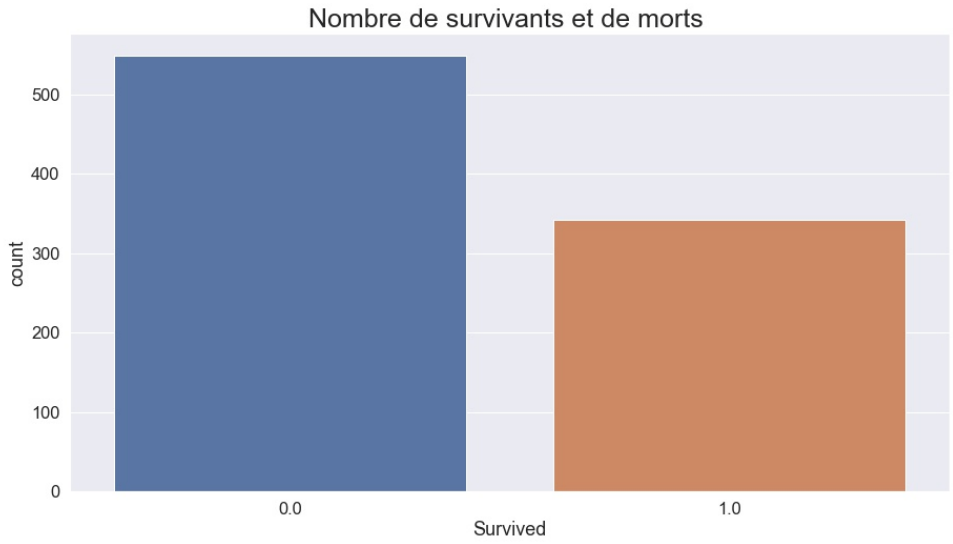
Nous allons à présent détailler les caractéristiques de l'ensemble de données d'entrainement et de test combinées, leurs statistiques et leur visualisation :

2.1.1 PassengerId

Cette caractéristique donne l'ID unique du passager. Elle va de 1 à 891.

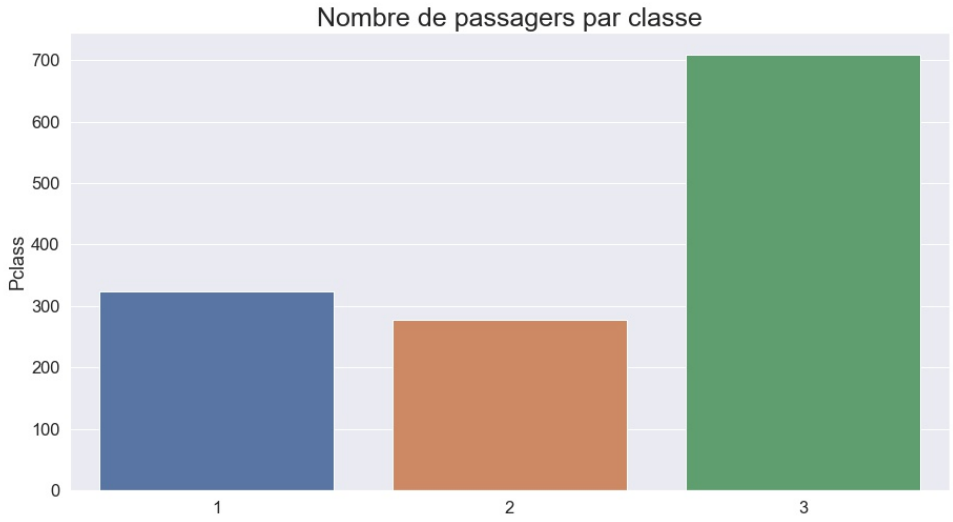
2.1.2 Survived

Cette caractéristique précise si le passager a survécu ou non. "0" signifie que le passager est mort et "1" signifie qu'il a survécu. Sur tous les passagers des données d'entrainement, 549 sont morts et 342 ont survécu. C'est ce que représente la figure suivante .

