TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS
PRUEBA DE EVALUACIÓN CONTINUA - PRÁCTICA 2
WILSON RODRIGO PÉREZ ROCANO

Contents

1 Presentación

En esta práctica se elabora un caso práctico 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. Para hacer esta práctica tendréis que trabajar en grupos de hasta 3 personas, o si preferís, también podéis hacerlo de manera individual. Tendréis que entregar un solo archivo con el enlace Github donde haya las soluciones incluyendo los nombres de los componentes del equipo.

1.1 Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

1. 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.
2. Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

1.2 Objetivos

Los objetivos concretos de esta práctica son:

1. 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.
2. Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
3. Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
4. Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
5. Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
6. Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
7. Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

2 Solución

Titanic dataset, es el conjunto de datos seleccionado para la práctica, disponible en la dirección web https://www.kaggle.com/c/titanic/data. Conformado por dos dataset, el primero train con 891 registros con información de entrenamiento para los algoritmos de minería de datos y/o machine rearning, y el segundo test dataset, que posee registros para la validación de los algoritmos creados, con 419 registros.

2.1 Descripción del dataset

El dataset de análisis describe la información de los pasajeros a bordo del Titanic, navío que sufrió su hundimiento el 14 de Abril de 1992 en el Océano Atlántico, donde fallecieron 1513 personas por ahogamiento o hipotermia. El conjunto de datos está conformado por los siguientes campos:

- PassengerId.- Identificación del pasajero
- Survived.- Contiene la información de si el pasajero del Titanic sobrevive o fallece en el naufragio (0 - No y 1 - Si).
- Pclass.- Define la clase del boleto que posee una determinada persona (1 - primera clase, 2 - segunda clase, 3 - tercera clase).
- Name.- El nombre del pasajero
- Sex.- Almacena la información relacionada al sexo del pasajero (Hombre, Mujer)
- Age.- describe la edad del pasajero
- SibSp.- Número de hermanos + el cónyuge en el caso de que tenga.
- Parch.- Número de padres de familia del pasajero + el número de hijos en el caso de que posea.
- Ticket.- El número del boleto con el que viajó el pasajero en el titanic.
- Fare.- Tarifa del pasaje.
- Embarked.- Puerto de embarcación del pasajero (C - Cherbourg, Q - Queenstown, S - Southampton).
- Cabin.- Identificador de la cabina en el que viajo el pasajero.

2.1.1 Objetivos de análisis

Nuestro objetivo de estudio es saber qué características poseen los pasajero que sobrevive y fallecen en el naufragio, por lo que se plantea la problemática de determinar qué variables influyen más para que un pasajero sobreviva.

Para dar solución a nuestra problemática, se plantea la creación de reglas que describa bajo qué características un pasajero sobrevive o fallece, obtenidas mediante árboles de decisión y reglas de agrupación. Del mismo modo, se relacionará las características similares de los pasajeros, mediante algoritmos de clusterización.

2.1.2 Importancia de análisis

Muchas hipótesis acerca de los sobrevivientes en el naufragio del Titanic se ha dicho que: i) las personas de primera clase, sobre todo mujeres, fueron las personas que más privilegios tuvieron al momento de abordar un bote salvavidas; ii) que las personas de la tercera clase, fueron los menos considerados y ellos representan el mayor porcentaje de fallecidos en el naufragio considerando la cantidad de personas a bordo. En este análisis se enfoca en establecer que características poseen las personas que sobrevivieron y fallecieron en el naufragio, analizando la edad de los pasajeros, la clase social y el lugar donde abordaron el Titanic y así apoyar las hipótesis que a simple vista son planteadas.

