

Pràctica 2: Neteja i anàlisi de les dades

Autors: Guillermo Camps Pons (individual)

1. Descripció

El *dataset* escollit per a realitzar la neteja i l'anàlisi de les dades és *Titanic: Machine Learning from Disaster*. El conjunt de dades descriu els passatgers del RMS Titànic, un creuer anglès que es fa enfonsar el 15 d'abril de 1912 després de xocar amb un iceberg i on l'escassa quantitat de bots salvavides va resultar en un gran nombre de morts. Aquest *dataset* conté les variables que podem observar a la Taula 1.

Taula 1. Descripció de les variables del *dataset* del Titànic.

Variable	Descripció	Tipus	Exemple
PassengerId	Nombre arbitrari ordinal que identifica una fila	Ordinal	7
Pclass	Classe del tiquet de viatge (on 1 és la millor classe)	Ordinal	1
Name	Nom del passatger	Cadena	'McCarthy, Mr. Timothy J'
Sex	Sexe del passatger (<i>male</i> o <i>female</i>)	Cadena (dicotòmica)	'male'
Age	Edat del passatger	Numèric (quantitatiu)	54.0
SibSp	Nombre de germans i esposes al creuer	Enter (quantitatiu)	0
Parch	Nombre de pares i fills	Enter (quantitatiu)	0
Ticket	Nombre del tiquet	Cadena	'17463'
Fare	Tarifa del tiquet	Numèric (quantitatiu)	51.8625
Cabin	Nombre de cabina	Cadena	'E46'
Embarked	Port d'embarcament (Q, C o S)	Cadena	'S'
Survived	Si el passatger va sobreviure (1) o no a l'accident (0)	Enter (dicotòmica)	0

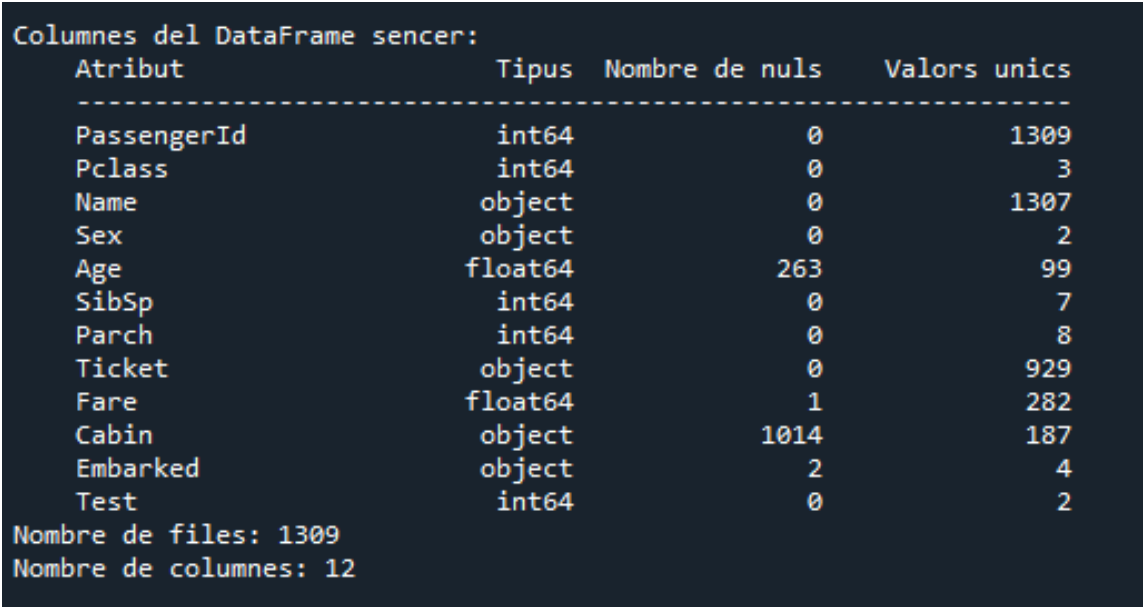
L'objectiu principal d'estudiar aquest *dataset* des del punt de vista de les dades és poder detectar patrons en quin tipus de passatgers van aconseguir sobreviure. Per exemple, se sol dir que es va donar prioritat d'embarcar a les dones i als nens i aquest *dataset* ens podria ajudar a veure si això es dona o no.

El *dataset* està disponible a Kaggle al següent enllaç: <https://www.kaggle.com/c/titanic> i forma part d'una competició on l'objectiu és aconseguir predir la variable *Survived* d'un conjunt (conjunt de testeig) dels passatgers a partir de l'altre conjunt (conjunt d'entrenament). Per a això, es donen dos arxius: un anomenat *train.csv* amb les dades de 891 passatgers i la variable *Survived* i un altre anomenat *test.csv* amb les dades de 418 passatgers i sense variable *Survived*.

La nostra feina serà llavors, netejar les dades d'aquests dos arxius, analitzar les diferències entre els passatgers que van sobreviure i els que no i preparar les dades per a construir un model de predicció. Per fer-ho utilitzarem *Python3*.

2. Integració i selecció de les dades

En primer lloc, per a facilitar el tractament de les dades, ens interessa ajuntar les dades dels dos arxius *train.csv* i *test.csv* en un mateix objecte (per exemple, un *DataFrame* de la llibreria *Pandas* de *Python*), afegint una variable (*Test*) que indica de quin arxiu prové. Un cop fet això, fem una ullada al contingut d'aquest a la Figura 2. Aquí es descriuen el nombre de valors nuls i el nombre de valors únics i diferents que conté cada variable. Els valors nuls els tractarem a l'apartat següent però de primeres, podem veure que la variable *Cabin* conté quasi tots els valors nuls i no ens pot aportar gaire informació ni tampoc podem omplir els valors. Per això, en lloc de tractar amb aquesta variable, serà més interessant saber si es coneix o no la cabina i creem una variable *Cabin_any*.



