Projet n°2 | Par : Karima HARET et Thibaut FIOLKA

Prédiction de survie sur le Titanic en s'appuyant sur les modèles linéaires de Machine Learning

Petite histoire sur le Titanic :

Le Titanic, qui n'a pas entendu parler du naufrage du Titanic en 1912, cette tragédie restée gravé dans les mémoires.

« *Un choc, une explosion, un incendie. Puis un bateau qui prend l'eau et bascule, des passagers qui s'accrochent aux chaloupes ou sautent dans les vagues, et une mer qui monte à n'en plus finir. Peu de circonstances possèdent l'intensité tragique des naufrages, capables de propulser en quelques secondes des êtres humains devant la perspective d'une mort imminente. Et de les projeter tout aussi vite devant les décisions les plus cruelles, lorsqu'il faut choisir qui doit être secouru en priorité et qui peut attendre au risque d'y laisser la vie*. »

Introduction :

Notre projet consiste à se replonger dans l'histoire et essayer de répondre, en s'appuyant sur des méthodes statistiques et des algorithmes de Machine Learning, à certaines questions évoquées concernant les décisions prises lors des évacuations des passagers.

- Les règles sociales résistent-elles au naufrage ?
- L'adage « les femmes et les enfants d'abord » a-t 'il été respecté lors du naufrage ?
- Un passager aurait-il pu prédire sa probabilité de survie en fonction du prix de son billet ?

Récupération-compréhension des données :

Pour mener à bien notre projet, nous avons récupéré les données, concernant les passagers, du site de compétition de data science « Kaggle » et nous les avons implémenté dans notre logiciel.

Mais, avant de commencer toute études, nous devons d'abord comprendre et interpréter nos tableaux, avoir des premières réponses que nous confirmerons après à l'aide des moyens statistiques et algorithmes que nous allons utiliser.

Description des données :

Tableau descriptif global :

Voici un aperçu de nos jeux de données :

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

L'âge est en années, le prix en Livre Sterling (£), la survie est égale à 1 si la personne a survécu, 0 sinon.

Sur la colonne « Cabin » il y'a beaucoup d'informations manquantes, notées « NaN », ainsi que sur la colonne « Age » et « Fare ».

De ce faites nous avons supprimé la colonne « Cabin », jugée non importante, remplacée les données manquantes de l'âge par sa moyenne, supprimé les NaN, nombre pas important, du « Embarked ».

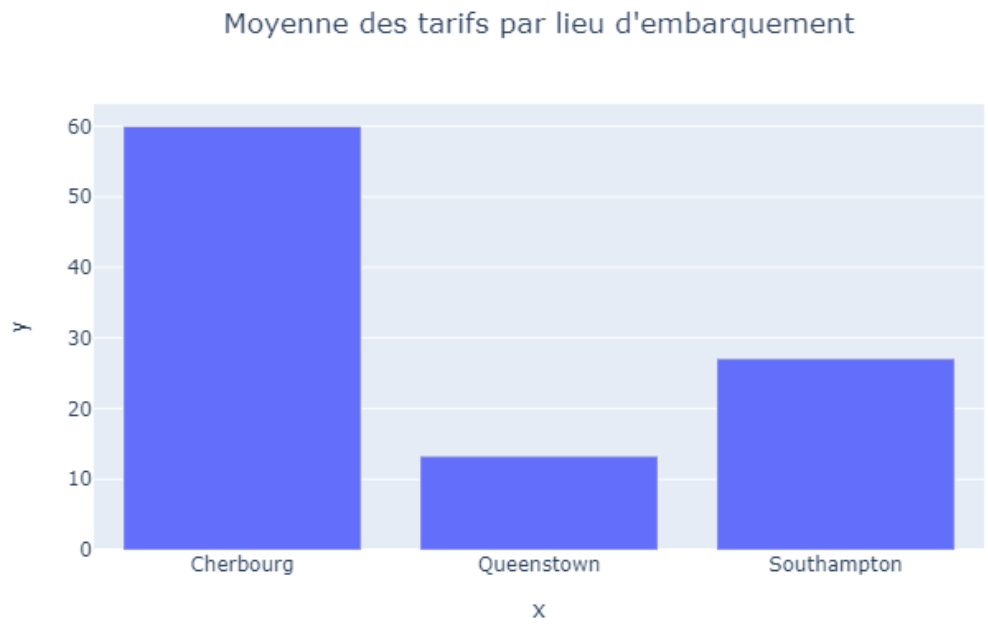
Ce qui réduit l'étude à un échantillon de 889 observations.

