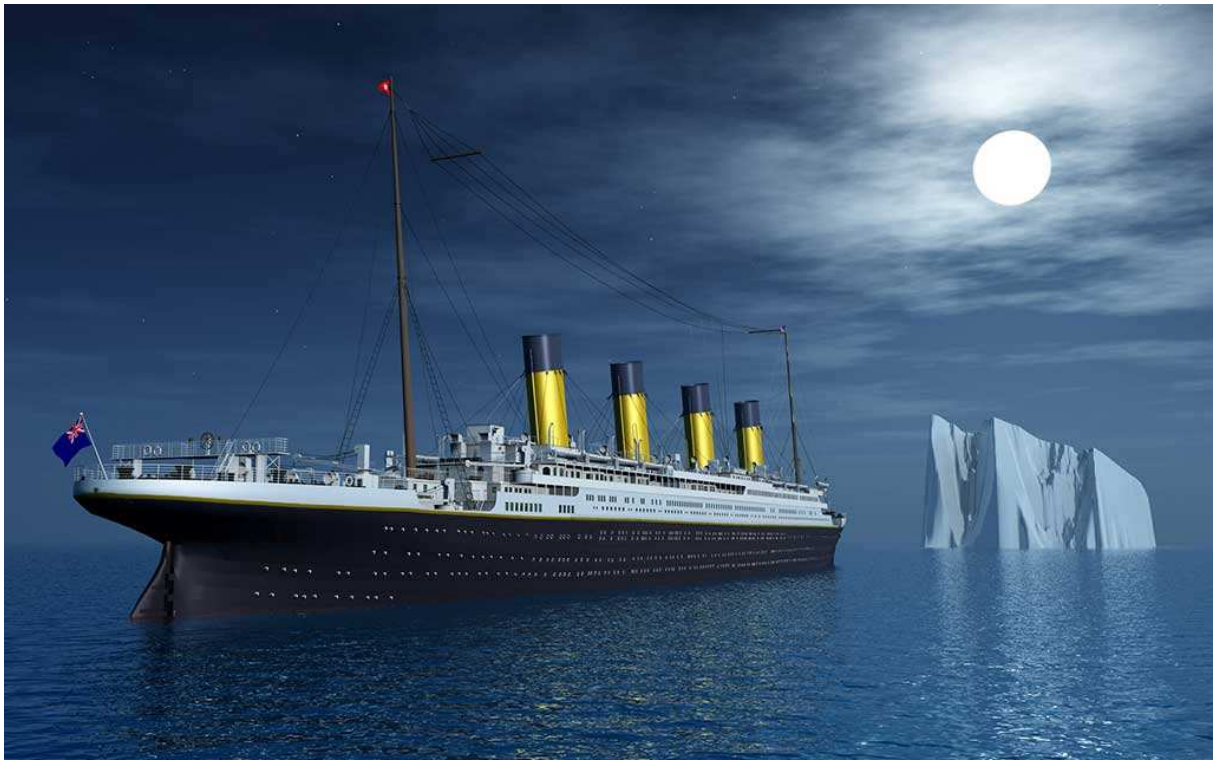
Projet Titanic



Introduction aux algorithmes de Machine Learning

Charles HAJJAR & Vincent PETEAU

INTRODUCTION de l'équipe pédagogique

L'objectif de ce brief projet est de retracer scientifiquement l'histoire du naufrage du Titanic en utilisant les données disponibles sur le Kaggle.

Il s'agit d'un jeu de données public très facile d'accès et qui possède plusieurs vertus pédagogiques. Bien que l'analyse des données liées au naufrage du Titanic n'ait aucun intérêt métier, les données sont riches pour pouvoir mettre en pratique les techniques et les modèles que nous avons abordés les dernières semaines.

Pour cela, il vous est demandé de mettre en œuvre une démarche complète d'exploitation de données allant de la compréhension du besoin jusqu'à l'évaluation des modèles élaborés en passant par une phase de préparation et d'analyse de données.

L'histoire du Titanic

C'est le 31 mars 1909 que débute la construction du Titanic et son histoire. La première se termine le 31 mars 1912, après avoir retardé sa mise ne service suite à un souci sur son frère, l'Olympic, entré en collision avec un autre bateau, le croiseur Hawke.

