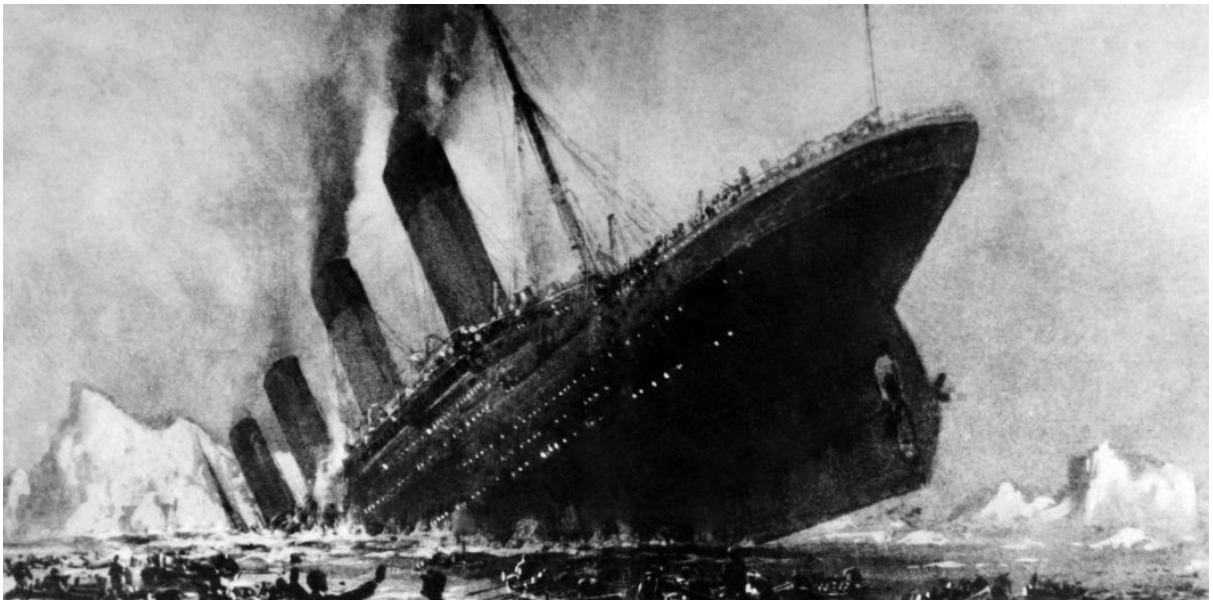
LE MAUFF Julien
NGUYEN Duong

04/12/2019

Master 2 - Economie Appliquée

Titanic : Machine Learning from Disaster - Logiciels Statistiques



Toulouse School of Economics

Année universitaire 2019-2020

Introduction

Le 15 avril 1912, à la suite d'une collision avec un iceberg, le Titanic, célèbre paquebot transatlantique qui réalisait alors son voyage inaugural en reliant Southampton à New York, sombra dans les profondeurs de l'océan Atlantique. Ce naufrage entraîna alors la mort de 1502 personnes (sur les 2224 présentes), et reste à ce jour l'une des plus grandes catastrophes maritimes de l'histoire. Ce grand nombre de décès est surtout dû au nombre limité d'embarcations de sauvetage présentes, mais également aux carences dans les procédures d'évacuation d'urgence. Mais alors avec les données concernant les passagers, pouvions-nous prédire leur chances de survie de manière individuelle ? C'est en tout cas ce que nous allons essayer de déterminer dans ce projet. Ainsi dans un premier temps, nous réaliserons une analyse descriptive des données que nous avons à disposition, puis dans un second temps, nous envisageons de déduire la probabilité de survie des passagers par des méthodes statistique d'apprentissage automatique.

I. Phase exploratoire

Nous commencerons notre étude par une phase d'exploration des données, étape essentielle afin de comprendre la signification des différentes variable ainsi que leur construction. Suite à cette étape, nous transformerons nos données dans le but de les rendre utilisables par les algorithmes statistiques de prédiction par apprentissage automatique. Enfin, nous réaliserons une phase d'analyse statistique descriptive qui nous permettra de mettre en avant les features les plus importants, ainsi que les différentes corrélation entre les différentes variables.

A. Analyse et transformation de la base de donnée

Lors de la phase d'exploration des données, nous remarquons certains points à améliorer ou à corriger pour la suite de notre étude. Dans un premier temps, en observant la variable *Name*, nous remarquons qu'il est possible d'extraire les titres de civilité des passagers indépendamment de leurs noms et prénoms. Au total, 18 titres de civilité différents peuvent être attribués: *Mr, Mrs, Miss, Master, Don, Rev, Dr, Mme, Ms, Major, Lady, Sir, Mlle, Col, Capt, the Countess, Jonkheer et Dona*. Toutefois aux vues de la diversité trop importante de ces titres, nous décidons de les regrouper sous 6 titres de civilité généralisés, à savoir: *Mr, Mrs, Miss, Master, Nobility et Officer*. Les titres de civilités correspondant à chaque passager sont ensuite regroupés dans la variable «*Title*».

Un travail similaire peut également être effectué en ce qui concerne la variable *Cabin* pour laquelle un deck peut être extrait indépendamment des numéros de cabines. Nous attribuons donc la valeur U lorsque la variable *Cabin* prend la valeur "Unknown" pour les passagers dont la dénomination du deck et les numéros de cabines sont manquants dans la base de donnée. Par ailleurs, nous considérons que les numéros des cabines des passagers ne sont pas des informations pertinentes, raison pour laquelle la variable

correspondant aux numéros de cabines pour chaque passager a été supprimée. Nous préférons en effet considérer que seule la dénomination du deck est relevante dans notre situation. Les decks correspondants à chaque passager sont donc stockés dans la variable "*Deck*".

