Universitat Oberta de Catalunya

uoc.edu

Tipologia i cicle de vida de les dades

PRÀCTICA 2

ALUMNES: CRISTIAN ALARCÓN SANABRIA i DANIEL VILASECA MIGUEL
ASSIGNATURA: TIPOLOGIA I CICLE DE VIDA DE LES DADES
CURS: MÀSTER DATA SCIENCE 2019-2020, SEGON SEMESTRE

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

Hem escollit el dataset del Titanic, per així no només fer la pràctica sinó veure una mica com funciona la competició a Kaggle.com

Aquest joc de dades ja ve separat per un conjunt de dades d'entrenament i un altre de testeig, en format csv. S'espera que donem resposta a la pregunta de si sobreviuen o no cada fila (registre) del conjunt de dades de test. Les files o observacions dels dos conjunts de dades representen cadascun dels passatgers del enfonsat Titanic, on cada fila disposa de les següents variables (separades per coma):

- survived: pot tenir els valors 0 o 1, sent 0 que no sobreviu i 1 que sí. Aquesta variable lògicament es troba únicament al conjunt de dades d'entrenament, faltant al conjunt de test (donat que és la variable que s'ha de predir mitjançant algoritmes de Machine learning), essent l'única diferència entre els dos fitxers.
- pclass: la classe de tiquet, pot tenir 3 valors, 1, 2 o 3, representant primera (alta), segona (mitjana) i tercera classe (baixa). Aquesta variable serveix per identificar la classe social a la qual pertany el passatger.
- sex: sexe del passatger, podent ser male (home) o female (dona) els únics valors possibles.
- Age: edat del passatger en valor numèric.
- sibsp: nombre de familiars (germans, dona, marit) en valor numèric a bord del vaixell.
- parch: nombre de familiars (pares/mares/fills) en valor numèric a bord del vaixell.
- ticket: número de tiquet (valor numèric). Serveix per identificar cada passatger.
- fare: preu del tiquet (valor numèric).
- cabin: número de cabina.
- Embarked: port des del que embarca el passatger (pot ser des de tres ports: Cherbourg (C), Queenstown (Q) o Southampton (S). Als datasets es mostra el valor entre parèntesis).
- Name: nom del passatger.
- PassengerId: identificador del passatger al dataset.

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

Per a tenir un conjunt de dades el més gran possible el que farem és ajuntar els registres o files de tots dos conjunts de dades (csv traint i test) en un de sol (integració vertical de les dues fonts de dades), conservant la columna dels supervivents (variable survived), per si volem en un futur omplir totes les columnes amb la predicció.

Com els dos conjunts de dades no coincideixen en columnes, la funció rbind no ens servia, en canvi bind_rows feia exactament el que volíem: per a cada fila dels dos datasets afegir les 12 variables, i si no hi havia dades a la variable survived (cas del dataset test) afegia un NA indicant que no tenim valor pel registre.

Gràcies a la funció distinct (retorna files no repetides per l'atribut indicat, en el nostre cas PassengerId que té un valor únic per cada observació) comprovem que la integració s'ha realitzat correctament, doncs obtenim el nombre total d'observacions del dataset al comptar les files amb nrow.

Selecció de dades: Fent la revisió global dels conjunts de dades (amb la funció str) podem comprovar que no tenim un dataset amb moltes observacions o atributs, i que per tant no serà necessari plantejar una reducció de la dimensionalitat o de la quantitat per millorar el temps d'execució dels nostres anàlisis posteriors.

3. Neteja de les dades.

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

A l'atribut d'edat hi ha moltes dades sense valor, així que el que és segur és que no podem esborrar aquestes fileres perquè afectaria al conjunt significativament i la futura predicció es veuria afectada. Per tant o bé les deixem o apliquem algun tipus de transformació. En el nostre cas la decisió ha estat aplicar la mediana, però no la mateixa a tots els registres sinó per classe (la variable pclass), o sigui, si és de classe 1 la mediana de la classe 1, si és classe 2 la mediana d'aquella classe i el mateix amb la tercera.

