Résumé des méthodes par règles d'associations :

Table des matières

1. Introduction

Le datascientist, dans son travail doit réaliser plusieurs types d'étude sur les jeux de données qui l'intéresse. Parmi elles, les règles d'associations permettent, d'associer des données qualitatives afin de déterminer si elles vont naturellement ensemble.

Nous verrons donc tout d'abord la réalisation de l'import des données puis la manière dont un data analyste réalise des règles d'associations et en tire des conclusions.

Dans le cadre de ce TP, les packages « Pandas », « sklearn », « Numpy », « Seaborn », « mlxtend » et « Matplotlib » ont été utilisés. Les jeux de donnée Titanic.csv, représentant les données sur les naufragées du Titanic, composera la base de notre étude.

2. Import des données

Afin d'apporter les données à étudier dans python, on utilise le package Pandas qui simplifie la création et la manipulation de dataframe.

Suite à l'import des packages,

```
# Import des packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
import seaborn as sns
```

Il suffit d'utiliser read_csv, une fonction de pandas, pour lire un csv et le récupérer sous forme de dataframe. Exemple :

```
# import des données
df = pd.read_csv('Data/Titanic.csv', sep=';')
df.head()
```

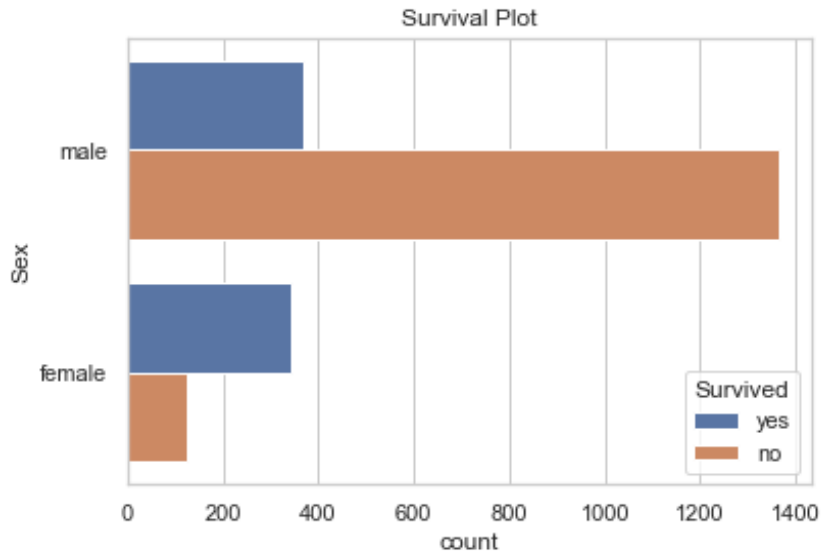
	Class	Age	Sex	Survived
0	first	adult	male	yes
1	first	adult	male	yes
2	first	adult	male	yes
3	first	adult	male	yes
4	first	adult	male	yes

3. Étude de cas : Naufrage du Titanic

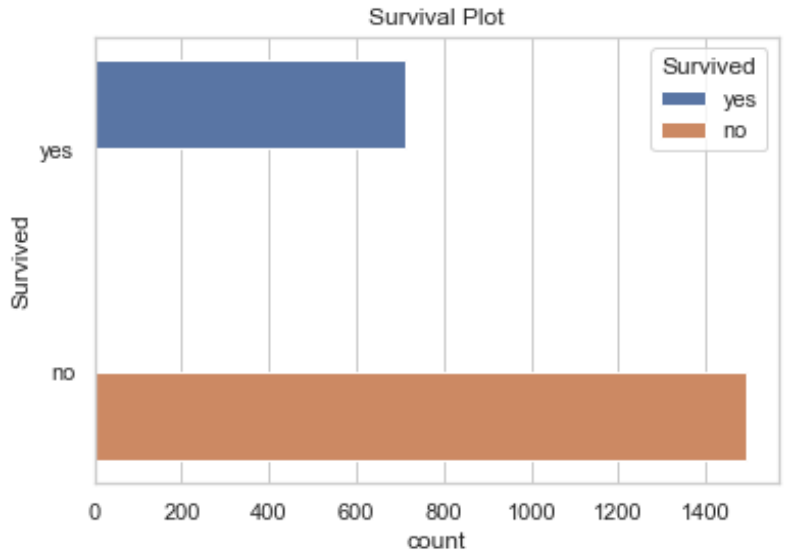
Nous étudions dans le cadre de cet exercice, les données sur les naufragés du Titanic (Classe, Sexe, Age, et si la personne à survécu). Toutes les données sont qualitatives.