Avec sa double coque en plaques d'acier rivetées et ses 16 compartiments séparés par 15 cloisons étanches, le Titanic offrait une sécurité maximale. Et au cas où, ses huit pompes offraient une capacité d'évacuation de 400 tonnes d'eau à l'heure. Deux compartiments pouvaient être inondés sans que le navire soit en danger, les autres assurant sa flottabilité.

L'appareillage a eu lieu le 10 avril 1912 à midi de Southampton (en Angleterre), risquant une première collision avec le paquebot le *New York*, aspiré par le Titanic, au point de rompre ses amarres… Direction Cherbourg (en Normandie) puis Queenstown (la pointe sud de l'Irlande), arrivée prévue à New York dans la matinée du 17 avril.

Le paquebot emmène 329 passagers en 1ère classe, 285 en 2ème, 710 en 3ème, soit un total de 1324 passagers, auxquels s'ajoutent les 899 membres d'équipage, soit 2223 personnes.

Le 14 avril 1912, seulement deux jours après le départ, le Titanic traverse l'Atlantique Nord à une vitesse de 22 noeuds, soit 700 m à la minute, et…

A 23h40 : la nuit de dimanche, les veilleurs Fleet et Lee, dans la hune du grand mat, se penchent en avant, les yeux écarquillés, un iceberg sort de la brume, droit devant, à environ 600 mètres de la coque. Fleet prend le téléphone et appelle l'officier Murdock : « Iceberg droit devant ». On connait la suite…

Des 868 survivants, 711 ont été rescapés par le Carpathia, arrivé sur les lieux du drame vers 5 h du matin. L'équipage de ce navire a mené une opération de sauvetage qui a duré sept heures.

Organisation du projet d'étude

Début du projet lundi 6 avril, restitution vendredi 10.

Nettoyage des données, combler les manquants ou supprimer des variables, ajouter des variables quantitatives ou qualitatives…

Affichage en mode graphique des variables afin de pouvoir les analyser rapidement, choisir lesquelles peuvent être les plus parlantes,

Utilisation d'algorithmes pour combler les valeurs manquantes (âge, prédire la survie), comparer les résultats et la pertinence des retours,

Ne pas oublier de noter les étapes de la progression du projet, les essais effectués (concluants ou pas), analyser les résultats et quelques graphiques,

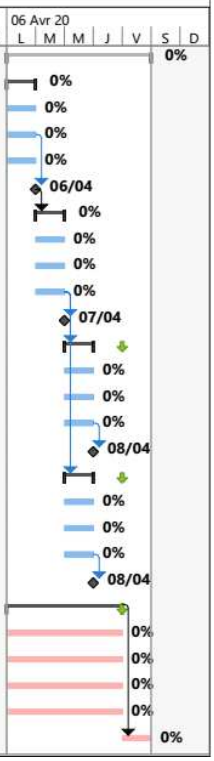
Et conclure ce que l'on a réussi à faire. Et les prochaines étapes de notre progression,

Rédaction de ce rapport tout au long du projet.

Ce qui donne le diagramme de Gantt suivant :

Titanic.mpp

N°	Mode Tâche	Nom de la tâche	Durée	Début	Fin	Prédécesseurs	Noms ressources
0		**Titanic**	**5 jrs**	**Lun 06/04/20**	**Ven 10/04/20**		
1		**Données**	**1 jr**	**Lun 06/04/20**	**Lun 06/04/20**		
2		Chargement	1 jr	Lun 06/04/20	Lun 06/04/20		
3		Analyse	1 jr	Lun 06/04/20	Lun 06/04/20		
4		Nettoyage	1 jr	Lun 06/04/20	Lun 06/04/20		
5		Résumé	0 jr	Lun 06/04/20	Lun 06/04/20	3	
6		**DataViz**	**1 jr**	**Mar 07/04/20**	**Mar 07/04/20**	**5**	
7		Choix des variables	1 jr	Mar 07/04/20	Mar 07/04/20		
8		Choix des graphes	1 jr	Mar 07/04/20	Mar 07/04/20		
9		Dashboard	1 jr	Mar 07/04/20	Mar 07/04/20		
10		Résumé	0 jr	Mar 07/04/20	Mar 07/04/20	9	
11		**Régression Lin.**	**1 jr**	**Mer 08/04/20**	**Mer 08/04/20**	**9**	
12		Choix variables	1 jr	Mer 08/04/20	Mer 08/04/20		
13		Algorythme	1 jr	Mer 08/04/20	Mer 08/04/20		
14		Analyse	1 jr	Mer 08/04/20	Mer 08/04/20		
15		Résumé	0 jr	Mer 08/04/20	Mer 08/04/20	14	
16		**Régression Log.**	**1 jr**	**Mer 08/04/20**	**Mer 08/04/20**	**9**	
17		Choix variables	1 jr	Mer 08/04/20	Mer 08/04/20		
18		Algorythme	1 jr	Mer 08/04/20	Mer 08/04/20		
19		Analyse	1 jr	Mer 08/04/20	Mer 08/04/20		
20		Résumé	0 jr	Mer 08/04/20	Mer 08/04/20	19	
21		**Rédac. Rapport**	**4 jrs**	**Lun 06/04/20**	**Jeu 09/04/20**		
22		Résumé	4 jrs	Lun 06/04/20	Jeu 09/04/20		
23		Graphiques	4 jrs	Lun 06/04/20	Jeu 09/04/20		
24		Analyses	4 jrs	Lun 06/04/20	Jeu 09/04/20		
25		Conclusions	4 jrs	Lun 06/04/20	Jeu 09/04/20		
26		Soutenance	1 jr	Ven 10/04/20	Ven 10/04/20	21	



