SANCHEZ_SIMBANA_PRAC2

Autores: Diana, Carlos

7 de January, 2020

Detalles de la actividad

El presente documento muestra un caso práctico orientado a identificar, limpiar, integrar, preparar, validar y analizar los datos dentro de un proyecto analítico en el cual se utiliza herramientas de código abierto para la consecución de los objetivos.

Descripción

Para la realización de la práctica, se han elegido los datos de la competencia "Titanic ML" la cual originalmente busca la realización de un modelo capaz de predecir los pasajeros que sobrevivieron al naufragio del 15 de abril de 1912. En esta actividad se realizará la limpieza de los datos para corregir valores erróneos que pudieren encontrarse; la validación de los mismos y el análisis descriptivo para proveer información de valor en relación a las probabilidades de supervivencia de los pasajeros del Titanic.

Objetivos

Los objetivos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y la 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 les permitan 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

En esta práctica se desarrollan las siguientes competencias:

- 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

El dataset contiene dos archivos referentes al naufragio del Titanic, train.csv (891 registros) test.csv (418 registros) y correspondientes a la competencia "Titanic ML" disponible en el enlace https://www.kaggle.com/c/titanic.

Los archivos de datos se encuentran constituidos de la siguiente manera:

- **PassengerId:** Es el identificador del pasajero.
- **Survived:** Indica si el pasajero sobrevivió o no. Este campo solo se encuentra en el fichero train.csv debido a que este se utilizará como archivo de entrenamiento para predecir la supervivencia de cada uno de los pasajeros del archivo test.csv. Cero (0) es igual a "No sobrevive". Uno (1) es igual a "Sobrevive".
- **Pclass:** Es el tipo de ticket que adquirió el pasajero. 1 es Upper, 2 es Middle y 3 es Lower.
- **Name:** Es el nombre del pasajero.
- **Sex:** Es el sexo del pasajero.
- **Age:** Es la edad del pasajero expresada en años.
- **SibSp:** Es el número de hermanos y/o esposas(os) a bordo, por pasajero.
- **Parch:** Es el número de padres y/o hijos a bordo, por pasajero.

- **Ticket:** Es el número de ticket del pasajero.
- **Fare:** Es la tarifa pagada por el pasajero.
- **Cabin:** Es el número de camarote del pasajero.
- **Embarked:** Es el Puerto de embarque del pasajero. C = Cherbourg, Q = Queenstown, S = Southampton.

Importancia y objetivos de los análisis

La importancia de este conjunto de datos radica en encontrar la relación entre las diferentes variables las cuales permiten que aumenten o disminuyan la probabilidad de los pasajeros de sobrevivir al naufragio. Para esto es importante realizar la creación y cambios que fueren necesarios en las variables utilizando métodos como la normalización, discretización y transformaciones necesarias para posteriormente llevar a cabo el análisis de los datos. Debido a que en total los datos de los dos archivos suman 1309 registros y el objetivo del análisis es el contar con la probabilidad individual de cada pasajero para sobrevivir, no se utilizará el método de reducción de la cantidad.

La realización de este análisis permitirá que se realicen otros modelos predictivos aplicados a otros naufragios con la finalidad que puedan establecerse nuevas normativas y planes prevención en temas de navegación fluvial en el mundo. No obstante, también sirve de motivación para que aplicando técnicas similares, se realicen modelos enfocados en la navegación terrestre, aérea y ferroviaria.

Limpieza de los datos

Previa a la ejecución de la limpieza de los datos, se requiere realizar un análisis visual que permita determinar la ruta a seguir para obtener la mejor calidad de datos posibles. Para esto se han realizado las siguientes observaciones:

