**PAC 3** TOMAS ARAGONES MERCADE **LINK GITHUB** : https://github.com/0tmsgit/Titanic/wiki

Tabla de contenidos

[Descripción 1](#_Toc534664390)

[Objetivos 1](#_Toc534664391)

[Competencias 2](#_Toc534664392)

[Resolución 2](#_Toc534664393)

[Descripción del dataset (apartado 1) 2](#_Toc534664394)

[Importancia y objetivos de los análisis (apartados 2 y 3) 4](#_Toc534664395)

[Limpieza de los datos (apartados 2 y 3) 4](#_Toc534664396)

[Selección de los datos de interés 4](#_Toc534664397)

[NA y elementos vacíos 4](#_Toc534664398)

[Valores extremos 5](#_Toc534664399)

[Exportación de los datos preprocesados 5](#_Toc534664400)

[Análisis de los datos (apartado 4 y 5) 5](#_Toc534664401)

[Método de árbol de decisión 5](#_Toc534664402)

[Método de Modelo de regresión 7](#_Toc534664403)

[Conclusiones (apartado 7) 12](#_Toc534664404)

[Recursos 13](#_Toc534664405)

[Otros graficos (apartado 5) 13](#_Toc534664406)

# Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), orientado a aprender a identiﬁcar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

# Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica son los siguientes:

* Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
* Saber identiﬁcar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico. Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
* Identiﬁcar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico. Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
* Desarrollar las habilidades de aprendizaje que permita continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo. Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

# Competencias

Así, las competencias del Máster en Data Science que se desarrollan son:

* Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
* Capacidad para aplicar las técnicas especíﬁcas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

# Resolución

## Descripción del dataset (apartado 1)

El conjunto de datos objeto de análisis se ha obtenido a partir de este enlace (https://www.kaggle.com/c/titanic) en Kaggle y está constituido por 12 características (columnas) que presentan 891 pasajeros (ﬁlas o registros) en el fichero train, en el fichero test hay 11 columnas – las mismas menos la survived , variable que se quiere predecir- y 418 registros. Entre los campos de este conjunto de datos, encontramos los siguientes:

PassengerId: entero correlativo que identifica al pasajero.

Survived : entero binario que identifica si sobrevivio – 1- o no – 0-.

Pclass : entero para la clase de pasajero: 1, 2 y 3.

Name : String con el nombre del pasajero.

Sex : string con el sexo del pasajero de tipo factor.”female” o “male”

Age : numerico con la edad.

SibSp : entero numero de hermanos o esposas/os a bordo.

Parch : entero numero de padres o hijos a bordo.

Ticket : numero de ticket de tipo factor.

Fare : numerico decimal tipo de tarifa.

Cabin : identificador de la cabina string de tipo factor.

Embarked : identificador de la localidad de embarque con un caracter de tipo factor.

'data.frame': 891 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...

$ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

> summary(train)

PassengerId Survived Pclass Name Sex

Min. : 1.0 Min. :0.0000 Min. :1.000 Abbing, Mr. Anthony : 1 female:314

1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Abbott, Mr. Rossmore Edward : 1 male :577

Median :446.0 Median :0.0000 Median :3.000 Abbott, Mrs. Stanton (Rosa Hunt) : 1

Mean :446.0 Mean :0.3838 Mean :2.309 Abelson, Mr. Samuel : 1

3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000 Abelson, Mrs. Samuel (Hannah Wizosky): 1

Max. :891.0 Max. :1.0000 Max. :3.000 Adahl, Mr. Mauritz Nils Martin : 1

(Other) :885

Age SibSp Parch Ticket Fare Cabin Embarked

Min. : 0.42 Min. :0.000 Min. :0.0000 1601 : 7 Min. : 0.00 :687 : 2

1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000 347082 : 7 1st Qu.: 7.91 B96 B98 : 4 C:168

Median :28.00 Median :0.000 Median :0.0000 CA. 2343: 7 Median : 14.45 C23 C25 C27: 4 Q: 77

