M2.851 - Tipología y ciclo de vida de los datos
PRAC 2
Diego Cheuquepán Maldonado

https://github.com/cheukepan/proyectoTitanic

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos se ha obtenido a partir de este enlace en Kaggle:

https://www.kaggle.com/c/titanic/data

Está constituido por 12 atributos (columnas) que presentan 891 pasajeros (registros).

Los campos del conjunto de datos los describí en términos del tipo de variable, de escala y de dato según se muestra a continuación:

Variable	Nombre	Descripción	Tipo de Variable	Tipo de Escala	Tipo de Dato
Atributo 1	PassengerId	Identificador del pasajero	Cuantitativa	Discreta	Entero
Atributo 2	Survived	Sobrevive	Cuantitativa	Discreta	Entero
Atributo 3	Pclass	Clase de pasajero	Cualitativa	Ordinal	Entero
Atributo 4	Name	Nombre del pasajero	Cualitativa	Nominal	Cadena de texto
Atributo 5	Sex	Sexo del pasajero	Cualitativa	Nominal	Cadena de texto
Atributo 6	Age	Edad del pasajero	Cuantitativa	Discreta	Número
Atributo 7	SibSp	Número de familiares	Cuantitativa	Discreta	Entero
Atributo 8	Parch	Número de familiares	Cuantitativa	Discreta	Entero
Atributo 9	Ticket	Número de ticket	Cualitativa	Nominal	Cadena de texto
Atributo 10	Fare	Tarifa del ticket	Cuantitativa	Continua	Número
Atributo 11	Cabin	Número de cabina del pasajero	Cualitativa	Nominal	Cadena de texto
Atributo 12	Embarked	Puerto de embarque	Cualitativa	Nominal	Cadena de texto

Respecto a la pregunta que se busca responder, la competencia de Kaggle indica lo siguiente:

> "Usar el aprendizaje automático para crear un modelo que prediga qué pasajeros sobrevivieron al naufragio del Titanic"

Por lo anterior, podemos afirmar que estamos frente a un problema de analítica predictiva donde se deben usar técnicas de aprendizaje automático que nos permitan predecir una variable objetivo, dado un conjunto de variables predictivas del set de datos.

Por lo anterior, el análisis realizado tiene puesto el foco en los siguientes aspectos:

- Determinar cuáles son las variables predictoras adecuadas.
- Definir un modelo predictivo, basado en un set de entrenamiento, que nos permita clasificar a los pasajeros en sobrevivientes y no sobrevivientes.
- Establecer, dado un set de prueba, cuál es la capacidad predictiva del modelo.

2. Integración y selección de los datos de interés a analizar.

En la pregunta anterior, presentamos la característica de los datos de acuerdo con su tipo. Esto nos permite definir una serie de técnicas a usar de acuerdo con las sugerencias propuestas por Reguant-Álvarez, Vilà-Baños y Torrado-Fonseca (2018).

De esta manera, el análisis de variables realizado seguirá la siguiente propuesta:

Tipo de escala	Procedimiento	Gráfico
2 nominales	Chi cuadrado	Mosaico
2 ordinales	Correlación de Spearman	Barra
1 ordinal y 1 escalar	Correlación de Spearman	De cajas y bigotes
2 escalares	Correlación de Pearson	Dispersión

Por otro lado, el análisis de las variables nos llevó a decidir que era necesario construir un nuevo atributo, que denominamos familia y que contiene a los datos *SibSp* y *Parch* que informan el número de hermanos/cónyuges y de padres/hijos, respectivamente.

Con este procedimiento hemos creado una nueva variable cuantitativa que indica el tamaño de la familia.

3. **Limpieza de los datos.**
   3.1. **¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?**

El análisis de elementos nulos o cero se realizó para cada una de las variables y se encontraron en tres:

- Age
- Cabin
- Embarked

Dado que estas variables son, cualitativas o cuantitativas, la estrategia de imputación de datos debe considerar este antecedente.

En el caso de la variable cuantitativa Age, se analizaron tres estrategias de imputación: basada en media, basada en mediana y basada en vecinos próximos.

La última estrategia, KNN, es más robusta que las dos primeras dado que hace uso de la técnica de vecinos próximos para establecer un valor. Sin embargo, fue descartada ya que no existían otras variables cuantitativas que tuvieran sentido para ser usadas como puntos sobre los que construir los vecinos cercanos.