Etude des chiffres à disposition pour notre projet

Nous avons 2 fichiers csv à notre disposition :

- train.csv pour entrainer votre modèle (celui-ci contient les libellés : Survived)
- test.csv pour calculer le résultat à partir du modèle choisi

Le fichier train.csv possède 12 Variables et 891 enregistrements.
Le fichier test possède 11 Variables et 418 enregistrements.

Contenu des Variables :

Variables	Informations	Structure
PassengerId	*N° identification du passager*	*De 1 à 891 dans le train.csv* *de 892 à 1309 dans le test.csv*
Survived	*Survivant ou non* *Non dispo sur le jeu de test*	*1 si le passager a survécu, 0 s'il est décédé*
Pclass	*Classe du passager*	*1 = 1ère classe, 2 = 2ème classe, 3 = 3ème classe*
Name	*Nom du passager*	*Style : Nom, titre. Prénoms*
Sex	*Sexe du passager*	*'male' ou 'female'*
Age	*Age du passager*	*Décimal si inférieur à 1, estimé si de la forme xx.5*
SibSp	*Nombre d'époux, de frères ou de sœurs présents à bord*	
Parch	*Nombre de parents ou d'enfants présents à bord*	
Ticket	*Numéro du ticket*	
Fare	*Prix des tickets*	*Le prix est indiqué en £ et pour un seul achat (peut correspondre à plusieurs tickets)*
Cabin	*Numéro de Cabine*	*Un ou plusieurs numéros de cabine, de la forme 'A123'*
Embarked	*Port d'embarcation*	*C = Cherbourg, Q = Queenstown, S = Southampton*

Descriptions issues d'Internet : *https://www.kaggle.com/roryhny/fr-predictions-sur-le-titanic#creation-des-nouvelles-variables-et-enrichissement*

Première analyse rapide des données

Fichier d'entrainement

Le type et le nombre de valeurs renseignées de la base train.csv :

Train	Column	Non-Null	Count	Dtype
0	PassengerId	891	non-null	int64
1	Survived	891	non-null	int64
2	Pclass	891	non-null	int64
3	Name	891	non-null	object
4	Sex	891	non-null	object
5	Age	714	non-null	float64
6	SibSp	891	non-null	int64
7	Parch	891	non-null	int64
8	Ticket	891	non-null	object
9	Fare	891	non-null	float64
10	Cabin	204	non-null	object
11	Embarked	891	non-null	object
12	nSex	891	non-null	int8
13	nEmbarked	891	non-null	int8

891 passagers répartis comme suit :

- 342 survivants – 549 décédés
- 577 hommes et 314 femmes (adultes et enfants)
- Tous les passagers ont un ticket, certains tickets servent pour plusieurs passagers, donc prix de la place en conséquence ?
- 687 passagers n'ont pas de numéro de cabine. On peut avoir plusieurs numéros pour un même passager : 3 cabines B51 B53 B55 utilisées pour 2 passagers et un accompagnant par exemple… comment est alors calculé le prix du billet ? nous n'avons pas accès à la grille tarifaire, donc cette information sera laissée de côté.
- Seules 2 passagères n'ont pas de port d'embarquement. Suppression ou ajout ?
- L'âge est inconnu pour 177 personnes (52 survivants et donc 125 décédés). On va devoir compléter cette variable