Columns del DataFrame sencer:

Atribut	Tipus	Nombre de nuls	Valors únics
PassengerId	int64	0	1309
Pclass	int64	0	3
Name	object	0	1307
Sex	object	0	2
Age	float64	263	99
SibSp	int64	0	7
Parch	int64	0	8
Ticket	object	0	929
Fare	float64	1	282
Cabin	object	1014	187
Embarked	object	2	4
Test	int64	0	2

Nombre de files: 1309
Nombre de columnes: 12

Figura 1. Valors nuls i únics de tot el conjunt de dades

Si ens fixem en els valors únics, veiem que *PassengerId*, *Name* i *Ticket*, que no són variables quantitatives, tenen molts valors únics, ja que no són categories. Llavors, de primeres, no tenen valor analític i per tant no les necessitem. Pel cas de la variable del nom, podem extreure, fent ús d'expressions regulars, els títols honorífics com *Mr.*, *Mrs.*, *Miss.* o *Master*. Per això creem una variable *Title* que contengui només aquests títols, fent que els valors únics siguin molt menors.

Per acabar, ens interessarà transformar les variables categòriques en variables binàries per cada categoria (també anomenades *dummy*) per a poder fer càlculs amb correlacions i regressions. Per això, cada una de les categories de *Sex* (*male*, *female*), *Embarked* (*S*, *Q*, *C*) i *Title* (*Mr*, *Mrs*, *Miss*, *Master*, *Other*) tindrà una columna de 0 o 1. Eliminarem una categoria per variable per a evitar problemes de multicolinealitat, és a dir, que una es pugui expressar com a combinació de les altres.

A la Figura 2, podem veure el resultat de crear aquestes noves variables, tot i eliminant les variables originals, conjuntament

```
Atribut
-----
Pclass
Age
SibSp
Parch
Fare
Title_Master
Title_Miss
Title_Mr
Title_Mrs
Cabin_any
Sex_male
Embarked_Q
Embarked_S
```

Figura 2. Variables d'anàlisi resultants de la selecció.

3. Neteja de les dades

3.1. Tractament de valors nuls

A la Figura 1, hem vist que *Fare* té 1 valor nul, *Embarked* (abans de transformar-la en variables *dummy*) en té 2 i *Age* en té 263. El tractament de les dues primeres quasi no esbiaixarà la mostra mentre que el de la tercera presenta un percentatge gran del conjunt de dades. En general, no ens interessa eliminar els registres amb valors nuls, ja que volem poder fer una predicció de tots els passatgers del conjunt de testeig.

Per *Fare*, realitzem una imputació amb el mètode dels *k* veïns més propers (KNN). Amb aquest mètode podem comparar el registre amb la resta de les dades i assignar un valor al valor buit que sigui similar als registres semblants. En aquest cas, fem ús de la variable *Pclass* (i la pròpia *Fare*) per a mirar la proximitat dels punts, ja que segons la matriu de correlació que veurem a l'apartat de l'anàlisi, aquestes dues estan una mica correlacionades.

Per *Embarked*, com que es tracta de una variable categòrica i té pocs valors nuls, imputem el valor segons el valor més freqüent. Com veiem a la Figura 3, el port d'embarcació S és el més freqüent amb diferència i, per tant, l'esbiaix serà mínim.

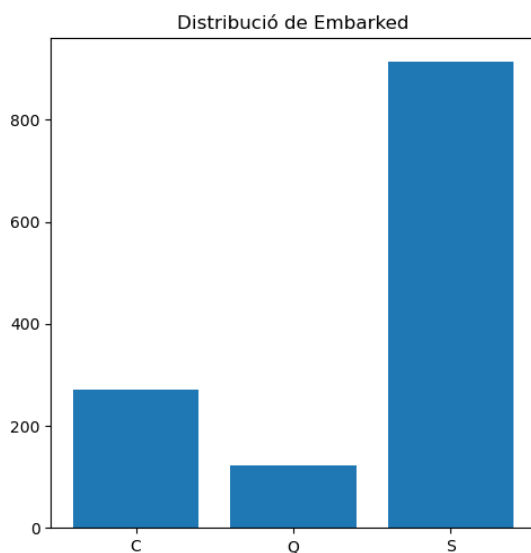


Figura 3. Distribució d'*Embarked*.

Per *Age*, realitzarem també una imputació amb el mètode KNN, però aquest cop amb totes les altres variables per a avaluar la proximitat. A la Figura 4, podem veure com fer aquest procés modifica la distribució de les dades de la variable *Age*. Veiem que, encara que el pic de la distribució ha canviat una mica la seva posició, la forma de la distribució no canvia molt, ja que els nous punts s'han distribuït sobre diverses edats (no com passaria per exemple si utilitzéssim l'estratègia emprada per *Embarked*). Per tant, decidim acceptar aquesta imputació.