Mean :29.70 Mean :0.523 Mean :0.3816 3101295 : 6 Mean : 32.20 G6 : 4 S:644

3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000 347088 : 6 3rd Qu.: 31.00 C22 C26 : 3

Max. :80.00 Max. :8.000 Max. :6.0000 CA 2144 : 6 Max. :512.33 D : 3

NA's :177 (Other) :852 (Other) :186

> summary(test)

PassengerId Pclass Name Sex Age

Min. : 892.0 Min. :1.000 Abbott, Master. Eugene Joseph : 1 female:152 Min. : 0.17

1st Qu.: 996.2 1st Qu.:1.000 Abelseth, Miss. Karen Marie : 1 male :266 1st Qu.:21.00

Median :1100.5 Median :3.000 Abelseth, Mr. Olaus Jorgensen : 1 Median :27.00

Mean :1100.5 Mean :2.266 Abrahamsson, Mr. Abraham August Johannes : 1 Mean :30.27

3rd Qu.:1204.8 3rd Qu.:3.000 Abrahim, Mrs. Joseph (Sophie Halaut Easu): 1 3rd Qu.:39.00

Max. :1309.0 Max. :3.000 Aks, Master. Philip Frank : 1 Max. :76.00

(Other) :412 NA's :86

SibSp Parch Ticket Fare Cabin Embarked

Min. :0.0000 Min. :0.0000 PC 17608: 5 Min. : 0.000 :327 C:102

1st Qu.:0.0000 1st Qu.:0.0000 113503 : 4 1st Qu.: 7.896 B57 B59 B63 B66: 3 Q: 46

Median :0.0000 Median :0.0000 CA. 2343: 4 Median : 14.454 A34 : 2 S:270

Mean :0.4474 Mean :0.3923 16966 : 3 Mean : 35.627 B45 : 2

3rd Qu.:1.0000 3rd Qu.:0.0000 220845 : 3 3rd Qu.: 31.500 C101 : 2

Max. :8.0000 Max. :9.0000 347077 : 3 Max. :512.329 C116 : 2

(Other) :396 NA's :1 (Other) : 80

## Importancia y objetivos de los análisis (apartados 2 y 3)

A partir de este conjunto de datos se plantea la problemática de determinar qué variables de las que se disponen juegan una influencia en la supervivencia en el Titanic.

Las variables que se excluyen por ser irrelevantes, son Name, PassagerId, Ticket. Las otras variables se incluyen en el analisis.

Así se utilizará el método del árbol de decisión para determinar que variables determinan la supervivencia de un pasajero en función de las variables: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked.

Asimismo, se realizará una regresión logística en función de las variables anteriores.

En un último caso se realizará una prueba de contraste respecto a la variable Edad.

Evidentemente puede ser que de resultas del análisis de los datos se concluya en eliminar una variable por el hecho que hay demasiados valores perdidos. Igualmente, el modelo determinara cuales son las variables relevantes para decidir sobre la supervivencia.

## Limpieza de los datos (apartados 2 y 3)

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura de los ﬁchero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame.

### Selección de los datos de interés

Se eliminan las columnas que no tienen interés para el análisis: Name, PassagerId, Ticket.

### NA y elementos vacíos

Los NA se sustituyen por su median en Age y por el valor más común en embarked.

table(train$Embarked)

C Q S

2 168 77 644

La variable Cabin se elimina por tener demasiados valores vacíos, en total 687 sobre los 891 del fichero train.

table(train$Cabin)

A10 A14 A16

687 1 1 1

### Valores extremos

Las variables restantes Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked no contienen valores extremos.