Par la suite, notre analyse exploratoire a également mis en évidence qu'une variable dérivée représentant la taille de la famille peut être créé afin d'apporter davantage de précision dans nos analyses statistiques et prédictives. Nous créons donc la variable dérivée «*Famsize*» à partir des variables *SibSp* (nombre de frères, soeurs et épouses à bord du Titanic) et *Parch* (nombre de parents et enfants à bord du Titanic). Nous définissons 3 catégories différentes pour cette variable:

- solo : individus voyageant seul à bord du titanic
- small family : individus voyageant avec deux à quatre personnes de sa famille à bord du titanic (frères, soeurs, enfants, parents et épouses/époux compris).
- big family : individus voyageant avec plus de quatre personnes de sa famille à bord du titanic (frères, soeurs, enfants, parents et épouses/époux compris)

Enfin, nous constatons la présence de valeurs manquantes pour plusieurs de nos variables, en particuliers pour l'âge (*Age*), le port d'embarcation (*Embarked*), le prix du ticket (*Fare*) et les cabines respectives des passagers (*Cabin*). Le traitement de ces données manquantes est alors une étape importante dans la modification de notre base de donnée, et doit être réalisée rigoureusement en appliquant des méthodes adaptées selon les variables. Ainsi en premier lieu, nous avons la variable *Fare* qui correspond au prix du ticket acheté par chaque individu. Il n'y a qu'une seule valeur manquante pour cette variable, ce pourquoi nous remplaçons simplement la valeur manquante par sa médiane. Ensuite concernant la variable "Embarked", il n'y a une nouvelle fois que 2 valeurs manquantes. Nous remplaçons donc simplement les ports manquant par le port le d'embarcation le plus commun, à savoir Southampton (S). Enfin pour la variable *Age,* qui quant à elle présente un nombre important de valeur manquantes, nous ne pouvons pas simplement remplacer les âges manquants de certains individus par la médiane ou la moyenne. D'autant plus que cette variable peut être considérée comme primordiale dans l'analyse statistique et prédictive. Par conséquent, plusieurs méthodes ont été testées afin de déterminer l'âge de la manière la plus précise possible. Ainsi nous regroupons dans un premier temps les individus par titres de civilité, classe et sex du passager (voir variable *Title* détaillée ci-après) et nous remplaçons l'âge manquant de chaque individu par la médiane des âges des personnes possédant le même titre, le même sex et avec le même niveau de classe (1er classe, 2nd classe ou 3eme classe). Cette méthode permet ainsi d'approximer l'âge des individus de manière plus précise. Dans le tableau ci-dessous, nous pouvons observer les moyennes et les médianes d'âge perçu par titre de civilité ainsi que le nombre d'individus par titre dont l'âge est manquant.

Titre	Sexe	Classe des passagers	Moyenne d'âge	Médiane d'âge
Master	male	1	7,0	6
Master	male	2	2,8	2
Master	male	3	6,1	6
Miss	female	1	30,1	30
Miss	female	2	20,9	20
Miss	female	3	17,4	18
Mr	male	1	41,5	42
Mr	male	2	32,3	30
Mr	male	3	28,3	26
Mrs	female	1	42,9	45
Mrs	female	2	33,5	31
Mrs	female	3	32,3	31
Nobility	female	1	40,0	39
Nobility	male	1	42,3	40
Officer	female	1	49,0	49
Officer	male	1	51,1	52
Officer	male	2	40,7	42

Les âges manquants qui ont été approximés par regroupement de titre ont donc été stockés au sein de la variable “Age_replace”

Dans un second temps nous testons une nouvelle méthode en essayant de retrouver les âges manquants non pas par approximation mais à l'aide de méthodes statistiques prédictives d'apprentissages automatiques. Les deux méthodes utilisés pour prédire l'âge sont random forest et svm regressor. Les âges manquants qui ont été prédits par ces deux méthodes sont stockés dans les variables “Age_RandomForest” et “Age_SVM” respectivement. Cependant avant d'entamer la phase de prédiction des âges manquants nous devons transformer certaines variables afin qu'elles puissent être utilisables par les méthodes de prédictions statistiques. Ainsi, nous avons alors transformé la variable “pclass” en 3 variables binaire “Pclass_1”, “Pclass_2” et “Pclass_3” qui prennent respectivement
0 ou 1 en fonction de si l'individus était en première, seconde ou troisième classe. Le même travail est effectué pour la variable “Sex” qui devient “male” et “female”, ainsi que les variables “Title”, “Famsize”, “Embarked” et “Deck”.

Le tableau ci-dessous nous donne une description statistique des trois différentes méthodes utilisées pour retrouver les âges manquants.

	Age_Randomorest	Age_SVM	Age_replace
mean	28,6	27,0	26,8
std	8,0	8,5	8,1
min	6,0	0	6,0
25%	26,0	23,0	26,0
50%	28,0	27,0	26,0

Enfin, nous décidons de créer une variable dérivée de l'âge sous forme catégorique afin de faire des groupes d'individus se trouvant à l'intérieur d'une même tranche d'âge. Cette variable est appelée "Age_group" est prend comme valeur:

- 0_16 : les jeunes jusqu'à 16 ans (enfants et adolescents)
- 17_30 : les jeunes adultes de 17 à 30 ans
- 31_40 : les adultes de 31 à 40 ans
- over_40 : les adultes de plus de 40 ans

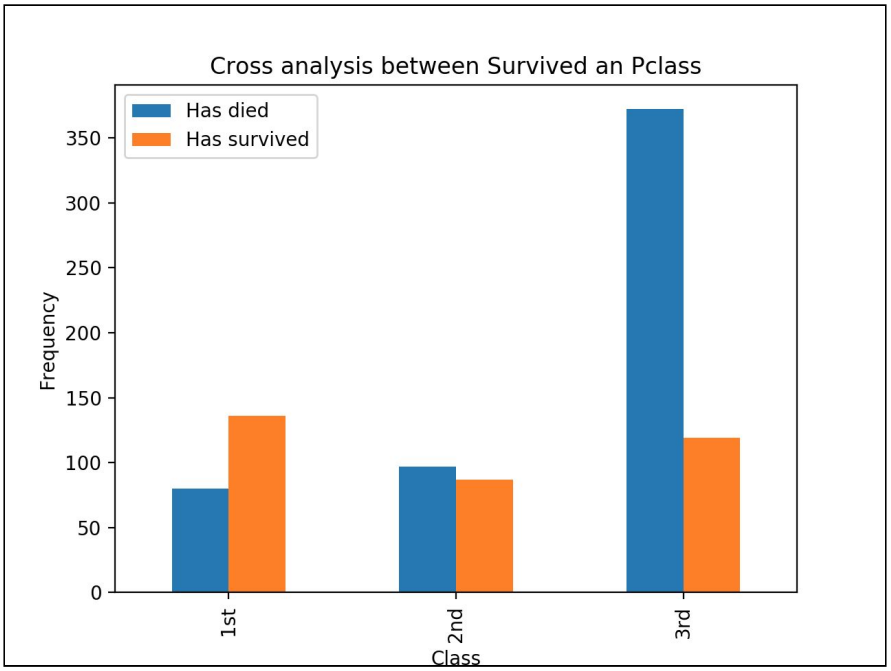
Notre dernière étape de cette phase d'exploration est le traitement de la variable *Ticket*.
Au premier abord, cette variable peut paraître peu importante puisque les valeurs sont très distinctes les une des autres. Cependant nous pouvons remarquer que certains tickets ont des préfix qui reviennent régulièrement en plus des différents nombres qui suivent ou précèdent ce préfix. De ce fait, nous allons procéder à une modification de la variable afin de la rendre utilisable par les algorithmes d'apprentissage automatique. Pour les tickets dont il n'y a pas de préfix, nous attribuons la valeur "XXX" par défaut.