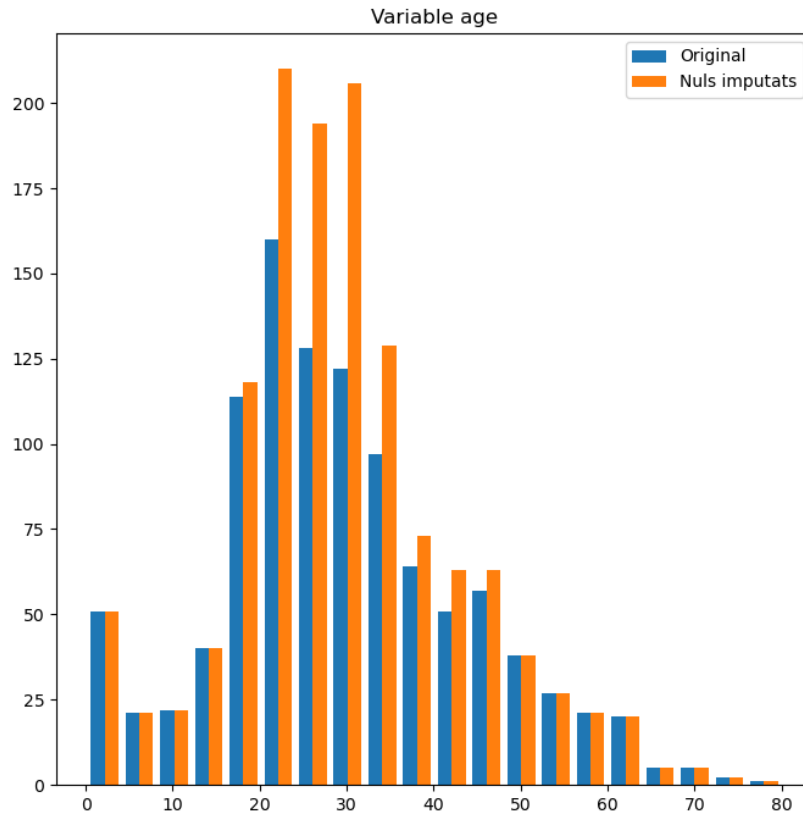


Figura 4. Distribució d'*Age* abans (blau) i després (d'aplicar) la imputació de valors.

3.2. Tractament de valors extrems

Pel tractament dels valors extrems, dibuixem un *boxplot* per les variables *Age* i *Fare*. Un *boxplot* és un diagrama on es dibuixa una caixa que representa el rang interquartil (IQR) d'una variable, això és, la distància entre el quartil 3 (Q3) i el quartil 1 (Q1). Dins d'aquesta caixa, es representa la mediana (Q2) i fora d'aquesta es representen uns 'bigotis' que arriben fins a 1.5 vegades el rang interquartil (o fins el màxim si no hi ha dades majors) per sota i per sobre de la caixa. Un cop fet això, les dades que no cauen dins el rang dels bigotis són identificades com valors extrems i representades al diagrama.

A la Figura 5, podem veure els diagrames de caixes per les variables *Age* i *Fare*. Per la primera, veiem que els valors extrems són edats que podem ser normals (80 anys) i, per tant, els deixem com estan. Per la segona, ens trobem amb 4 punts amb un *Fare* de més de 500. Tot i que aquests valors poden ser legítims, són molt llunyans fins i tot als altres valors extrems. Per tant, els substituïm pels següents màxims (propers a 300).

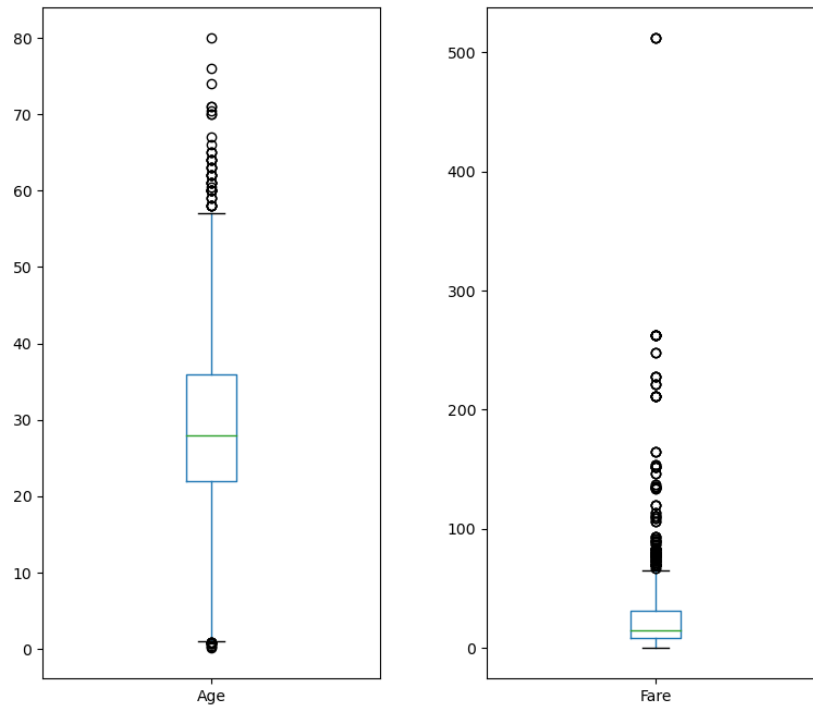


Figura 5. Diagrames de caixes per les variables *Age* (esquerra) i *Fare* (dreta).