Tableau récapitulatif :

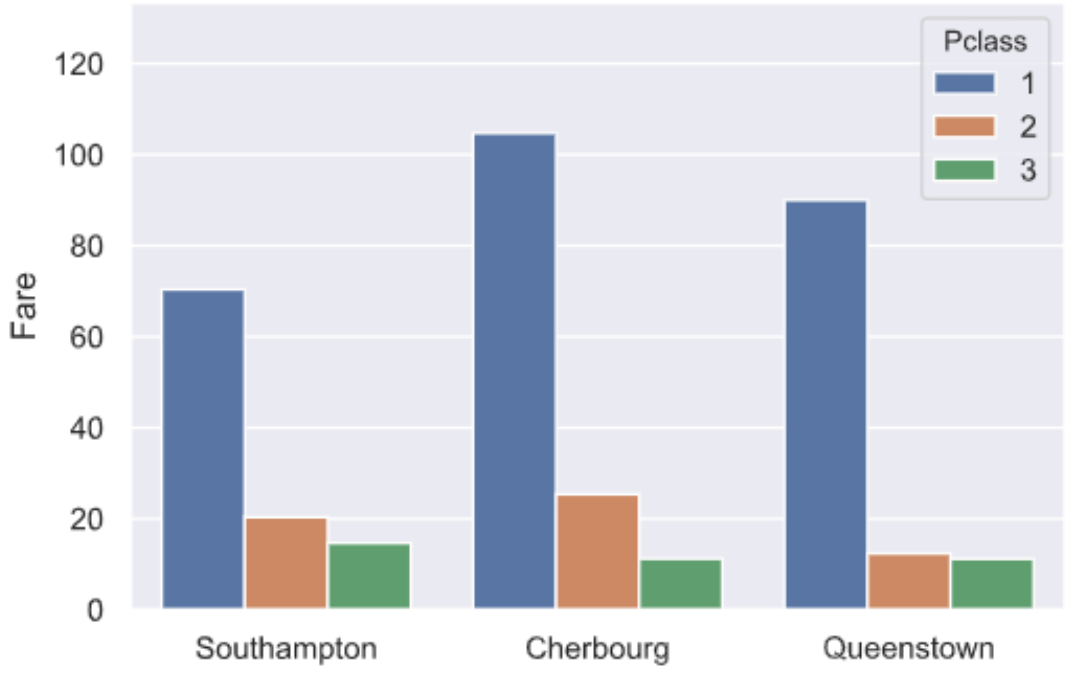
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Visualisation des données :

A l'embarquement des 889 passagers que nous avons sélectionnés, 644 ont embarqués à Southampton (Angleterre), 168 à Cherbourg (France), et 77 à Queenstown (Irlande).