Campo	Tipo de variable	Observaciones
PassengerId	Descriptiva	Es un número secuencial de cada pasajero. En el archivo train.csv inicia en 1 y finaliza en 891. En el archivo test.csv inicia en 892 y finaliza en 1309.
Survived	Categórica	No existen valores en blanco.
Pclass	Categórica	No existen valores en blanco.
Name	Cualitativa	Contiene las abreviaturas Mr. (hombres adultos), Mrs. (mujeres adultas casadas), Miss. (mujeres solteras), Master. (hombres jóvenes); datos que pueden ser de utilidad para determinar la edad en caso de no encontrarse en los

		archivos iniciales.
Sex	Categórica	No existen valores en blanco.
Age	Cuantitativa	Existen las siguientes cantidades: Valores menores a uno train.csv (7), test.csv (5), Valores con decimales xx,5 train.csv (18), test.csv (15), Campos en blanco train.csv (177), test.csv (86).
SibSp	Cuantitativa	No existen valores en blanco.
Parch	Cuantitativa	No existen valores en blanco.
Ticket	Categórica	No existen valores en blanco.
Fare	Cuantitativa	No existen valores en blanco.
Cabin	Categórica	Existen valores en blanco: test.csv (327), train.csv (687)
Embarked	Categórica	Existen valores en blanco: train.csv (2)

Selección de los datos de interés

Los atributos presentes en el análisis contienen información de cada pasajero del Titanic en los cuales se pueden encontrar características físicas, emocionales y restricciones que pudieron afectar a cada pasajero al momento del naufragio y por ende afectaron en su supervivencia. Por ejemplo, en el caso de tener niños, seguramente los padres, trataron de salvarlos primero incluso poniendo en riesgo su propia supervivencia.

De las columnas provistas en los conjuntos de datos, se puede obviar: PassengerId, Ticket, Fare, Cabin y Embarked en vista que son variables que no tienen relación con la probabilidad de supervivencia.

```
if(!require(dplyr)){
    install.packages('dplyr', repos='http://cran.us.r-project.org')
    library(dplyr)
}
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#Carga de Datos desde el directorio actual del archivo MarkDown
test <- read.csv('./titanic-test.csv', header = TRUE)
train <- read.csv('./titanic-train.csv', header = TRUE)

# Eliminar las columnas que no utilizaremos para el analisis
train <- train[, -1]
train <- train[, -(8:11)]

test <- test[, -1]
test <- test[, -(7:10)]
```

Ceros y elementos vacíos

```r
#Vista de la estructura básica de Titanic
str(train)
```

```
## 'data.frame':    891 obs. of  7 variables:
##  $ 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 ...
```

```r
summary(train)
```

```
##     Survived          Pclass                                          Name
##  Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                  :
1
##  1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :
1
##  Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :
1
##  Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :
1
##  3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):
1
##  Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :
1
##                                   (Other)
:885
##      Sex           Age             SibSp           Parch
##  female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  male  :577   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##               Median :28.00   Median :0.000   Median :0.0000
```

```
##                Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                NA's   :177
```

```
# Estadísticas de valores vacíos
colSums(is.na(train))
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch
##        0        0        0        0      177        0        0
```

```
colSums(train=="")
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch
##        0        0        0        0       NA        0        0
```

```
# Estadisticas de valores vacios
colSums(is.na(test))
```

```
## Pclass   Name    Sex    Age  SibSp  Parch
##      0      0      0     86      0      0
```

```
colSums(test=="")
```

```
## Pclass   Name    Sex    Age  SibSp  Parch
##      0      0      0     NA      0      0
```

```
#TRAIN
# Obtenemos la media de las edades de acuerdo al nombre de la persona Mr,
Mrs, Miss, Master
meanmr <- round(mean(train$Age[regexpr("Mr.", train$Name) > 0 &
!is.na(train$Age)]), 0)
meanmrs <- round(mean(train$Age[regexpr("Mrs.", train$Name) > 0 &
!is.na(train$Age)]), 0)
meanmiss <- round(mean(train$Age[regexpr("Miss.", train$Name) > 0 &
!is.na(train$Age)]), 0)
meanmaster <- round(mean(train$Age[regexpr("Master.", train$Name) > 0 &
!is.na(train$Age)]), 0)

# Tomamos la media para valores vacíos de la variable "Age" y los
reemplazamos por la media
train$Age[regexpr("Mr.", train$Name) > 0 & is.na(train$Age)] <- meanmr
train$Age[regexpr("Mrs.", train$Name) > 0 & is.na(train$Age)] <- meanmrs
train$Age[regexpr("Miss", train$Name) > 0 & is.na(train$Age)] <- meanmiss
train$Age[regexpr("Master", train$Name) > 0 & is.na(train$Age)] <- meanmaster
train$Age[is.na(train$Age)] <- meanmr

#TEST
# Obtenemos la media de las edades de acuerdo al nombre de la persona Mr,
Mrs, Miss, Master
meanmr <- round(mean(test$Age[regexpr("Mr.", test$Name) > 0 &
!is.na(test$Age)]), 0)
```

```
meanmrs <- round(mean(test$Age[regexpr("Mrs.", test$Name) > 0 &
!is.na(test$Age)]), 0)
meanmiss <- round(mean(test$Age[regexpr("Miss.", test$Name) > 0 &
!is.na(test$Age)]), 0)
meanmaster <- round(mean(test$Age[regexpr("Master.", test$Name) > 0 &
!is.na(test$Age)]), 0)

# Tomamos la media para valores vacíos de la variable "Age" y los
reemplazamos por la media
test$Age[regexpr("Mr.", test$Name) > 0 & is.na(test$Age)] <- meanmr
test$Age[regexpr("Mrs.", test$Name) > 0 & is.na(test$Age)] <- meanmrs
test$Age[regexpr("Miss", test$Name) > 0 & is.na(test$Age)] <- meanmiss
test$Age[regexpr("Master", test$Name) > 0 & is.na(test$Age)] <- meanmaster
test$Age[is.na(test$Age)] <- meanmr
```