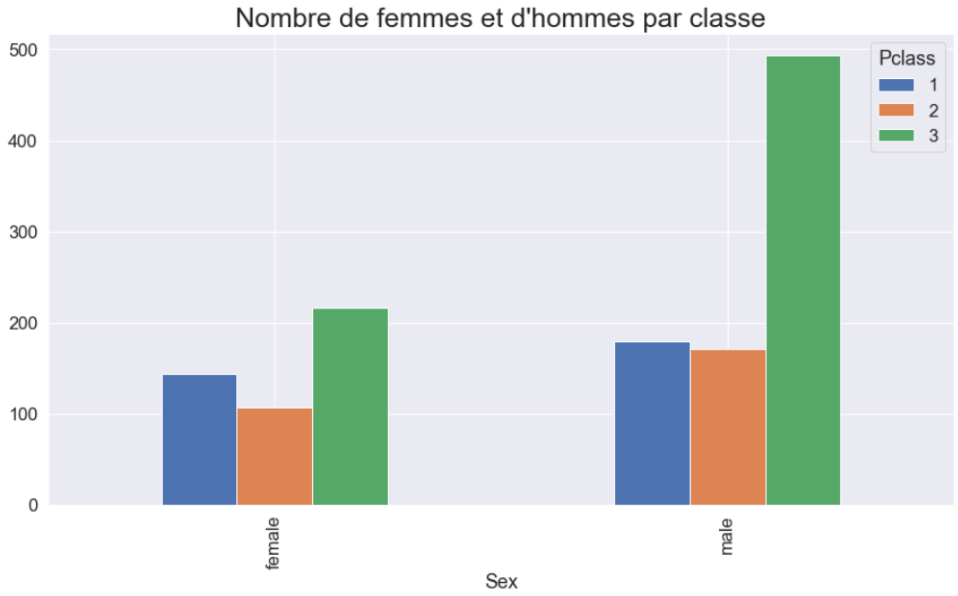


2.1.3 Pclass

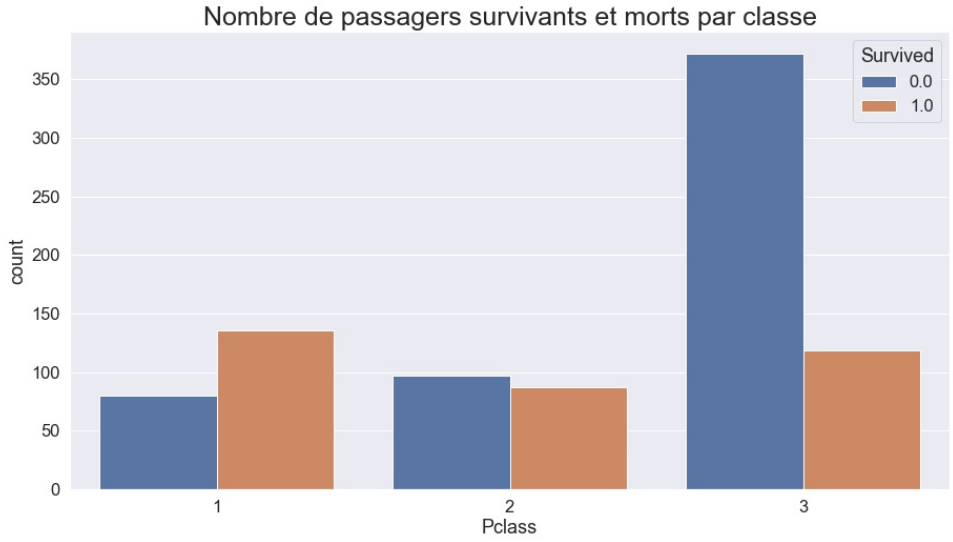
Cette caractéristique donne la classe du passager. "1" signifie que le passager était en 1ère classe (323 passagers), "2" en 2e classe (277 passagers) et "3" en 3e classe (709 passagers). C'est un indicateur du statut socio-économique du passager. La figure suivante représente le nombre de passagers par classe.



On représente également le nombre de femme et d'homme par classe.



Les passagers avec le taux de survie le plus élevé sont les passagers les plus riches c'est à dire les passagers de la première classe , c'est ce que la figure en dessous représente.