2.2 Integración y selección de los datos de interés a analizar

Nuestro proceso inicia con la lectura del fichero CVS de entrenamiento, paso seguido se analiza visualiza las características y formatos que posee cada variable del dataset. Haciendo uso de los comandos srt() y summary () se visualiza la información.

```
#Librerias a usar
library (rpart.plot);

## Warning:  package 'rpart.plot' was built under R version 3.4.4
## Loading required package:  rpart
## Warning:  package 'rpart' was built under R version 3.4.4

library (rpart);
library(car);

## Warning:  package 'car' was built under R version 3.4.4
## Loading required package:  carData
## Warning:  package 'carData' was built under R version 3.4.4

library('cluster');
library('fpc');
```

```
## Warning:  package 'fpc' was built under R version 3.4.4

library('NbClust');
library('ggplot2');

## Warning:  package 'ggplot2' was built under R version 3.4.4

library('factoextra');

## Warning:  package 'factoextra' was built under R version 3.4.4
## Welcome!  Related Books:  'Practical Guide To Cluster Analysis in
R' at https://goo.gl/13EFCZ

library('stringr');

## Warning:  package 'stringr' was built under R version 3.4.4

library('discretization');

# Lectura de datos desde el archivo
df <- read.csv("C:/Users/Usuario-03/train.csv");

# ====CARACTERISTICAS DE LOS DATOS ===
# Dimensión del dataset
dim (df);

## [1] 891  12

# Tipo de dato asignado a cada campo
sapply(df[,1:4], function(x) class(x))

## PassengerId    Survived      Pclass        Name
##   "integer"   "integer"   "integer"    "factor"

sapply(df[,5:8], function(x) class(x))

##       Sex       Age     SibSp     Parch
##  "factor" "numeric" "integer" "integer"

sapply(df[,9:12], function(x) class(x))

##    Ticket      Fare     Cabin  Embarked
##  "factor" "numeric"  "factor"  "factor"

# Resumen de datos
summary(df[,1:4]);
```

```
##   PassengerId       Survived          Pclass
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000
##  Median :446.0   Median :0.0000   Median :3.000
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##                                     Name
##  Abbing, Mr. Anthony                   :  1
##  Abbott, Mr. Rossmore Edward           :  1
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1
##  Abelson, Mr. Samuel                   :  1
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1
##  Adahl, Mr. Mauritz Nils Martin        :  1
##  (Other)                               :885
```

```
summary(df[,5:12]);
```

```
##      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
##       Ticket         Fare                Cabin     Embarked
##  1601    :  7   Min.   :  0.00              :687   C   :168
##  347082  :  7   1st Qu.:  7.91   B96 B98    :  4   Q   : 77
##  CA. 2343:  7   Median : 14.45   C23 C25 C27:  4   S   :644
##  3101295 :  6   Mean   : 32.20   G6         :  4   NA's:  2
##  347088  :  6   3rd Qu.: 31.00   C22 C26    :  3
##  CA 2144 :  6   Max.   :512.33   D          :  3
##  (Other) :852                    (Other)    :186
```

Para el estudio no se consideran las variables sibSp, Parch, Ticket, Fare, Cabin, Name y PassengerId, debido a que no aportan significativamente al análisis.

```
# Eliminación de campos
df <- df[, -(7:11)];  # campos: SibSp,Parch,Ticket,Fare,Cabin
df <- df[, -(4)];     # campo Name
df <- df[, -(1)];     # campo PassengerId
dfaux=df;
```

2.3 Limpieza de los datos

En este paso se procede a realizar la limpieza de la información, donde se analizará los datos nulos y los datos outliers.

2.3.1 Tratamiento de ceros y elementos vacíos

Los valores vacíos hacen valores de variables que no poseen información, debido a que no se ha registrado o no se ha definido. El dataset de análisis posee valores nulos solo para los datos de la variable Age y Embarked.

```
# Datos Nulos
sapply(df, function(x) sum(is.na(x)));

## Survived   Pclass      Sex      Age Embarked
##        0        0        0      177        2

df$Age[is.na(df$Age)];

##   [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [24] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [47] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [70] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [93] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [116] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [139] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [162] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA

df$Embarked[is.na(df$Embarked)];

## [1] <NA> <NA>
## Levels: C Q S
```

Ahora, es necesario decidir qué proceso realizar con los mismos, Ignorar tuplas, definir un valor manualmente o usar una medida de tendencia central. En nuestro caso se usa la función KNN, donde se agrupa los elementos y se mide las distancias con el objetivo de clasificar los datos y buscar los vecinos más próximos. De este modo, los valores nulos son sustituidos por el valor más cercano, sin despreciar el registro con el valor vacío.

```
# Inputación de valore nulos
suppressWarnings(suppressMessages(library(VIM)));
df$Age<-kNN(df)$Age;
df$Embarked<-kNN(df)$Embarked;
```

2.3.2 Identificación y tratamiento de valores extremos

Los extreme scores son observaciones extremas, distante al resto de los datos, que no coinciden con el resto de observaciones, valores extrañamente grandes o pequeños. Estas observaciones pueden influir considerablemente en el análisis y afectar a la capacidad de previsión del modelo. Haciendo uso de la función boxplots.stats() se obtiene los valores extremos para cada variable. En nuestro dataset los únicos valores extremos son para la variable Age.

