PRAC2: Neteja i validació de dades

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Alguns exemples de dataset amb els que podeu treballar són:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic). L'últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició.

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset escollit per a la pràctica és l'obtingut a partir de la pàgina web de la competició de *Kaggle* "Titanic: Machine Learning from Disaster". Aquest dataset és un clàssic del camp del Machine Learning que proposa intentar predir qui va sobreviure o no a l'enfonsament del Titanic a partir de les dades disponibles dels viatgers, i per tant, suposant que la sort no va ser l'únic factor important. Algunes de les variables que van poder ser determinants són: l'edat, el sexe, el nombre d'acompanyants o el preu del tiquet. Per tant, a partir d'una possible classificació de supervivència o no dels passatgers, també es podria inferir quins factors van ser els més importants en aquesta situació (si el classificador escollit ho permet). A més, aquest dataset suposa un dels primers reptes que la comunitat de Kaggle ofereix per a que la gent es familiaritzi amb la plataforma i les seves competicions, per tant, es podria dir que és d'aquells problemes que tard o d'hora s'ha d'intentar resoldre.

El dataset està format per 1309 files (891 de train i 418 de test) que representen viatgers del Titanic i 12 columnes amb informació que els defineix. Les variables que trobem són:

- **PassengerId**: número identificador del viatger
- **Survival**: supervivència (1) o no (0) del viatger a l'enfonsament.
- **Pclass:** classe de ticket segons el preu (alta=1, mitja=2, baixa=3).
- **Name**: nom complet del viatger
- **Sex**: sexe del viatger ("male" o "female")
- **Age**: edat del passatgers. Si la persona era menor a 1 any es presenta en forma de fracció.
- **Sibsp**: nombre de germans, germanastres i/o marit/muller que el viatger tenia a bord. No es consideren per a aquest camp ni els amants ni els promesos.
- **Parch**: nombre de pares, fills i/o fillastres que el viatger tenia a bord. Alguns nens viatjaven només amb la mainadera, per tant, parch = 0.
- **Ticket**: número de tiquet del viatger.
- **Fare**: preu del ticket.

- **Cabin**: cabina assignada al viatger.
- **Embarked**: port a on es va embarcar el viatger (C = Cherbourg, Q = Queenstown, S = Southampton).

2. Integració i selecció de les dades d'interès a analitzar

En aquest cas eliminarem les tres variables que permeten identificar cada registre de manera unívoca (un amb un nombre enter i dos amb un string): "PassengerId", "Ticket" i "Name". Això és degut a que a l'hora de crear un classificar no aporten informació que permeti discernir entre la supervivència o no del viatger. Per tant, ens quedarem amb 8 variables independents per a intentar predir la variable "Survived".

3. Neteja de les dades

Per a la neteja de dades primer de tot ens centrarem en el training set, d'on extraurem els procediments i paràmetres necessaris per a normalitzar tant el training set com el testing set. La raó per a fer-ho d'aquesta manera és perquè els registres del test representen dades que serveixen per a comprovar el rendiment d'un classificador, i per tant, s'obtenen després d'haver realitzat la normalització del training set i la creació de models de predicció.

1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Variable	NaN	Percent
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22
Survived	0	0.00
Pclass	0	0.00
Sex	0	0.00
SibSp	0	0.00
Parch	0	0.00
Fare	0	0.00

Com es pot veure a la taula, en el training set trobem tres variables amb valors buits: "Cabin", "Age" i "Embarked".

- Com la columna "Cabin" té valors buits en més del 70% del registres eliminarem la columna.
- "Age" presenta 177 valors buits, per tant no eliminarem la columna i intentarem reemplaçar aquests valors amb la mediana de les observacions que formen part del mateix grup que la observació que volem arreglar. Les variables que considerarem per a decidir si forma part del mateix grup o no seran "Pclass", "Sex" i "Parch". Per tant, les observacions que tindrem en consideració per a reemplaçar el valor buit seran de la mateixa classe econòmica, el mateix sexe i el mateix nombre de pares i/o fills. A més, la raó per la que fem servir la mediana en la imputació és per a evitar l'error ocasionat pels outliers presents en els diferents grups.
- Com "Embarked" només té 2 valors buits reemplaçarem els valors manualment utilizant la informació present a internet sobre els viatgers del Titanic (al jupyter notebook estan els links).

Variable	NaN	Percent
Cabin	327	78.23
Age	86	20.57
Fare	1	0.24
Pclass	0	0.00
Sex	0	0.00
SibSp	0	0.00
Parch	0	0.00
Embarked	0	0.00

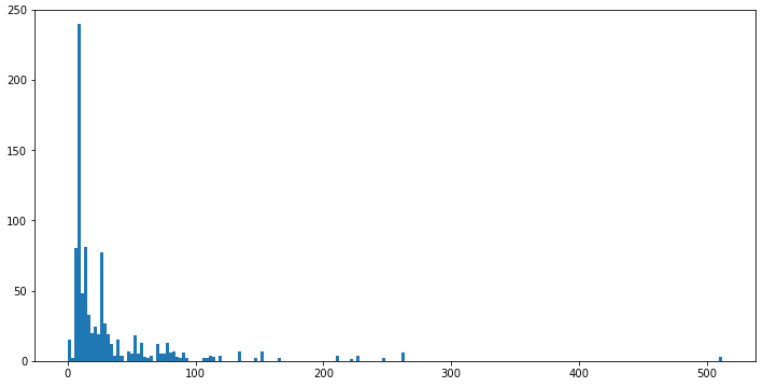
Un cop substituïts tots els valors buits del training set passem a avaluar el test set. Com es pot veure en la segona taula, té un comportament molt semblant a les dades d'entrenament. La columna "Cabin" també s'haurà d'eliminar degut a que falten el 78% dels valors i la variable "Age" té un 20%, per tant, els missing values seran

reemplaçats per la mediana de les observacions de la mateixa classe econòmica, el mateix sexe i el mateix nombre de pares i/o fills del training set. A més, veiem que "Fare" té un missing value que correpon a un registre del que no trobem aquesta informació a internet. Per aquesta raó, el seu valor serà substituït per la mediana de les observacions del set d'entrenament que formin part de la mateixa classe social, del mateix sexe i amb el mateix nombre de pares i/o fills i de germans i/o marit/muller.