Fichier de test pour les algorithmes, 418 passagers :

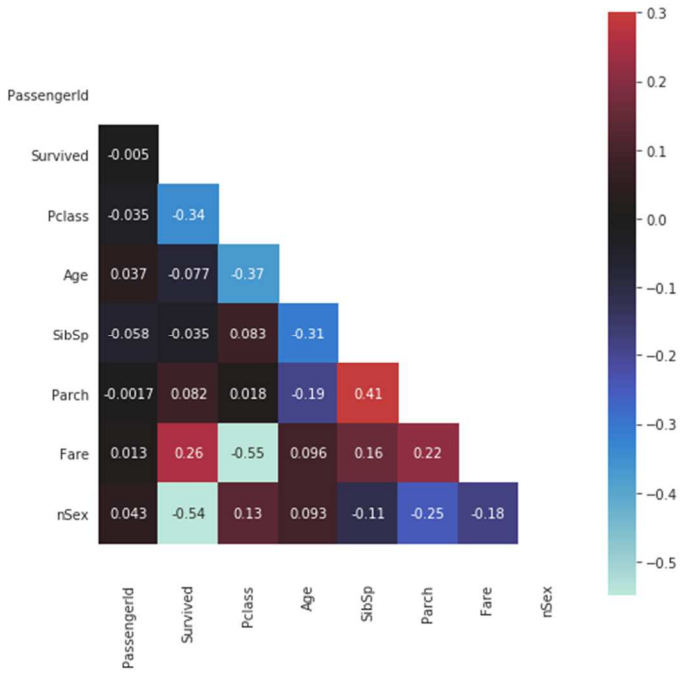
Le type et le nombre de valeurs renseignées de la base test.csv :

Test	Column	Non-Null	Count	Dtype
0	PassengerId	418	non-null	int64
1	Pclass	418	non-null	int64
2	Name	418	non-null	object
3	Sex	418	non-null	object
4	Age	332	non-null	float64
5	SibSp	418	non-null	int64
6	Parch	418	non-null	int64
7	Ticket	418	non-null	object
8	Fare	417	non-null	float64
9	Cabin	91	non-null	object
10	Embarked	418	non-null	object

418 passagers répartis comme suit :

- 86 sans âge indiqué
- 1 sans prix du billet
- 327 sans numéro de cabine
- Aucune indication de survie ou pas…

Nous avons recherché une corrélation entre les informations de la table « train » :



En clair, nous remarquons rapidement des corrélations entre :

- Très forte corrélation entre la survie des personnes et
    - Le sexe (vert pale : -.054)
    - Le prix (en rouge = 0.26)
    - La classe (en bleu : -0.34)
- La classe du passager et
    - le prix (vert pale : -.055)
    - l'âge (en bleu = -0.34)
- Les accompagnants de la famille avec l'âge, donc les enfants et parents (SibSp : -0.31 en bleu, Parch avec SibSp : 0.41 en rouge) ainsi qu'entre
- Bien entendu, aucune corrélation entre le numéro d'identification (PassengerID), le nom et le numéro de ticket des passagers avec les autres informations…

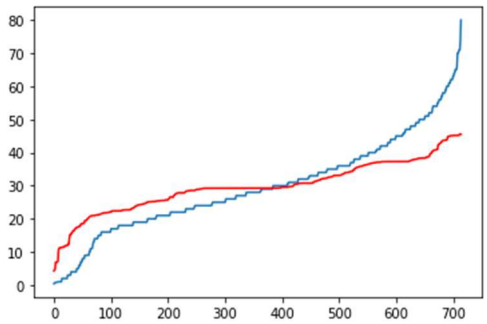
Pour répondre à la corrélation entre âge et classe, voici les moyennes d'âge pour chacune d'elles :

- 1ère classe : 32,9 ans
- 2ème classe : 28,1 ans
- 3ème classe : 18,4 ans (10 ans d'écart avec la 2ème classe)

Nous avons donc une corrélation entre âge et classe : les plus jeunes étant en 3ème (faibles moyens financiers), les plus âgés en 1ère, en moyenne d'âge.

Afin de sortir les premiers graphiques, nous avons complété les âges manquants. Nous avions plusieurs solutions à disposition : mettre une valeur fixe pour tous en fonction de l'âge moyen de l'ensemble des passagers, mettre une valeur aléatoire ou suivre une règle de régression linéaire.