4. Anàlisi de les dades

4.1. Selecció dels grups de dades de l'anàlisi

Un cop realitzat el procés de neteja, les variables resultants per a l'anàlisi queden com veiem a la Figura 6. Per l'anàlisi, ens interessarà estudiar els dos grups per veure si presenten característiques diferents: *Survived* i *no Survived*. Per tant, no podrem tractar amb les dades de l'arxiu *test.csv*.

```
Columnes del DataFrame d'anàlisi:
  Atribut      Tipus  Nombre de nuls  Valors únics
-----
Pclass      int32      0              3
Age         float64    0             171
SibSp       int32      0              7
Parch       int32      0              7
Fare        float64    0             247
Title_Master int32      0              2
Title_Miss  int32      0              2
Title_Mr    int32      0              2
Title_Mrs   int32      0              2
Cabin_any   int32      0              2
Sex_male    int32      0              2
Embarked_Q  int32      0              2
Embarked_S  int32      0              2
Survived    int64      0              2
Nombre de files: 891
Nombre de columnes: 14
```

Figura 6. Resultats del procés de neteja.

4.2. Comprovació de la normalitat i homogeneïtat de la variància

En aquest apartat, en primer lloc, volem veure si les variables quantitatives segueixen una distribució normal. Per fer-ho realitzarem els test de *Shapiro-Wilk* i de *Kolmogorov-Smirnov* de sobre les diferents variables i per cada un dels dos grups. Els tests estan implementats per la llibreria *Scipy.stats* amb el nom *shapiro* i *kstest*, respectivament. Els dos tests tenen com a hipòtesi nul·la que les dades segueixen una distribució normal. Si el p-valor és menor al nivell de significació, que agafem com $\alpha = 0.05$, es rebutja la hipòtesi nul·la. Si es dona el cas contrari ($p \geq \alpha$) no es pot rebutjar la hipòtesi nul·la i s'assumeix que les dades segueixen una distribució normal.

A la Figura 7, podem veure el resultat d'aplicar aquests tests. Veiem que en tots els casos rebutgem la hipòtesi nul·la i, per tant, no podem dir que les variables segueixen una distribució normal.

```
-----Anàlisi Normalitat-----
```

--Test Shapiro-Wilk:

H0: les dades segueixen una distribució normal
H1: les dades no segueixen una distribució normal

	Variable	Survived	Estadistic	p-valor	H0
0	Age	0	0.959063	0.000000	False
1	Age	1	0.982878	0.000431	False
2	Fare	0	0.513037	0.000000	False
3	Fare	1	0.702672	0.000000	False
4	SibSp	0	0.484180	0.000000	False
5	SibSp	1	0.654768	0.000000	False
6	Parch	0	0.458816	0.000000	False
7	Parch	1	0.638870	0.000000	False

--Test Kolmogorov-Smirnov:

H0: les dades segueixen una distribució normal
H1: les dades no segueixen una distribució normal

	Variable	Survived	Estadistic	p-valor	H0
0	Age	0	0.982257	0.0	False
1	Age	1	0.954790	0.0	False
2	Fare	0	0.974469	0.0	False
3	Fare	1	0.997076	0.0	False
4	SibSp	0	0.500000	0.0	False
5	SibSp	1	0.500000	0.0	False
6	Parch	0	0.500000	0.0	False
7	Parch	1	0.500000	0.0	False

Figura 7. Resultats de la comprovació de la normalitat.

A continuació, volem comprovar l'homogeneïtat de la variància (o homoscedasticitat) entre els diferents grups. Com que hem vist que les dades no segueixen una distribució normal ens convé utilitzar el test de *Fligner-Killeen*, que es una alternativa no paramètrica a altres test com el de *Levene*. La hipòtesi nul·la d'aquest test és que la variància és igual en els dos grups. Si el p-valor és menor al nivell de significació, que agafem com $\alpha = 0.05$, es rebutja la hipòtesi nul·la. Si es dona el cas contrari ($p \geq \alpha$) no es pot rebutjar la hipòtesi nul·la i s'assumeix que hi ha homoscedasticitat.

A la Figura 8, podem veure el resultat d'aplicar aquest test. Veiem que rebutgem la hipòtesi nul·la en tots els casos menys en *SibSp*, és a dir, que podem assumir que les variàncies són diferents pels dos grups per totes les variables menys *SibSp*.

```

-----Anàlisi Homoscedasticitat-----

--Test Fligner-Killeen:
  H0: la variància és igual en ambdós grups (homoscedasticitat)
  H1: la variància no és igual entre els grups (heteroscedasticitat)
  Variable  Estadistic  p-valor  H0
0    Age    7.650857  0.005675  False
1    Fare   94.230832  0.000000  False
2    SibSp   1.251390  0.263287  True
3    Parch  11.253329  0.000795  False

```

Figura 8. Resultats de la comprovació de la homoscedasticitat.

4.3. Altres proves estadístiques

4.3.1. Comparació de la tendència central

En aquest apartat, realitzem dos tests per a comparar la tendència central dels dos grups. En primer lloc, realitzem un test *t de Student* per a comparar les mitjanes entre grups. Això ho podem fer perquè tenim unes mostres suficientment grans ($N > 30$) i la distribució de la mitjana s'acosta a una normal, però no coneixem les variàncies poblacionals de les variables. El cas de variàncies diferents també es coneix com test *t de Welch* i s'avalua el següent estadístic t:

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

on \bar{X} són les mitjanes mostrals i s són les variàncies mostrals. Si les variàncies es poden assumir com iguals, com passa amb *SibSp*, l'estadístic es pot calcular com:

$$t = \frac{\bar{X}_2 - \bar{X}_1}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

on

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

La hipòtesi nul·la d'aquest test és que la mitjana és igual en els dos grups. Si el p-valor és menor al nivell de significació, que agafem com $\alpha = 0.05$, es rebutja la hipòtesi nul·la. Si es dona el cas contrari ($p \geq \alpha$) no es pot rebutjar la hipòtesi nul·la i s'assumeix que les mitjanes són iguals.