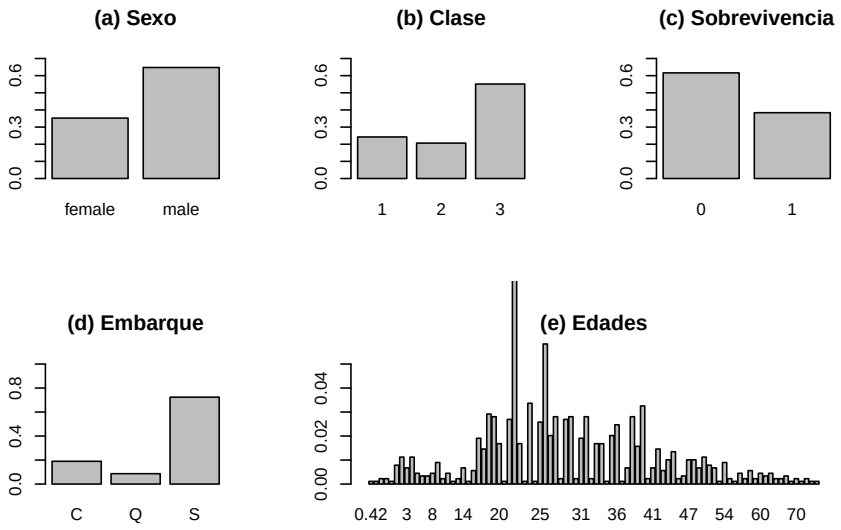
```
# Valores extremos
boxplot.stats(df$Survived)$out;

## integer(0)

boxplot.stats(df$Pclass)$out;

## integer(0)

boxplot.stats(df$Sex)$out;

## Warning in Ops.factor(x[floor(d)], x[ceiling(d)]):  '+' not meaningful
for factors

## factor(0)
## Levels: female male

boxplot.stats(df$Age)$out;

##  [1] 66.0 65.0 71.0 70.5 61.0 62.0 63.0 65.0 61.0 60.0 64.0 65.0 63.0 71.0
## [15] 64.0 62.0 62.0 60.0 61.0 80.0 70.0 60.0 60.0 70.0 62.0 74.0

boxplot.stats(df$Embarked)$out;

## Warning in Ops.factor(x[floor(d)], x[ceiling(d)]):  '+' not meaningful
for factors

## factor(0)
## Levels: C Q S
```

Dado que los valores extremos identificados para la variable Age son valores mayores a 60 años, no son considerados datos erróneos, debido a que los pasajeros del Titanic si pudieron tener edades entre 60 y 80 años. Bajo este pequeño análisis, no se procesa los datos extremos y se trabajaba tal y como fueron recogidos, sin cambio alguno.

2.4 Análisis de los datos

En este paso se analizará el dataset con el objetivo de analizar el comportamiento y las características que posee. Se empieza con graficar las frecuencias de cada variable de análisis.

```
# Gráfica de las Frecuencias de cada una de las variables
nf<-layout(matrix(c(1,2,3,4,5,5), 2, 3, byrow=TRUE),respect=TRUE);
barplot(prop.table(table(df$Sex)),ylim=c(0,0.7),
        main="(a) Sexo");
barplot(prop.table(table(df$Pclass)),ylim=c(0,0.7),
        main="(b) Clase");
barplot(prop.table(table(df$Survived)),ylim=c(0,0.7),
        main="(c) Sobrevivencia");
barplot(prop.table(table(df$Embarked)),ylim=c(0,1),
        main="(d) Embarque");
barplot(prop.table(table(df$Age)),ylim=c(0,0.05),
        main="(e) Edades");
```



2.4.1 Comprobación de la normalidad y homogeneidad de la varianza

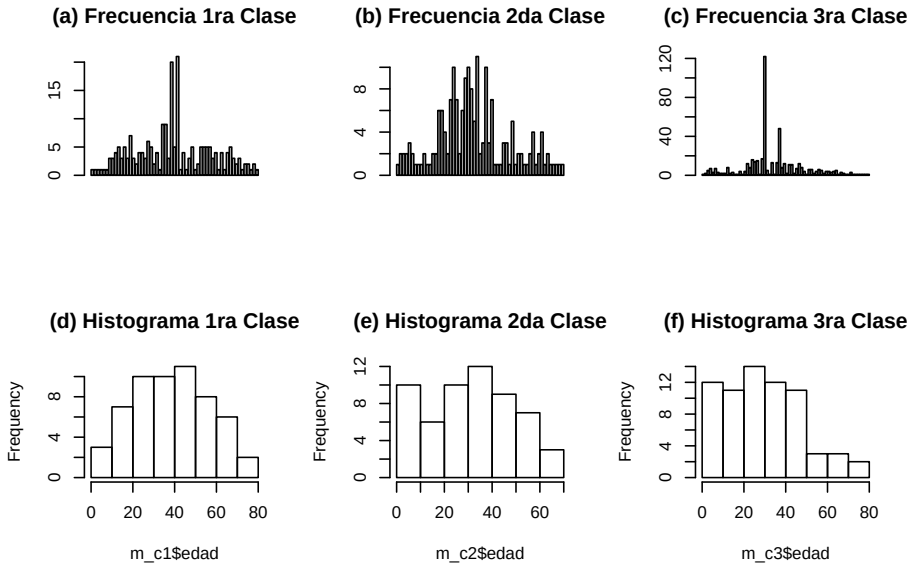
Luego de tener los datos agrupados y haciendo uso de la función fligner.test(), que hace uso del test de Fligner-Killeen que compara la homogeneidad de las varianzas del conjunto. El valor obtenido es de 2.2e-16 por lo que se concluye que las varianzas de las variables son totalmente distintas.

```
# Comparación de varianzas
var <- fligner.test(df);
var;

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df
## Fligner-Killeen:med chi-squared = 2545.8, df = 4, p-value <
## 2.2e-16
```