Veiem també que a la variable fare hi ha una única filera sense informació, en aquest cas substituirem aquest valor nul per la mitjana de les tarifes.

L'altre variable amb moltes fileres amb elements buits és Cabin, però en aquest cas no actuarem, ja que no podem predir la cabina en base a res i al haver-hi moltes no es poden treure aquestes fileres. El nom de la cabina tampoc és la variable més interessant d'aquest conjunt de dades.

Per últim, a la columna Embarked tenim un parell de files sense valor, les omplirem amb el valor més comú: S.

Les dades perdudes de la variable Survived no les tractarem, doncs l'objectiu és predir-ne el valor.

3.2. Identificació i tractament de valors extrems.

Comprovem els possibles valors extrems a totes les dades numèriques mitjançant gràfiques boxplot. A Age trobem unes poques edats avançades, però és plausible que gent de gairebé 80 anys viatgi en vaixell. També comprovem que a les variables SibSp i Parch (que ens diuen el nombre de germans/esposa/marit o fills/pares/mares) hi han alguns nombres força alts, però res poc realista per a l'època, on les famílies eren més nombroses. Per últim, en quant a les tarifes (Fare) hi ha valors altíssims que en un principi ens plantejàvem treure, com 500 per exemple, però informant-nos una mica hem vist que els preus podien pujar fins a 870 lliures, per tant són tots valors vàlids i no els traurem.

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

El que farem serà veure si hi ha relació entre diferents variables i els supervivents, començarem per uns anàlisis de correlació entre algunes de les variables del dataset. També analitzarem la relació, aquest cop gràfica, entre la resta de variables.

Com el nostre objectiu és poder predir si un passatger sobreviu o no, farem models que ens permetin fer prediccions sobre la variable survived.

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Hem realitzat el test de Shapiro-Wilk per a les variables Age i Fare del conjunt de dades. En ambdós casos el p-value ha donat un valor molt proper a zero, donant a entendre que cap dels atributs segueix una distribució normal. Hem representat les dades gràficament amb un histograma per poder-ho comprovar (podem veure que l'edat s'apropa més a una distribució normal, mentre que la variable Fare divergeix molt més).

En aquest cas no modificarem les dades per a que segueixin una distribució normal, ja que llavors es perdria sentit a les dades, doncs les tarifes i l'edat són números amb sentit propi i si els canviem no funcionarien igual. Igualment, per a normalitzar només caldria utilitzar alguna funció com normalize.

Pel que fa a la comprovació de l'homogeneïtat de la variància (homoscedesticitat), tant pel cas de la variable Age com per Fare (i utilitzant els tests de Levene i Fligner-Killeen) el p-value es troba per sota del nivell de significació (0,05), per tant les variàncies de les variables Age i Fare són estadísticament diferents de la variància de la variable amb la qual voldrem fer les comparacions (Survived).

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

S'han aplicat el model de regressió logístic, el model gaussià de Support Vector Machines i el mètode de Random Forests que surt als apunts per trobar una forma de predir la variable survived en noves observacions.

A més s'han aplicat correlacions de Pearson en tots els paràmetres numèrics i als factors s'han aplicat anàlisis visuals per a veure les relacions. També hem aplicat el test de la Chi Squared per veure la relació entre les variables pclass i survived.

La regressió logística múltiple ens ajuda a predir el valor de la variable dicotòmica dependent (Survived), utilitzant les variables independents Pclass, Sex, Age, Fare, SibSp i Parch (fem servir un conjunt d'entrenament per executar el model). Amb la funció summary podem consultar els estimadors i p-values i veure quines variables expliquen més la variació de Survived. En el nostre cas les variables independents que són estadísticament significatives per explicar el valor que pren Survived (p-value més proper a 0) són pclass, sex i age (sent les 2 primeres les que tenen un valor absolut de l'estimador més alt, i que per tant més influeixen en el valor de la variable dependent).