Seguidament s'ha passat a comprovar si hi ha columnes amb valors 0. Com es pot veure a la tercera taula, les variables "Survived", "SibSp", "Parch" i "Fare" en tenen, però en totes elles el 0 té un significat coherent. A "Survived" el 0 senyala la no supervivència i a "SibSp" i "Parch" la absència de familiars en el vaixell.

	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.018148	0.523008	0.381594	32.204208
std	0.486592	0.836071	13.577341	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	21.500000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	26.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	36.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

En el cost del tiquet el 0 es podria considerar com a un missing value o com a un error, ja que en principi significaria que la persona va viatjar gratis. Una altra possibilitat però, podria ser que es tractés de tiquets que corresponguessin a treballadors del Titanic, idea que es troba reforçada amb la poca freqüència amb la que es troben aquests costs:

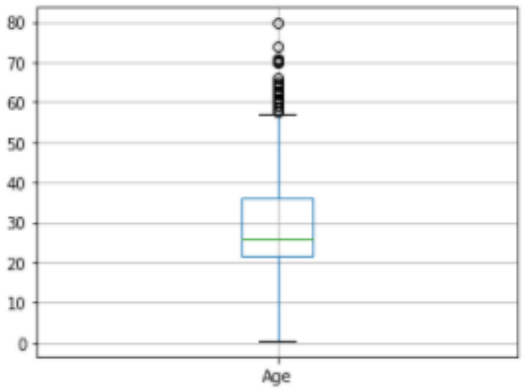


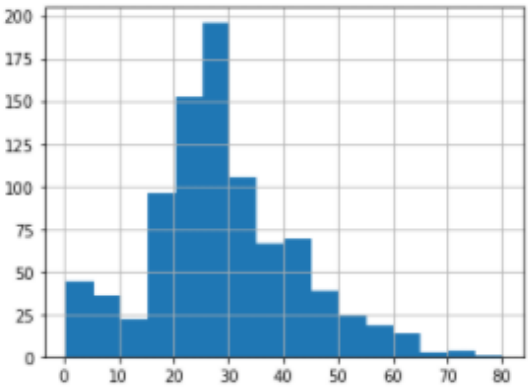
Per tant, considerarem com a bons els valors 0 en la variable "Fare".

2. Identificació i tractament de valors extrems

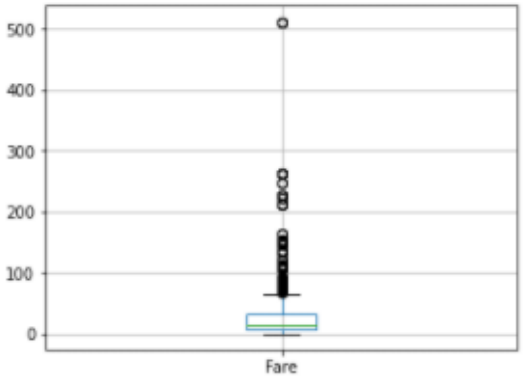
En aquest apartat ens centrarem en les variables "Age" i "Fare", i en principi consideraríem com a outliers aquells valors que sobrepassin el 1.5 del Rang Interquartil de les variables.

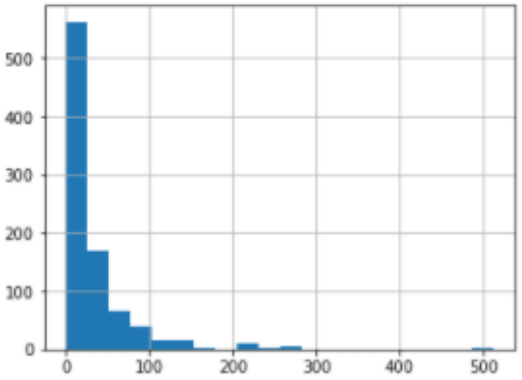
"Age" està entre els valors de 0.42 i 80 anys d'edat, i té una mitjana de 29 i una mediana de 26, per tant els outliers no tindrien un gran impacte en aquesta variable. Per sobre dels 55 anys trobem 33 valors extrems, però degut a que una edat entre 55 i 80 és coherent amb l'esperança de vida dels humans, a que cada cohort és menor en freqüència que l'anterior, i a que aquestes edats no difereixen molt de la resta, en aquesta variable no considerarem cap valor com a outlier.



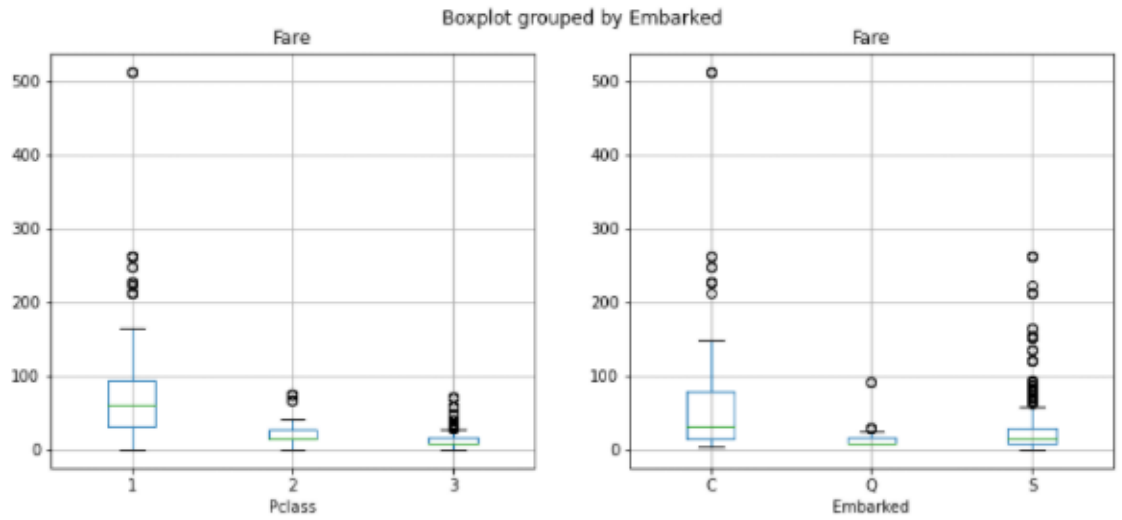


En quant a “Fare” trobem valors entre 0 i 512, però aquest cop els outliers si difereixen molt de la resta de valors. Això es veu reflectit en la mitjana i la mediana, que tenen uns valors de 32 i de 14. Els 116 outliers estan tots per sobre del 65, però com ja hem comentat, pot arribar fins al 512.