Même si celle-ci n'est pas parfaite, c'est ce choix qui a été choisi :



Après avoir complété les informations manquantes, créé une variable supplémentaire pour la catégorie d'âge, nous avons sorti rapidement quelques graphiques du fichier de « train » : le « 1 » sur la droite indique que le passager a survécu à l'accident.

Les catégories d'âge :

- Bébé : moins de 3 ans
- Enfant : moins de 15 ans – *les jeunes pouvaient déjà travailler à cet âge , d'où cette valeur**
- Adulte : moins de 60 ans
- Vieux : les autres…

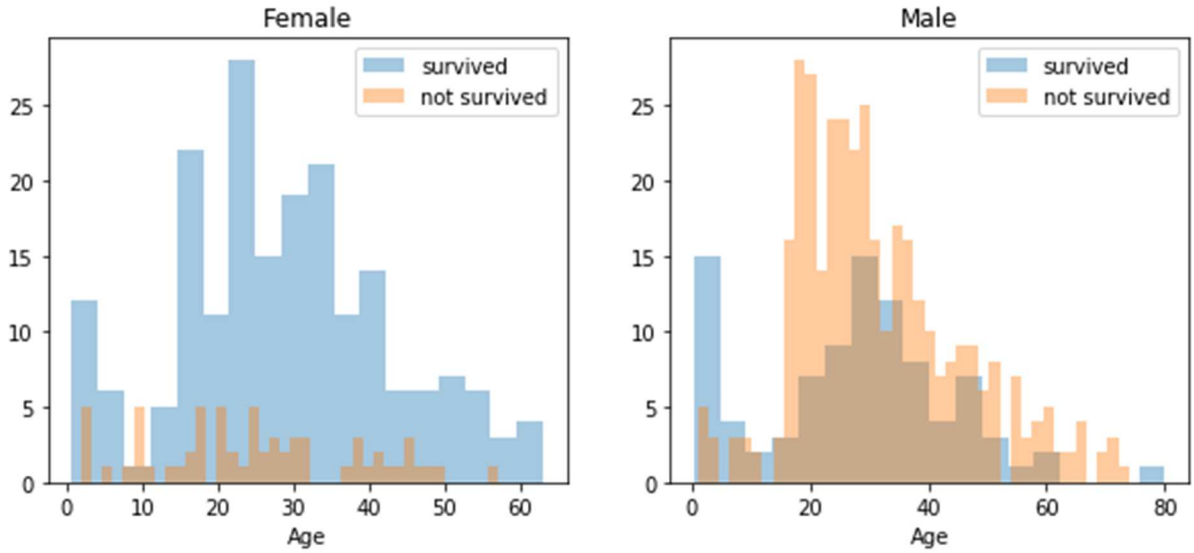
* : Le nombre d'enfants au travail au 19ème siècle : *http://www.droitsenfant.fr/travail_histoire.htm*

Nous avons voulu connaître quelles catégories de personnes allaient le plus décéder. L'épaisseur des traits représente le nombre de passagers concernés :

- Par classe : le nombre de passager réparti suivant la classe – 1ère, 2ème, 3ème
- Par catégorie d'âge
- Par port d'embarquement

Hélas, les graphiques représentent les passagers en nombre et non par ratio. Pas les suivants.

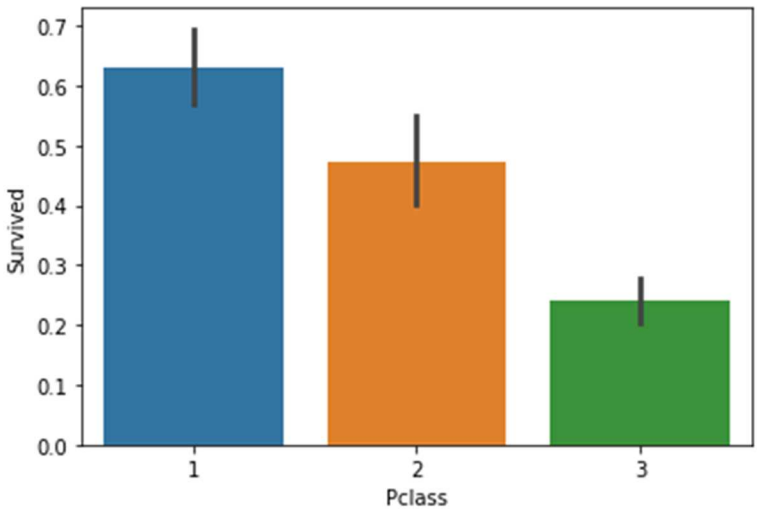
Analyse des survivants par âge et sexe :



Le graphique montre que les femmes ont plus survécu au drame que les hommes : le partie orange du graphique (survivants) recouvrant le bleu (disparus) chez les hommes, contrairement à le gente féminine.