Utilitzem el conjunt de test per comprovar la eficàcia del model.

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = ent1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4145  -0.6096  -0.4202   0.6268   2.4169
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.015804   0.673077   5.966 2.43e-09 ***
## Pclass2     -1.126586   0.428557  -2.629 0.008569 **
## Pclass3     -2.191255   0.437521  -5.008 5.49e-07 ***
## Sexmale     -2.765054   0.286245  -9.660  < 2e-16 ***
## Age         -0.043791   0.011450  -3.824 0.000131 ***
## Fare         0.005683   0.003850   1.476 0.139945
## SibSp       -0.364677   0.155605  -2.344 0.019098 *
## Parch       -0.279208   0.193634  -1.442 0.149320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 590.92  on 444  degrees of freedom
## Residual deviance: 393.10  on 437  degrees of freedom
## AIC: 409.1
##
## Number of Fisher Scoring iterations: 5
```

Amb el mètode de classificació random forest també busquem un model que ens ajudi a predir si un passatger sobreviu o no.
El random forest es construeix a partir de diferents arbres de decisió utilitzant el dataset d'entrenament. Per predir la variable survived d'una nova observació, es mira el resultat (sobreviu o no) de cada arbre de decisió del random forest, quedant-se amb l'opció més repetida.

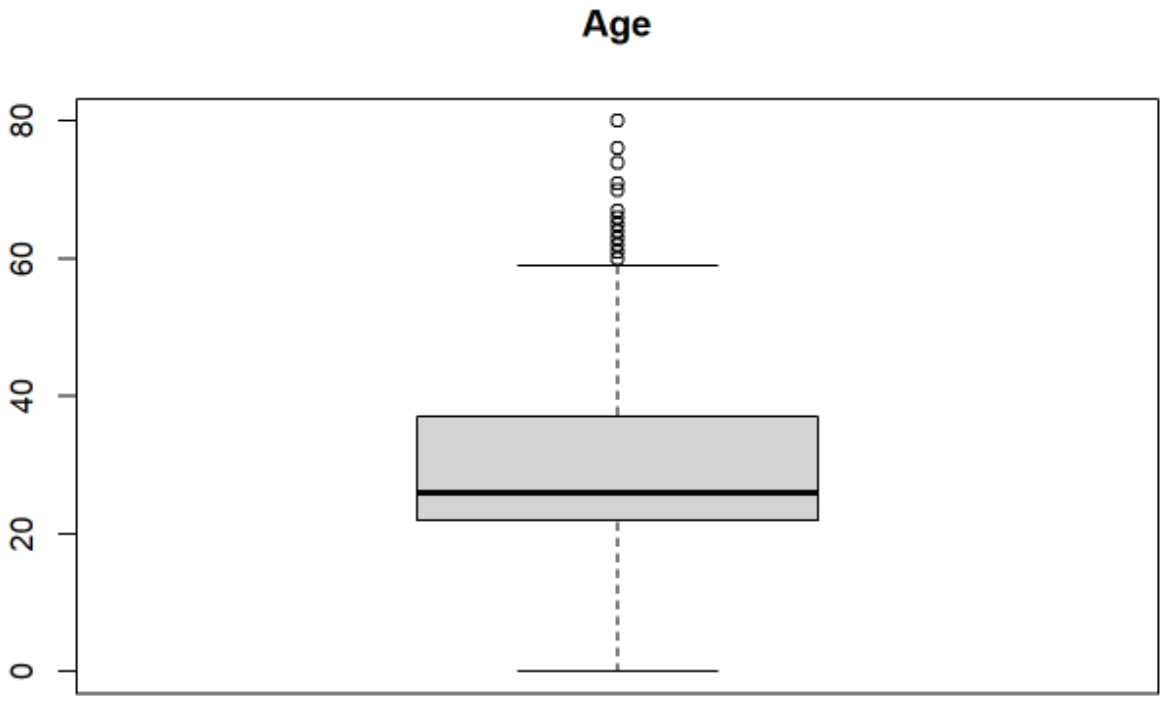
Utilitzem el conjunt de test per comprovar la eficàcia del model. Per cada registre del conjunt de test, es mira per cada arbre de decisió del random forest quin valor pren survived; l'opció més repetida entre els arbres serà la predicció del registre; la predicció de survived del random forest serà correcta si coincideix amb el valor de survived del dataset de test.

El test de la Chi Squared entre la variable de classe (pclass) i la variable survived ha donat un p-value de 0 (arrodonint). Això indica que aquestes dues variables estan correlacionades, la classe social del passatger (pclass) influeix en les possibilitats de sobreviure.

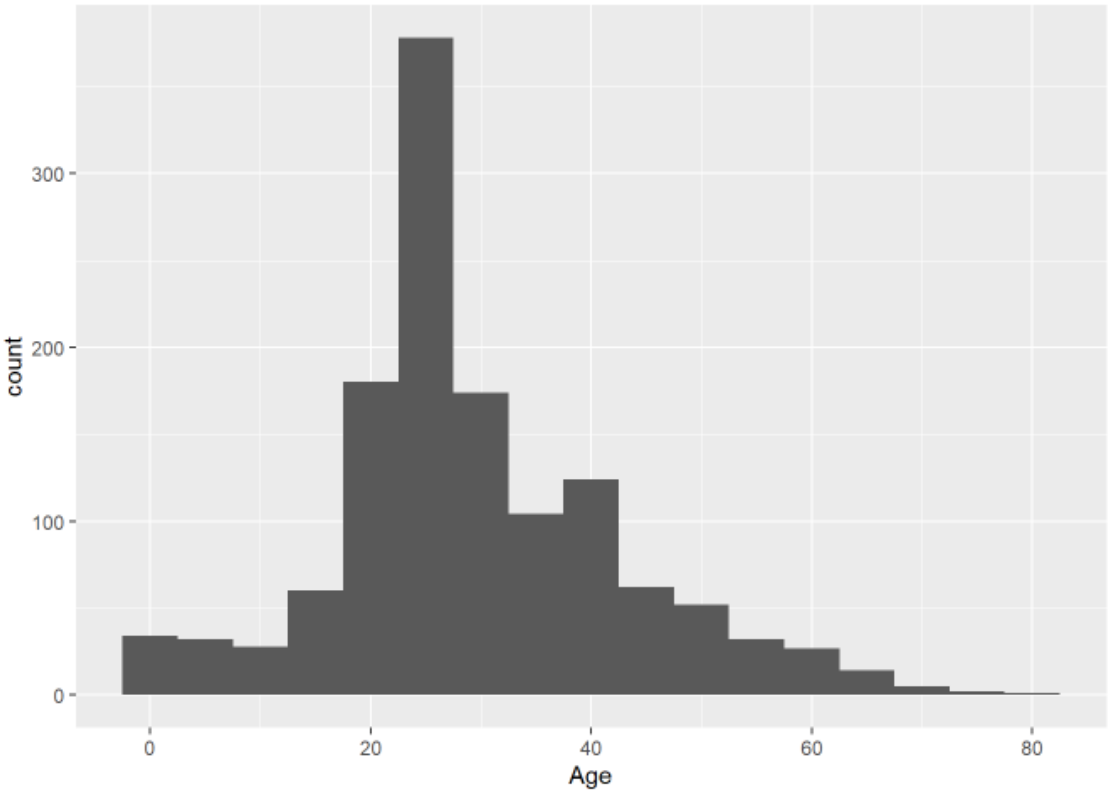
5. Representació dels resultats a partir de taules i gràfiques.