Per intentar entendre aquesta variabilitat s’han creat els diagrames de caixes separant els valors de “Fare” segons la classe econòmica i el lloc d’embarcament:



Com es pot veure en el gràfic de la dreta, els tiquets més cars tindrien coherència, ja que corresponen a la 1a classe i per tant a priori no serien deguts a errors humans. A més, sembla que els tiquets més cars estan repartits entre Cherbourg i Southampton, potser degut a que són les ciutats amb el nivell adquisitiu més alt. Per últim, apuntar que els 3 passatgers que superen el valor de 500 van sobreviure tots (a diferència dels de 200), potser degut a que van rebre un tractament especial durant l'enfonsament, això significaria que mantenir aquesta informació podria ser interessant.

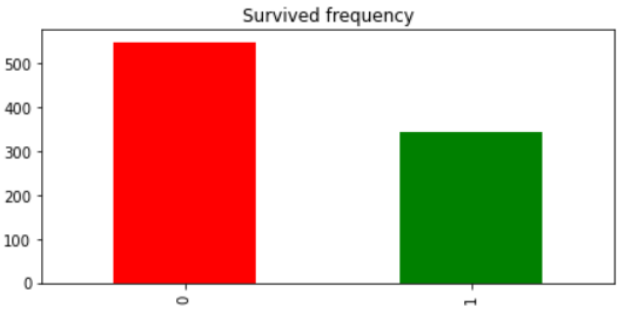
Per tant, tenint en consideració que aquests valors tenen coherència i poden aportar informació a un possible classificador, no modificarem els valors extrems de la variable "Fare".

4. Anàlisi de les dades

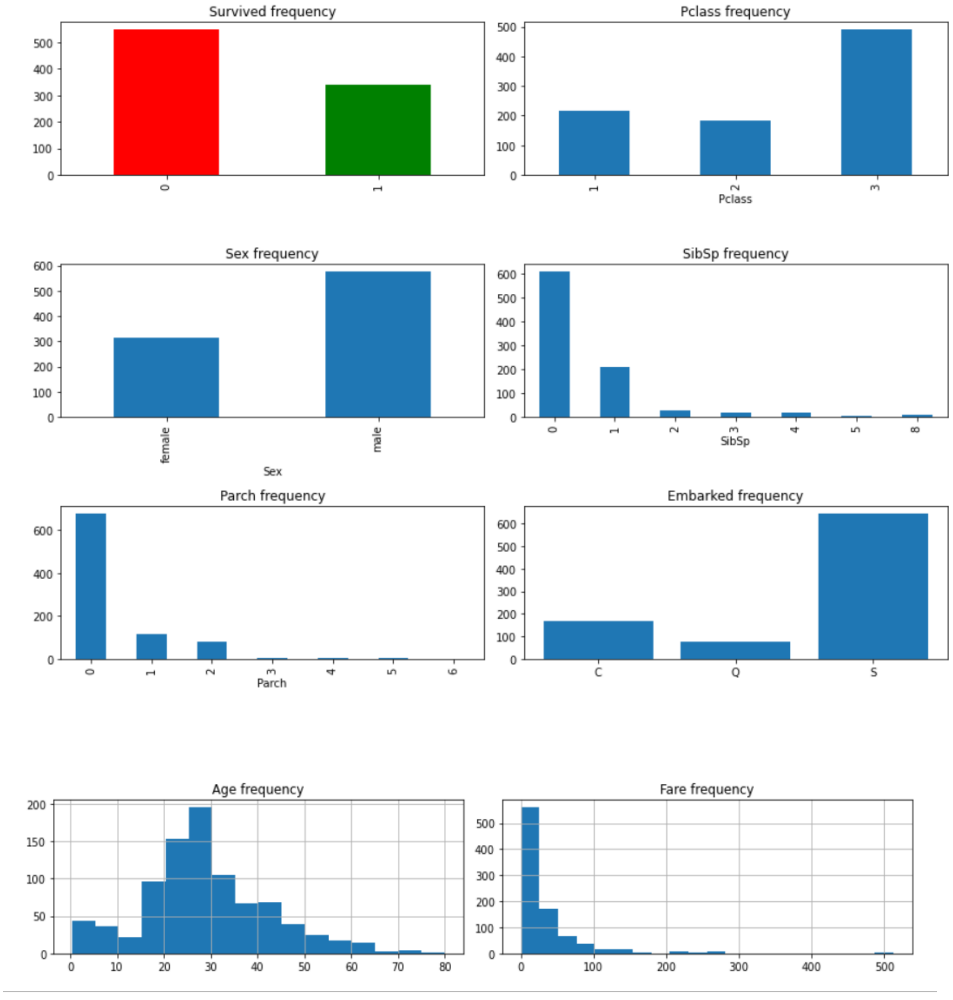
Tot i que el primer que se'ns demani en aquest apartat sigui determinar quins grups de dades es volen analitzar o comparar, trobo interessant acabar el que hem iniciat en l'apartat interior: l'anàlisi de la resta de variables.

Farem dos tipus d'anàlisis: univariant i multivariant. En el primer mirem la distribució d'aquestes variables en sí mateixes, mentre que en el segon busquem veure la distribució/possible relació d'aquestes entorn a la nostra variable objectiu: la supervivència dels passatgers.