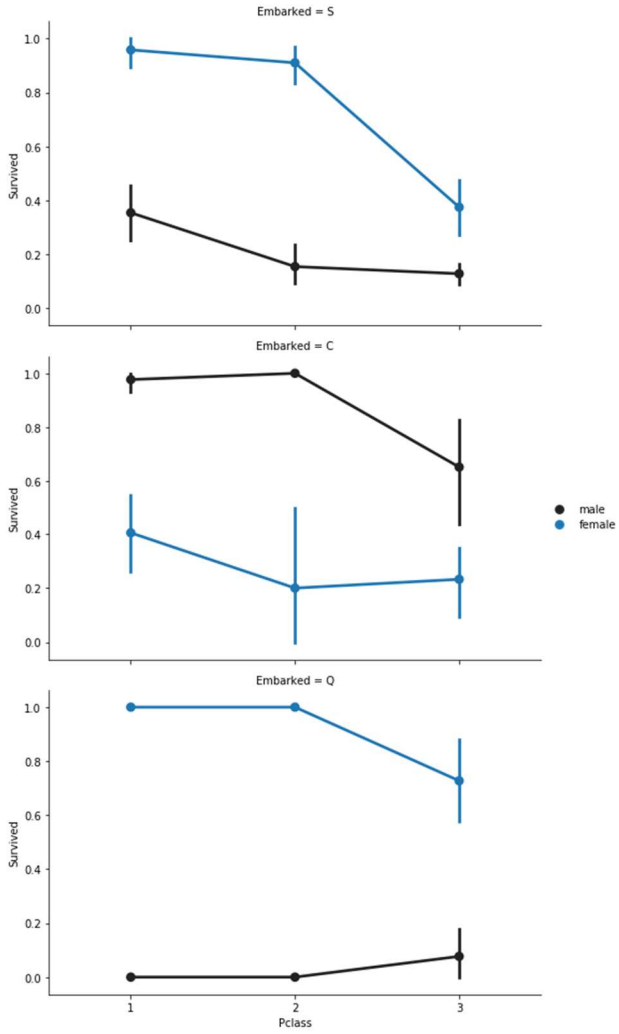
Autres informations : l'âge maximum des femmes est plus petit que celui des hommes..

Analyse par classe :



Le graphique est sans équivoque : les passagers de 1ère classe ont plus survécu que les autres classes.

Analyse par port d'embarquement :



On s'est demandé si le fait de monter dans un port ou un autre pouvait jouer sur la survie des passagers…

Port de Southampton :

La logique des classes et sexe est respectée…

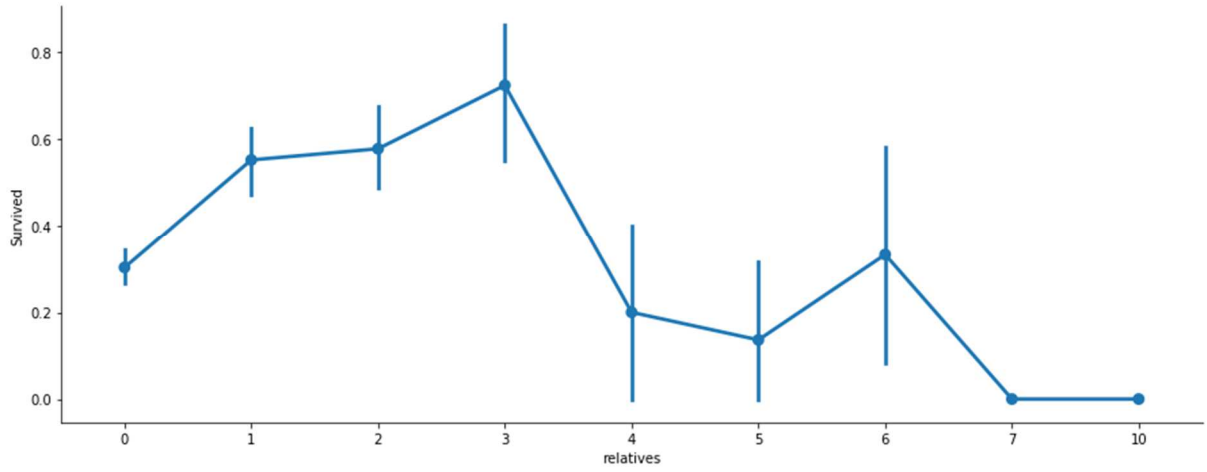
Port de Cherbourg :