- Como otro ejemplo de gestión, en la columna Age, a las filas vacías se le asigna el promedio de la edades de los pasajeros registrados en el dataset

---

Valores extremos

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de dos vías: (1) representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja) o (2) utilizar la función boxplots.stats() de R, la cual se emplea a continuación.

Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
boxplot.stats(train$Survived)$out

## integer(0)

boxplot.stats(train$Pclass)$out

## integer(0)

boxplot.stats(train$Age)$out

##  [1]  2.00 58.00 55.00  2.00 66.00 65.00  0.83 59.00 71.00 70.50  2.00
55.50
## [13]  1.00 61.00  1.00 56.00  1.00 58.00  2.00 59.00 62.00 58.00 63.00
65.00
## [25]  2.00  0.92 61.00  2.00 60.00  1.00  1.00 64.00 65.00 56.00  0.75
2.00
## [37] 63.00 58.00 55.00 71.00  2.00 64.00 62.00 62.00 60.00 61.00 57.00
80.00
## [49]  2.00  0.75 56.00 58.00 70.00 60.00 60.00 70.00  0.67 57.00  1.00
0.42
## [61]  2.00  1.00 62.00  0.83 74.00 56.00
```

```r
boxplot.stats(train$SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3
5 4 3
## [39] 4 8 4 3 4 8 4 8
```

```r
boxplot.stats(train$Parch)$out
```

```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2
1 2 1
##  [38] 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1 1
2 1 2
##  [75] 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2 2 3 4
1 2 1
## [112] 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1 2 1 1 2 5
2 1 1
## [149] 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1 1 2 1 2 3 1 2
1 2 2
## [186] 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1 1 5 2
```

Observamos que tanto como para la variable Survived y Pclass los valores extremos son normales ya que tenemos solo 2 estados, sobrevive o no y 3 clases baja, media y alta.

Para las variables de Age, SibSp y Parch observamos mas variedad de datos ya que existen muchas personas de diferentes edades en el barco ademas de el numero de hermanos, esposasm padres e hijos a bordo es muy distinto entre cada pasajero por lo tanto a continuacion tenemos los valores maximo y minimos de cada una de estas variables:

```r
# Edad Máxima y Mínima
min(train$Age)
```

```
## [1] 0.42
```

```r
max(train$Age)
```

```
## [1] 80
```

```r
# SibSp Máxima y Mínima
min(train$SibSp)
```

```
## [1] 0
```

```r
max(train$SibSp)
```

```
## [1] 8
```

```r
# Parch Máxima y Mínima
min(train$Parch)
```

```
## [1] 0
```

```r
max(train$Parch)
```

```
## [1] 6
```

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 un nuevo fichero denominado train_data_clean.csv:

```
# Exportación de los datos limpios en .csv
write.csv(train, "train_data_clean.csv")
```

Análisis de los datos

Selección de los grupos de datos a analizar

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

```
# Agrupación por alta, media y baja clase
train.upperclass <- train[train$Pclass == 1,]
train.middleclass <- train[train$Pclass == 2,]
train.lowerclass <- train[train$Pclass == 3,]