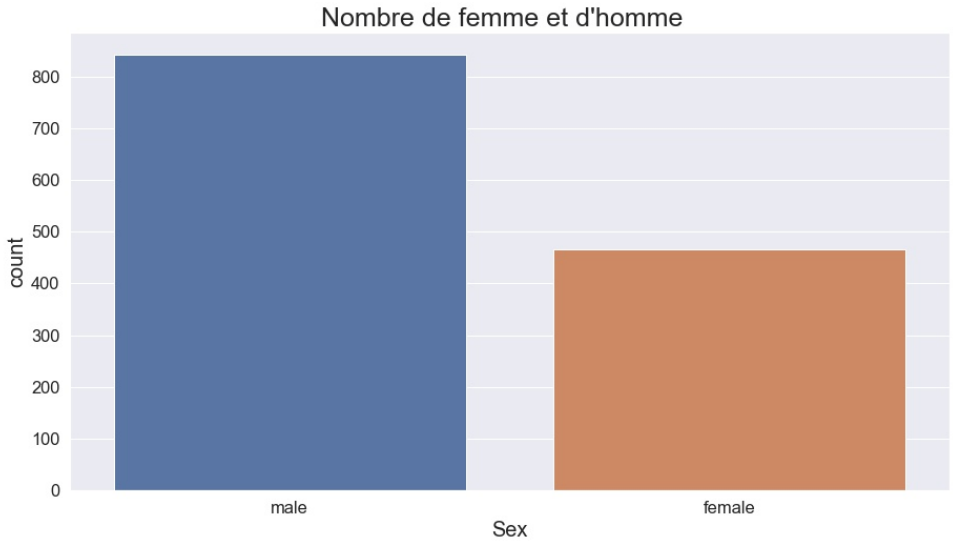


2.1.4 Name

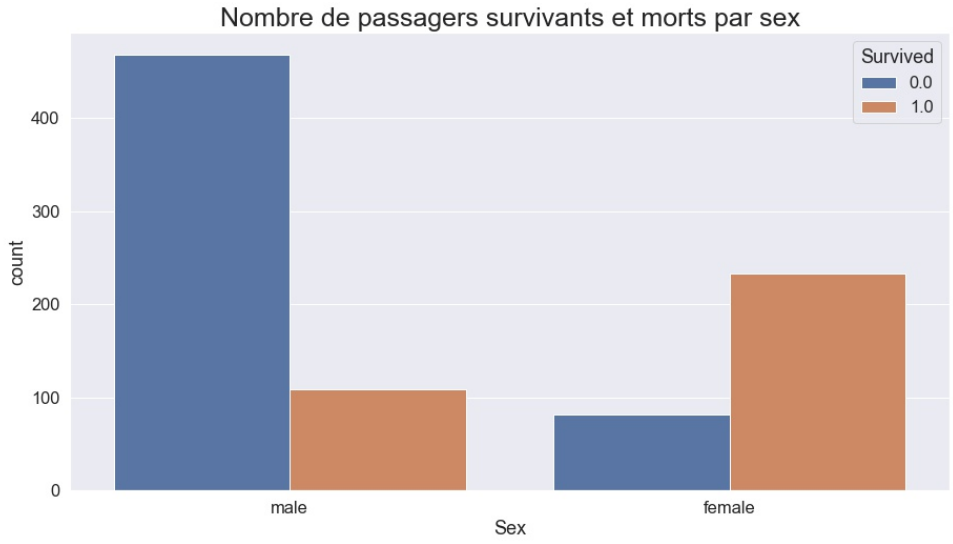
Cette caractéristique donne le nom de chaque passager. Ce nom est unique.

2.1.5 Sex

Cette caractéristique donne le genre du passager. Il y a deux catégories "Male" et "Female". 466 passagers sont des femmes (35,59 % des passagers) et 843 sont des hommes (64,40 % des passagers).