El primer que mirem és la distribució de la supervivència, i observem que no ens trobem davant d'un dataset gaire des-balancejat:



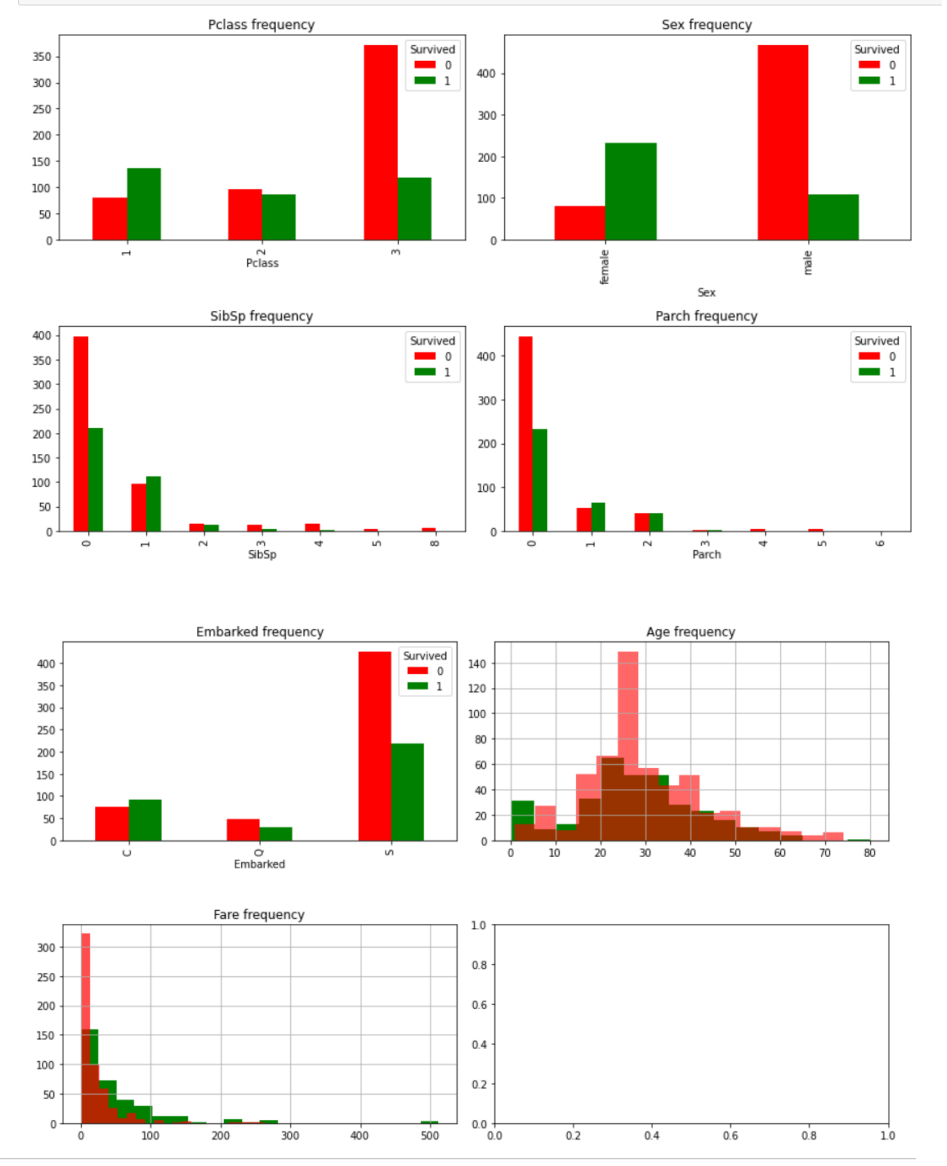
A més a més, de les dades podem veure que quasi la meitat de tripulants eren de la classe més baixa (mentre que les dues primeres classes es reparteixen prou equitativament), que hi havia el doble d'homes que de dones, que més de la meitat de la gent embarcava a Southampton i que el nombre de germans/esposos i de pares/fills segueixen una distribució molt similar a una logarítmica.



Les variables edat i preu del tiquet (variables numèriques) es poden trobar explicades més detalladament en l'apartat anterior.

És interessant veure, però, com es relacionen aquestes dades amb la variable objectiu per saber, així amb una ullada ràpida, si existeix la relació amb aquesta.

Quan fem una primera inspecció visual veiem com moren moltíssimes més persones proporcionalment de la classe més baixa que de les altres dues, i que sobreviuen més dones que homes. Addicionalment, també veiem que, proporcionalment, es tendeix a salvar les persones que tenen mínim un germà/germana/marit-esposa i les persones que tenen mínim un fill/filla/pare/mare a bord. Finalment veiem també que proporcionalment se salva menys gent que va embarcar a Southampton que a Queenstown, i que els embarcats a Cherbourg sobreviuen més que moren. Finalment sembla intuir-se de la distribució del preu de tiquets que, també proporcionalment, tendeixen a sobreviure més les persones que van pagar un ticket més car (les persones de classe més alta), com ja hem pogut comprovar en la distribució de Pclass. També observem que la gent més jove tendeix a sobreviure més.



1. Selecció de grups de dades a comparar/analitzar

Així doncs, podem preveure més o menys quines són les dades o grups de dades que estudiarem/analitzarem/compararem: bàsicament analitzarem els grups de supervivència i no supervivència, però sub-dividits (o més ben dits, estudiats/analitzats) segons els grups de sexe,

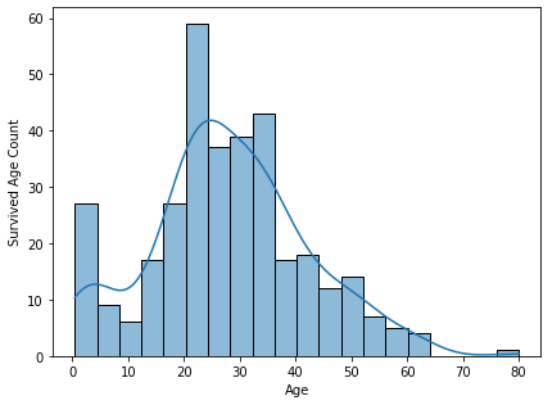
classes, edats i preu de tiquet. Ja anticipant-nos una mica, presentarem que és el que realitzarem en l'últim apartat al voltant d'aquests grups.

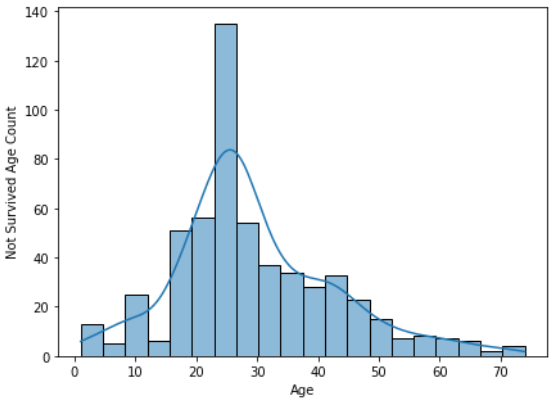
La nostra intenció és, com a objectiu principal, entendre qui sobreviu i qui no i arribar també a crear un model de predicció. Per fer això usarem estudis de correlacions entre variables numèriques i testos estadístics per estudiar relacions/dependències entre les variables i per veure si existeixen diferències significatives.