# Agrupación por sexo
train.male <- train[train$Sex == "male",]
train.female <- train[train$Sex == "female",]
```

Comprobación de la normalidad y 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 AndersonDarling.

Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado $\alpha = 0,05$. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```r
if(!require(nortest)){
    install.packages('nortest', repos='http://cran.us.r-project.org')
    library(nortest)
}
```

```
## Loading required package: nortest
```

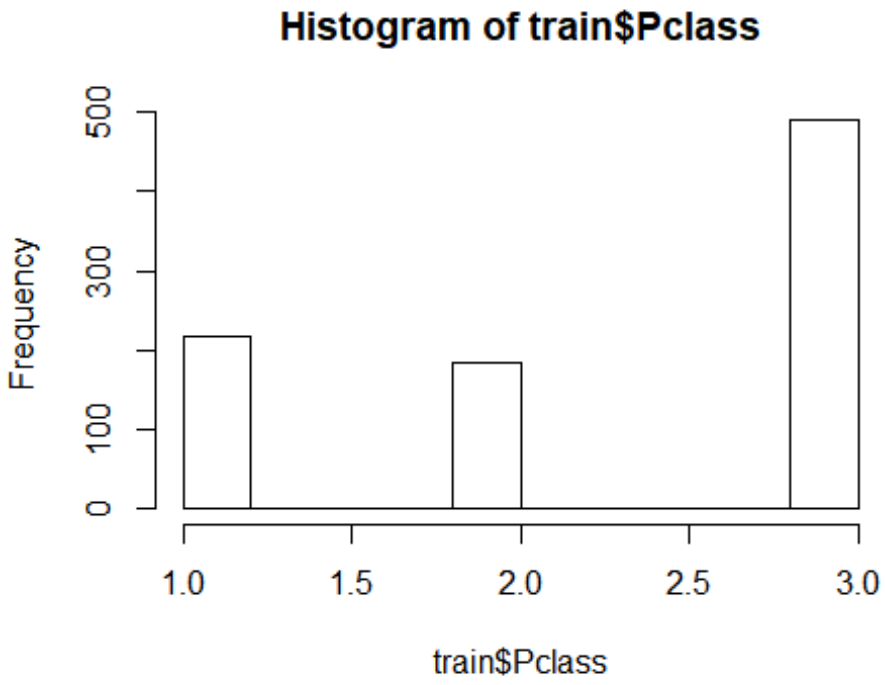
```
## Warning: package 'nortest' was built under R version 3.5.2
```

```r
alpha = 0.05
col.names = colnames(train)
for (i in 1:ncol(train)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
    if (is.integer(train[,i]) | is.numeric(train[,i])) {
      p_val = ad.test(train[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(train) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

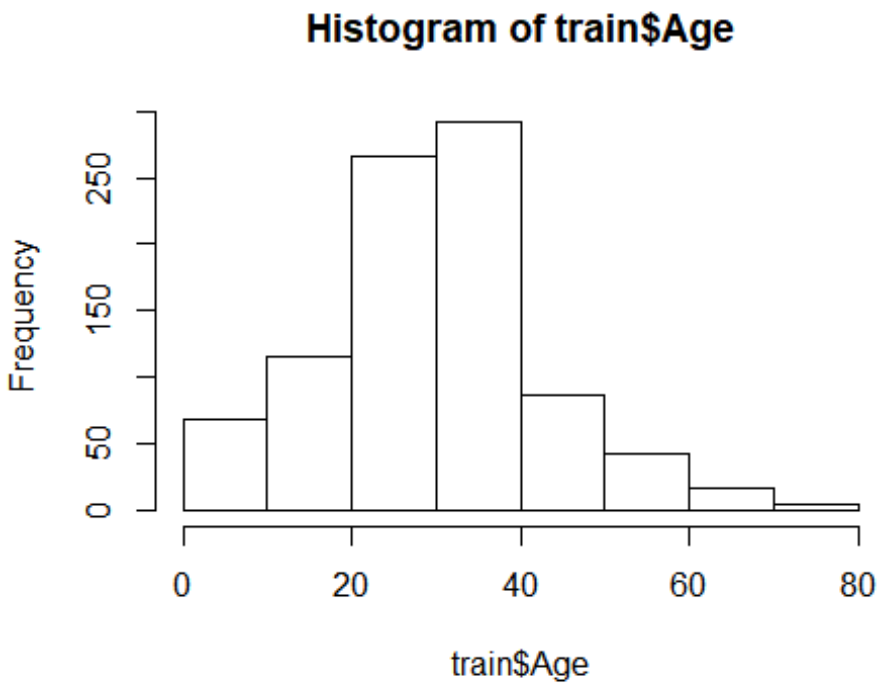
```
## Variables que no siguen una distribución normal:
## Survived, Pclass, Age, SibSp
## Parch
```