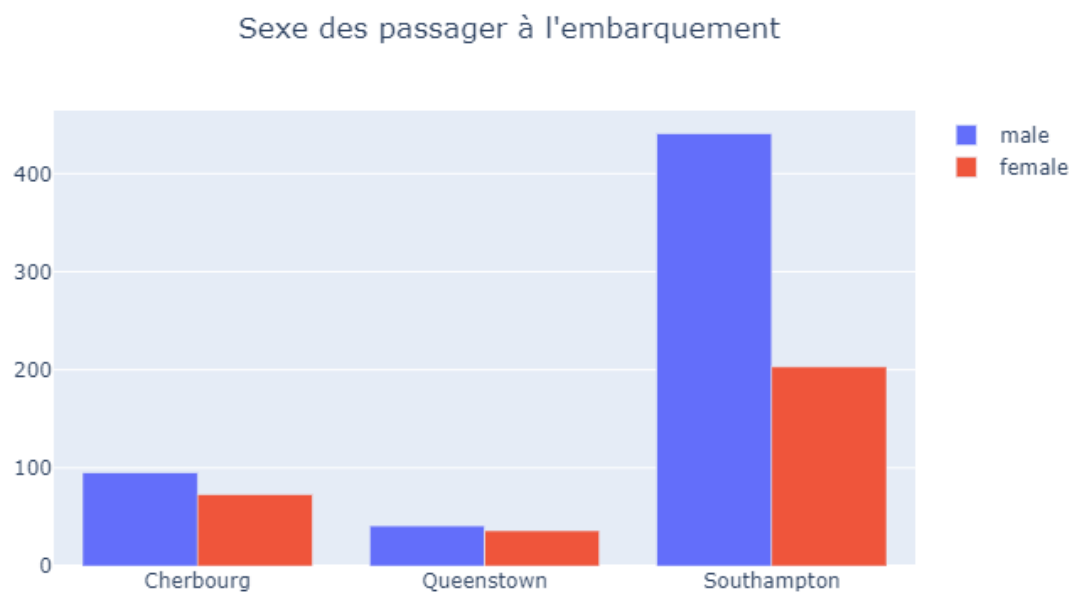


Nous pouvons voir ici que les passagers qui ont embarqué à Cherbourg ont payé plus cher que les autres.

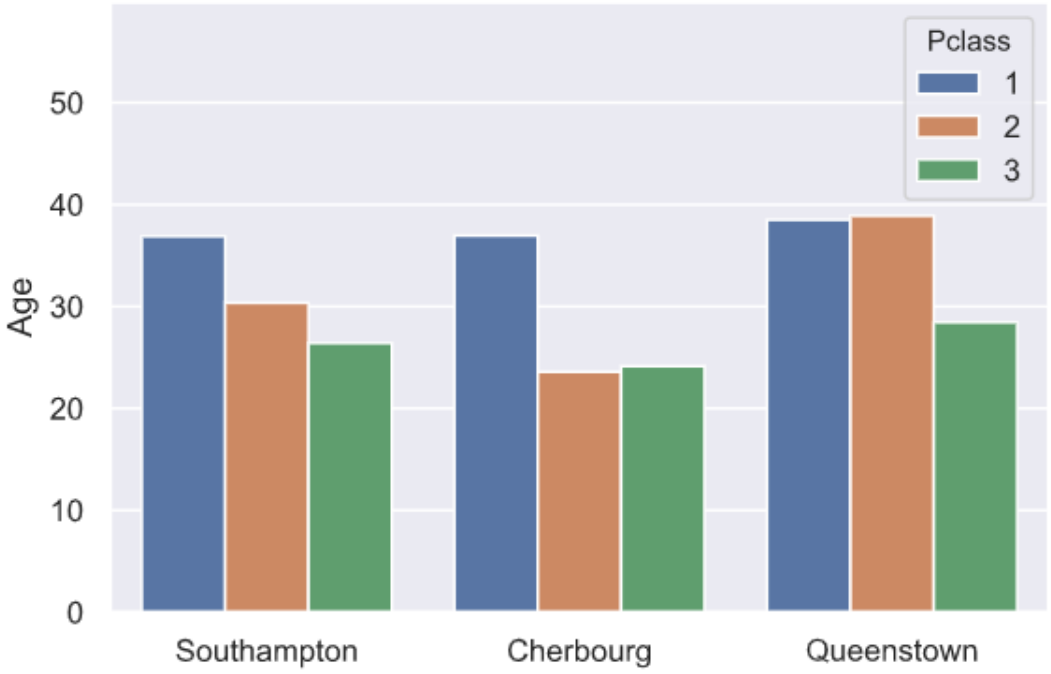


Nous pouvons également voir que la 3eme classe était majoritaire dans chaque lieu d'embarquement.

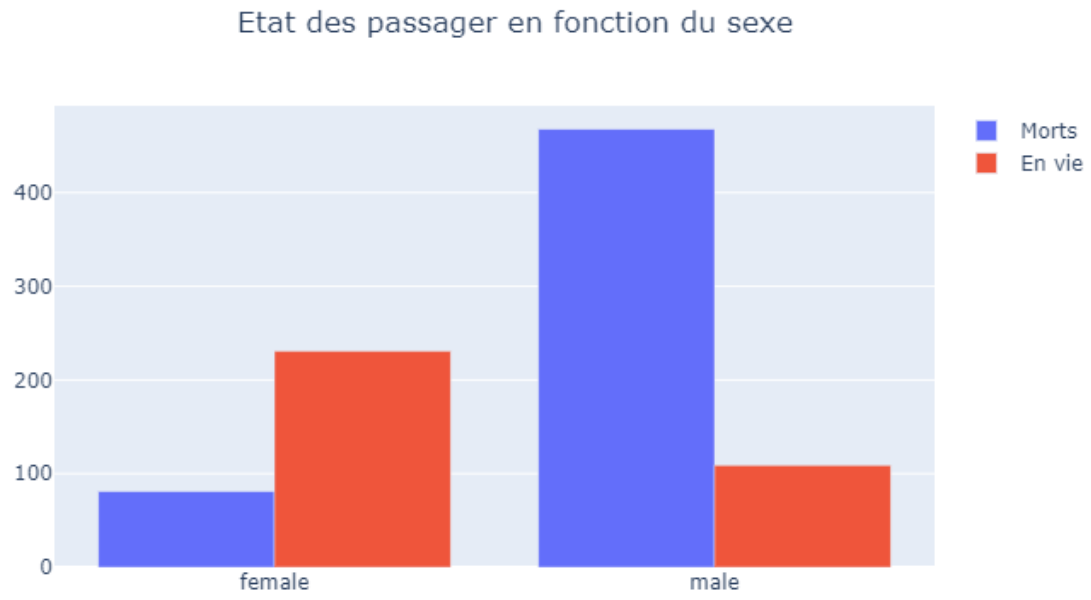
Voici la répartition des genres pour chaque lieu d'embarquement :