Totes les taules i gràfiques estan incloses al document html adjunt a aquest fitxer. Aquí farem un resum de les principals visualitzacions que podem trobar al treball.

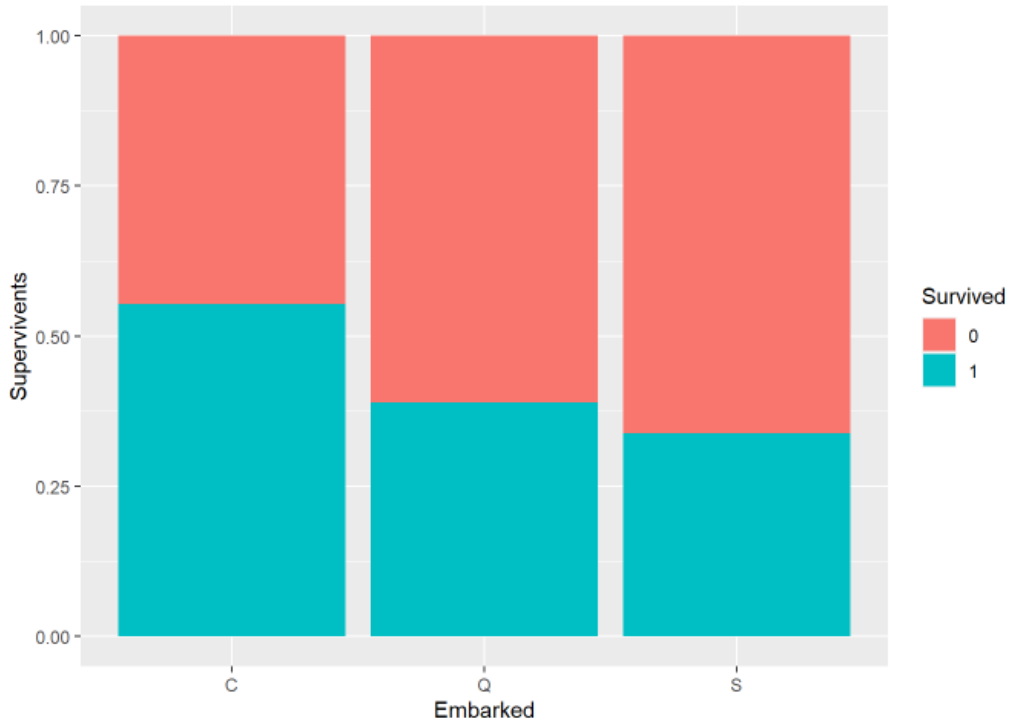
Hem utilitzat boxplots per identificar outliers visualment:



Hem utilitzat histogrames per complementar els tests de normalitat i poder veure gràficament la distribució de les variables:

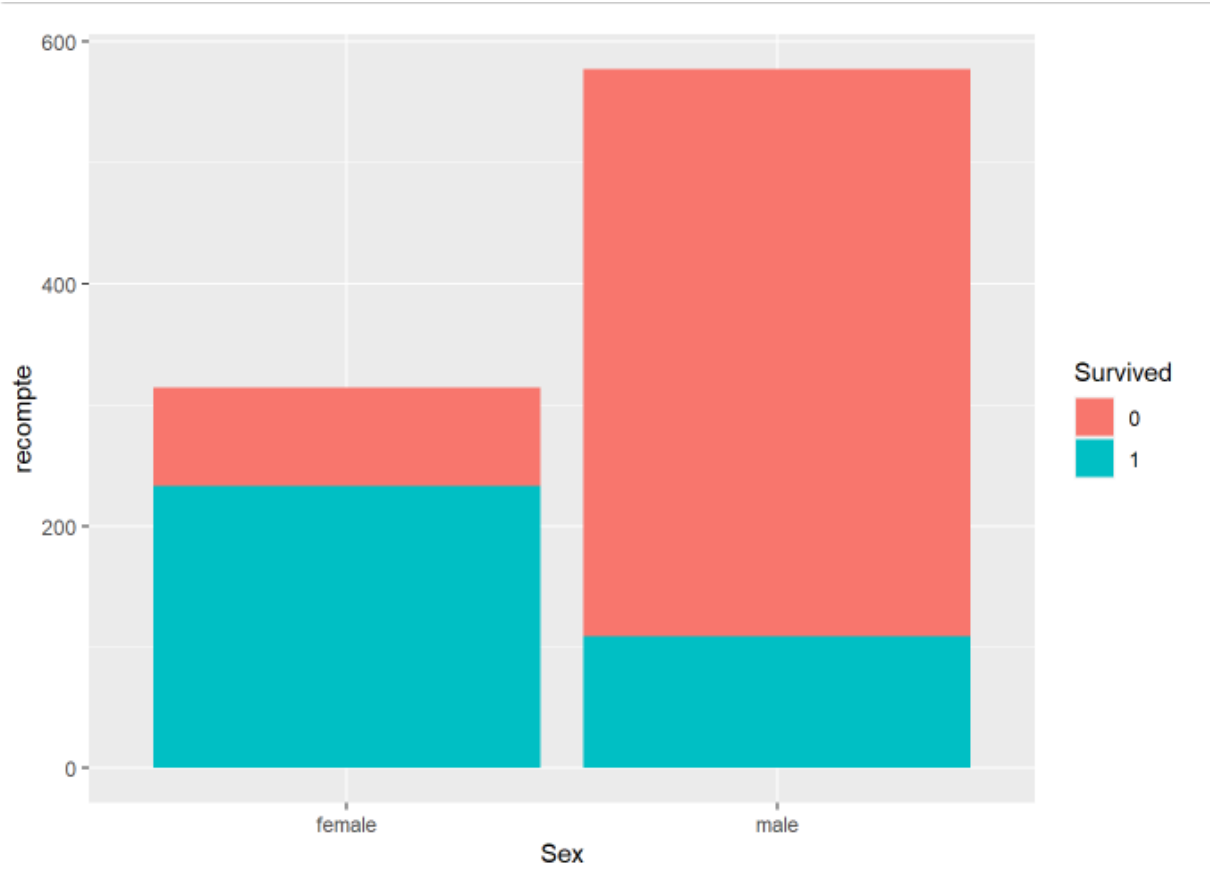


Hem utilitzat gràfiques de barres per veure relacions de variables categòriques:

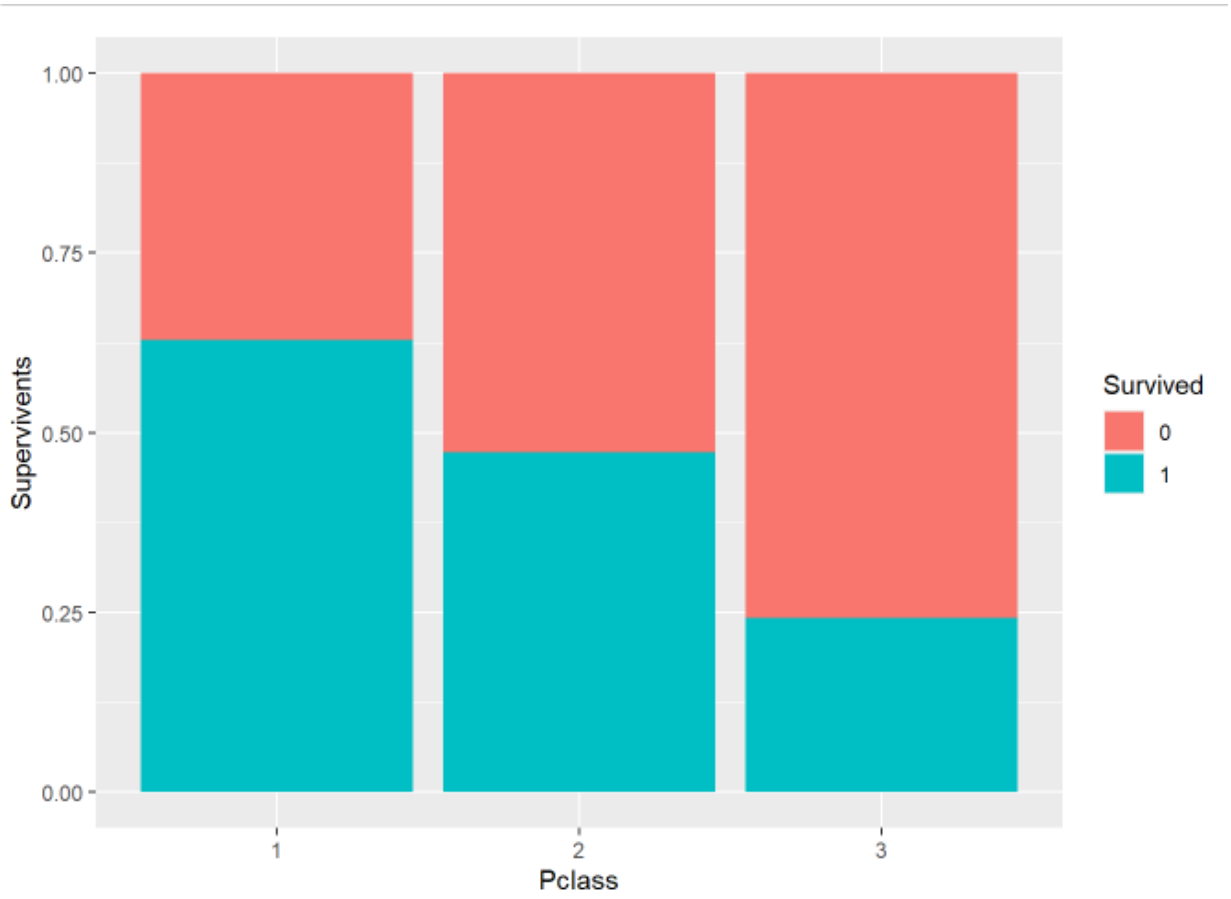


6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Una de les conclusions que podem extreure dels anàlisis visuals realitzats és que el sexe del passatger sembla influir en la probabilitat de supervivència. El percentatge de dones que sobreviu és molt major.



Alguns dels anàlisis ens indiquen que les persones de baixa classe social tenen menys probabilitats de supervivència. El gràfic de sota indica que per passatgers de menor classe social, les