#valors extrems, tots els valors semblen possibles, per tant no seliminen

boxplot(train$Age, main="Age")

boxplot(train$Pclass, main="PClass")

boxplot(train$SibSp, main="SibSp")

boxplot(train$Parch, main="Parch")

boxplot(train$Fare, main="fare")

### Exportación de los datos preprocesados

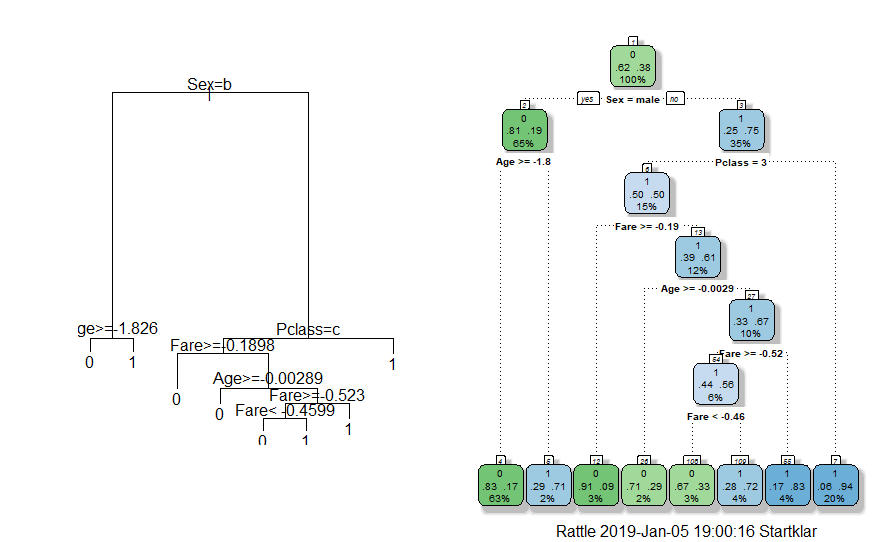
Una vez que hemos acometido sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, procedemos a guardar estos en unos nuevos ﬁcheros denominados training\_set\_clean.csv, validation\_set\_clean.csv, test\_clean.csv, final\_training\_set\_clean.csv .

## Análisis de los datos (apartado 4 y 5)

Se aplicarán dos métodos distintos, primer un árbol de decisión y en segundo término un modelo de regresión.

### Método de árbol de decisión

Se utiliza la función Dtree para generar un modelo tipo árbol de decisión en base al fichero de datos limpios con los registros train, training\_set\_clean.



Como se ve en el gráfico, las variables decisorias en el modelo son por este orden: el sexo, la edad y fare.

library(rpart)

start.time <- Sys.time()

Dtree <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,

data=training\_set\_clean,

method="class")

Dtree\_pred = predict(Dtree, newdata = validation\_set\_clean[-1], type = 'class')# eliminem amb -1 allo que volem predir, la variable survived.

end.time <- Sys.time()

time.takenDT <- end.time - start.time

par(mfrow=c(1,2)) # imprimeix dos caracter horitzontalment

plot(Dtree)

text(Dtree)

#### Validación del modelo

Ahora se testea el modelo con el set de validación y creamos, la matriz de confusión para definir la validez del modelo. La matriz de confusión nos posibilita un análisis del éxito de la predicción. El eje horizontal contiene las clases y el vertical tiene las clases predichas.

El ratio de éxito de la predicción es del 77.53% , la predicción exitosa de la supervivencia es del 63.24% mientras que el éxito de predicción de la no supervivencia es del 86.36%. El modelo ha tardado 0.08 segundos en realizar la predicción.

#creracio de la matriu de confusio

Dtree\_cm = table(t(validation\_set\_clean[, 1]), Dtree\_pred)

# calculem el rati dexit total

Dtree\_success = sum(diag(Dtree\_cm))/sum(Dtree\_cm)\*100

# calculem el rati dexit per predir una supervivencia

Dtree\_success\_survived = Dtree\_cm[2,2]/(Dtree\_cm[2,1]+Dtree\_cm[2,2])\*100

# calcul del rati d'exit per predir la casualitat.

Dtree\_success\_notsurvived = Dtree\_cm[1,1]/(Dtree\_cm[1,2]+Dtree\_cm[1,1])\*100

# prdit en vertical i actual en horitzontal ( 68 sobrevivents en el set de validacio )

Dtree\_cm

#El rati dexit de prediccio es de 77.53% , la prediccio exitosa de la supervivencia es del 63.24% mentres l'exit de prediccio per la no supervivencia es del 86.36%. El model ha tardat 0.03 segons en fer la prediccio.