En segon lloc, realitzem un test *U de Mann-Whitney* per a comparar les medianes entre grups. Aquest mètode és més robust que l'anterior donat a que tracta amb medianes. Aquest test funciona assumint que, si les medianes fossin iguals, la probabilitat de que un element aleatori del grup 1 sigui major (o menor) a un element aleatori del grup 2 ha de ser 50%.

La hipòtesi nul·la d'aquest test és que la mediana és igual en els dos grups. Si el p-valor és menor al nivell de significació, que agafem com $\alpha = 0.05$, es rebutja la hipòtesi nul·la. Si es dona el cas contrari ($p \geq \alpha$) no es pot rebutjar la hipòtesi nul·la i s'assumeix que les medianes són iguals.

A la Figura 9, veiem els resultats d'aquests dos tests. Veiem que el test *t de Student* diu que podem assumir les mitjanes de *SibSp* iguals pels dos grups mentre que el test *U de Mann-Whitney* diu que podem assumir les medianes de *Age* iguals.

```
-----Anàlisi Tendència Central-----

--Test t de Student:
  H0: les mitjanes són iguals
  H1: les mitjanes no són iguals
  Variable Estadistic p-valor H0
0 Age 2.104672 0.035694 False
1 Fare -7.471253 0.000000 False
2 SibSp 1.053837 0.292244 True
3 Parch -2.478908 0.013395 False

--Test U de Mann-Whitney:
  H0: és igual de probable que un element aleatori de la mostra 1
  sigui menor o major a un element aleatori de la mostra 2
  H1: les medianes de les dues mostres són diferents
  Variable Estadistic p-valor H0
0 Age 88076.5 0.060174 True
1 Fare 57809.5 0.000000 False
2 SibSp 85775.0 0.004008 False
3 Parch 82385.0 0.000019 False
```

Figura 9. Resultats de la comparació de la tendència central.

4.3.2. Correlacions

En aquest apartat, mesurarem les correlacions entre les parelles de variables a través de dos coeficients de correlació: el de *Pearson* i el de *Spearman*. El primer mesura la linealitat que presenten les relacions per parelles i el segon com a que de monòtona són aquestes relacions. Un coeficient de 1 o -1 indica correlació màxima mentre que un coeficient de 0 indica la independència entre les variables.

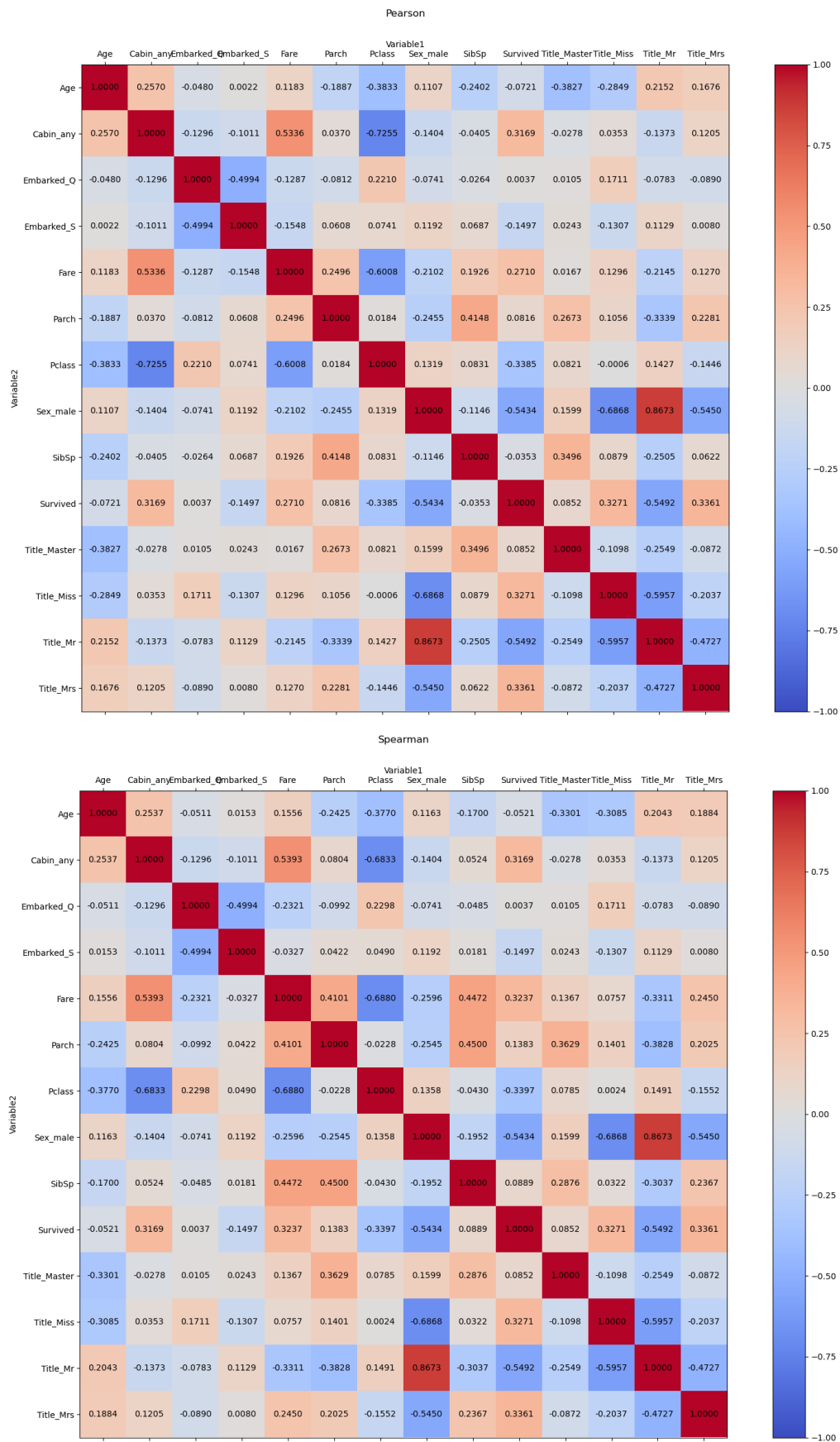
La hipòtesi nul·la de mesurar les correlacions és que les variables són independents (no estan correlacionats). Si el p-valor és menor al nivell de significació, que agafem com $\alpha = 0.05$, es rebutja la hipòtesi nul·la. Si es dona el cas contrari ($p \geq \alpha$) no es pot rebutjar la hipòtesi nul·la i s'assumeix que les variables són independents.