possibilitats de sobreviure disminueixen. Tal i com hem comentat anteriorment, el resultat del test Chi Squared ens indica que aquestes dues variables estan correlacionades.



Els tres models predictius (regressió logística, random forest i el support vector machines) són capaços de fer una predicció sobre qui mor o qui sobreviu amb els paràmetres que hem estudiat.

Els models de regressió logística i random forest són els que millor classifiquen les observacions del dataset de test, amb un percentatge d'encerts superior al 80%.

El model de support vector Machines és el de menor qualitat, doncs només classifica correctament el 60% dels registres.

Aquests tres models ens permeten complir l'objectiu inicial del nostre estudi: elaborar un model que ens permeti predir el valor de la variable Survived en noves observacions d'altres passatgers (on disposem de la resta de variables i s'hagi de predir el valor de Survived en base a aquestes).

BIBLIOGRAFIA

https://www.tinsa.es/blog/curiosidades/cuanto-costo-el-titanic/
https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/
https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/distinct
https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/nrow
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/shapiro.test
https://datasharkie.com/how-to-do-levene-test-in-r/
https://www.rdocumentation.org/packages/DescTools/versions/0.99.32/topics/LeveneTest
http://www.sthda.com/english/wiki/compare-multiple-sample-variances-in-r
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fligner.test
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test
http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r
https://www.tutorialspoint.com/r/r_chi_square_tests.htm
http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence
https://rpubs.com/Cristina_Gil/Regresion_Logistica
https://www.youtube.com/watch?v=J4Wdy0Wc_xQ
https://datatofish.com/export-dataframe-to-csv-in-r/

TAULA DE CONTRIBUCIONS

Contribucions	Firma
Investigació prèvia	DV, CA
Redacció de les respostes	DV, CA
Desenvolupament codi	DV, CA