Luego, se analizó la estrategia basada en media y mediana. Dado que la media se ve afectada por valores extremos y la mediana no lo es, se decidió por esta última, de manera que el tratamiento de valores nulos para el caso de Age fue basado en mediana.

```
# Imputación por mediana
datos$Age[is.na(datos$Age)] <- median(datos$Age, na.rm = TRUE)

# Verificamos variable Age
colSums(is.na(datos))
```

Luego, se siguió con las variables cualitativas Cabin y Embarked fueron analizadas de manera que Cabin no fue tratada, pues no es una variable seleccionada en el estudio.
En tanto Embarked, se asignaron los dos valores nulos a Cherbourg.
Esta decisión se debe a la idea de valor medio, pues al analizar la distribución de los datos la mayoría de los pasajeros embarcó en Southampton, en tanto, en Queenston subió la menor cantidad de pasajeros.

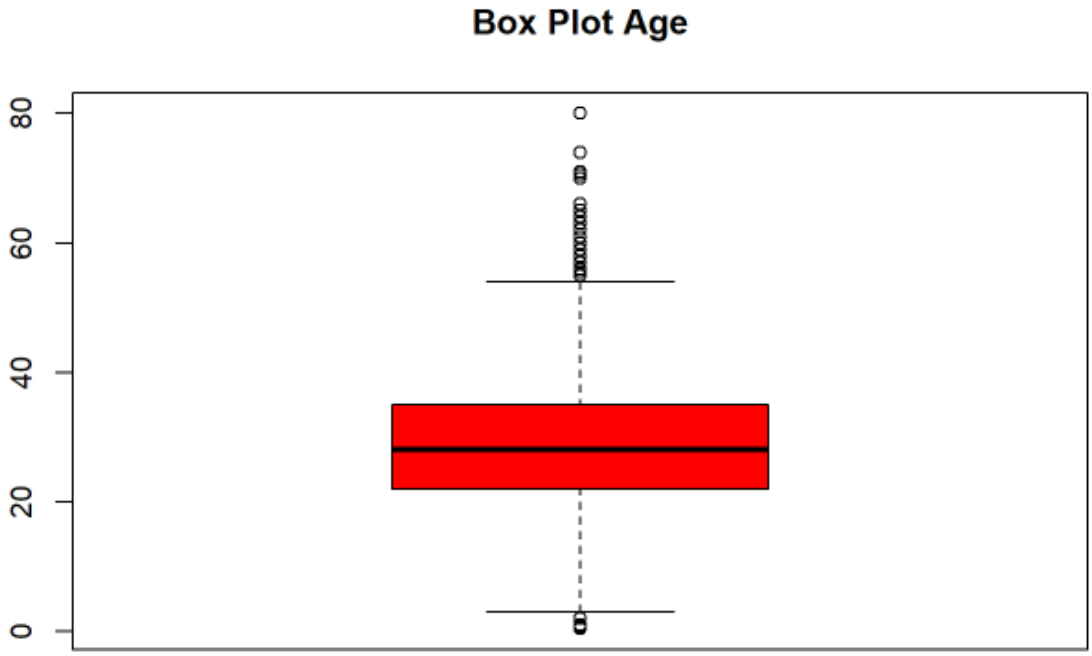
```
# Recodificamos variable Embarked
datos$Embarked[datos$Embarked=="C"]="Cherbourg"
datos$Embarked[datos$Embarked=="Q"]="Queenston"
datos$Embarked[datos$Embarked=="S"]="Southapmtpon"
datos$Embarked[datos$Embarked==""]="Cherbourg"
```

3.2. Identificación y tratamiento de valores extremos.

Luego, se analizaron los datos atípicos haciendo uso de los boxplot y revisando los datos outlier.

Todas las variables cuantitativas analizadas presentaban datos outlier en gran cantidad. Dado de que, en general, no se recomienda eliminar los datos atípicos; se siguió ese criterio y se mantuvieron dentro del set de datos.