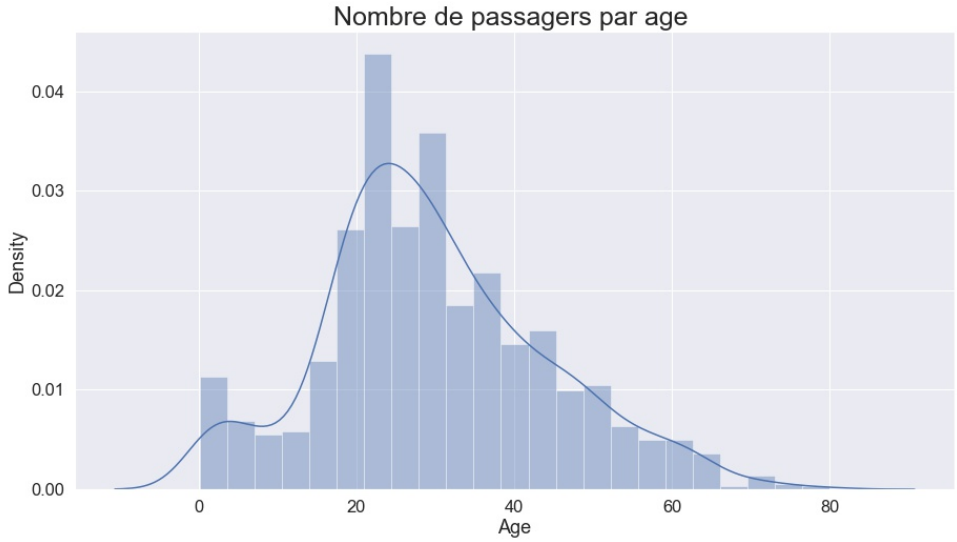
Cela nous permet d'obtenir le taux de survie pour les données d'entrainement 50% des femmes ont survécu (233 passagères), contre seulement 12.9% des hommes (109 passagers).



2.1.6 Age

Cette caractéristique précise l'âge du passager. L'âge des passagers va de 0 à 80 ans et l'âge moyen est de 29.88 ans. La majorité des passagers ont entre 20 et 40 ans, c'est ce que représente la boite à moustache en dessous.

On visualise également le nombre de passager par âge.

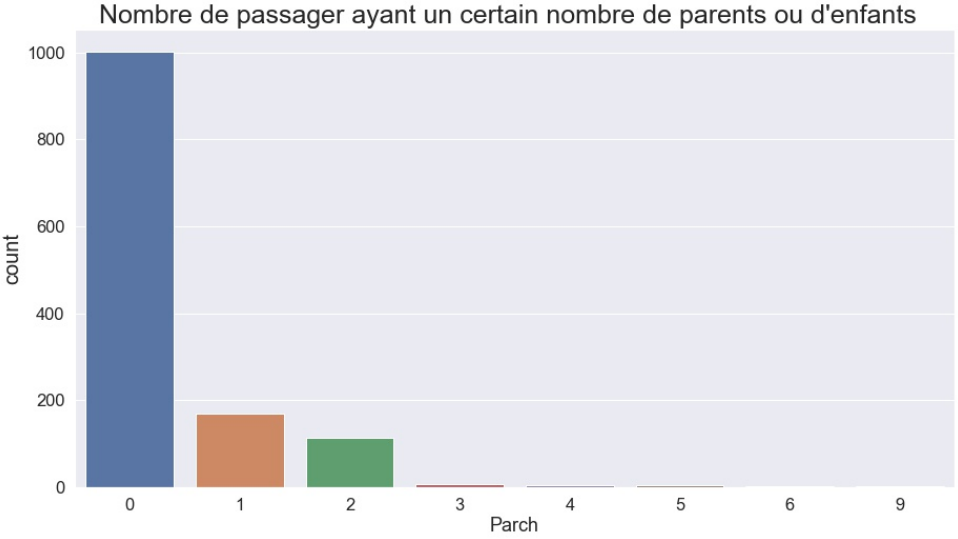


2.1.7 SibSp

Cette caractéristique donne le nombre de frères et sœurs (les demi-frères et demi-sœurs sont pris en compte) et de conjoints (les fiancés sont ignorés) à bord du Titanic. Elle va de 0 à 8. 891 passagers n'ont pas de frères et sœurs ou de conjoints à bord, et 418 en ont au moins un. Ceci est montré dans la figure ci-dessous.

2.1.8 Parch

Cette caractéristique donne le nombre de parents/enfants du passager à bord, les belles-filles et les beaux-fils sont compris dans ce nombre. Ce nombre va de 0 à 9. 1002 passagers n'ont pas de parents ou d'enfants à bord.