A la Figura 10, es pot veure un llistat de les variables amb correlació superior a 0.5. A la Figura 11, es poden veure mapes de calor amb les totes les parelles i pels dos coeficients. Veiem resultats obvis com que la variable de sexe (*Sex_male*) està correlacionada amb les variables de títols. A part d'això, veiem que les variables *Pclass*, *Cabin_any* i *Fare* estan una mica correlacionades entre elles per parelles. També veiem que *Survived* està relacionat amb el sexe (i el títol *Mr.*) negativament, és a dir, que si el passatger és home, aquest té menys probabilitats de sobreviure.

```
--Coeficient de correlació de Pearson:
  H0: la correlació és nul·la
  H1: les variables no són independents linealment
  Valors amb correlació major o igual a 0.5:
  Variable1 Variable2 Correlacio Abs_corr p-valor H0
5 Title_Mr Sex_male 0.867334 0.867334 0.0 False
1 Pclass Cabin_any -0.725541 0.725541 0.0 False
4 Title_Miss Sex_male -0.686808 0.686808 0.0 False
0 Pclass Fare -0.600839 0.600839 0.0 False
3 Title_Miss Title_Mr -0.595692 0.595692 0.0 False
6 Title_Mr Survived -0.549199 0.549199 0.0 False
7 Title_Mrs Sex_male -0.545050 0.545050 0.0 False
8 Sex_male Survived -0.543351 0.543351 0.0 False
2 Fare Cabin_any 0.533565 0.533565 0.0 False

--Coeficient de correlació de Spearman:
  H0: la correlació és nul·la
  H1: les variables mostren un comportament monoton
  Valors amb correlació major o igual a 0.5:
  Variable1 Variable2 Correlacio Abs_corr p-valor H0
5 Title_Mr Sex_male 0.867334 0.867334 0.0 False
0 Pclass Fare -0.688032 0.688032 0.0 False
4 Title_Miss Sex_male -0.686808 0.686808 0.0 False
1 Pclass Cabin_any -0.683291 0.683291 0.0 False
3 Title_Miss Title_Mr -0.595692 0.595692 0.0 False
6 Title_Mr Survived -0.549199 0.549199 0.0 False
7 Title_Mrs Sex_male -0.545050 0.545050 0.0 False
8 Sex_male Survived -0.543351 0.543351 0.0 False
2 Fare Cabin_any 0.539321 0.539321 0.0 False
```

Figura 10. Parelles de variables més correlacionades.

Figura 11. Mapes de calor dels coeficients de correlació de *Pearson* (dalt) i de *Spearman* (baix).

No obstant, un punt de vista més interessant d'estudi és el de les correlacions de les variables amb la variable *Survived*. A la Figura 12, podem veure com cada variable es relaciona amb aquesta. Veiem un altre cop que el sexe i el títol tenen una mica de correlació (>50%) amb la variable. Amb unes correlacions una mica menors (>30%) trobem també les variables *Pclass*, *Cabin_any* i *Fare*. Les variables *Age* i *Embarked_Q* són significativament independents amb *Survived*.

	Variable1	Variable2	Correlacio	Abs_corr	p-valor	H0
195	Survived	Survived	1.000000	1.000000	0.000000	False
189	Survived	Title_Mr	-0.549199	0.549199	0.000000	False
192	Survived	Sex_male	-0.543351	0.543351	0.000000	False
182	Survived	Pclass	-0.339668	0.339668	0.000000	False
190	Survived	Title_Mrs	0.336074	0.336074	0.000000	False
188	Survived	Title_Miss	0.327093	0.327093	0.000000	False
186	Survived	Fare	0.323709	0.323709	0.000000	False
191	Survived	Cabin_any	0.316912	0.316912	0.000000	False
194	Survived	Embarked_S	-0.149683	0.149683	0.000007	False
185	Survived	Parch	0.138266	0.138266	0.000035	False
184	Survived	SibSp	0.088879	0.088879	0.007941	False
187	Survived	Title_Master	0.085221	0.085221	0.010932	False
183	Survived	Age	-0.052072	0.052072	0.120377	True
193	Survived	Embarked_Q	0.003650	0.003650	0.913353	True

Figura 12. Correlacions de les diferents variables amb *Survived*.