Exemple de de ticket après traitement:

Avant traitement de ticket	Après traitement de ticket
A/4 48871	A4
330911	XXX

B. Analyse statistique & descriptive des données

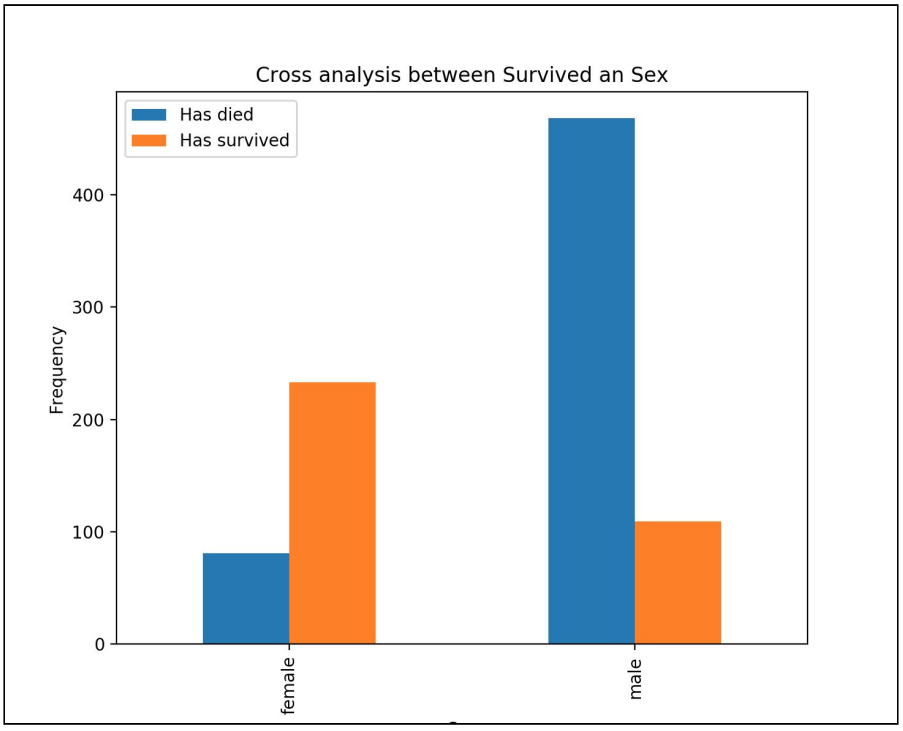
1. Analyse croisée entre la survie et le niveau de classe des passagers



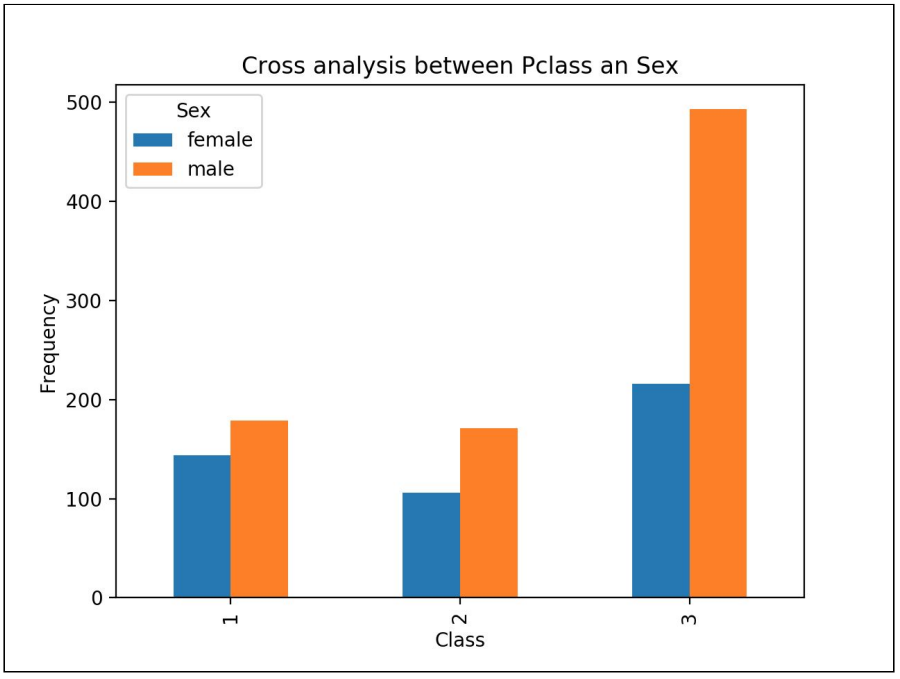
Dans un premier temps nous pouvons voir que parmi les individus en première classe, le nombre de personnes ayant survécu est supérieur au nombre de personnes décédées. En revanche s'agissant des individus appartenant aux secondes et troisièmes classes, nous observons le phénomène inverse. En particulier en ce qui concerne la troisième classe, le nombre de personnes décédées est plus de trois fois supérieur à celui des personnes ayant survécues. Nous pouvons également noter que la majorité des individus faisaient parti de la troisième classe (≈ 55%).

2. Analyse croisée entre la survie et le sexe des passagers

Le graphique ci-après nous permet d'observer que parmi les hommes et les femmes à bord du titanic, le nombre d'hommes morts est nettement supérieur à ceux ayant survécus, tandis qu'à l'inverse le nombre de femmes ayant survécues est nettement supérieur à celles qui sont décédées. Par ailleurs nous remarquons que la répartition homme/femme n'est pas égale, il y a davantage d'hommes que de femmes à bord du titanic.