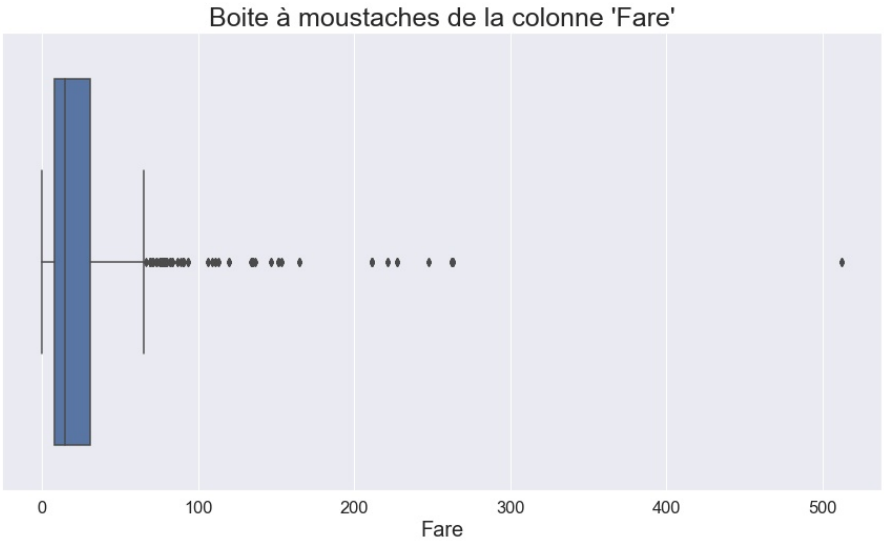
2.1.9 Ticket

Cette caractéristique précise le numéro de billet du passager. Ticket a 681 valeurs uniques.

2.1.10 Fare

Cette caractéristique précise le prix payé par chaque passager pour son ticket et ce montant est compris entre 0 et 512. La majorité des passagers ont payés entre 7.89 et

31.27, c'est ce que la boite à moustaches en dessous montre .

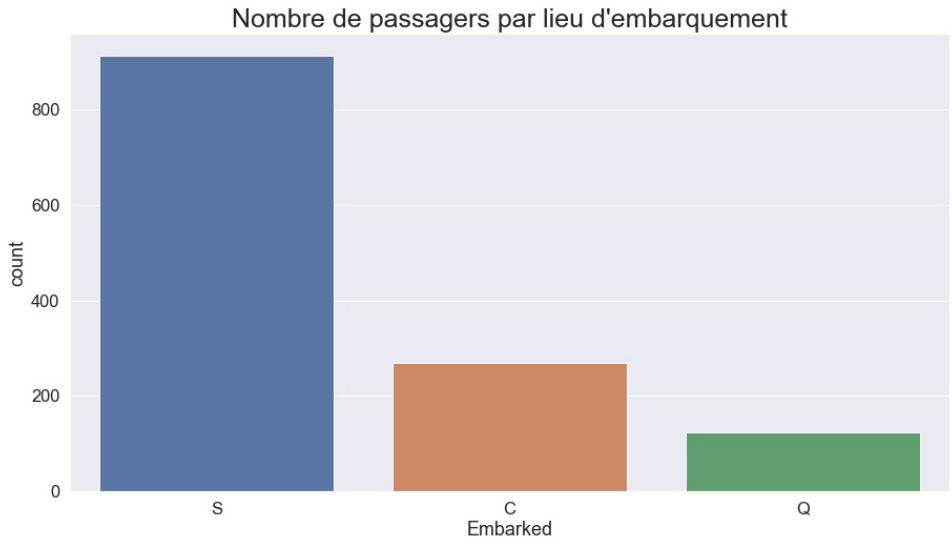


2.1.11 Cabin

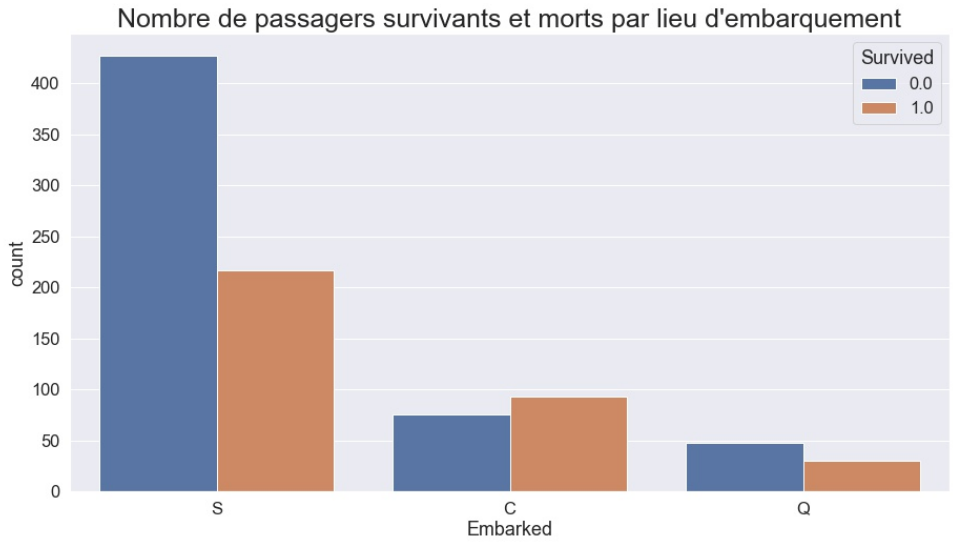
Cette caractéristique donne le numéro de cabine du passager. On ne connaît pas ce numéro pour 1014 passagers.