4.3.3. Regressió logística

Un cop vistes les correlacions de *Survived* amb la resta de variables, podem construir una regressió que predigui la variable *Survived* a partir de la resta de variables. Donat que la variable dependent (*Survived*) és dicotòmica, la regressió adequada a ajustar és una regressió logística. La regressió logística agafa com a input una combinació lineal de les variables explicatives i retorna un valor comprès entre 0 i 1, que es pot entendre com la probabilitat de sobreviure. Es pot expressar de la següent manera:

$$Y = \frac{e^{\beta_i X_i}}{1 + e^{\beta_i X_i}}$$

on X_i són les variables explicatives, β_i són els coeficients de la regressió i Y és la variable dependent.

Per a avaluar el rendiment de la regressió logística, separem el conjunt de dades, que prové de l'arxiu *train.csv*, en dos conjunts: un per crear la regressió i l'altre per comparar les prediccions amb els valors reals de *Survived*. Utilitzarem les variables que abans han mostrat les majors correlacions amb *Survived* però sense posar alhora aquelles relacionades entre sí per evitar problemes de multicollinearitat, per exemple, no posarem *Sex_male* i *Title_Mr* alhora.

Per comparar com diversos conjunts de variables creen un model pitjor o millor, dibuixarem les corbes ROC de cada model. Per fer-ho, crearem diferents llindars de probabilitat d'on escollirem quines dades són predites com *Survived* o no. Per exemple, si tenim un llindar del 50% de probabilitat, aquells majors a 50% seran predits com sobrevivents mentre que la resta no. Un cop fet això, representarem el ràtio de positius vertaders (TPR) contra el ràtio de falsos positius (FPR). Una corba ajustada a l'esquerra serà millor i cobrirà més àrea sota aquesta (AUC) mentre que un model aleatori és una línia recta i conté una àrea de 0.5.

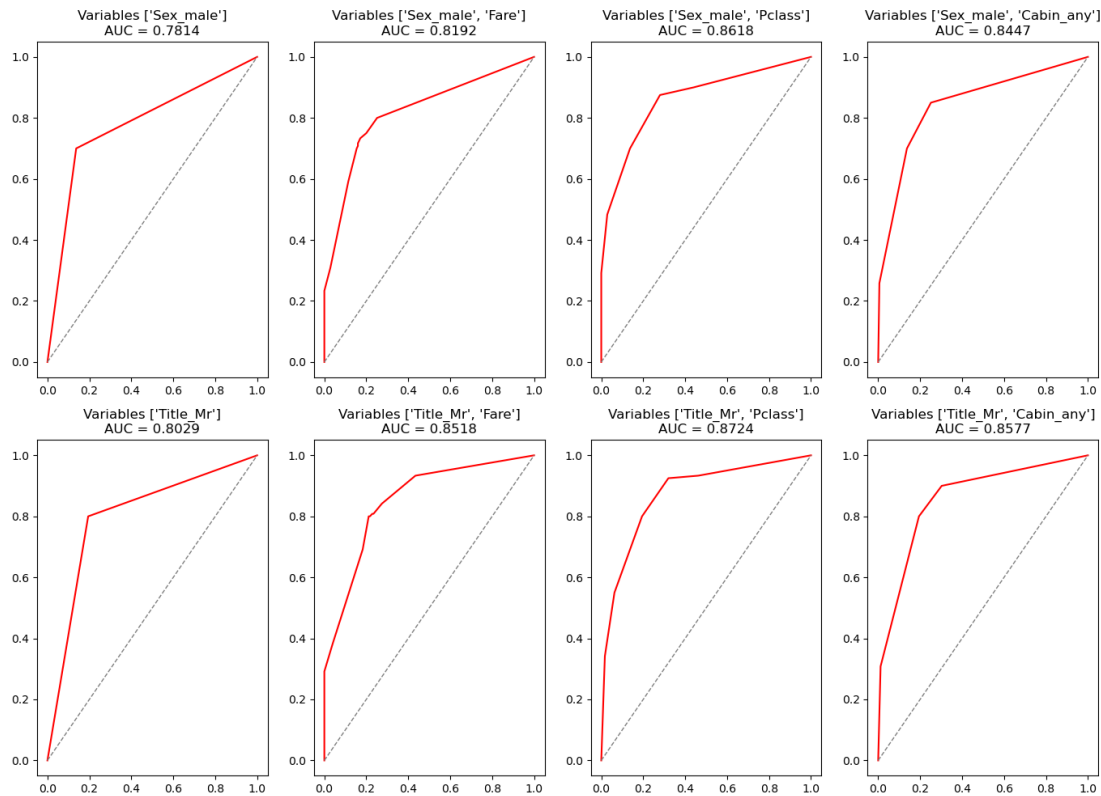


Figura 13. Corbes ROC de les diferents regressions.

A la Figura 13, podem veure les corbes ROC amb els diferents conjunts de variables provats. Veiem que els models tenen una àrea sota la corba (AUC) semblant però la que guanya és el que empra les variables *Title_Mr* i *Pclass*. El millor llindar per a aquest model és el de 50% i presenta una precisió de predicció del 80.32%.

4.3.4. Random Forest

Pot ser interessant, crear un altre model diferent per comparar-lo amb el rendiment de la regressió logística de l'apartat anterior. *Random Forests* és un algorisme de tipus *bagging* que combina diferents arbres de decisió amb mostreigs aleatoris i dóna una predicció basant-se en un sistema de votació de les prediccions de tots els arbres.

Podem escollir diversos paràmetres per tal de reduir el sobreentrenament dels arbres, com la profunditat màxima de l'arbre o el mínim de valors per separar un node. Per a escollir els millors paràmetres, avaluem el conjunt de test amb una validació creuada. A la Figura 14, avaluem com cada paràmetre afecta al rendiment del model en un mapa de calor. Veiem que, en general, els paràmetres són millors paràmetres si tenim una profunditat no molt petita i si el mínim per separar un node és baix.