### Método de Modelo de regresión

Utilizando los mismos sets de datos del caso anterior ya limpios, training\_set\_clean y el validation\_set\_clean se realiza un modelo de regresión logística.

train1<-training\_set\_clean # Utilitzem el mateix set per entrenar el model que amb el model darbre

train2<-validation\_set\_clean # Utilitzem el mateix set per validar el model que amb el model darbre

> # Generación de un modelo de regresión logit

>

> model <- glm(Survived ~.,family=binomial(link='logit'),data=train1)

> summary(model)

Call:

glm(formula = Survived ~ ., family = binomial(link = "logit"),

data = train1)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5315 -0.6218 -0.4181 0.6180 2.4585

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.07402 0.36235 8.484 < 2e-16 \*\*\*

Pclass2 -0.89844 0.33284 -2.699 0.00695 \*\*

Pclass3 -2.10402 0.33340 -6.311 2.77e-10 \*\*\*

Sexmale -2.72914 0.22279 -12.250 < 2e-16 \*\*\*

Age -0.50377 0.11465 -4.394 1.11e-05 \*\*\*

SibSp -0.27578 0.11522 -2.394 0.01669 \*

Parch -0.04967 0.13450 -0.369 0.71190

Fare -0.01282 0.12670 -0.101 0.91942

EmbarkedQ -0.02165 0.44578 -0.049 0.96127

EmbarkedS -0.60229 0.27299 -2.206 0.02736 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 946.06 on 710 degrees of freedom

Residual deviance: 628.56 on 701 degrees of freedom

(2 observations deleted due to missingness)

AIC: 648.56

Number of Fisher Scoring iterations: 5

Ver resultados anteriores en forma de tabla, el modelo resultante permite la predicción pero hay ciertas variables que no son relevantes, por tanto las vamos a proceder a eliminar mejorando el modelo.

#Parch, Fare, EmbarkedQ no tenen prou rellevancia >0.05 les eliminem del model

> stepmodel = step(model, direction="both")

Start: AIC=648.56

Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked

Df Deviance AIC

- Fare 1 628.57 646.57

- Parch 1 628.70 646.70

<none> 628.56 648.56

- Embarked 2 634.86 650.86

- SibSp 1 635.11 653.11

- Age 1 649.39 667.39

- Pclass 2 675.80 691.80

- Sex 1 816.57 834.57

Step: AIC=646.57

Survived ~ Pclass + Sex + Age + SibSp + Parch + Embarked

Df Deviance AIC

- Parch 1 628.73 644.73

<none> 628.57 646.57

+ Fare 1 628.56 648.56

- Embarked 2 634.92 648.92

- SibSp 1 635.38 651.38

- Age 1 649.47 665.47

- Pclass 2 694.34 708.34

- Sex 1 816.84 832.84

Step: AIC=644.73

Survived ~ Pclass + Sex + Age + SibSp + Embarked

Df Deviance AIC

<none> 628.73 644.73

+ Parch 1 628.57 646.57

+ Fare 1 628.70 646.70

- Embarked 2 635.20 647.20

- SibSp 1 637.52 651.52

- Age 1 649.59 663.59

- Pclass 2 694.99 706.99

- Sex 1 824.72 838.72

> formula(stepmodel)

Survived ~ Pclass + Sex + Age + SibSp + Embarked

>

> summary(stepmodel)

Call:

glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Embarked,

family = binomial(link = "logit"), data = train1)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5609 -0.6211 -0.4162 0.6258 2.4604

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.032749 0.326824 9.279 < 2e-16 \*\*\*