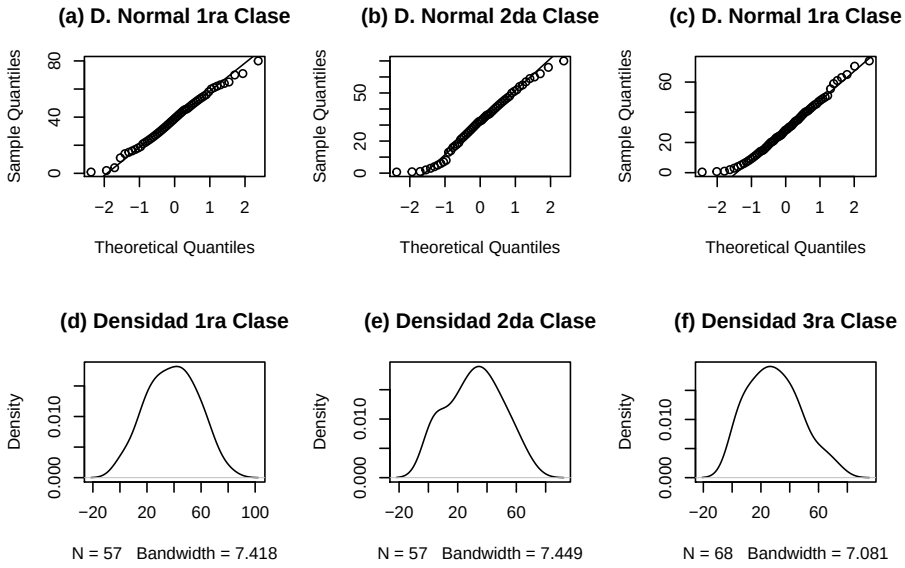
Para el análisis de la normalidad se toma dos variables que influyen directamente en que si un pasajero sobrevive o no, la edad y la clase. Se inicia con la agrupación de los datos del dataset de la edad y el sexo dada la clase del pasajero.

```
# Se estudia el dataset
total_clase=aggregate(df$Pclass, by=list(clase=df$Pclass),
                      FUN=function(x){NROW(x)});
total_sexo=aggregate(df$Pclass, by=list(sex=df$Sex,
  clase=df$Pclass),FUN=function(x){NROW(x)});
total_edad=aggregate(df$Pclass, by=list(edad=df$Age,
  clase=df$Pclass), FUN=function(x){NROW(x)});
m_Hombres<-subset(total_sexo, sex=='male');
m_Mujeres<-subset(total_sexo, sex=='female');

m_Hombres$porcentaje <- round(prop.table(m_Hombres$x),4)*100;
m_Mujeres$porcentaje <- round(prop.table(m_Mujeres$x),4)*100;

m_c1<-subset(total_edad, clase=='1');
m_c2<-subset(total_edad, clase=='2');
m_c3<-subset(total_edad, clase=='3');
```

```
nf<-layout(matrix(c(1,2,3,4,5,6), 2, 3, byrow=TRUE),respect=TRUE);
barplot(m_c1$x, main="(a) Frecuencia 1ra Clase");
barplot(m_c2$x, main="(b) Frecuencia 2da Clase");
barplot(m_c3$x, main="(c) Frecuencia 3ra Clase");

hist(m_c1$edad, main="(d) Histograma 1ra Clase");
hist(m_c2$edad, main="(e) Histograma 2da Clase");
hist(m_c3$edad, main="(f) Histograma 3ra Clase");
```

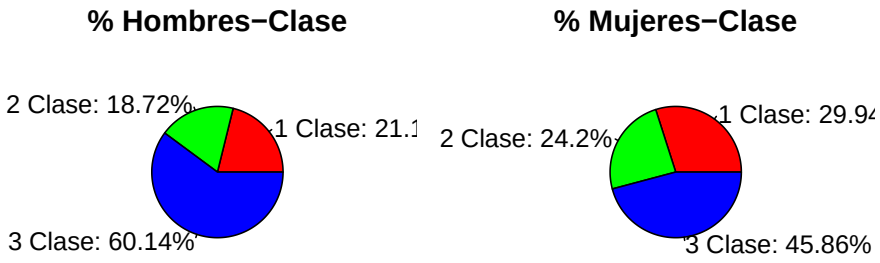



```
nf<-layout(matrix(c(1,2,3,4,5,6), 2, 3, byrow=TRUE),respect=TRUE);
qqnorm(m_c1$edad, main="(a) D. Normal 1ra Clase");
qqline(m_c1$edad);
qqnorm(m_c2$edad, main="(b) D. Normal 2da Clase");
qqline(m_c2$edad);
qqnorm(m_c3$edad, main="(c) D. Normal 1ra Clase");
qqline(m_c3$edad);

plot(density(m_c1$edad), main="(d) Densidad 1ra Clase");
plot(density(m_c2$edad), main="(e) Densidad 2da Clase");
plot(density(m_c3$edad), main="(f) Densidad 3ra Clase");
```