2. Comprovació normalitat i homogeneïtat de la variança

En el nostre cas, només es pot comprovar normalitat i homocedasticitat en les variables *Age* i *Fare* (les numèriques). De totes maneres, només veient la distribució de l'histograma del preu del tiquet per a supervivents i no supervivents (últim gràfic del punt 4) veiem que ens podem estalviar el test de normalitat, ja que té un clar comportament no normal.

Quan procedim a l'estudi de la normalitat segons la variable edat, decidim dividir la nostra població d'estudi en dos grups: els supervivents i els no supervivents. En ambdós sub-poblacions podem rebutjar, després d'executar un test de Shapiro-Wilk, la hipòtesi nul·la de normalitat de les dades. A més a més, després d'aplicar un test de Fligner-Killeen (test no paramètric), veiem que podem rebutjar la hipòtesi nul·la d'igualtat de variàncies.

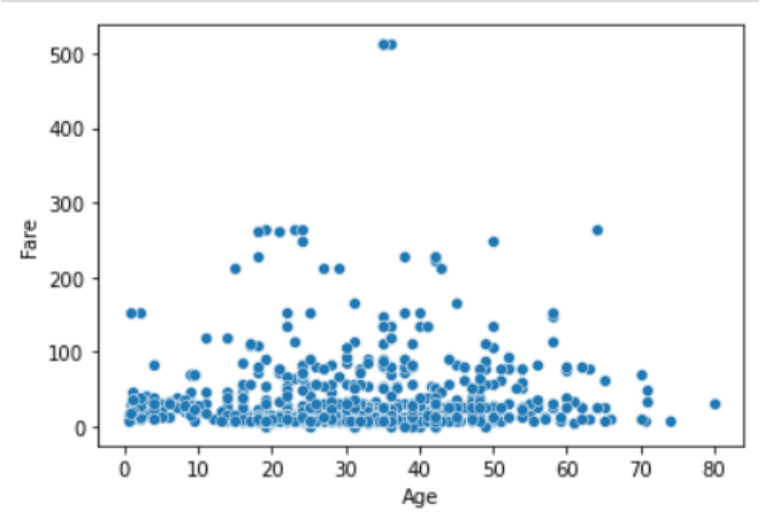




3. Anàlisi estadístic de les dades

3.1. Correlació edat-preu

El primer anàlisi estadístic que fem és la comprovació de correlació entre les dues úniques variables numèriques que tenim: edat i preu del tiquet. Primer de tot en representem la dispersió, a partir de la qual ja veiem que no existeix correlació. Tot i això, calculem el coeficient de Spearman obtenint un valor de 0.15:



3.2. Mann-Whitney test: survived vs not-survived

El següent estudi que fem és l'equivalent dels no paramètrics al *t-test*: el test de Mann-Whitney. Aquest test parteix de la hipòtesi nul·la que la probabilitat de que la observació d'una població sigui superior a la observació de l'altra és del 50 % (és a dir, que la probabilitat d'observacions és igual per a ambdues poblacions). Mitjançant aquest test descobrim que no hi ha diferències significatives entre l'edat dels que sobreviuen i els que no, obtenint un *p-value* de 0.15.

3.3. Mann-Whitney test: male survived vs male not-survived; female survived vs female not-survived

Immediatament després, decidim repetir aquest test, però, aquest cop, ho farem per cada un dels sexes: separem el nostre conjunt de dades en supervivents i no supervivents del sexe masculí i el mateix per al sexe femení. Seguidament n'analitzem la normalitat i veiem que aquestes dues noves *sub-poblacions* tampoc es troben distribuïdes normalment. És per això que tornem a aplicar un test de *Mann-Whitney* entre supervivents i no supervivents (per a cada sexe); obtenint un *p-value = 0.18* (per al cas dels homes) i un *p-value = 0.001* per al cas del sexe femení. Què significa això? Doncs que la supervivència no es veu afectada per l'edat en el cas dels homes (no podem rebutjar *H0*), però sí es veu afectada quan estem parlant de dones (*p-value > 0.05* → Rebutgem *H0*). Finalment calculem les mitjanes i intervals de confiança per cada un d'aquests 4 grups (home supervivent, home no supervivents, dones supervivents, dones no supervivents) i obtenim el següent:

```
Male Survived CI: [ 0.83 ,  57.2 ]
Male Survived mean:  27.69

Male Non-Survived CI: [ 9.0 ,  62.65 ]
Male Non-Survived mean:  30.74

Female Survived CI: [ 2.8 ,  58.0 ]
Female Survived mean:  28.28

Female Non-Survived CI: [ 2.0 ,  48.0 ]
Female Non-Survived mean:  23.0
```

3.4. Chi Squared: Survived vs Sex

Un cop vist que la supervivència guarda relació significativa amb l'edat en el cas en què parlem del sexe femení, ens plantegem si, més que l'edat, aquesta última troballa podria tractar-se degut al fet que la supervivència guardi relació significativa amb pertànyer a un sexe o a un altre. Decidim doncs, estudiar la relació entre aquestes dues variables categòriques mitjançant el test no paramètric *Chi-Squared*; que planteja com a hipòtesi nul·la la independència de les dues variables categòriques. Quan realitzem aquets test per a les variables supervivència i sexe, obtenim un estadístic *chi2* de 260.71 i un *p-value* de 1.19·10^(-58); valors amb els quals podem rebutjar la hipòtesi nul·la de independència i afirmar que existeix una relació significativa entre aquestes dues variables:

```
The contingency table for the Chi-Squared test is:

 Sex       female  male
Survived
0             81   468
1            233   109

Chi-Squared results:
Statistic, p-value:  260.72 1.20E-58
Critical:  3.84
```

Abans d'entrar més en detall, anem a veure què és el que estem veient en aquest test. D'entrada, i tal com ja s'ha comentat (i crec que ha quedat prou clar), el test en sí medeix la possible dependència entre dues variables categòriques; i ho fa mitjançant proporcions.