Pclass2 -0.887719 0.299357 -2.965 0.00302 \*\*

Pclass3 -2.092034 0.280044 -7.470 8.0e-14 \*\*\*

Sexmale -2.708208 0.216401 -12.515 < 2e-16 \*\*\*

Age -0.501988 0.114154 -4.397 1.1e-05 \*\*\*

SibSp -0.293125 0.107614 -2.724 0.00645 \*\*

EmbarkedQ 0.008495 0.439026 0.019 0.98456

EmbarkedS -0.594440 0.269922 -2.202 0.02765 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 946.06 on 710 degrees of freedom

Residual deviance: 628.73 on 703 degrees of freedom

(2 observations deleted due to missingness)

AIC: 644.73

Number of Fisher Scoring iterations: 5

Esta regresión logística permite la perdición de la supervivencia según sus características, ver resultados anteriores en forma de tabla. Seguidamente veremos su validación, la precisión y sensibilidad del modelo.

#### Validación del modelo

El modelo tiene una mediano de predicción acertada del 77.53%, que es idéntica al modelo Dtree.

> #validacion del model amb el set de validacio

>

> pred.train <- predict(model,train2)

> pred.train <- ifelse(pred.train > 0.5,1,0)

> # Mitjana de prediccio encertada

> mean(pred.train==train2$Survived)

[1] 0.7752809

#### Precisión y sensibilidad del modelo

En el reconocimiento de patrones, recuperación de información y clasificación binaria, la precisión (también llamada valor predictivo positivo) es la fracción de instancias relevantes entre las instancias recuperadas, mientras que la recuperación (también conocida como sensibilidad) es la fracción de instancias relevantes que se han recuperado sobre el total Cantidad de instancias relevantes. Tanto la precisión como el recuerdo se basan, por lo tanto, en una comprensión y medida de relevancia. En este modelo la precisión es del 0.77 y la sensibilidad es del 0.91.

t1<-table(pred.train,train2$Survived)

> # Presicio i sensibilidad del model

> presicion<- t1[1,1]/(sum(t1[1,]))

> recall<- t1[1,1]/(sum(t1[,1]))

> presicion

[1] 0.7692308

> recall

[1] 0.9090909

#### Prueba estadística de la Edad

Visto la edad es una variable decisoria en el modelo de árbol, se analiza si realmente se puede corroborar que la edad determina la no supervivencia.

Los grupos serán aquellos que son más jóvenes comparados con los más ancianos. Así se comprobara si los supervivientes tenían una edad más joven estadísticamente.

En un primer lugar se analiza la variable Age sobre el evento. Veamos las edades medias de los supervivientes y los muertos.

#Edat mitja dels que van sobreviure i dels que van morir

aggregate(Age~Survived, data=training\_set\_clean, mean)

> t.test(Age ~ Survived, data=training\_set\_clean)

Welch Two Sample t-test

data: Age by Survived

t = 1.7499, df = 552.36, p-value = 0.0807

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.01672108 0.28964990

sample estimates:

mean in group 0 mean in group 1

0.05244214 -0.08402226

El p-value es menor a 0.05 por tanto se rechaza la hipótesis nula y se confirma la hipótesis de que los jóvenes tuvieron un mayor supervivencia.

##### Comprobación de la normalidad y la homogeneidad de la varianza.

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson Darling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de signiﬁcación preﬁjado α = 0,05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal. Para el set training\_set\_clean y validation\_set\_clean se cumple que todas las variables son normales.

> library(nortest)

> alpha = 0.05

> col.names = colnames(training\_set\_clean)

> for (i in 1:ncol(training\_set\_clean)) {

+ if (i == 1) cat("Variables que no siguen una distribución normal:\n")

+ if (is.integer(training\_set\_clean[,i]) | is.numeric(training\_set\_clean[,i])) {

+ p\_val = ad.test(training\_set\_clean[,i])$p.value

+ if (p\_val < alpha) {

+ cat(col.names[i])

+ # Formateig de la sortida

+ if (i < ncol(training\_set\_clean) - 1) cat(", ")

+ if (i %% 3 == 0) cat("\n")

+ }

+ }

+ }

Variables que no siguen una distribución normal:

Seguidamente, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los pasajeros del set de train son iguales que los del set de validación. En el siguiente test, la hipótesis nula consiste en que las varianzas son iguales.