Finalmente se presenta gráficamente los porcentajes de Hombres y Mujeres por la clase a la que pertenecen, concluyendo que en el Titanic, del total de los Hombres, el 60% eran de la 3ra clase, seguido por los de la primera clase con el 21%; del mismo modo para las Mujeres, el 45% eran de tercera clase y el 24% de segunda clase.

```
lblsH <- paste(m_Hombres$clase, " Clase: ",m_Hombres$
                  porcentaje,"%",sep="");
lblsM <- paste(m_Mujeres$clase, " Clase: ",m_Mujeres$
                  porcentaje,"%",sep="");

#Se dibuja los cada uno de las subconjuntos obtenidos
nf<-layout(matrix(c(1,2), 1, 2, byrow=TRUE),respect=TRUE);
pie(m_Hombres$x,labels = lblsH, col=rainbow(length(lblsH)),
    main="% Hombres-Clase");
pie(m_Mujeres$x,labels = lblsM, col=rainbow(length(lblsM)),
    main="% Mujeres-Clase");
```



2.5 Aplicación de pruebas estadísticas para comparar los grupos de datos

Nuestro objetivo de estudio es saber qué características poseen los pasajero que sobrevive y fallecen en el naufragio, por lo que se plantea la problemática de determinar qué variables influyen más para que un pasajero sobreviva. Para lo cual se plantea implementar un árbol de decisión y clustering.

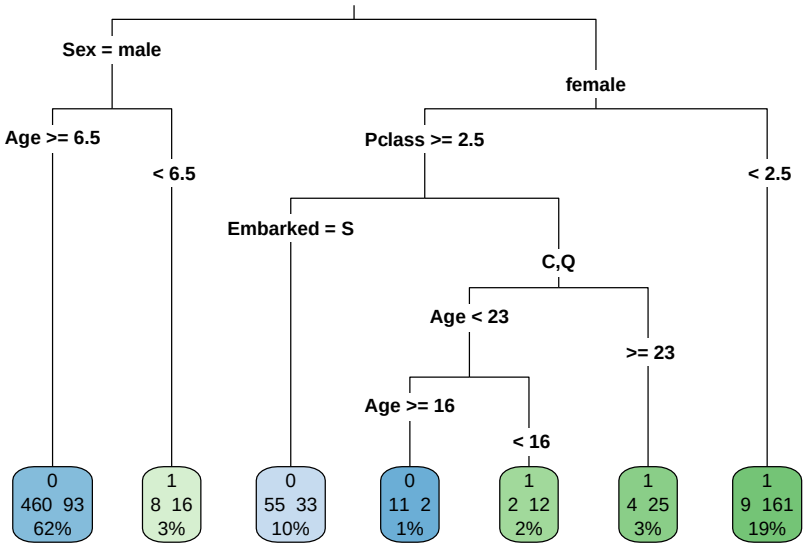
2.5.1 Árbol de decisión y representación de los resultados

Los árboles de decisión o clasificación son modelos de predicción, donde dado un conjunto de datos se fabrican diagramas lógicos, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.
El proceso inicia con la transformación de la variable Survived a tipo factor, la cual es el campo categórico para la generación del árbol.

```
#==CREACION DEL ARBOL==
nf<-layout(matrix(c(1), 1, byrow=TRUE),respect=TRUE);
datos2p=as.data.frame(df);
datosp=datos2p[,2:5];
# El campo SURVIVED se convierte a tipo factor
datosp$Survived=as.factor(datos2p$Survived);
# Creación el árbol
arbol<- rpart(Survived ~., data=datosp, method="class");
```