Al final, per cada un dels sexes, mirem la diferència de proporcions entre supervivents i no supervivents (en el cas dels homes, per exemple,

homes supervivents/total supervients - homes no supervivents/total no supervivents)

i en calculem l'interval de confiança (que ens indica com podria variar la nostra **diferència** de proporcions. Remarco que seguim parlant de diferències de proporcions).

Anem ara a calcular l'interval de confiança de la diferència entre proporcions de supervivents i no supervivents (per cada sexe). Per fer-ho, primer calculem l'error estàndard d'una diferència de proporcions (que segueix la mateixa lògica que l'error estàndard de dues mitjanes), calculat mitjançant:

$$SE(p1 - p2) = \sqrt{(p1(100 - p1)/n1) + (p2(100 - p2)/n2)}$$

On una de les proporcions (*p1*) és la proporció d'un sexe supervivent en el total de la població supervivent, i, l'altre proporció (*p2*), el total de no supervivents del mateix en el total de l'altra població (és a dir, *homes supervivents/total supervients* seria *p1*, i *homes no supervivents/total no supervivents* seria *p2*). Finalment definim el nostre interval de confiança com:

$$CI = [(p1 - p2) - 1.96 * SE\ ,\ (p1 - p2) + 1.96 * SE]$$

(ambdues fórmules extretes dels webs *6. Differences between percentages and paired alternatives* i *8. The Chi squared tests*).

Quan ho fem obtenim que les diferències de les proporcions entre els homes que sobreviuen i els que no és de -0.53; mentre que la diferència per a les dones és de 0.53; amb uns intervals de confiança de [-0.59, -0.47] i de [0.47, 0.59] respectivament.

Recordem que, al estar parlant de diferències entre dues proporcions, és possible obtenir un nombre negatiu al fer la resta. I justament això és el que obtenim per al cas dels homes:

```
Male Confidence Interval is: [ -0.59 ,  -0.48 ],
Proportions differnece =  -0.53
```

```
Female Confidence Interval is: [ 0.48 ,  0.59 ]
Proportions difference  = 0.53
```

Què significa tot això? Doncs d'entrada, que les variables supervivència i sexe guarden relació; i que la diferència entre la proporció d'homes supervivents i no supervivents sigui negativa significa que la proporció d'homes morts és major a la de supervivents (concretament la proporció de no supervivents és la proporció de supervivents més 0.53), mentre que per a les dones és just al contrari: la proporció de dones que sobreviuen és major a la que no ho fa (concretament, la proporció de dones supervivents és major que la de no supervivents per 0.53). Així doncs, la conclusió del test és que les variables supervivència i sexe guarden relació en tant que sobreviuen més dones que homes.

3.5. Chi-Square: Survived vs Class

El penúltim anàlisi que fem torna a ser un test Chi-Square entre les variables supervivència i classe. Executant el test obtenim un estadístic de 102.89 i un *p-value* de 4.54*10^(-23); valors amb els quals rebutgem la hipòtesi nul·la d'independència entre aquestes dues variables categòriques:

```
The contingency table for the Chi-Squared test is:

 Pclass      1   2    3
Survived
0           80  97  372
1          136  87  119

Chi-Squared results:
Statistic, p-value:  102.89 4.55E-23
Critical:  5.99
```

Calculem també les diferències entre les proporcions de supervivents i no supervivents (i els intervals de confiança) per cada una d'aquestes classes:

```
Class 1: difference of proportions:  0.25
Class 2: difference of proportions:  0.08
Class 3: difference of proportions:  -0.33


Class 1 Confidence Interval is: [ 0.19 ,  0.31 ]
Class 2 Confidence Interval is: [ 0.02 ,  0.13 ]
Class 3 Confidence Interval is: [ -0.39 ,  -0.27 ]
```

Observem aquí que les diferències entre les proporcions de supervivents i no supervivents són positives només a primera i segona classe, mentre que per la tercera és negativa. Això significa que la proporció d'homes que sobreviuen a les classes 1 i 2 és major que la dels que no ho fan (tot i que la diferència entre la proporció de supervients i no supervivents de la classe 1 és 3

cops més gran que la de la classe 2!). Per a la classe 3, veiem que la proporció de no supervivents és major en 0.33.

3.6. Chi-Square: Survived vs Sex in Classes

L'últim anàlisi estadístic que fem consisteix en incorporar la variable sexe en l'estudi realitzat anteriorment. Així doncs, per a cada una de les classes 1, 2 i 3, crearem una taula de contingència amb la qual aplicarem un test *Chi-Square* i comprovarem si el sexe afecta en la supervivència dins de cada una de les classes. Quan ho fem, veiem que obtenim, per cada un dels testos, valors *p* inferiors al 0.05; amb els quals rebutgem la hipòtesi nul·la d'independència i establim que existeix una relació significativa entre la supervivència i el sexe per cada una de les classes:

```
Les taules de contingència són, per a les classes 1, 2 i 3 respectivament:

 Survived   0   1
Sex
female     3  91
male      77  45
 Survived   0   1
Sex
female     6  70
male      91  17
 Survived    0   1
Sex
female     72  72
male      300  47


Pclass == 1:
        Chi-Squared results:
        Statistic, p-value:  79.2 5.60E-19
        Critical:  3.84
Pclass == 2:
        Chi-Squared results:
        Statistic, p-value:  101.32 7.82E-24
        Critical:  3.84
Pclass == 3:
        Chi-Squared results:
        Statistic, p-value:  71.68 2.53E-17
        Critical:  3.84
```

Després, calculem les diferències de les proporcions entre supervivents i no supervivents de cada sexe (per cada classe) i els seus intervals de confiança, obtenint:

```
        --------------------
        CLASS 1
        --------------------
        Class 1 Male proportions difference:  -0.63
        Class 1 Female proportions difference:  0.63
        Class 1 Male Confidence Interval is: [ -0.72 ,  -0.54 ]
        Class 1 Female Confidence Interval is: [ 0.57 ,  0.69 ]

        --------------------
        CLASS 2
        --------------------
        Class 2 Male proportions difference:  -0.74
        Class 2 Female proportions difference:  0.74
        Class 2 Male Confidence Interval is: [ -0.84 ,  -0.65 ]
        Class 2 Female Confidence Interval is: [ 0.69 ,  0.8 ]

        --------------------
        CLASS 3
        --------------------
        Class 3 Male proportions difference:  -0.41
        Class 3 Female proportions difference:  0.41
        Class 3 Male Confidence Interval is: [ -0.51 ,  -0.31 ]
        Class 3 Female Confidence Interval is: [ 0.35 ,  0.47 ]
```

Recuperem raonaments similars als vistos en l'apartat 3.4; en què la proporció de supervivents és inferior a la de no supervivents per a homes, i a l'inrevés per a les dones.