```r
#Para comprobarlo graficamente utilizamos histograma para Survived
hist(train$Survived)
```

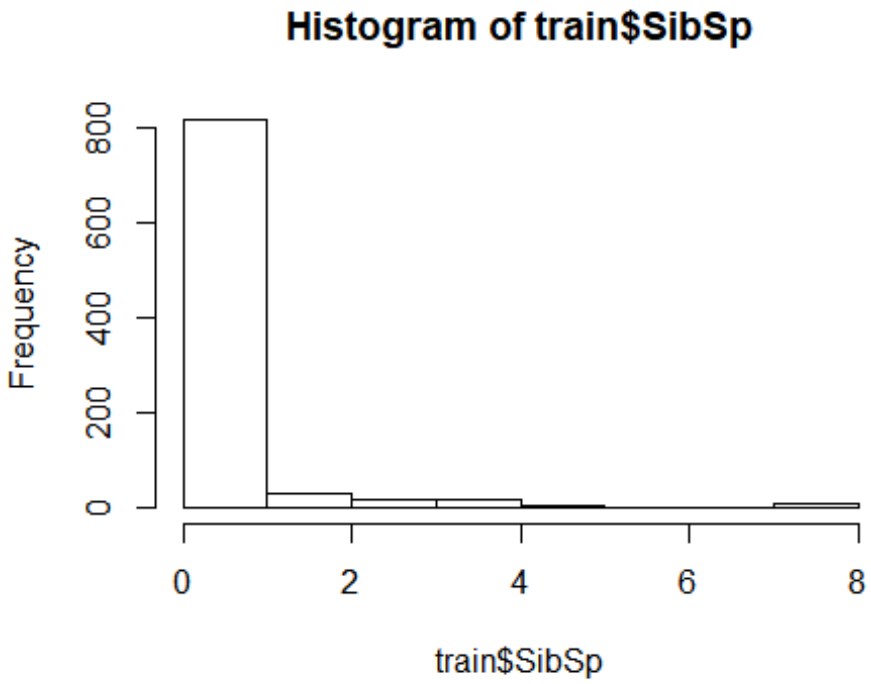

```r
#Para comprobarlo graficamente utilizamos histograma para Pclass
hist(train$Pclass)
```



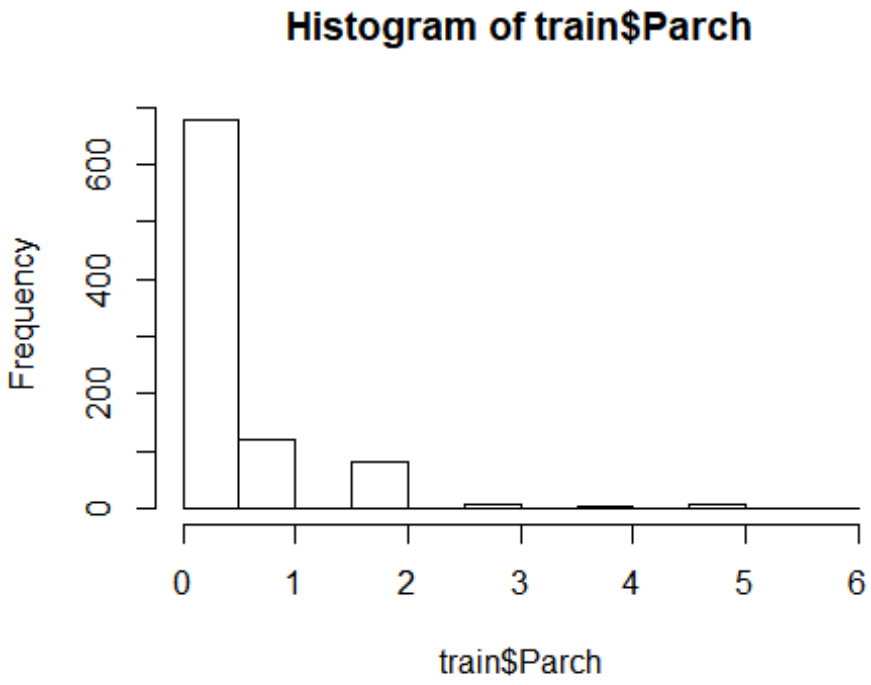
```r
#Para comprobarlo graficamente utilizamos histograma para Age
hist(train$Age)
```