Etonnant, mais ce sont les hommes qui, cette fois, ont le plus survécu. Les français ne seraient pas si respectueux galants ? Et la 2ème classe « homme » s'en ait mieux sortie que la 1ère ?

Port de Queenstown :

L'inverse du port français… mais même ratio entre 1ère et 2ème classe par contre ici.

Analyse suivant la taille de l'accompagnement :



D'après ce graphique, vous aviez plus de chance de survivre en ayant une petite famille de 2 à 4 personnes (1 à 3 + vous), puis personne célibataire (0). Par-contre, les familles plus nombreuses n'ont pas été favorisées. Relation avec la classe des gens ?

Analyse suivant la famille et la classe :

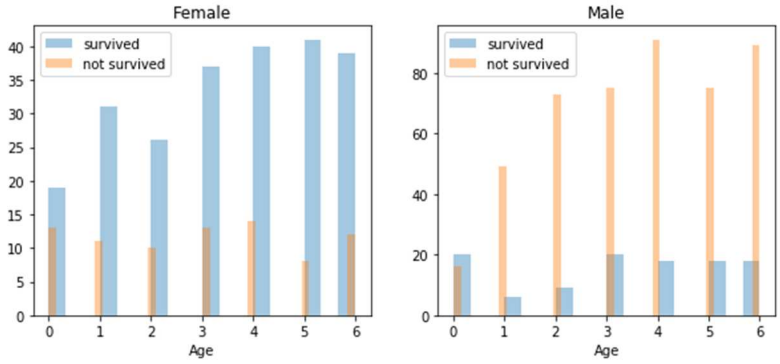
relatives	Pclass	Survécu		Total	Ratio survécu	
		Non	Oui		Non	Oui
0	1	51	58	109	0.47	0.53
0	2	68	36	104	0.65	0.35
0	3	255	69	324	0.79	0.21
1	1	19	51	70	0.27	0.73
1	2	16	18	34	0.47	0.53
1	3	37	20	57	0.65	0.35
2	1	6	18	24	0.25	0.75
2	2	10	21	31	0.32	0.68
2	3	27	20	47	0.57	0.43
3	1	2	5	7	0.29	0.71
3	2	3	10	13	0.23	0.77
3	3	3	6	9	0.33	0.67
4	1		2	2	-	1.00
4	2		1	1	-	1.00
4	3	12		12	1.00	-
5	1	2	2	4	0.50	0.50
5	2		1	1	-	1.00
5	3	17		17	1.00	-
6	3	8	4	12	0.67	0.33
7	3	6		6	1.00	-
10	3	7		7	1.00	-

Les célibataires (relatives à 0) ont plus de chance de s'en sortir en étant en 1ère classe (53%) contre une 2ème (35%) ou 3ème classe (21%), ce qui confirme les informations précédentes, la classe est importante. Et plus le nombre de personnes augmente, plus le risque de ne pas survivre augmente, d'autant plus en descendant en classe.

Résolution des hypothèses du sujet :

1. Faire un test d'hypothèse pour savoir si oui ou non, les enfants (moins de 11 ans dans notre test) ont été privilégiés lors du naufrage. Une autre utilisation de l'instruction crosstab :

Survived Age	0	1
0	29	39



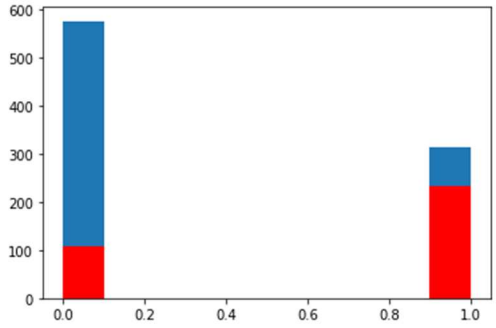
Le test du chi2 donnera les résultats suivants :

- Chi2 = 0
- Pvalue = 1.0

L'hypothèse H0 est rejetée, le fait d'être un enfant a bel et bien « joué » en leur faveur pour survivre.