2.1.12 Embarked

Cette caractéristique donne le port d'embarquement du passager. "C" signifie Cherbourg, "S" signifie Southampton et "Q" signifie Queenstown. 914 des passagers ont embarqué à Southampton contre 270 à Cherbourg et 123 à Queenstown .

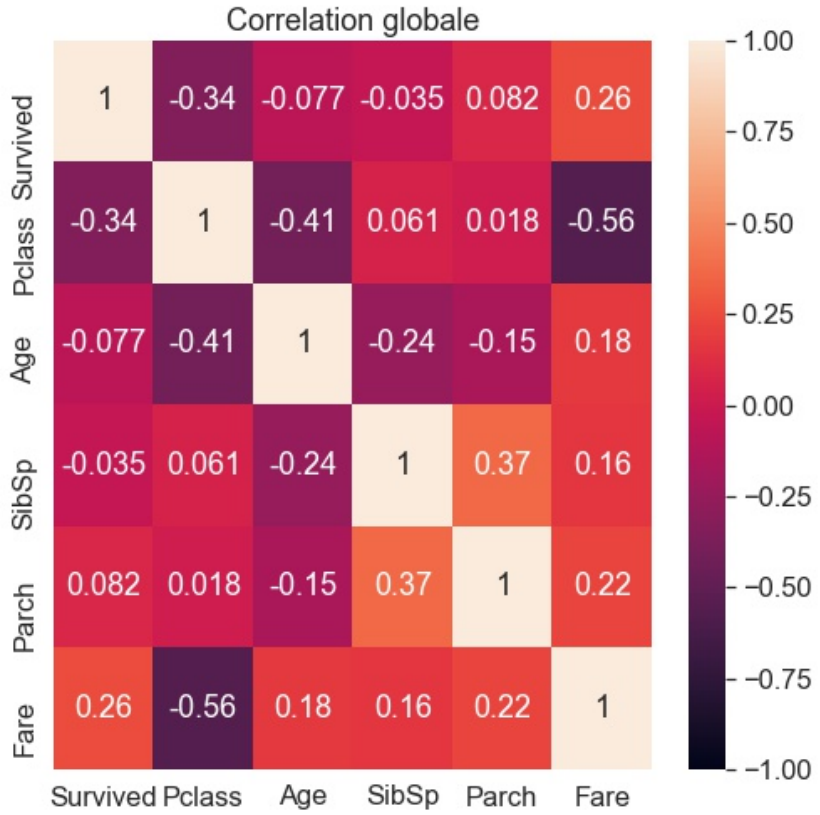


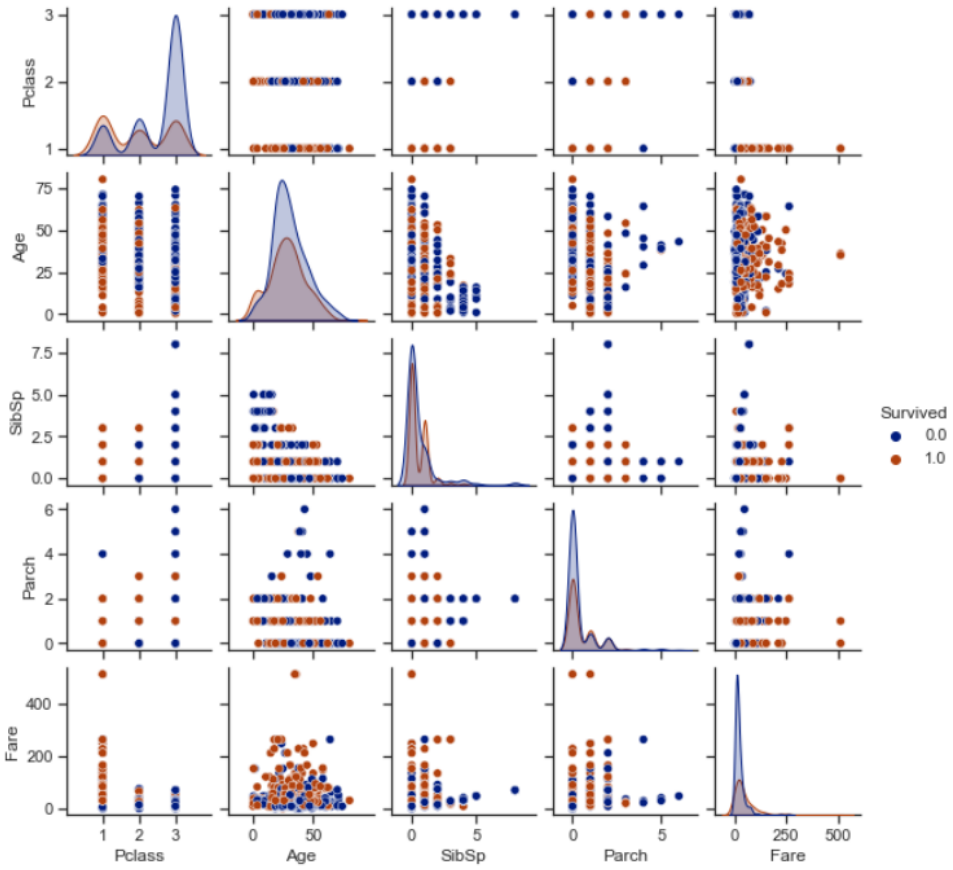
La figure suivante montre que les personnes qui ont embarqué à Cherbourg ont le taux de survie le plus élevé.



2.2 Visualisation Globale

Pour finir, nous allons vous présenter une visialisation globale des données avec une matrice de corrélation.





Grace au premier graphique, nous seront en mesure de savoir quels attributs privilégier dans la partie 2 par rapport aux autres. Typiquement *Pclass* semble avoir un impact bien plus important sur la survie que *SibSp*.

Le second graphique, le pairplot, permet également (entre autres) de mettre en lumiere l'importance de la classe du passager dans sa survie. En effet la colonne "classe 3" est majoritairement de couleurs bleue, indiquant que les personnes moins favorisées économiquement étaient également moins susceptibles de survivre.

3 Historique des travaux et développements faits en rapport avec le sujet

Le dataset Titanic est un Dataset classique des sciences de données. On peut s'en rendre compte facilement en regardant le nombre de résultat fourni par Google Scholar avec '*dataset Titanic*' : environ 10000 ! Nous avons isolé 3 papiers de recherche sur le sujet que vous pourrez retrouver dans la section References :