3.1. Graphiques

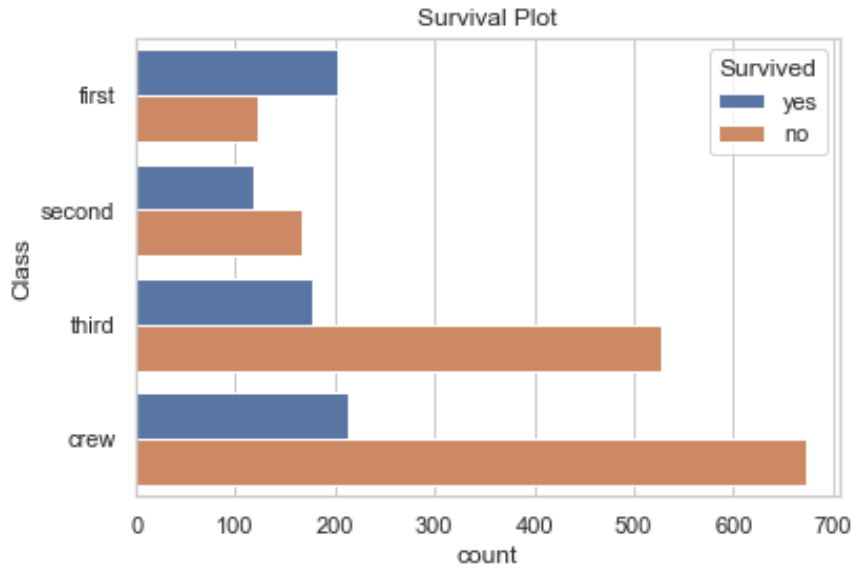
Nous commençons par réaliser un rendu visuel des données avec une analyse bivarié :



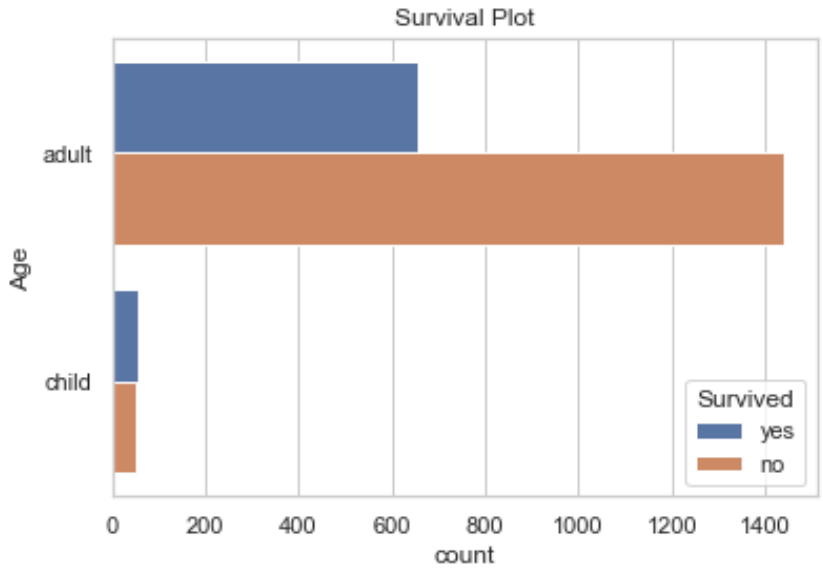
Ce graphique nous montre que plus trois quarts des femmes ont survécus lorsque cinq septièmes des hommes ont péris.



En totalité seulement le quart des membres du paquebot ont survécus.



On peut voir ici qu'une majorité des membres d'équipages n'ont pas survécus au naufrage lorsque plus de la moitié des premières classes ont survécus.



D'après ce diagramme, la moitié des enfants ont pu survivre lorsque les trois quarts des adultes sont morts

3.2. Règles d'associations

Le package mlxtend nous permet de réaliser les règles d'associations rapidement :

```
# Convertion du dataset en One_hot_encoding (0,1 pour les datas)
df_encoded = pd.get_dummies(df)

# application de l'algo apriori pour obtenir les itemsets fréquents
frequent_itemsets = apriori(df_encoded, min_support=0.5, use_colnames=True)

# Dérivé des règles d'association à partir des itemsets fréquents
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
print(rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']])
```

Nous obtenons cette sortie python comme résultat :

```
                 antecedents                 consequents   support  confidence  \
0                  (Sex_male)                (Age_adult)  0.757383    0.963027
1                (Age_adult)                 (Sex_male)  0.757383    0.796845
2              (Survived_no)                (Age_adult)  0.653339    0.965101
3                (Age_adult)              (Survived_no)  0.653339    0.687380
4                  (Sex_male)              (Survived_no)  0.619718    0.787984
5              (Survived_no)                (Sex_male)  0.619718    0.915436
6    (Sex_male, Survived_no)                (Age_adult)  0.603816    0.974340
7      (Sex_male, Age_adult)              (Survived_no)  0.603816    0.797241
8   (Survived_no, Age_adult)                (Sex_male)  0.603816    0.924200
9                  (Sex_male)  (Survived_no, Age_adult)  0.603816    0.767764
10             (Survived_no)    (Sex_male, Age_adult)  0.603816    0.891946
11               (Age_adult)   (Sex_male, Survived_no)  0.603816    0.635277

        lift
0   1.013204
1   1.013204
2   1.015386
3   1.015386
4   1.163995
5   1.163995
6   1.025106
7   1.177669
8   1.175139
9   1.175139
10  1.177669
11  1.025106
```

Ce tableau nous permet d'affirmer que les hommes sont présents dans 75% des cas. De plus si le passager est un homme, il y a 96% de chance que ce soit un adulte. Le lift est très proche de 1 mais supérieur donc le conséquents et l'antécédents sont un peu plus susceptible d'être rencontré ensemble plutôt que séparément.