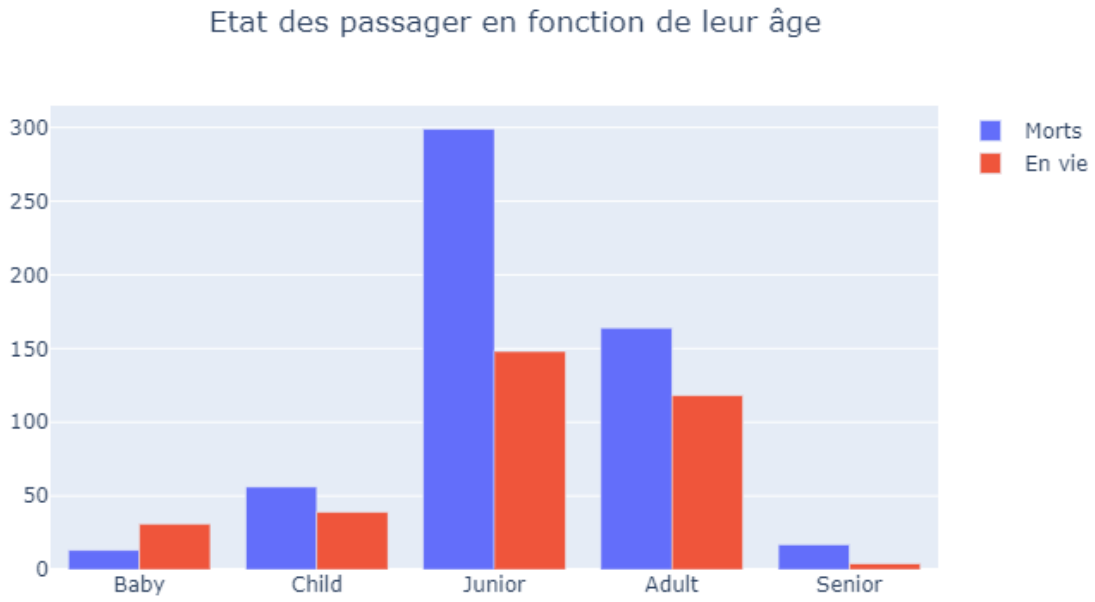


Ainsi que la répartition des âges :



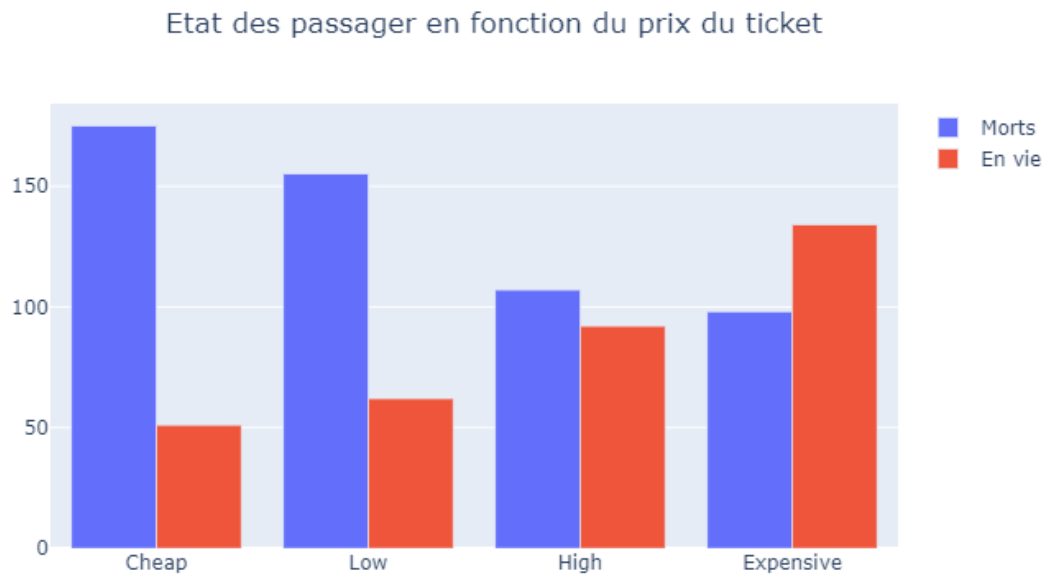
Puis, il y a eu le naufrage, et certaines personnes ont survécu, d'autres non. Quels sont les critères pour survivre dans une telle situation ? Les femmes et les enfants ont-ils été privilégiés ?