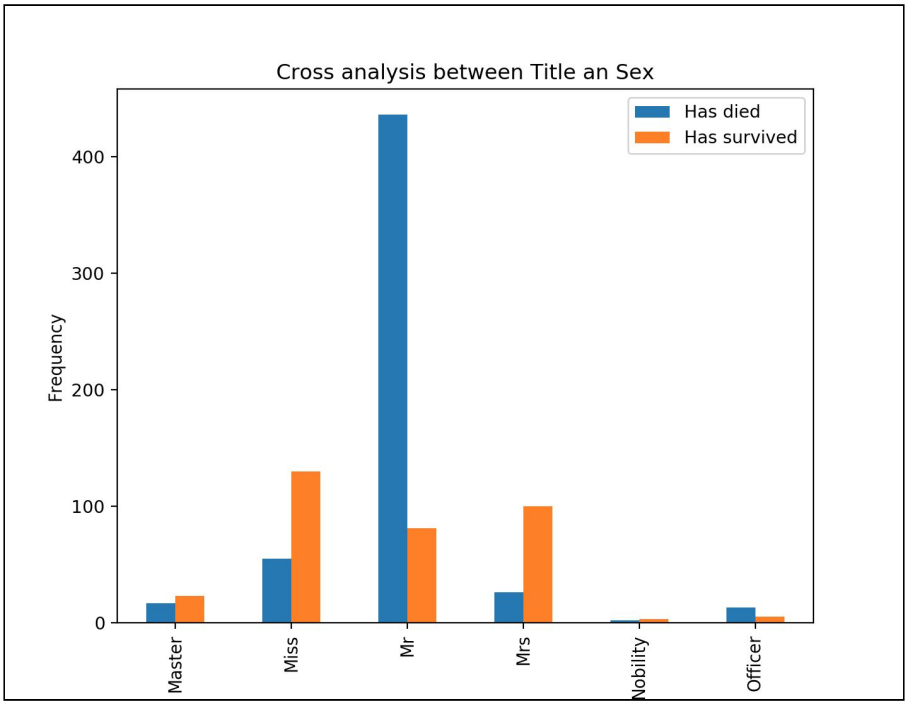


3. Analyse croisée entre le sexe et le niveau de classe des passagers



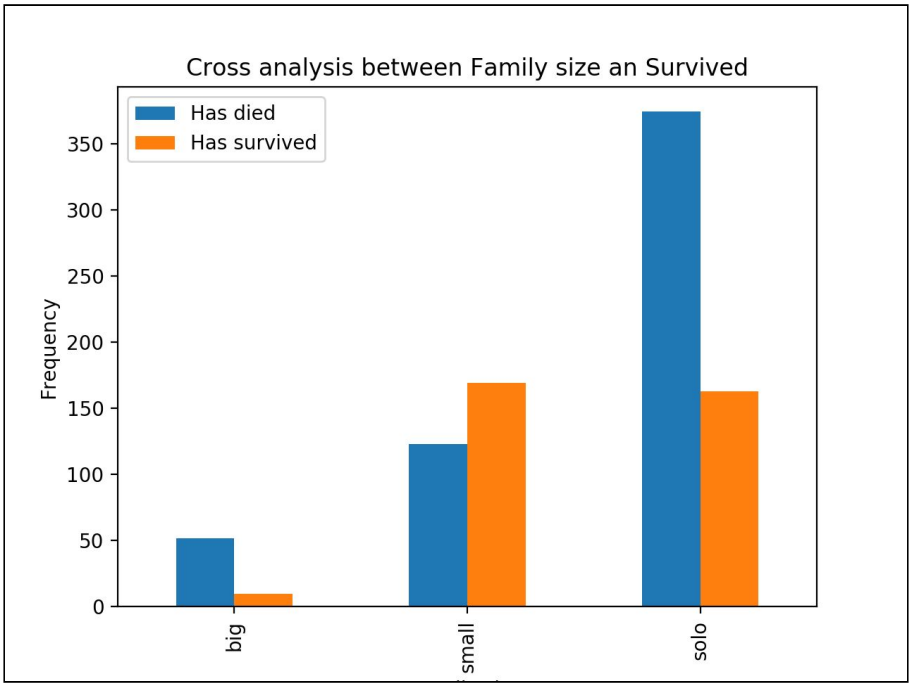
On observe ici que près de la moitié des hommes appartiennent à la 3ème classe (classe la plus présente à bord du Titanic). A l'inverse, la répartition des femmes au sein des trois différentes classe est plutôt équilibrée.

4. Analyse croisée entre la survie et les titres des passagers



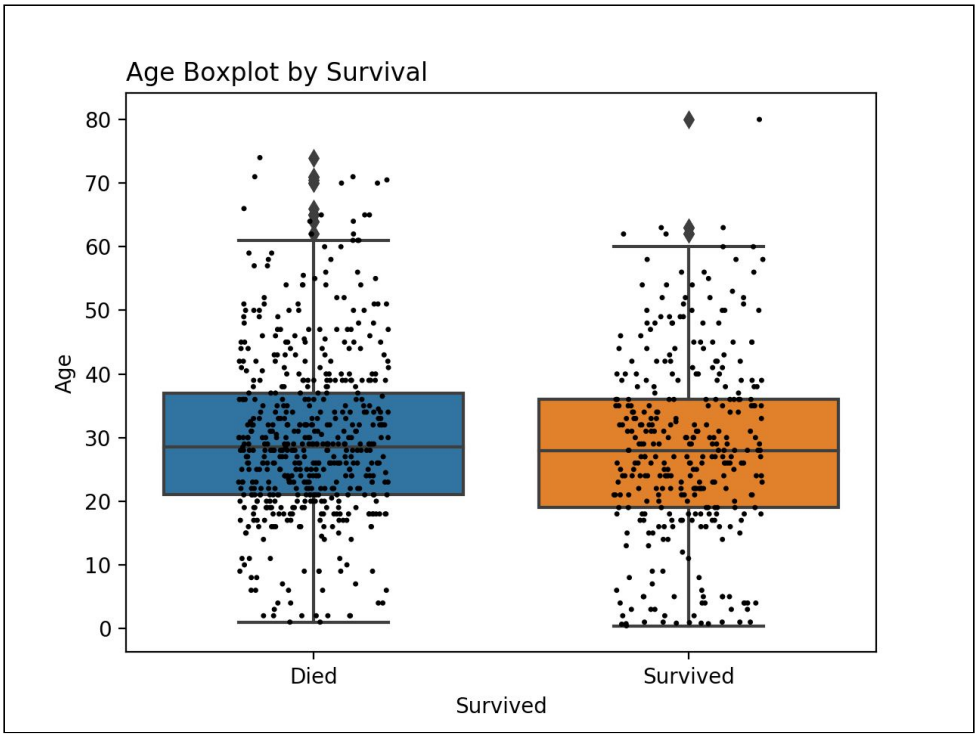
Le point marquant concernant ce graphique est que la répartition des femmes est plutôt bien établie entre les différents titres de civilité auxquels elles peuvent appartenir (Miss ou Mrs), à l'inverse de celle des hommes pour lesquels la quasi-totalité d'entre eux se retrouvent au sein du titre Mr. Il n'y a en effet que peu d'individu possédant le titre de noble, maître ou officier. Par ailleurs, nous remarquons une nouvelle fois ici que la majorité des femmes ont survécu tandis que la majorité des homme (hormis ceux ayant un titre de maître et qui par conséquent sont des enfants) ont succombé lors du naufrage.