Una vez generado el algoritmo se procede a graficar el mismo haciendo uso de la función rpart.plot(). El árbol resultante está formado por 5 niveles, con un total de 7 reglas resultantes, las que se presentan en las secciones finales del documento.

```
# Gráfico del árbol
rpart.plot(arbol, type=3, extra=101, fallen.leaves=T);
```




```
# Resumen del árbol
arbol;

## n= 891
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 891 342 0 (0.61616162 0.38383838)
##    2) Sex=male 577 109 0 (0.81109185 0.18890815)
##      4) Age>=6.5 553  93 0 (0.83182640 0.16817360) *
##      5) Age< 6.5 24   8 1 (0.33333333 0.66666667) *
##    3) Sex=female 314  81 1 (0.25796178 0.74203822)
##      6) Pclass>=2.5 144  72 0 (0.50000000 0.50000000)
##       12) Embarked=S 88  33 0 (0.62500000 0.37500000) *
##       13) Embarked=C,Q 56  17 1 (0.30357143 0.69642857)
##         26) Age< 23 27  13 1 (0.48148148 0.51851852)
##           52) Age>=16.5 13   2 0 (0.84615385 0.15384615) *
```

```
##           53) Age< 16.5 14   2 1 (0.14285714 0.85714286) *
##         27) Age>=23 29   4 1 (0.13793103 0.86206897) *
##      7) Pclass< 2.5 170   9 1 (0.05294118 0.94705882) *
```

2.5.2 Clustering y representación de los resultados

Un algoritmo de agrupamiento es un procedimiento de agrupar una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia.
El proceso inicia con la eliminación de la variable Survived que para la agrupación no aporta significativamente. Paso seguido la variable Sex se convierte a una variable categórica, donde:

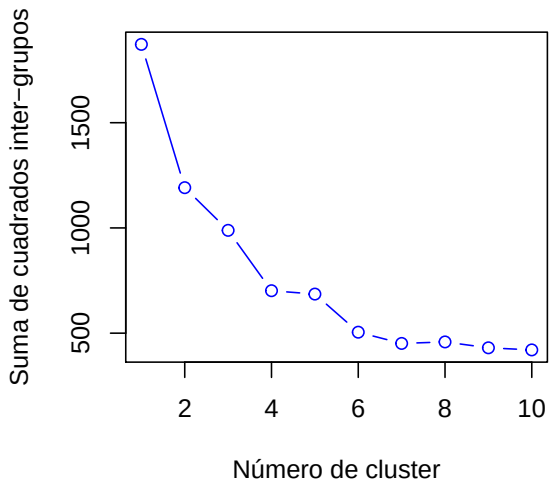
1. Las edades de 0.40 a 12 se les denomina niño representado por el valor de 1.
2. Las edades de 12 a 18 se les denomina adolecentes representado por el valor de 2.
3. Las edades de 18 a 30 se les denomina joven representado por el valor de 3.
4. Las edades de 30 a 50 se les denomina adulto representado por el valor de 4.
5. Las edades de 50 a 81 se les denomina mayor representado por el valor de 5.

```
dfc=df;
dfc <- dfc[, -5];
dfc$Sex=str_replace_all(dfc$Sex,"female","1");
dfc$Sex=str_replace_all(dfc$Sex,"male","2");
dfc$Sex=as.integer(dfc$Sex);
#niño=1, adolecente=2, joven=3, adulto=4, mayor=5
dfc$Age=cut(dfc$Age, breaks = c(0.40,12,18,30,50,81),
    labels = c(1, 2, 3, 4, 5))
dfc$Age=as.integer(dfc$Age);
summary(dfc);
```

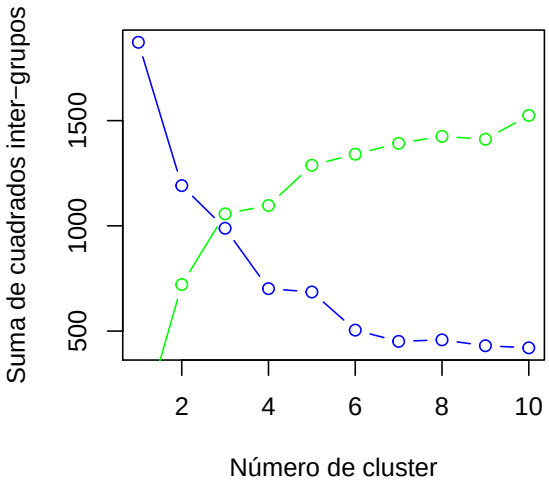
```
##     Survived          Pclass           Sex             Age
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000
##  Median :0.0000   Median :3.000   Median :2.000   Median :3.000
##  Mean   :0.3838   Mean   :2.309   Mean   :1.648   Mean   :3.222
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :3.000   Max.   :2.000   Max.   :5.000
```

Un paso importante en los algoritmos de cluster es definir el número de grupos que poseerá nuestro resultado, para la cual se analiza la distancia inter-grupos y la distancia entre-grupos. Con esta información se genera un vector de distancias, con el objetivo encontrar el número correcto de grupos donde la distancia inter-grupos sea la mínima y la distancia entre-cluster sea la máxima, donde a un mayor número de grupos, mejor son los resultados obtenidos.

```
#Se selecciona el número adecuado de grupos
vec_distancias=kmeans(dfc,centers=1)$betweenss;
vec_distancias2=kmeans(dfc,centers=1)$tot.withinss;
for (i in 1:10) {
  vec_distancias[i]<-kmeans(dfc,centers=i)$betweenss;
  vec_distancias2[i]<-kmeans(dfc,centers=i)$tot.withinss;
}
nf<-layout(matrix(c(1), 1, byrow=TRUE),respect=TRUE);
plot (vec_distancias2, type='b',xlab='Número de cluster',
      ylab='Suma de cuadrados inter-grupos',col = 'blue');
```




```
lines(vec_distancias, type='b',col = 'green');
```