A modo de ejemplo se presenta el análisis de la variable Age.




```
boxplot.stats(datos$Age)$out
```

```
##  [1]  2.00 58.00 55.00  2.00 66.00 65.00  0.83 59.00 71.00 70.50  2.00
## [12] 55.50  1.00 61.00  1.00 56.00  1.00 58.00  2.00 59.00 62.00 58.00
## [23] 63.00 65.00  2.00  0.92 61.00  2.00 60.00  1.00  1.00 64.00 65.00
## [34] 56.00  0.75  2.00 63.00 58.00 55.00 71.00  2.00 64.00 62.00 62.00
## [45] 60.00 61.00 57.00 80.00  2.00  0.75 56.00 58.00 70.00 60.00 60.00
## [56] 70.00  0.67 57.00  1.00  0.42  2.00  1.00 62.00  0.83 74.00 56.00
```

4. **Análisis de los datos.**
   4.1. **Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).**

   En el punto 2 de este documento, se presentó la estrategia para analizar los datos, de acuerdo con su tipo y escala de medición.

   Considerando las características de cada variable, las que finalmente se analizaron fueron las siguientes:

   - Fare
   - Age
   - SibSp
   - Parch
   - Familia
   - Pclass
   - Sex
   - Embarked
   - Survived

   4.2. **Comprobación de la normalidad y homogeneidad de la varianza.**

   Para comprobar la normalidad de las variables cuantitativas se aplicó el test shapiro-wilk, dado que el dataset tiene menos de 5.000 registros. En caso contrario, se recomienda aplicar el test de Anderson-Darling.

   A modo de ejemplo se presenta el análisis de la variable Age.

```
# Variable Age
shapiro.test(datos$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Age
## W = 0.9541, p-value = 4.651e-16
```

   El análisis realizado nos permite concluir que nuestras variables cuantitativas no se distribuyen normalmente.

   Para analizar la homogeneidad de la varianza se aplicó el test de Fligner-Killeen. La decisión se debe a que no se cumple la condición de normalidad en las muestras.

```
fligner.test(Age ~ familia, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by familia
## Fligner-Killeen:med chi-squared = 47.344, df = 8, p-value =
## 1.318e-07
```

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

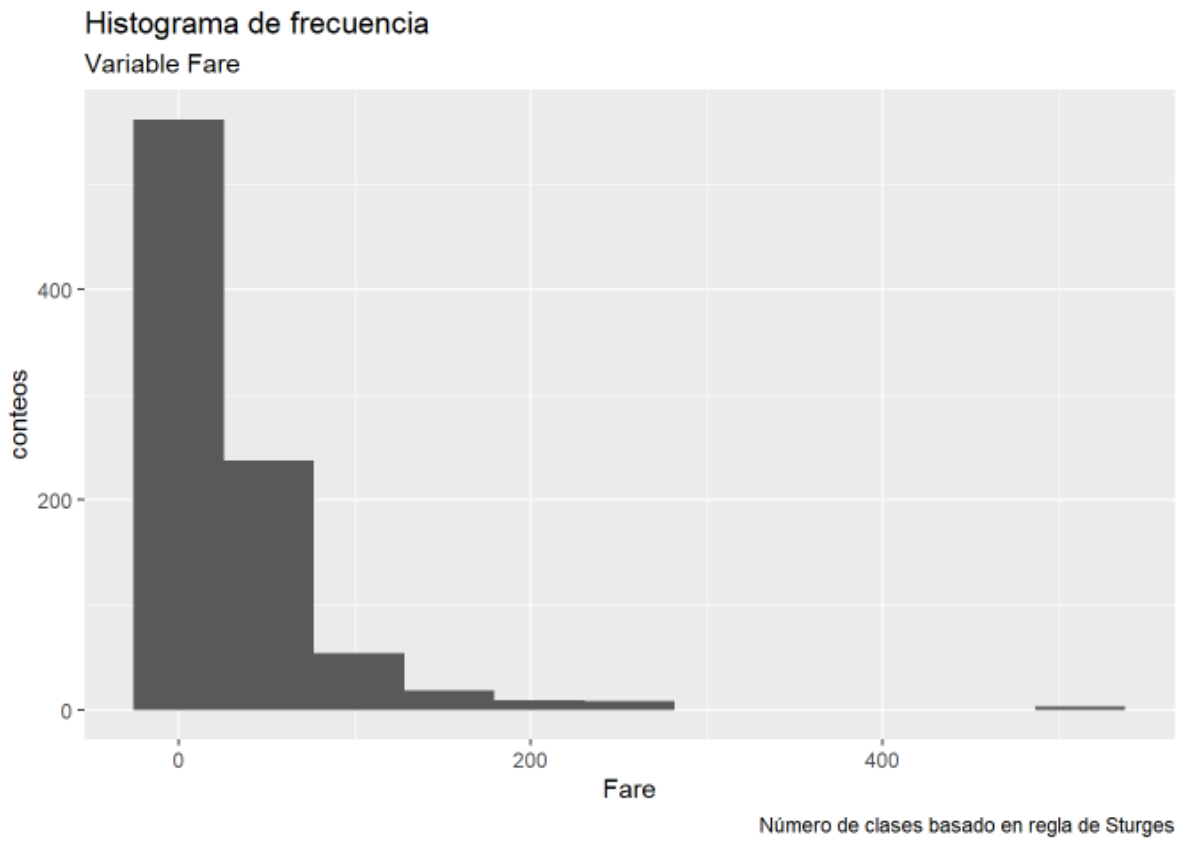
Llegados a este punto, se realizaron análisis de tres tipos, según el tipo de variables:

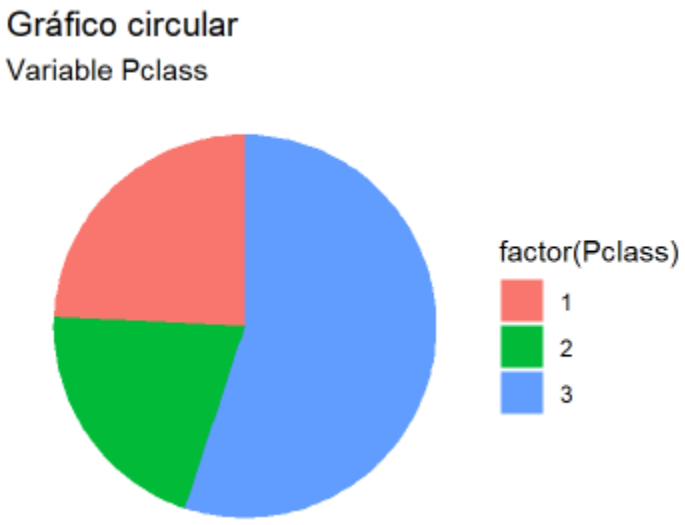
- R Pearson: Para variables continuas.
- Spearman: Para variables ordinales y ordinales/continuas.
- Chi-cuadrado: Para variables nominales.

5. Representación de los resultados a partir de tablas y gráficas.

A lo largo de la prueba se mostraron los resultados mediante gráficos y tablas, considerando siempre el tipo de variable.

Las variables continuas se mostraron mediante gráficos de barra.
Las variables ordinales mediante gráficos circulares.





6. **Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

   Con los análisis preliminares realizados no se puede concluir el problema inicial, sin embargo, en esta prueba procedí a hacer un análisis de reducción de dimensionalidad para determinar el conjunto de variables que podrían usarse en la construcción de un modelo.

   Para ello, procedí con:

   - Análisis de Componentes Principales (PCA).
   - Análisis de importancia de variables mediante Random Forest.

   El análisis de PCA nos permitió concluir que, examinando los pesos, vemos que los dos primeros componentes principales miden el equilibrio principalmente entre:

   - Familia (pesos positivos grandes) versus (2) edad (pesos negativos grandes).
     Los puntajes altos en el componente principal 1 significan que los pasajeros con alta cantidad de miembros de familia tienen baja edad.
   - Clase de pasajero (pesos positivos grandes) versus (2) precio del pasaje (pesos negativos grandes).
     Los puntajes altos en el componente principal 2 significan que los pasajeros con número alto de clase (3), tienen bajo precio de boleto.

   Luego, en el caso de Random Forest observamos que las variables Sex, Fare y Age, son buenos predictores. Luego le siguen clase de boleto y tamaño de familia.

   Llegados a este punto, emprendimos la tarea de resolver el problema de inicio, este es, establecer un modelo de aprendizaje automático para predecir si un pasajero es sobreviviente o no.
   Para ello elaboramos tres modelos, usando el ser de datos de prueba provistos para el problema:

   - Árbol de decisión.
   - Bosque aleatorio.
   - Regresión logística.