3.7. Regressió logística

Per últim, intentarem predir la supervivència o no dels passatgers mitjançant un model de regressió logística. Per a fer-ho, primer dividirem les columnes categòriques amb més de dos nivells en diferents columnes mitjançant la funció pd.get_dummies(). El dataset de training final és el següent:

	Pclass	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S
0	3	22.0	1	0	7.2500	1	0	1
1	1	38.0	1	0	71.2833	0	0	0

Un cop preparades les dades, mitjançant validació creuada amb 4 grups, escollirem la millor combinació d'hiperparàmetres per al model (considerant penalty, C i solver).

	param_C	param_penalty	param_solver	mean_test_score
46	0.1	l2	lbfgs	0.801373
45	0.1	l2	newton-cg	0.801373
20	0.01	none	newton-cg	0.799120
80	10	none	newton-cg	0.799120
72	1	l1	liblinear	0.799120

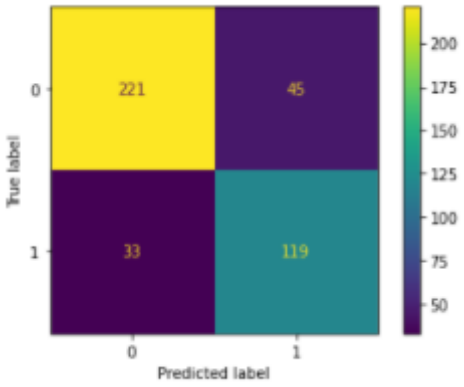
Com es pot veure a la taula anterior, els millors paràmetres han permès obtenir un 0.8 d'accuracy. Per tant, 8 de cada 10 observacions del training dataset s'han predit correctament. Anem a veure els coeficients del model:

```
Pclass: -0.938
Age: -0.036
SibSp: -0.301
Parch: -0.028
Fare: 0.004
Sex_male: -1.981
Embarked_Q: 0.107
Embarked_S: -0.224
```

Com quasi tots els coeficients tenen signe negatiu, es pot afirmar que a més valor d'aquestes variables, menys probable és sobreviure. En són excepcions però, els diners que costa el tiquet i el fet d'haver embarcat a Queenstown. Per tant, un home de tercera classe, d'edat avançada, sense família abord que va embarcar a Southampton tindria una probabilitat alta de no sobreviure en el Titanic. S'ha de tenir en compte però, que el valor absolut dels coeficients no mostren la magnitud de la relació d'una manera intuitiva, és per a això que aplicarem la funció np.exp() als coeficients per a obtenir els odds-ratio (OR):

```
Pclass: 0.391
Age: 0.964
SibSp: 0.74
Parch: 0.972
Fare: 1.004
Sex_male: 0.138
Embarked_Q: 1.113
Embarked_S: 0.799
```

Observant els OR, podem veure per exemple, que la probabilitat de sobreviure sent home és 0.138 vegades menys probable comparat amb les dones. Finalment, passem a veure les prediccions en les dades de test mitjançant una matriu de confusió:



El model té una accuracy del 0.81 ((221+119)/total), per tant la predicció és una mica millor que la del training set i podem concloure que el model és capaç de generalitzar i no hi hagut overfitting. A més, com es pot veure en la especificitat de 0.83 (221/(221+45)) i la sensibilitat de 0.78 (119/(119+33)), s'ha obtingut un rendiment molt semblant a l'hora de predir els passatgers que sobreviviuen i els que no.

5. Representació dels resultats a partir de gràfiques i taules

Atribut	**Classe**	**Test**	**Valor (sobreviu)**	**Valor (no sobreviu)**	**p-value**	**CI (sobreviu)**	**CI (no sobreviu)**
Edat Homes	Numèrica	Mann–Whitney	27.69	30.73	0.1766	[0.83 , 57.20]	[9.0 , 62.65]
Edat dones	Numèrica	Mann-Whitney	28.28	23	0.0010	[2.8, 58.0]	[2.0 , 48.0]
Sexe: dona	Categòrica	Chi-Squared	0.53	###	1.19*10^(-58)	[0.4, 0.59]	###
Sexe: Home	Categòrica	Chi-Squared	-0.53	###	1.19*10^(-58)	[-0.59, -0.4]	###
Classe: 1	Categòrica	Chi-Squared	0.25	###	4.55*10^(-23)	[0.19, 0.31]	###
Classe: 2	Categòrica	Chi-Squared	0.08	###	4.55*10^(-23)	[0.02 , 0.13]	###
Classe: 3	Categòrica	Chi-Squared	-0.33	###	4.55*10^(-23)	[-0.39 , -0.26]	###
Classe:1 Sexe: home	Categòrica	Chi-Squared	-0.63	###	5.60*10^(-19)	[-0.72 , -0.54]	###
Classe:1 Sexe: dona	Categòrica	Chi-Squared	0.63	###	5.60*10^(-19)	[0.57 , 0.69]	###
Classe:2 Sexe: home	Categòrica	Chi-Squared	-0.74	###	7.82*10^(-24)	[-0.84, -0.65]	###

Classe: 2 Sexe: dona	Categòrica	Chi-Squared	0.74	###	7.82*10^(-24)	[0.68 , 0.80]	###
Classe: 3 Sexe: home	Categòrica	Chi-Squared	-0.41	###	2.53*10^(-17)	[-0.51 , -0.31]	###
Classe: 3 Sexe: dona	Categòrica	Chi-Squared	0.41	###	2.53*10^(-17)	[0.35 , 0.47]	###

Què és el que estem veient en aquesta taula? Donat que estic barrejant testos no paramètrics numèrics i categòrics, però he volgut juntar els resultats en una sola taula, s'hauria d'explicar què és què. Bàsicament, les columnes que poden diferir més són valor (en el test de Mann-Whitney es refereixen a mitjana; mentre que en els Chi-Square es refereixen a la diferència entre proporcions de supervivents i no supervivents).