- [2] qui applique 14 techniques de machine Learning au dataset.
- [3] qui utilise des outils d'exploration de données modernes (Weka) pour trouver une relation convaincante entre la survie des passagers et leurs caractéristiques.
- [4] qui est un récent article de 2020 reprenant les techniques classiques de ML et visant une fois de plus à prédire le sort des passagers basé sur leurs caractéristiques.

Nous allons vous présenter une liste non exhaustive des techniques et algorithmes présentés dans les articles ci-dessus. Pour chacune d'entres elles nous feront une brève descrition explicative ainsi que la liste des article mentionnant cette technique.

- Regression logistique : C'est un algorithme de classification qui fonctionne sur des données discrètes. Il est basé sur l'équation de la sigmoid :

  $$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

  Cette approche est utilisée par [4]

- Machine à Support de Vecteur (SVM) : C'est un algorithme qui permet de séparer les données grace à un hyperplan. Cette technique fonctionne mieux sur les petits ensembles de données. Cette approche est utilisée par [2] et [4]

- Arbre décisionnel : génère un arbre de décision suggérant un chemin de classification et donc les caractéristiques les plus significatives en ce qui concerne la survie. L'intéret majeur de cette approche est qu'elle a une très bonne interprétabilité. Cette approche est utilisée par [2], [3] et [4].