5. Analyse croisée entre la survie et la taille de la famille des passagers



Le point marquant concernant ce graphique est que seul les individus ayant voyagés avec seulement 2 à 3 membres de leurs famille à bord du titanic possède un niveau de survie supérieur à ceux ayant succombé lors du naufrage. De plus, on constate que s'agissant des individus voyageant seuls la probabilité de survie est deux fois inférieur et avec plus de 4 membres de leurs famille (*Famsize* = big), la probabilité de survie est cinq fois inférieure.

6. Boxplot de l'âge par niveau de survie



Nous pouvons voir dans ce graphique que la médiane de l'âge et le 3ème quartile des 2 groupes est relativement proche, seul le premier quartile des individus ayant survécu est légèrement inférieur. Enfin, on constate la présence d'individus aliens dans nos 2 échantillons, le plus marquant restant celui dont l'âge est égal à 80 ans dans le groupe des individus ayant survécus. Ce dernier peut potentiellement avoir un effet significatif dans nos prédictions futures, une des solutions serait de l'enlever de notre base de données.

7. Statistiques descriptives

Classe des passagers		
1ère classe	2nde classe	3ème classe
491	216	184
55%	24%	21%

Sexe	
Homme	Femme
577	314
65%	35%

Statu de survie	
Décédé	A survécu
549	342
62%	38%

Port d'embarquement		
Southampton	Cherbourg	Queenstown
646	168	77
73%	19%	9%

Taille de la famille		
Seul	Petite	Grande
537	292	62
60%	33%	7%

Deck des passagers								
Inconnu	A	B	C	D	E	F	G	T
687	15	47	59	33	32	13	4	1
77%	2%	5%	7%	4%	4%	1%	0,4%	0,1%

II. Règles d'association et fouille de séquences

A. Mise en forme du Dataframe

Une fois nos données nettoyées, nous réalisons une sélection afin de retenir uniquement les variables qui seront utiles dans la réalisation du Pattern Mining. Pour ce faire, nous ne retenons tout d'abord que les variables textuelles, et éliminons (ou transformons) les variables de type numérique. En particulier, nous retirons les *dummies variables* que nous avions créés au préalable et dont nous aurons besoin pour la seconde partie de notre projet, à savoir :

- *female; male*
- *Pclass_1; Pclass_2; Pclass_3*
- *Master; Miss; Mr; Mrs; Nobility; Officer*
- *big_family; small_family; solo*
- *Embarked C; Embarked_Q; Embarked_S*

Les autres variables que nous retirons sont donc les suivantes :

- *PassengerId*
- *Name*
- *Age; Age_Randomforest; Age_SVM; Age_replace*
- *Sex*
- *Deck* (et toutes les variables dummies créer à partir de Deck)
- *Fare*

- *Embarked*
- *Parch; SibSp*

Nous retirons tout d'abord les variables PassengerId et Name tout simplement car elles sont propres à chacun des individus composant la base de donnée. De la même manière, nous ne considérons pas dans cette partie les individus selon leur âge mais plutôt selon la tranche d'âge à laquelle ils appartiennent comme nous l'avons vu précédemment. La variable *Fare* est également enlevée pour notre analyse car nous suspectons que le prix du billet peut déjà être indiqué à travers la variable *Pclass* qui correspond à la catégorie du ticket. Par conséquent, regrouper les prix du billets en groupe de prix comme nous l'avons fait pour l'âge risquerait d'être fortement corrélé avec la variable *Pclass.*

Par la suite, nous choisissons de prendre en compte le titre de l'individu et non simplement son sexe, ce pourquoi la variable Sex a été retirée. En effet, la variable Title que nous avons créé nous donne à la fois une indication sur le statut social de l'individu, mais également sur son sexe. Nous retirons également la variable Deck en raison du nombre trop important d'informations incomplètes, ainsi que la variable Embarked étant donné que la grande majorité des individus ont embarqué depuis le port de Southampton. Après avoir été utilisées pour la constitution de la variable Famsize, qui est quant à elle gardée pour la suite de cette partie, les variables Parch et SibSp ont elles aussi été enlevées. Enfin nous avons transformé les variables Survived et Pclass de manière à n'avoir plus que des variables de type texte dans la base de donnée, dont nous présentons ci-après les 5 premières lignes afin d'avoir la représentation finale des variables que nous allons utiliser.

	Survived	Pclass	Title	Famsize	Age_group
0	Died	Third_class	Mr	small_family	17_30
1	Survived	First_class	Mrs	small_family	31_40
2	Survived	Third_class	Miss	solo	17_30
3	Survived	First_class	Mrs	small_family	31_40
4	Died	Third_class	Mr	solo	31_40

Maintenant que nous disposons de notre dataframe, nous le transformons en «liste» afin de pouvoir appliquer les différentes fonctions du package mlxtend de Python.

	0_16	17_30	31_40	Died	Dr	First_class	Master	Miss	Mr	Mrs	Nobility	Officer	Second_class	Survived
0	False	True	False	True	False	False	False	False	True	False	False	False	False	False
1	False	False	True	False	False	True	False	False	False	True	False	False	False	True
2	False	True	False	False	False	False	False	True	False	False	False	False	False	True
3	False	False	True	False	False	True	False	False	False	True	False	False	False	True
4	False	False	True	True	False	False	False	False	True	False	False	False	False	False

Ce n'est qu'à partir de cette transformation que nous pouvons procéder à la recherche de motifs fréquents et donc de règles d'association.