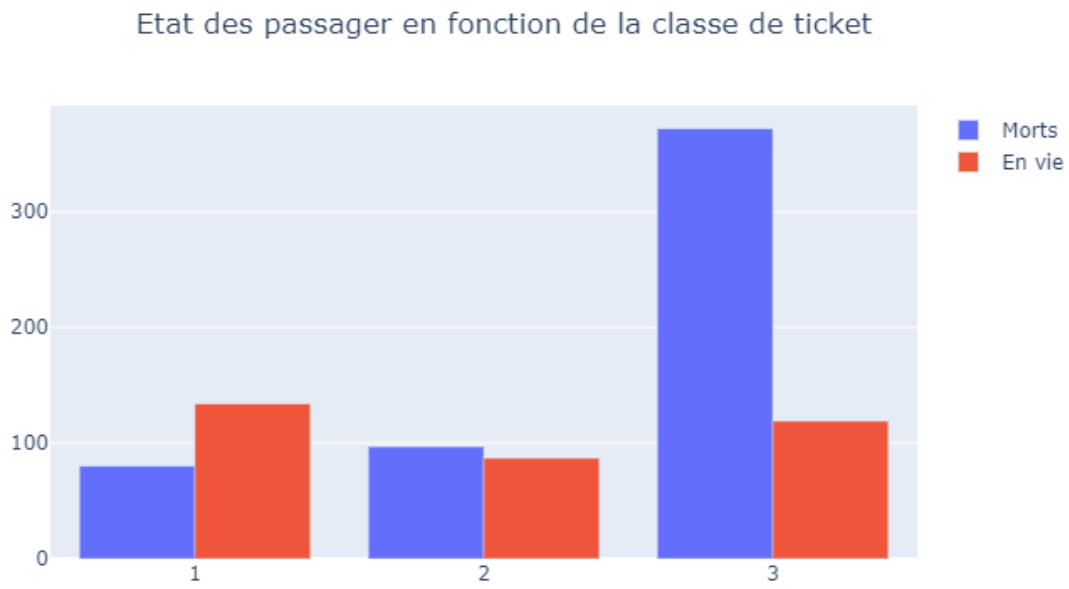
On peut voir qu'en effet, les femmes et les enfants ont été privilégiés.

Mais y'a-t-il d'autres critères ?