Primer de tot, veiem els resultats per als tests no paramètrics de l'edat. En aquests casos veiem, tant per a homes com per a dones, la mitjana de les edats per a homes/dones supervivents i homes/dones no supervivents (i els seus intervals de confiança); a més a més dels valors *p* de cada un dels testos realitzats sobre la població masculina i la femenina. Observem que, en el cas dels homes, la mitjana d'edat dels supervivents és més petita, metre que per les dones és més gran; concretament vora 5 anys més gran. Aquest resultat, acompanyat del *p-value* inferior a 0.05 ens permet confirmar que hi ha diferències significatives entre les edats de les dones supervivents i no supervivents. Ara bé, faltaria plantejar un altre test per comprovar si la tendència era salvar dones joves o grans.

Seguidament veiem els resultats del Chi-Square aplicat entre les variables supervivència i sexe. Recordem que d'aquest test havíem obtingut que existia depndència entre les variables supervivència i sexe, i que a aquesta dependència ho feia en favor a les dones (tal i com es pot veure en la columna ***Valor (sobreviu)***. Comentar també, que en els testos Chi-Square, al ser la nostra variable de referència una diferència entre proporcions de supervivents i no supervivents, només tindrem un valor observable i un interval de confiança (és per això que les columnes ***Valor (no sobreviu*)** i ***CI (no sobreviu*)** es troben en blanc).

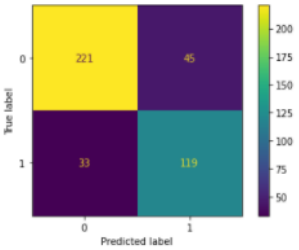
Les 3 files següents fan referència al test Chi-Square que engloba les 3 diferents classes, amb la qual cosa els registres *classe: 1, classe: 2* i *classe: 3* resulten del mateix test i comparteixen *p-value*. En aquests registres podem observar com les diferències de proporcions entre les classes 1, 2 i 3 van reduint-se a mesura que baixem un esglaó en el status social, fins a assolir una diferència negativa per l'última classe.

Finalment, els últims 6 registres tornen a anar emparellats dos a dos ja que ara estem fent testos Chi-Square entre la supervivència d'homes i dones per a cada una de les classes.

Finalment dir que, en el cas del test de Mann-Whitney, l'interval de confiança és el conjunt de valors entre el qual es trobarà un valor concret (amb un 95%) de confiança, mentre que en el cas dels testos Chi-Square serà l'interval entre el que podrem trobar les proporcions sobreviu/no sobreviu.

Per altra banda, hem vist que el millor model per a la regressió logística és el que té els hiperparàmetres = {C=0.1, penalty=l2, solver=lbfgs} i que el seu rendiment en el test dataset ha estat d'una accuracy de 0.81, una especificitat de 0.83 i una sensibilitat de 0.78.

	param_C	param_penalty	param_solver	mean_test_score
46	0.1	l2	lbfgs	0.801373
45	0.1	l2	newton-cg	0.801373
20	0.01	none	newton-cg	0.799120
80	10	none	newton-cg	0.799120
72	1	l1	liblinear	0.799120



6. Conclusions

En quant a la part d'anàlisi descriptiu, hem pogut observar com l'edat mitjana de les dones que sobreviuen és major a la dels homes, i que la probabilitat de supervivència d'una persona es veu afectada per l'edat (si i només si, aquesta és una dona). Sembla ser, però que es prioritzen les dones majors a les més joves. A banda d'això, podem confirmar que el sexe influeix en la decisió de salvar la vida d'una persona: la diferència de proporcions de dones salvades enfront de les que no es del 0.53 (mentre que per als homes és just al contrari: -0.53).

En el següent test *Chi-Square* calculem els valors esperats de les proporcions de la gent que es salva enfront dels que no per cada una de les classes; obtenint un balanç positiu per les dues primeres classes (0.25 i 0.08 respectivament) i un de negatiu per a la tercera (-0.33). Ens trobem davant d'una situació en què sembla que el status social marcava la prioritat de salvament.

Ja per acabar, comprovem que la regla de que les dones es salven, es compleixi per cada una de les classes amb un altre test *Chi-Square*; amb el qual comprovem que efectivament això és així. Aquí veiem que, tot i que els homes de les classes majors semblin ser més sacrificats (les diferències de proporcions d'homes són, per a les classes 1, 2, i 3, -0.63, -0.74 i -0.41), es segueixen salvant, proporcionalment, més dones de classes altes que baixes; cosa que té tot el sentit del món ja que si hem vist que es prioritza el salvament de dones abans que el d'homes, i el de classes altes al de classes baixes, aquest comportament en el salvament de sexes s'hauria d'estendre dins de cada una de les classes. De totes maneres, la diferència entre homes no supervivents i homes supervivents de la classe 2 és major que la diferència entre homes de la classe 1.

Per tant, les conclusions d'aquest anàlisi estadístic han estat és que, en el moment de produir-se l'accident, es van prioritzar les vides de les persones de classe més alta, de les dones i, dintre d'aquestes, les dones una mica més grans que joves.

Fora d'aquest "protocol d'operació" que hem deduït, trobem interessant ressaltar la diferència de proporcions sobreviu-no sobreviu entre les diferents classes; arribant quasi a una diferència de 1 entre les classes 1 i 3 (és a dir que es va salvar, proporcionalment, quasi el doble de gent de primera classe que de última). A més a més, destaca el fet de que es salven, proporcionalment, més dones de la classe 2 que de la 1.

Per últim, s'ha creat un classificador amb un model de Regressió Logística que ha aconseguit una accuracy de 0.81, una especificitat de 0.83 i una sensibilitat de 0.78. Per tant, la capacitat de predicció per a les dues classes és raonablement bona.

Taula contribucions

Contribucions	Signatures
Investigació prèvia	AGT, JGE
Redacció de les respostes	AGT, JGE
Desenvolupament del codi	AGT, JGE