B. Analyse des résultats

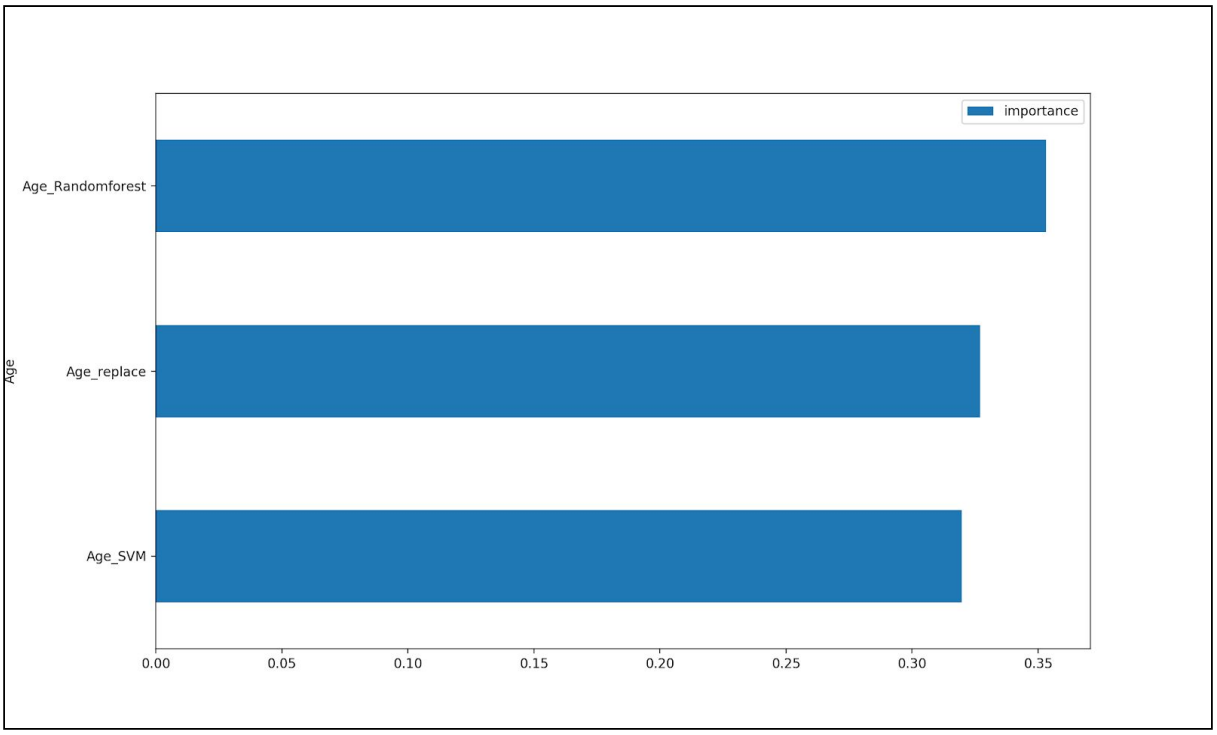
Les règles d'associations que nous avons obtenues (voir annexe 1) sont alors en accord avec les statistiques descriptives que nous avons présenté auparavant, et ont également permis de mettre en avant les différents profils susceptibles de mourir lors du naufrage. En l'occurrence nous pouvons nous apercevoir que les hommes seuls âgés entre 17 et 30 ans, et qui appartiennent à une catégorie sociale plutôt modeste (Third_class) seraient typiquement les profils pour lesquels les chances de survies seraient les plus faibles. Plus précisément, 84% des hommes ayant pour titre "Mr" sont décédés, un profil (Mr n'ayant pas survécu) qui compose 49% de notre échantillon. De la même manière, trois quarts des individus de troisième classe sont morts ce 15 avril 1912, et 9 d'entre eux étaient des hommes. Toutefois ces informations sont à nuancer, et ce pour plusieurs raisons. En effet comme nous l'avons vu précédemment lors de l'analyse de nos différentes variables, il apparaît clairement que notre base de donnée est principalement composée de jeunes hommes (de 17 à 30 ans) au titre de Mr. Ce profil étant le plus répandu à bord du Titanic, il n'est pas surprenant de le retrouver fréquemment parmi les victimes du naufrage. Ce qui est intéressant en revanche, c'est d'observer que plus de la moitié (55%) des jeunes de moins de 16 ans ont survécu, alors même que nous savons que plus de 6 personnes sur 10 sont décédées dans notre échantillon. De la même manière, les trois quarts des femmes ont survécu. Nous pouvons donc supposer avec ces résultats que lors de la procédure d'évacuation d'urgence, la priorité pour accéder aux canaux de sauvetage a été donnée aux femmes et aux enfants. Enfin, nous remarquons que parmi les survivants, 40% appartenaient à la première classe. Les chances de survies ne sont en effet pas les même selon la catégorie du billet : 63% des individus de première ont survécu contre 25% pour les individus de troisième classe. En faisant l'hypothèse que des individus de classes sociales plus modestes étaient en troisième classe, nous pourrions supposer au travers de ce résultat que les individus issus de classes sociales plus aisées ont été prioritaires pour accéder aux canaux de sauvetages.

III. Analyse prédictive par apprentissage automatique

A. Choix des variables à utiliser lors de la prédiction

L'analyse prédictive par apprentissage automatique se fera en deux partie en utilisant deux méthodes d'apprentissage automatique supervisé, à savoir random forest classifier (forêts aléatoires/arbres décisionnels) et support-vector machines (SVM) classifier (machine à vecteur de support).

Lors de notre analyse prédictive, nous commencerons tout d'abord par déterminer parmi les 3 variables dérivés de l'âge, laquelle aura le plus d'impact dans les prédictions. Pour rappel, les 3 variables en question sont l'âge prédit par random forest, l'âge prédit par SVM et le dernier qui comprend l'âge manquant des passagers remplacé par la médiane d'âge par titre de civilité.