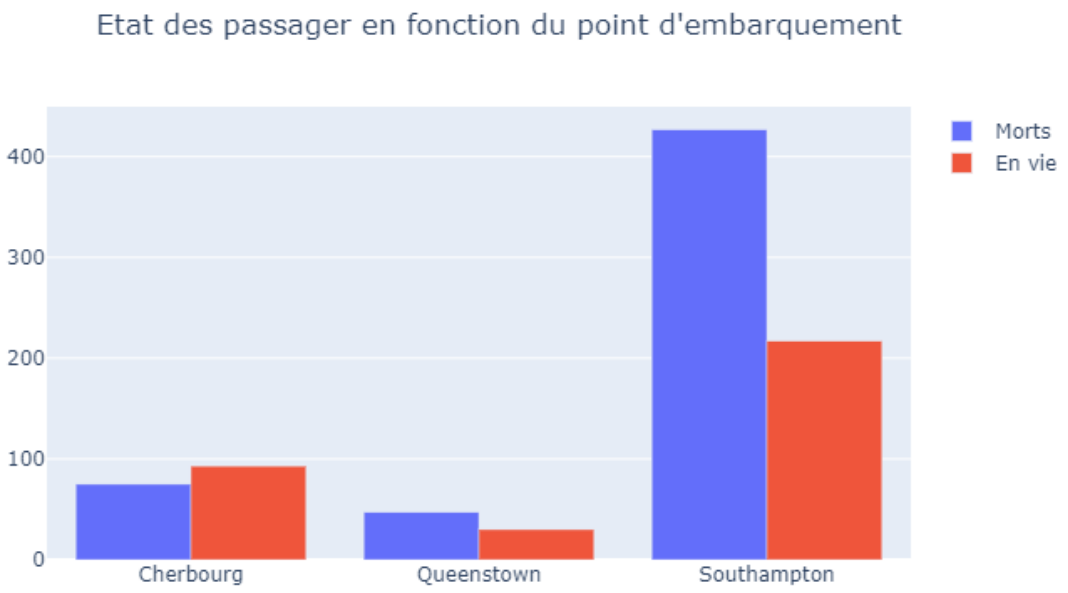


Sur ce graphe, on peut voir que plus le prix du ticket est élevé, plus les chances de survie sont hautes.

Même constatation au sujet des classes :



C’est ainsi que les passagers français, qui ont embarqués à Cherbourg et qui, rappelez-vous, ont payé leur ticket plus cher, comptent un meilleur taux de survie que les autres points d’embarquement :



Tests statistiques :

D’après les graphes précédents, nous pouvons dégager plusieurs hypothèses :

- 1 : Les femmes ont été privilégiées durant le sauvetage
- 2 : Les enfants (passagers étant âgés de moins de 18 ans) ont été privilégiés durant le sauvetage
- 3 : Les passagers ayant payé un ticket plus cher ont été privilégiés durant le sauvetage
- 4 : Les passagers ayant une meilleure classe ont été privilégiés durant le sauvetage
- 5 : Les passagers ayant embarqué à Cherbourg ont été privilégiés durant le sauvetage

Afin de vérifier ces hypothèses, nous allons mettre en place des tests statistiques, en particulier le test du khi-2.

Mais avant ça, nous allons simplifier les hypothèses à l’aide de la matrice de corrélation :

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
PassengerId	1.0	-0.005	-0.035	0.03	-0.058	-0.0017	0.013	-0.043	0.043	-0.0012	-0.034	0.022
Survived	-0.005	1.0	-0.34	-0.075	-0.034	0.083	0.26	0.54	-0.54	0.17	0.0045	-0.15
Pclass	-0.035	-0.34	1.0	-0.33	0.082	0.017	-0.55	-0.13	0.13	-0.25	0.22	0.076
Age	0.03	-0.075	-0.33	1.0	-0.23	-0.18	0.089	-0.089	0.089	0.034	-0.013	-0.022
SibSp	-0.058	-0.034	0.082	-0.23	1.0	0.41	0.16	0.12	-0.12	-0.06	-0.027	0.069
Parch	-0.0017	0.083	0.017	-0.18	0.41	1.0	0.22	0.25	-0.25	-0.012	-0.082	0.062
Fare	0.013	0.26	-0.55	0.089	0.16	0.22	1.0	0.18	-0.18	0.27	-0.12	-0.16
Sex_female	-0.043	0.54	-0.13	-0.089	0.12	0.25	0.18	1.0	-1.0	0.085	0.075	-0.12
Sex_male	0.043	-0.54	0.13	0.089	-0.12	-0.25	-0.18	-1.0	1.0	-0.085	-0.075	0.12
Embarked_C	-0.0012	0.17	-0.25	0.034	-0.06	-0.012	0.27	0.085	-0.085	1.0	-0.15	-0.78
Embarked_Q	-0.034	0.0045	0.22	-0.013	-0.027	-0.082	-0.12	0.075	-0.075	-0.15	1.0	-0.5
Embarked_S	0.022	-0.15	0.076	-0.022	0.069	0.062	-0.16	-0.12	0.12	-0.78	-0.5	1.0

Sur cette matrice, nous pouvons voir que le prix du ticket et la classe du passager sont corrélés, et que le prix du ticket et le lieu d'embarquement sont également corrélés. Nous pouvons donc conclure que les hypothèses 4 et 5 seront vérifiées par l'hypothèse 3.

Pour chaque test, nous prenons une marge d'erreur alpha = 5%.