#Seguidament sestudia la homegeneitat dels sets train i validacio

library(car)

fligner.test(Survived ~ Age, data = training\_set\_clean)

Fligner-Killeen test of homogeneity of variances

data: Survived by Age

Fligner-Killeen:med chi-squared = 1.1626, df = 1, p-value = 0.21

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de Age son homogéneas.

##### Prueba de contraste para edad

Vamos a mirar de responder la siguiente pregunta mediante el t-test: ¿Los supervivientes eran mas jóvenes?

La prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la supervivencia es superior dependiendo si se trata de un joven o no.

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:

H0 : µ1−µ2 =0 H1 : µ1−µ2 < 0 donde µ1 es la media de la población de la que se extrae la primera muestra y µ2 es la media de la población de la que extrae la segunda. Así, tomaremos α =0,05.

Prueba de contraste de hipótesis, hipótesis nula: la diferencia de edad entre los supervivientes y los no supervivientes es nula. H1: Hay una diferencia de edad entre los supervivientes y los no supervivientes.

> #Edat mitja dels que van sobreviure i dels que van morir

> aggregate(Age~Survived, data=training\_set\_clean, mean)

Survived Age

1 0 0.05244214

2 1 -0.08402226

A simple vista se ve en los resultados de la tabla anterior que la mediana de los valores normalizados de edad son mayores en los no supervivientes que en los supervivientes, 0.052>-0.08. Para realizar la prueba de contraste utilizaremos t.test.

>

> t.test(Age ~ Survived, data=training\_set\_clean)

Welch Two Sample t-test

data: Age by Survived

t = 1.7499, df = 552.36, p-value = 0.0807

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.01672108 0.28964990

sample estimates:

mean in group 0 mean in group 1

0.05244214 -0.08402226

Se confirma que los que sobrevivieron eran más jóvenes, ya que se puede descartar la hipótesis nula p-value 0.807>0.05.

# Conclusiones (apartado 6)

Como se ha visto, se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes variables relativas a pasajeros con motivo de cumplir en la medida de lo posible con el objetivo que se planteaba al comienzo. Para cada una de ellas, hemos podido ver cuáles son los resultados que arrojan (entre otros, mediante tablas) y qué conocimientos pueden extraerse a partir de ellas. Así, el análisis de por árbol Dtree (relacionando las variales sexo, Age y Fare) y la prueba sobre la edad nos ha permitido conocer cuáles de estas variables ejercen una mayor inﬂuencia sobre la supervivencia del pasajero, mientras que el modelo de regresión logística obtenido resulta de utilidad a la hora de realizar predicciones para esta variable dadas unas características concretas (utilizando todas las variables significativas). Por este motivo se considera haber cumplido el objetivo inicial de predecir mediante distintas aproximaciones la supervivencia en función de las variables de nuestros datos. Previamente, se han sometido los datos aun preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (outliers). Para el caso del primero, se ha hecho uso de un método de imputación de valores de tal forma que no tengamos que eliminar registros del conjunto de datos inicial y que la ausencia de valores no implique llegar a resultados poco certeros en los análisis. Para el caso del segundo, el cual constituye un punto delicado a tratar, se ha detectado que no hay valores extremos.

# Recursos

1. Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.

2. Vegas, E. (2017). Preprocesamiento de datos. Material UOC.

3. https://stats.idre.ucla.edu/r/dae/logit-regression/

4. Rovira, C. (2008). Contraste de hipótesis. Material UOC.

5. Test for homogeneity of variances - Lavene’s test and the Fligner Killeen test (2016) [en línea]. bioSt@TS. [Consulta: 26 de diciembre de 2017] https://biostats.w.uib.no/ test-for-homogeneity-of-variances-levenes-test/

# Otros graficos (apartado 5)