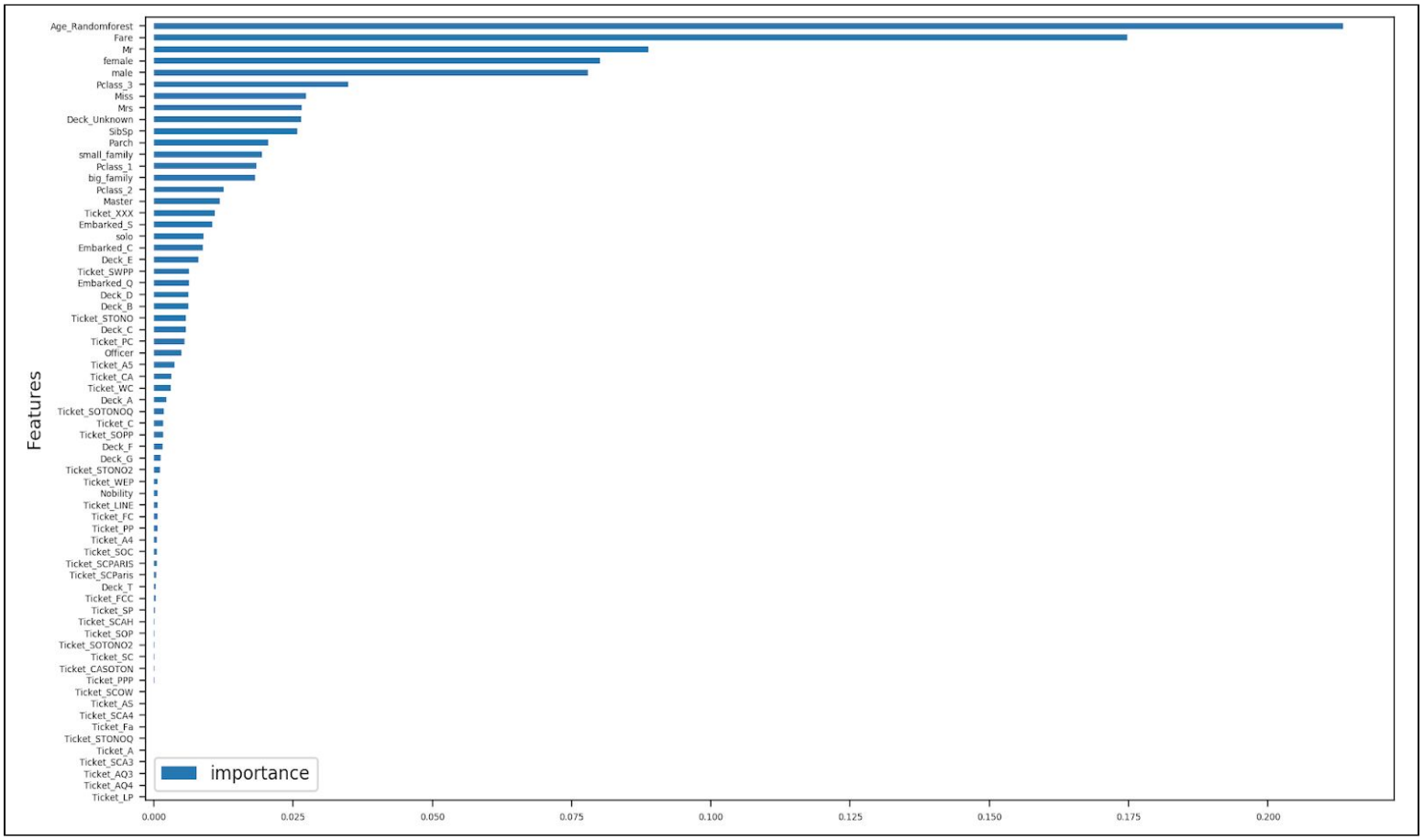
Nous pouvons voir dans le tableau ci-dessus que l'âge prédit par random forest est légèrement plus important que l'âge remplacé par les médianes regroupés par titre, sexe et classe du passager ainsi que l'âge prédit par SVM. Nous allons donc continuer nos prédictions en utilisant la valeur de l'âge prédit par random forest.

Les variables utilisées dans nos différents modèle sont donc :

- Age_Randomforest: âge avec les valeurs manquantes prédit par random forest
- SibSp: nombre de frères/soeurs et époux/épouses à bord du titanic
- Parch: nombre d'enfants et de parents à bord du titanic
- Fare: Prix du ticket
- female: variable binaire, 1 si la personne est une femme et 0 sinon
- male: variable binaire, 1 si la personne est un homme et 0 sinon
- Pclass_1: variable binaire, 1 si la personne était en première classe et 0 sinon
- Pclass_2: variable binaire, 1 si la personne était en seconde classe et 0 sinon
- Pclass_3: variable binaire, 1 si la personne était en troisième classe et 0 sinon
- Master: variable binaire, 1 si la personne avait le titre *Master* et 0 sinon
- Miss: variable binaire, 1 si la personne avait le titre *Miss* et 0 sinon
- Mr: variable binaire, 1 si la personne avait le titre *Mr* et 0 sinon
- Mrs: variable binaire, 1 si la personne avait le titre *Mrs* et 0 sinon
- Nobility: variable binaire, 1 si la personne avait le titre *Nobility* et 0 sinon
- Officer: variable binaire, 1 si la personne avait le titre *Officer* et 0 sinon
- big_family: variable binaire, 1 si la personne voyageait en grande famille et 0 sinon
- small_family: variable binaire, 1 si la personne voyageait en petite famille et 0 sinon
- solo: variable binaire, 1 si la personne voyageait seul et 0 sinon
- Embarked_C: variable binaire, 1 si le port d'embarcation était *Cherbourg* et 0 sinon
- Embarked_Q: variable binaire, 1 si le port d'embarcation était Queenstown et 0 sinon

- Embarked_S: variable binaire, 1 si le port d'embarcation est *Southampton* et 0 sinon
- Deck_A: variable binaire, 1 si le deck était *A* et 0 sinon
- Deck_B: variable binaire, 1 si le deck était *B* et 0 sinon
- Deck_C: variable binaire, 1 si le deck était *C* et 0 sinon
- Deck_D: variable binaire, 1 si le deck était *D* et 0 sinon
- Deck_E: variable binaire, 1 si le deck était *E* et 0 sinon
- Deck_F: variable binaire, 1 si le deck était *F* et 0 sinon
- Deck_G: variable binaire, 1 si le deck était *G* et 0 sinon
- Deck_T: variable binaire, 1 si le deck était *T* et 0 sinon
- Deck_Unknown: variable binaire, 1 si le deck était *inconnu* et 0 sinon
- Les 37 variables binaires dérivées de la variable Ticket après transformation

Dans ce graphique ci-dessous nous pouvons voir l'importance de chaque variable lors de notre prochaine étape de prédiction de la survie des passagers à bord du titanic.



B. Explication des paramètres utilisés pour chaque modèle

Paramètres	Explication	Modèle
n_estimators	Nombre d'arbre de décision dans la forêt	Random forest
min_samples_split	Nombre minimum d'observations requises pour séparer un noeud interne	Random forest
min_samples_leaf	Nombre minimum d'observations requises dans une feuille	Random forest
max_features	Nombre d'échantillon maximal de variables à considérer pour chaque split	Random forest
max_depth	Profondeur maximale de chaque arbre	Random forest
bootstrap	Si oui ou non l'échantillon bootstrap est utilisé lors de la construction de l'arbre. (Échantillon aléatoire avec remise de la base de donnée initial)	Random forest
random_state	Etat aléatoire prédéfini	Random forest
Kernel	Type de kernel utilisé dans l'algorithme	SVM classifier

C. Comparaison des modèles

Le tableau ci-après résume les différents modèles testés dans le cadre de nos prévisions. Dès lors après avoir testé 5 différents modèles random forest et 2 différents modèle SVM, nous pouvons conclure que le modèle de prédiction random forest numéro 5 est le plus précis car le score attribué par l'algorithme de vérification kaggle lui a attribué 0.80861, ce qui est légèrement plus élevé que le modèle numéro 4 (0.80382). Nous pouvons cependant remarquer que sa précision moyenne des 10 fold cross validation (validation croisé) est inférieurs aux autres modèles de prédiction random forest (prédiction 1, 2, 3 et 4). De manière générale, on peut également voir que les modèles random forest atteignent un plus haut niveau de précision vis-à-vis des modèles SVM. Notons enfin que le second modèle SVM est celui qui possède le plus petit écart-type.