Vérification de l'hypothèse 1 :

Avec le test du khi-2, nous obtenons une p-value bien inférieure à alpha, ce qui indique que l'hypothèse 1 est validée.

```
pvalue = 3.7799096665574906e-58 alpha = 0.05
Les deux variables ne sont pas indépendantes.
```

Vérification de l'hypothèse 2 :

Avec le test du khi-2, nous obtenons une p-value bien inférieure à alpha, ce qui indique que l'hypothèse 2 est validée.

```
pvalue = 0.00034261369274988666 alpha = 0.05
Les deux variables ne sont pas indépendantes.
```

Vérification de l'hypothèse 3 :

Avec le test du khi-2, nous obtenons une p-value bien inférieure à alpha, ce qui indique que l'hypothèse 3 est validée.

```
pvalue = 1.553812876621213e-11 alpha = 0.05
Les deux variables ne sont pas indépendantes.
```

Conclusion :

Les femmes, les enfants, et les passagers ayant payé cher leur billet ont été privilégiés durant le sauvetage.

Régression linéaire

Régression multiple sur toutes les variables :

Dans cette partie, nous cherchons à savoir si la survie peut s'exprimer en fonction de l'âge, la classe, le prix du billet et le sexe. On suppose qu'il existe une relation linéaire du type :

$$survived = \alpha_0 + \alpha_1.Age + \alpha_2.classe + \alpha_3.Fare + \alpha_4.Sex$$

Pour cela, nous décomposons notre jeu de données en deux parties Train et Test. La partie Train nous servira à faire l'apprentissage de notre modèle et Test pour tester sa performance en s'appuyant sur l'erreur des moindres carrés.

Nous estimons les paramètres $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ en important le module « LinearRegression » de la librairie « sklearn » de python, prédire si un passager survie ou pas avec notre modèle et puis calculer l'erreur moyenne quadratique en important le module « mean_squared_error ».

Les résultats obtenus sont résumés dans le tableau ci-dessus :

α_0	α_1	α_2	α_3	α_4	R^2	erreur
0.8375	-0.0056	-0.1924	0.00004	0.4612	0.46	0.46

Ce qui est de loin insuffisant car le coefficient de corrélation R^2 est éloigné de 1. Nous allons donc faire des régressions linéaires en ne prenant en compte que certaines variables.

Régression linéaire une à une :

Voici les R^2 obtenus pour chaque variable, prise individuellement :

Variable	Modèle	R^2
Age	$Survived = 0.5 - 0.004.Age$	-0.029
Classe	$Survived = 0.509 - 0.004.Classe$	0.097
Fare	$Survived = 0.509 - 0.004.Fare$	0.015
Sexe	$Survived = 0.509 - 0.004.Sexe$	0.39

Nous remarquons la variable « Sexe » à le plus grand R^2 .

Nous allons faire une régression avec des combinaison de la variable « Sexe » et les autres variables. Les résultats se résument ci-dessous :

Variable	Modèle	R^2
Sexe + Classe	$Survived = 0.59 - 0.48.Sexe - 0.16.Classe$	0.45
Sexe+ Classe+ Fare	$Survived = 0.58 - 0.48.Sexe - 0.16.Classe + 0.0002.Fare$	0.46

Nous remarquons que le R^2 ne change pas même si on ajoute d'autres variables après la variable « Sexe » et reste loin de 1.

Conclusion :

La régression linéaire n'est pas envisageable sur notre jeu de données.

Régression logistique

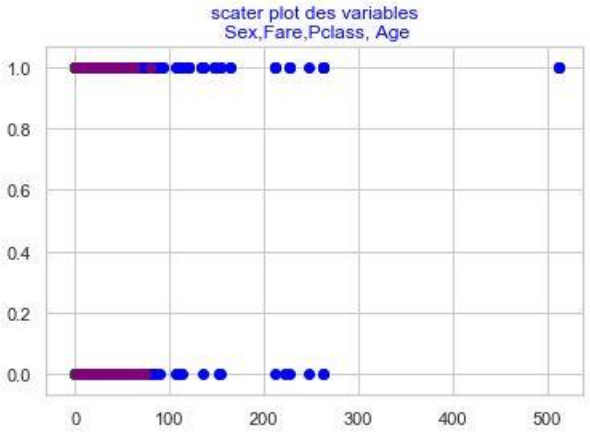
L'intérêt de la régression logistique est de caractériser les relations entre une variable dépendante (ou variable à expliquer) et une seule (régression logistique simple) ou plusieurs variables prises en compte simultanément (régression logistique multiple). Il s'agit donc d'un modèle permettant de relier la variable dépendante (Y), qui est qualitative, aux variables explicatives (X_1, X_2, X_3, …, X_n).