Para nuestro análisis se considera cuatro grupos, con lo que se procede a crear el cluster, haciendo uso del algoritmo Kmeans.

```
#Se crea el cluster
cl <- kmeans(dfc, 4);
#Se presenta las caracteristicas del cluster
cl$centers;

##    Survived   Pclass      Sex      Age
## 1 0.1045576 2.793566 2.000000 3.340483
## 2 0.5035971 2.561151 1.510791 1.503597
## 3 0.6260163 1.186992 1.540650 4.036585
## 4 0.5939850 2.759398 1.000000 3.180451
```

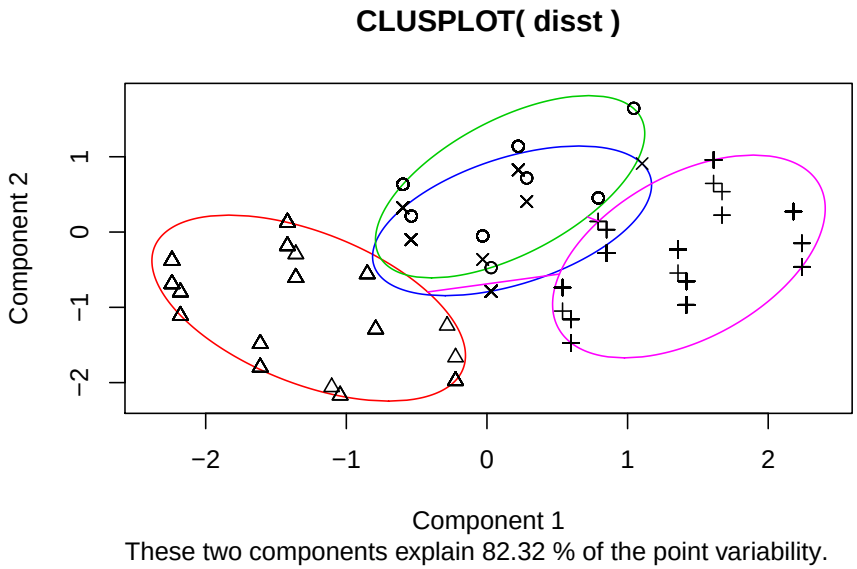
Una vez aplicado el algoritmo de clustering se debe evaluar la calidad del modelo y evitar conclusiones erróneas de agrupación que no se corresponden con la realidad. Para lo cual se analiza el promedio de la distancia interna de los clusters y la distancia externa de los clusters.

```
cl$iter #Número de iteraciones para generar los grupos

## [1] 2

cl$size #Tamaño de los grupos

## [1] 373 139 246 133

disst <- daisy(dfc); #distancia entre observaciones
```

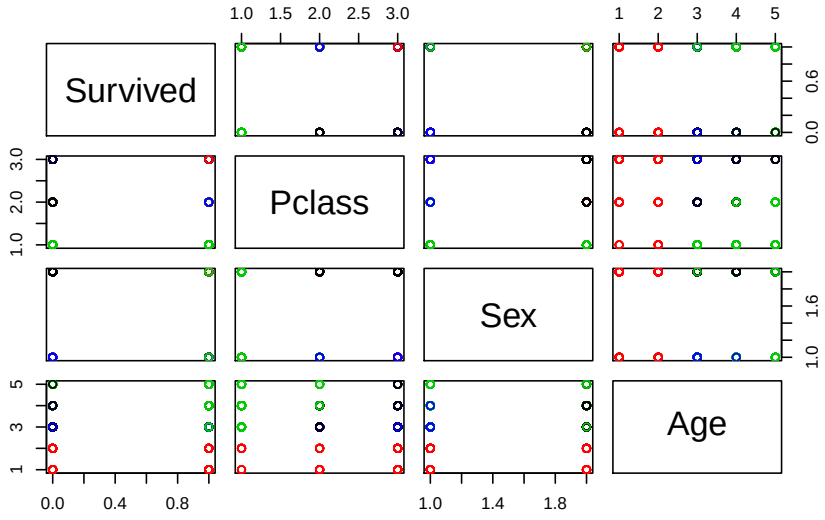
```
## Warning in daisy(dfc):  binary variable(s) 1, 3 treated as interval
scaled

#Distancia inter-grupos
cluster.stats(disst,cl$cluster)$average.between;

## [1] 2.207948

#Distancia entre-grupos
cluster.stats(disst,cl$cluster)$average.within;

## [1] 1.012923
```