```r
#Para comprobarlo graficamente utilizamos histograma para SibSp
hist(train$SibSp)
```




```r
#Para comprobarlo graficamente utilizamos histograma para Parch
hist(train$Parch)
```



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 las clases de pasajeros que se encontraban en el barco. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
fligner.test(Survived ~ Pclass, data = train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Pclass
## Fligner-Killeen:med chi-squared = 35.766, df = 2, p-value = 1.712e-08
```

Los valores de p inferiores a 0,05 sugieren que las variaciones son significativamente diferentes y se ha violado el supuesto de homogeneidad de la varianza.

Pruebas estadísticas

¿Qué variables cuantitativas influyen más en la probabilidad de supervivencia al naufragio?

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre la supervivencia. Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "precio"
for (i in 2:ncol(train)) {
  if (is.integer(train[,i]) | is.numeric(train[,i])) {
    spearman_test = cor.test(train[,i],
                             train[,length(train)],
                             method = "spearman")

    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value

    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
```

```
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(train)[i]
  }
}
```

```
## Warning in cor.test.default(train[, i], train[, length(train)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(train[, i], train[, length(train)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(train[, i], train[, length(train)], method =
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(train[, i], train[, length(train)], method =
## "spearman"): Cannot compute exact p-value with ties
```

```
print(corr_matrix)
```

```
##           estimate      p-value
## Pclass -0.02280134 4.966663e-01
## Age    -0.24574357 1.011225e-13
## SibSp   0.45001397 1.226002e-45
## Parch   1.00000000 0.000000e+00
```

Así, identificamos cuáles son las variables más correlacionadas con la supervivencia en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, queda patente cómo la variable más relevante en la fijación de la supervivencia es Es el número de padres y/o hijos a bordo (Parch).

Nota. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

¿Es mayor la probabilidad de supervivencia para las mujeres y niños?

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la probabilidad de sobrevivir es mayor para las mujeres y niños. Para ello, tendremos dos muestras: la primera de ellas se corresponderá a los pasajeros mujeres y ninos, la segunda, con aquellos pasajeros de sexo masculino.

```
train.female.sobrev <- train[train$Sex == "female" & train$Survived == 1,]
train.male.sobrev <- train[train$Sex == "male" & train$Survived == 1,]

#Calculos mujeres
nrow(train.female.sobrev)
```

```
## [1] 233

#Calculos hombres
nrow(train.male.sobrev)

## [1] 109

t.test(train.female.sobrev$Age, train.male.sobrev$Age, alternative =
"greater")

##
##  Welch Two Sample t-test
##
## data:  train.female.sobrev$Age and train.male.sobrev$Age
## t = 0.48948, df = 182.66, p-value = 0.3125
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2.016578       Inf
## sample estimates:
## mean of x mean of y
##  28.45064  27.60248
```

¿Cuáles son las variables dependientes en el análisis?

Para nuestro caso la variable dependiente es Survived, ya que necesitamos conocer de que manera afectan las demas variables para determinar las probabilidades de sobrevivir.

Modelo de regresión lineal

```
#Estimacion del modelo de acuerdo a la clase y al sexo
Model.1.1<- lm(Survived~Pclass+Sex, data=train)
summary(Model.1.1)

##
## Call:
## lm(formula = Survived ~ Pclass + Sex, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9252 -0.2505 -0.0925  0.2328  0.9075
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08327    0.04028   26.89   <2e-16 ***
## Pclass      -0.15803    0.01567  -10.09   <2e-16 ***
```

```
## Sexmale     -0.51667    0.02740  -18.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3874 on 888 degrees of freedom
## Multiple R-squared:  0.3677, Adjusted R-squared:  0.3663
## F-statistic: 258.2 on 2 and 888 DF,  p-value: < 2.2e-16
```