Paramètres		10 Fold CV mean score accuracy	10 Fold CV standard deviation	Kaggle score
1 RF	n_estimators = 2000 , min_samples_split = 5 min_samples_leaf = 5 , max_features = sqrt max_depth = 5 , bootstrap = True	0.8317	0.0309	0.78947
2 RF	n_estimators = 1000, min_samples_split = 10 min_samples_leaf = 4 , max_features = auto max_depth = 5 , bootstrap = False	0.8328*	0.0316	0.79425
3 RF	n_estimators = 1000, min_samples_split = 5 min_samples_leaf = 4 , max_features = auto max_depth = 5 , bootstrap = False	0.8317	0.0309	0.79425
4 RF	n_estimators = 5000, min_samples_split = 10 min_samples_leaf = 4 , max_features = sqrt max_depth = 10 , bootstrap = True	0.8317	0.034	0.80382
5 RF	n_estimators = 5000, min_samples_split = 10 min_samples_leaf = 5 , max_features = sqrt max_depth = 10 , bootstrap = True	0.8305	0.0335	0.80861*
1 SVM	Kernel = rbf	0.7678	0.0369	0.66985
2 SVM	Kernel = linear	0.8216	0.0247*	0.77990

Conclusion

A travers le challenge Titanic que nous avons relevé, nous avons pu mettre en évidence les profils types des individus ayant survécus au drame, à savoir majoritairement des femmes et des enfants, et des personnes que l'on pourrait supposer de classes sociales supérieures. Les personnes voyageant seules présentaient également de faibles chances de survie. Ainsi être une femme voyageant en famille en première classe offrait de meilleures chances de survie qu'être un jeune homme voyageant seul en troisième classe. Par la suite, nous avons mis en place un algorithme permettant de prédire la survie suite au naufrage d'un individu selon les caractéristiques qu'il présentait, et il s'est avéré que dans 8 cas sur 10 notre algorithme prédisait juste. Toutefois nous pourrions éventuellement aller plus loin dans la précision de notre modèle, notamment en y intégrant des réseaux de neurones (deep learning). Avoir plus de variables, comme par exemple avoir une indication sur le statut de l'individu à bord du navire (voyageur ou employé), ou moins de valeurs manquantes en particulier concernant l'âge, aurait également pu améliorer la précision du modèle que nous avons mis en place.

Annexe 1 :

	antecedents	consequents	support	confidence
0	frozenset({'Died'})	frozenset({'Mr'})	0.489337822671156	0.7941712204007286
1	frozenset({'Mr'})	frozenset({'Died'})	0.489337822671156	0.8433268858800773
2	frozenset({'Mr'})	frozenset({'solo'})	0.44556677890011226	0.7678916827852998
3	frozenset({'solo'})	frozenset({'Mr'})	0.44556677890011226	0.7392923649906891
4	frozenset({'Died', 'Mr'})	frozenset({'solo'})	0.3771043771043771	0.7706422018348624
5	frozenset({'Died', 'solo'})	frozenset({'Mr'})	0.3771043771043771	0.8983957219251337
6	frozenset({'Mr', 'solo'})	frozenset({'Died'})	0.3771043771043771	0.8463476070528967
7	frozenset({'Died'})	frozenset({'Mr', 'solo'})	0.3771043771043771	0.6120218579234973
8	frozenset({'Mr'})	frozenset({'Died', 'solo'})	0.3771043771043771	0.6499032882011605
9	frozenset({'solo'})	frozenset({'Died', 'Mr'})	0.3771043771043771	0.6256983240223464
10	frozenset({'Third_class'})	frozenset({'Died'})	0.4175084175084175	0.7576374745417515
11	frozenset({'Died'})	frozenset({'Third_class'})	0.4175084175084175	0.6775956284153005
12	frozenset({'Third_class'})	frozenset({'Mr'})	0.35802469135802467	0.6496945010183298
13	frozenset({'Mr'})	frozenset({'Third_class'})	0.35802469135802467	0.6170212765957446
14	frozenset({'Third_class'})	frozenset({'solo'})	0.36363636363636365	0.6598778004073319
15	frozenset({'solo'})	frozenset({'Third_class'})	0.36363636363636365	0.6033519553072626
16	frozenset({'Third_class', 'Died'})	frozenset({'Mr'})	0.3176206509539843	0.760752688172043
17	frozenset({'Third_class', 'Mr'})	frozenset({'Died'})	0.3176206509539843	0.8871473354231976
18	frozenset({'Died', 'Mr'})	frozenset({'Third_class'})	0.3176206509539843	0.6490825688073395
19	frozenset({'Third_class'})	frozenset({'Died', 'Mr'})	0.3176206509539843	0.5763747454175152
20	frozenset({'Died'})	frozenset({'Third_class', 'Mr'})	0.3176206509539843	0.5154826958105647
21	frozenset({'Mr'})	frozenset({'Third_class', 'Died'})	0.3176206509539843	0.5473887814313346
22	frozenset({'Third_class'})	frozenset({'17_30'})	0.34231200897867564	0.6211812627291242
23	frozenset({'17_30'})	frozenset({'Third_class'})	0.34231200897867564	0.6869369369369369
24	frozenset({'Died'})	frozenset({'17_30'})	0.3389450056116723	0.5500910746812386
25	frozenset({'17_30'})	frozenset({'Died'})	0.3389450056116723	0.6801801801801802
26	frozenset({'Mr'})	frozenset({'17_30'})	0.33221099887766553	0.5725338491295938
27	frozenset({'17_30'})	frozenset({'Mr'})	0.33221099887766553	0.6666666666666666
28	frozenset({'solo'})	frozenset({'17_30'})	0.35353535353535354	0.5865921787709497
29	frozenset({'17_30'})	frozenset({'solo'})	0.35353535353535354	0.7094594594594594
30	frozenset({'Died'})	frozenset({'solo'})	0.41975308641975306	0.6812386156648451
31	frozenset({'solo'})	frozenset({'Died'})	0.41975308641975306	0.696461824953445