Haciendo uso de la función clusplot y de la distancia entre las observaciones, se genera el cluster, donde se observa que la correspondencia de cada observación a cada grupo.

```
#Se grafica el cluster generado
clusplot(disst, cl$cluster,diss = TRUE,color=TRUE,col.p="black");
```




```
#Se grafica el cluster con la combinacion de sus variables
plot(dfc, col = cl$cluster);
```



2.6 Resolución del problema. A partir de los resultados obtenidos

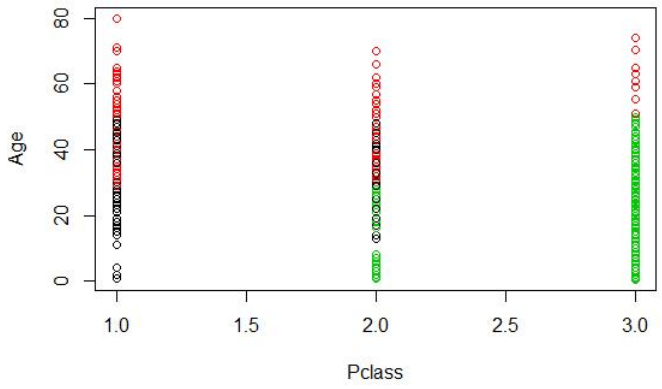
En base a los resultados obtenidos por parte de los algoritmos en la sección 2.5.1.1 y la sección 2.5.2.1, podemos concluir que:
1.- El árbol de decisión resultante está formado por 5 niveles, con un total de 7 reglas resultantes, las que se presentan a continuación.

1. Si (Sex='Male') (Age>=6.5) entonces muere.
2. Si (Sex='Male') (Age<6.5) entonces sobrevive.
3. Si (Sex='Female') (Pclass>=3) (Embarked='S') entonces muere.
4. Si (Sex='Female') (Pclass>=3) (Embarked='C' or Embarked='Q') (16<=Age<23) entonces muere.
5. Si (Sex='Female') (Pclass>=3) (Embarked='C' or Embarked='Q') (Age<16) entonces sobrevive.
6. Si (Sex='Female') (Pclass>=3) (Embarked='C' or Embarked='Q') (Age>=23) entonces sobrevive.
7. Si (Sex='Female') (Pclass<3) entonces sobrevive.

De lo que podemos concluir que ...
Al obtener el modelo usando Kmeans, se obtuvo un promedio de 2.25 para la distancia entre-grupos y el valor de 1.02 para la distancia inter-grupos. Luego del análisis se puede concluir lo siguiente:

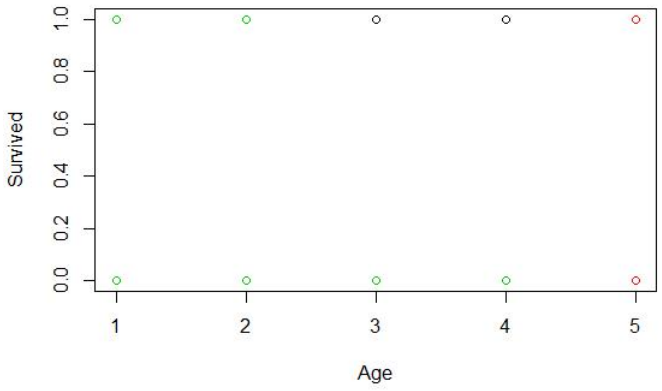
1. La mayor parte de las personas de la 2da y 3ra clase, y que poseen edades menores poseen características similares y forman el cluster 2.

```
#Se grafica el cluster de la Edad VS la clase del pasajero
plot(dfc$Pclass, df$Age,col=cl$cluster,xlab='Pclass',ylab='Age');
```



2. Las personas mayores representadas por la categoría 5 poseen características similares y conforman el cluster 3.

```
#Se grafica el cluster de la Edad VS si sobrevive o no
plot(dfc$Age, df$Survived,col=cl$cluster,xlab='Age',ylab='Survived');
```



En base a los resultados obtenidos se puede observar que el 73De esta forma se concluye que la variable que más influye en los sobrevivientes es la edad y el sexo, seguidos por la clase social, confirmando la hipótesis de "Mujeres y niños primero". Del mismo modo, la mayoría de sobrevivientes abordaron el Titanic en Cherbourg y Queenstown. Fuese interesante analizar el porcentaje de sobrevivientes varones, cuántos de ellos eran miembros de la tripulación y así tener

conclusiones más certeras.

Finalmente el resultado de los datos procesados se almacenan el dataset Titanic_data_clean.csv.

```
#Se almacena el archivo de salida
write.csv(dfc, "Titanic_data_clean.csv");
```

3 Recursos Bibliográficos

- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science Business Media.