Dans notre cas, il s'agit d'expliquer la « survie » en fonction des autres variables comme précédemment mais en utilisant la fonction « logit » suivante :

$$F(x) = \frac{e^x}{1+e^x} \quad \text{avec} \quad x = \alpha_0 + \alpha_1.Age + \alpha_2.classe + \alpha_3.Fare + \alpha_4.Sex$$

Logit(x) = $\alpha_0 + \alpha_1.Age + \alpha_2.classe + \alpha_3.Fare + \alpha_4.Sex$

Nous avons tracé le nuage de points de la variable survie et fonction des 4 autres, le résultat est dans le graphe ci-dessus :



Nous avons utilisé le jeu de données cité dans le chapitre ci-dessus, nous avons importé le module « **LogisticRegression** » de « **sklearn** », faire l'entrainement de notre modèle sur l'ensemble Train, ensuite prédire la survie sur l'ensemble « Test ».

Pour tester la performance du modèle, nous avons importé le deux module « **roc_curve** » et « **auc** » qui visualise et calcule le ratio entre les faux positifs sur la somme totale des individus.

Paramètres estimés :

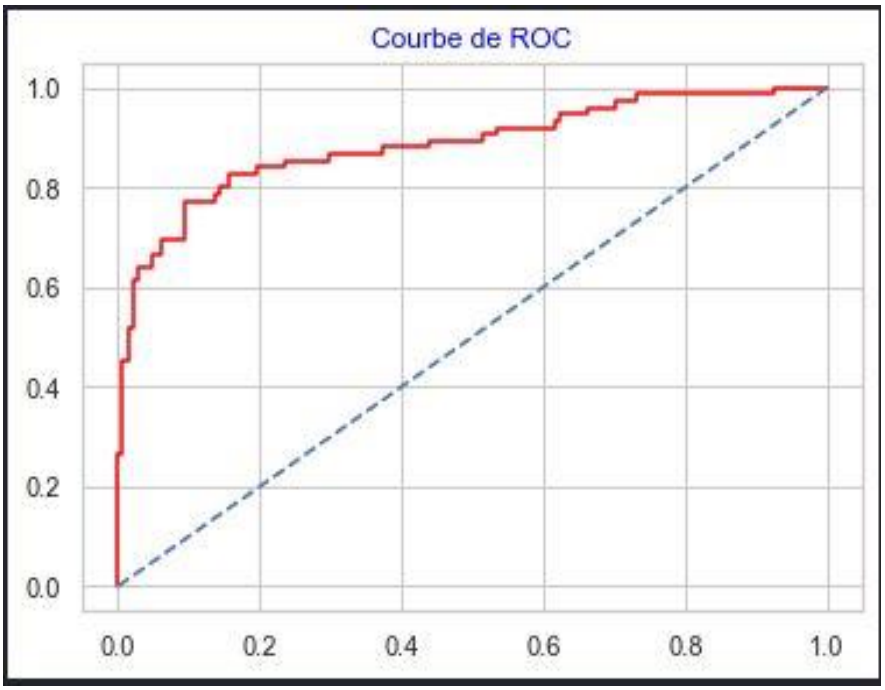
$$\alpha = \begin{pmatrix} 2.1445 \\ -1.100 \\ 2.272 \\ -0.003 \\ 0.0008 \end{pmatrix}$$

AUC = 0.89

Notre modèle de régression logistique s'écrit :

Logit(x) = $2.145 + -1.100\,Age + 2.272\,classe + -0.003\,Fare + 0.0008\,Sex$

La courbe de Roc obtenue :



Auc = 0.89 très proche de 1, nous jugeons que notre modèle est performant.

Conclusion

Ce projet nous aura permis de lire scientifiquement des données historiques liées à un évènement

tragique. Nous sommes donc maintenant en mesure de dire que l'adage dont il a été question dans ce dossier a été respecté. De même, nous avons démontré que la survie d'un passager était très liée à sa richesse. Cette dernière conclusion nous prouve que des privilèges étaient encore valables sur ce paquebot titanesque.

Ainsi nous avons pu comparer la performance de deux algorithmes de Machine Learning qui sont la régression logistique et linéaire. L'algorithme de régression logistique est le plus adapté dans notre cas d'étude.