- Simple K Means Cluster Analysis : regroupe les données à partir d'associations simples et permet d'analyser visuellement les clusters au sein de l'ensemble de données. Cette approche est utilisée par [3].

- K plus proches voisins (KNN) : un des algorithmes de classification les plus courants, utilise les métriques de distance pour mesurer la proximité entre les échantillons d'apprentissage et l'échantillon de test. Il attribue à l'échantillon de test la classe de ses k échantillons d'apprentissage les plus proches. Il est principalement basé sur la distance euclidienne. Cette approche est utilisée par [2] et [4].

- Bayes Naïf (NB) : basé sur le théorème de Bayes en supposant que toutes les caractéristiques sont indépendantes compte tenu de la valeur de la variable de classe, NB permet une classification efficace et rapide. NB fonctionne bien sur des ensembles de données complexes de grande dimension. Cette approche est utilisée par [2].

- Bagging : technique pour créer un ensemble de classificateurs, améliore la précision en rééchantillonnant l'ensemble d'apprentissage. Un classificateur unique de base est appliqué aux ensembles d'apprentissage générés puis les modèles de classification générés sont combinés en fonction du vote à la majorité. Cela permet de réduire la variance et le biais. Cette approche est utilisée par [2].

- AdaBoost : est une méthode d'apprentissage d'ensemble. La notion de base est qu'un classificateur fort peut être créé en combinant linéairement un certain nombre de classificateurs faibles. AdaBoost augmente le poids des points de données mal classés tout en diminuant les poids des points de données correctement classés ce qui permet de repondérer toutes les données d'entraînement à chaque itération. Les classificateurs faibles sont appliquées en série, puis les modèles de classification générés sont combinés en fonction du vote à la majorité pondérée. Cette approche est utilisée par [2].

- Extra Trees : méthode de classification d'ensemble d'arbres de décision basée sur la randomisation. Pour chaque noeud de l'arbre, des règles de fractionnement sont tirées au hasard, puis la règle la plus performante est associée à ce noeud. Cette approche est utilisée par [2].

- Forêt aléatoire : algorithme de classification qui utilise un ensemble de prédicteurs d'arbres, chaque arbre est construit en amorçant les données d'apprentissage et, pour chaque fractionnement, un sous-ensemble de caractéristiques sélectionné au hasard est utilisé. Cette méthode estime les données manquantes tout en conservant l'exactitude. Cette approche est utilisée par [2].

- Gradient Boosting : Cette méthode utilise le bossting pour estimer des fonctions. Cette approche est utilisée par [2].

- ANN et MLP. Les réseaux de neuronnes artificiels (ANN) et les Perceptrons multicouches (MLP) permettent de résoudre des problèmes de classification non linéaires avec une très bonne présision. Cette approche est utilisée par [2].

- Vote : On peut aggreger ensembles de nombreux modèles différents pour obtenir une classification de meilleure qualité. On peut faire voter les modèles uniformément. On peut également pondérer l'impact de certains modèles si on juge l'importance de ces derniers non équilibrée. Cette approche est utilisée par [2].

References

[1] Walid AIT LHAJ, Kaoutar BOUHAMIDI EL ALAOUI, Andrianihary RAZAFINDRAMISA, and Eliott THOMAS. Dépot git du projet. https://github.com/bouhamidi/projet_fdd.

[2] Neytullah Acun Ekin Ekinci, Sevinç İlhan Omurca. A comparative study on machine learning techniques using titanic dataset. https://www.researchgate.net/profile/Neytullah-Acun/publication/324909545_A_Comparative_Study_on_Machine_Learning_Techniques_Using_Titanic_Dataset/links/607533bc299bf1f56d51db20/A-Comparative-Study-on-Machine-Learning-Techniques-Using-Titanic-Dataset.pdf.

[3] Lauren Clarke Shawn Cicoria John Sherlock, Manoj Muniswamaiah. Classification of titanic passenger data and chances of surviving the disaster. https://arxiv.org/ftp/arxiv/papers/1810/1810.09851.pdf.

[4] Lauren Clarke Rajni Sehga Karman Sing, Renuka Nagpa. Exploratory data analysis and machine learning on titanic disaster dataset. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9057955.