```
#Estimacion del modelo de acuerdo a la edad y al sexo
Model.1.2<- lm(Survived~Sex+Age, data=train)
summary(Model.1.2)
```

```
##
## Call:
## lm(formula = Survived ~ Sex + Age, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7676 -0.1949 -0.1829  0.2585  0.8601
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.769566   0.036786  20.920   <2e-16 ***
## Sexmale     -0.549620   0.028896 -19.021   <2e-16 ***
## Age         -0.001001   0.001042  -0.961    0.337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4087 on 888 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.2944
## F-statistic: 186.6 on 2 and 888 DF,  p-value: < 2.2e-16
```

```
#Estimacion del modelo de acuerdo a la clase, Es el número de hermanos y/o
esposas(os) a bordo y Es el número de padres y/o hijos a bordo
Model.1.3<- lm(Survived~SibSp+Parch+Pclass, data=train)
summary(Model.1.3)
```

```
##
## Call:
## lm(formula = Survived ~ SibSp + Parch + Pclass, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8685 -0.2746 -0.2355  0.3966  0.8346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.82237    0.04537  18.127  < 2e-16 ***
## SibSp       -0.02336    0.01529  -1.528  0.12683
## Parch        0.06628    0.02085   3.179  0.00153 **
## Pclass      -0.19561    0.01835 -10.660  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.456 on 887 degrees of freedom
## Multiple R-squared:  0.1246, Adjusted R-squared:  0.1216
## F-statistic: 42.08 on 3 and 887 DF,  p-value: < 2.2e-16
```

Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeficiente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeficiente presente.

```
# Tabla con los coeficientes de determinación de cada modelo
tabla.coeficientes <- matrix(c(1, summary(Model.1.1)$r.squared,
2, summary(Model.1.2)$r.squared,
3, summary(Model.1.3)$r.squared),
ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

```
##      Modelo       R^2
## [1,]      1 0.3676802
## [2,]      2 0.2959624
## [3,]      3 0.1245982
```

En este caso, tenemos que el primer modelo es el más conveniente dado que tiene un mayor coeficiente de determinación. Ahora, empleando este modelo, podemos proceder a realizar de superviviencia como la siguiente:

```
# Predecir la capacida de supervivencia del dataset de pruebas agregando una
nueva columna, mientras mas se acerquen los valores predecido a 1 mayor
posibilidad de sobrevivir.
test$Survived <- predict(Model.1.1, test)

#Predecir si un pasajero hombre de 28 anos que compra un ticket de clase
media podria sobrevivir
newdata1 <- data.frame(
  Pclass = 2,
  Sex = "male",
  Age = 28
)

predict(Model.1.1, newdata1)
```

```
##         1
## 0.2505332
```

```
#Predecir si un pasajero mujer de 30 anos que compra un ticket de clase alta
podria sobrevivir
newdata2 <- data.frame(
  Pclass = 1,
  Sex = "female",
```

```
  Age = 30
)

predict(Model.1.1, newdata2)

##        1
## 0.925237

#Predecir si un pasajero mujer de 10 anos que compra un ticket de clase alta
podria sobrevivir
newdata3 <- data.frame(
  Pclass = 1,
  Sex = "female",
  Age = 10
)

predict(Model.1.1, newdata3)

##        1
## 0.925237
```

Conclusiones

Durante el análisis de los datos, hemos podido observar que quienes tienen mayor probabilidad de sobrevivir son las mujeres y niños de clase alta. En el estudio de regresión lineal las variables de mayor influencia son el sexo, la clase de ticket adquirido y en menor grado si el pasajero tiene hijos a bordo, por tal motivo se pudo determinar que influye mucho las clase social, es decir el tipo de ticket que se haya adquirido en este caso de primera clase son quienes tuvieron mayor posibilidad de sobrevivir entre estos mujeres y niños. En los modelos también se pudo observar que los pasajeros de sexo masculino tuvieron menos posibilidad de sobrevivir con un coeficiente negativo y aquellos con hijos a bordo tuvieron una mayor posibilidad de sobrevivir.

Contribuciones

Investigación Previa: DS / CS

****

Redacción de las respuestas: DS / CS

****

Desarrollo código: DS / CS