2. Faire un test d'hypothèse pour vérifier si oui ou non, les femmes ont été privilégiées lors du naufrage. Nous utiliserons simplement la fonction croostab pour répondre à cette question.

Survived	0	1
Sex		
0	468	109
1	81	233



Le test du chi2 donnera les résultats suivants :

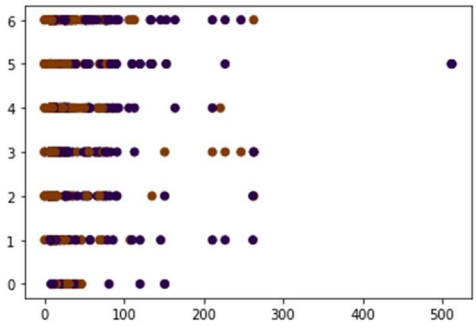
- Chi2 = 260.717
- Pvalue = 1.197

L'hypothèse H0 est rejetée, le fait d'être une femme a bel et bien « joué » en leur faveur pour survivre.

3. Faire un test d'hypothèse pour savoir si oui ou non, le prix du billet a une influence sur la survie d'un passager. Le crosstab donne cette fois :

Survived Fare	0	1
0	14	1
4	1	0
5	1	0
6	10	1
7	163	50

Ce graphique indique le nombre de survivants en fonction des tranches d'âge et valeur du billet. En violet, le nombre de survivants.



Le test du chi2 donnera les résultats suivants :

- Chi2 = 222
- Pvalue = 3.77 e-13

le pvalue est très largement inférieur à alpha, on peut accepter l'hypothèse zéro disant que le prix des billets a une influence sur la survie.

Passage au fichier de test :

Afin de compléter les informations manquantes dans le fichier de test (âge et survécu), nous allons tester plusieurs algorithmes.

Par facilité, nous allons mettre le même âge à tous les passagers dont la valeur n'est pas renseignée, l'âge moyen des autres passagers sur le fichier « train », une valeur aléatoire dans le fichier « test ».

Nous attribuons le port d'embarquement pour les 3 passagers non renseignés : Southampton, le plus présent dans les jeux de données

Pour faciliter le traitement des algorithmes, nous allons attribuer 8 casses d'âge.
De même pour les tarifs, 6 classes.

1er algorithme testé : la **descente de gradient**. Nous obtenons

- $R^2 = 0.675$
- RMSE = 0.569

2ème algorithme testé : le **random forest**. Nous obtenons

- $R^2 = 1$
- RMSE = 0.0

3ème algorithme testé : la **régression logistique**. Nous obtenons

	prédit 0	**prédit 1**
vrai 0	549	0
vrai 1	0	342

4ème algorithme testé : le **KNN**. Nous obtenons

- $R^2 = 0.919$
- RMSE = 0.284

En résumé, nous obtenons :

Model	**Score**
Logistic Regression	100.00
Random Forest	100.00
KNN	91.92
Gradient Decent	67.56

Nous garderons le « Random Forest » pour le jeu de données.

Le choix des variables qui serviront au test sont :

- Le sexe
- L'âge
- La classe
- Le prix du billet

Les autres variables n'entrant que faiblement dans le calcul des probabilités :

Variable	Importance
Survived	0.697
Sex	0.090
Title	0.088
Fare	0.032
Pclass	0.030
relatives	0.016
Age_Class	0.015
Age	0.010
SibSp	0.007
Embarked	0.006
Parch	0.004
not_alone	0.004

Pour jouer avec notre modèle, nous nous sommes « amusés » à tester différents cas de figures de passagers virtuels, pour savoir s'ils allaient survivre ou non.

	prédit 0	**prédit 1**
vrai 0	468	81
vrai 1	108	234

Nous avons $R^2 = 0.7878$

Pour conclure :

- Nous commençons à connaître le contenu des 2 jeux de données
- Nous savons choisir quelles variables utiliser après avoir vérifié leurs corrélations
- Nous extrayons les informations et en sortons des graphiques les plus parlants possibles
- Nous avons quelques algorithmes qui nous permettent de compléter les informations manquantes (âge, survie ou non)

Plus on met de variables, plus les prédictions s'améliorent.
Et inversement…