A la Figura 15, veiem que el resultat de predicció del model final escollit és del 81.69%. Aquest resultat, encara fent ús de totes les variables, dóna un resultat molt aproximat al de la regressió logística de l'apartat anterior.

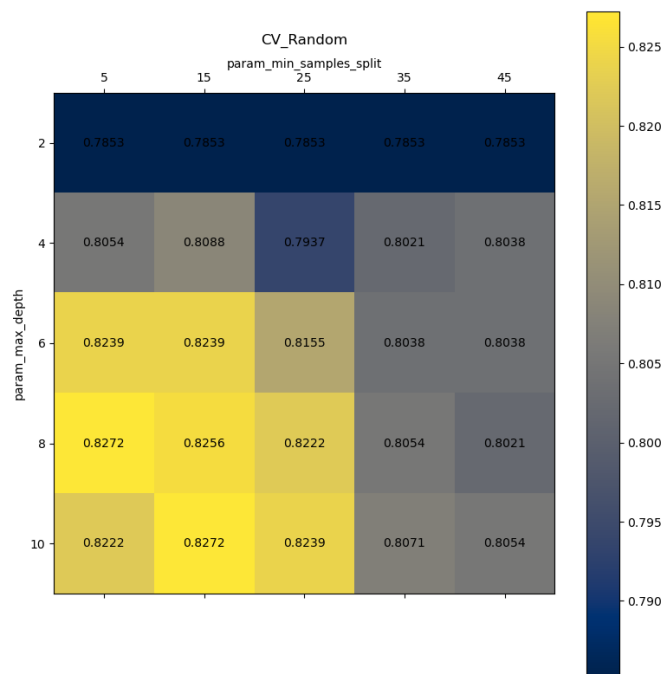


Figura 14. Mapa de calor per la selecció dels millors paràmetres.

5. Resum de resultats

Taula 2. Resum dels resultats de l’anàlisi.

Apartat	Resum resultats
Comprovació normalitat	Cap variable quantitativa en cap grup mostra seguir una distribució normal.
Comprovació homoscedasticitat	Només <i>SibSp</i> mostra homoscedasticitat.
Comparació tendència central	Només es poden assumir mitjanes iguals per <i>SibSp</i> i medianes iguals per <i>Age</i> .
Correlacions	<i>Survived</i> és independent de <i>Age</i> i <i>Embarked_Q</i> i mostra una correlació negativa major al 50% amb <i>Sex_male</i> i <i>Title_Mr</i> .
Regressió logística	El millor model per explicar els diferents grups (<i>Survived</i>) mostra ser la combinació de les variables <i>Title_Mr</i> i <i>Pclass</i> i amb una precisió màxima del 80.32%.
<i>Random Forest</i>	Combinant totes les variables, l’algorisme de <i>Random Forest</i> aconsegueix una precisió del 81.69%.

6. Conclusions

Dels anàlisi hem vist que, en general, els dos grups presenten distribucions diferents en cada una de les variables. Només la variable quantitativa *SibSp* sembla mostrar una distribució semblant per cada grup, amb mitjanes i variàncies que podem assumir iguals, és a dir, que el nombre de germans i esposes no afecta a la condició de supervivència.

La variable *Age* podria semblar en un principi que podria afectar a la supervivència per si es prioritzaven primer els nens i els joves, però la variable es pot assumir independent de *Age* per

no mostrar una correlació significativament diferent a 0 i per tenir una mediana semblant pels dos grups. Potser el fet rau en que els nens són minoria en el creuer.

També hem vist de l'anàlisi de correlacions que hi ha variables que aporten informació molt semblants i que poden ser redundants alhora de predir la variable *Survived*. Hem vist que, òbviament, els títols honorífics extrets del nom estan relacionats amb el sexe del passatger. També hem vist que *Fare* està relacionat amb *Cabin_any*, és a dir, que si al *dataset* original es mostra un nombre de cabina, això està relacionat amb que han pagat més. Addicionalment, la variable *Pclass* està correlacionada negativament amb aquestes dues últimes, donat que la 1a classe és millor que la 2a i la 3a, és a dir, que una classe millor pot indicar que s'ha pagat més i que té una cabina associada.

Finalment, les variables que millor semblen influenciar en la supervivència dels passatgers són *Title_Mr* (o *Sex_male*), amb una correlació negativa, i *Pclass* (o *Cabin_any* o *Fare*) amb una correlació negativa. És a dir, que, en general, els homes i les persones de classe més baixa (*Pclass* més alt) tenen menys probabilitats d'haver sobreviscut a l'incident del Titànic. La capacitat de predicció amb només aquestes dues variables es demostra al ajustar una regressió logística. La precisió de predicció amb *Title_Mr* i *Pclass* arriba al 80.32% amb aquesta regressió, un valor semblant (81.89%) a l'obtingut per l'algorisme *Random Forest* fent ús de totes les variables.

7. Codi

Tant el codi (*data_clean.py*) com els outputs d'aquest, que inclouen els *datasets* nets) es poden trobar al següent enllaç de GitHub:

https://github.com/guille97/titanic_clean

8. Extra

Addicionalment, amb els models creats de regressió logística i de *Random Forest*, s'ha intentat predir els passatgers de *test.csv*. Els arxius generats pel codi es poden carregar a la web de la competició de *Kaggle* per a veure la precisió de les prediccions. La predicció de la regressió logística i de *Random Forest* obtenen, respectivament, unes precisions de 75.598% i 78.468%.