   Los resultados obtenidos muestran que el modelo de regresión logística es capaz de predecir de mejor manera si un pasajero es sobreviviente o no, que los otros dos modelos. Sin embargo, tanto árbol de clasificación y bosque aleatorio permiten la elaboración de modelos bastante adecuados, siendo capaces de clasificar correctamente sobre el 90% de los casos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Acceso al código: https://github.com/cheukepan/proyectoTitanic

Código

```
---
title: 'Tipología de los datos: PRAC2'
author: "Autor: Diego Cheuquepán M"
date: "Junio 2020"
output:
  html_document:
    highlight: default
    number_sections: yes
    theme: cosmo
    toc: yes
    toc_depth: 2
    includes:
      in_header: M2-851-PRAC2-header.html
  pdf_document:
    highlight: zenburn
    toc: yes
  word_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

******

> Desarrollo PRAC

******

```{r, include=FALSE}
library(GGally) #ggpairs()
library(ggplot2)
library(fBasics) #kurtosis y skewness
library(cowplot) #plot_grid()
library(corrplot)
library(randomForest)
library(rpart)
library(rpart.plot)
library(caret) #ConfusionMatrix()
#library(Hmisc)
#library(VIM)
#library(stringr)
library(vcd) #assocstats()
#library(car)
#library(lmtest)
#library(ResourceSelection)
#library(ROCR)
#library(mice)
```

Preprocesamiento de datos

Presentación del set de datos

Para desarrollar esta práctica se usará el dataset Titanic, tomado de la competencia de kaggle.

La competencia consiste en usar el aprendizaje automático para crear un modelo que prediga qué pasajeros sobrevivieron al naufragio del Titanic.

Se puede acceder a este mediante el siguiente enlace:

https://www.kaggle.com/c/titanic/data

Carga de los datos

Iniciamos cargando los datos del problema.

```{r}
# Cargamos el juego de datos
datos <- read.csv('D:/Descargas/UOC/S1/Tipología de los datos/Pruebas/PRAC 2/titanic/train.csv', header=TRUE, sep=",", stringsAsFactors = FALSE)
```

Caracterización de los datos

Verificamos la estructura del conjunto de datos.
Observamos que el dataset posee 11 variables que posteriormente trataremos de acuerdo a su tipo, ya sea, cuantitativo o cualitativo.
```

```{r}
# Verificamos la estructura
str(datos)
```

El set de datos tiene una serie de atributos con las siguientes características:
```{r, echo=FALSE}
df <- data.frame(
  "Variable" = c("Atributo 1",
                 "Atributo 2",
                 "Atributo 3",
                 "Atributo 4",
                 "Atributo 5",
                 "Atributo 6",
                 "Atributo 7",
                 "Atributo 8",
                 "Atributo 9",
                 "Atributo 10",
                 "Atributo 11",
                 "Atributo 12"
                ),
    "Nombre" = c("PassengerId",
                 "Survived",
                 "Pclass",
                 "Name",
                 "Sex",
                 "Age",
                 "SibSp",
                 "Parch",
                 "Ticket",
                 "Fare",
                 "Cabin",
                 "Embarked"
                ),
  "Descripción" = c("Identificador del pasajero",
                 "Sobrevive",
                 "Clase de pasajero",
                 "Nombre del pasajero",
                 "Sexo del pasajero",
                 "Edad del pasajero",
                 "Número de familiares",
                 "Número de familiares",
                 "Número de ticket",
                 "Tarifa del ticket",
                 "Número de cabina del pasajero",
                 "Puerto de embarque"
                ),
  "Tipo de Variable" = c("Cuantitativa",
                 "Cuantitativa",
                 "Cualitativa",
                 "Cualitativa",
                 "Cualitativa",
                 "Cuantitativa",
                 "Cuantitativa",
                 "Cuantitativa",
                 "Cualitativa",
                 "Cuantitativa",
                 "Cualitativa",
                 "Cualitativa"
                ),
   "Tipo de Escala" = c("Discreta",
                 "Discreta",
                 "Ordinal",
                 "Nominal",
                 "Nominal",
                 "Discreta",
                 "Discreta",
                 "Discreta",
                 "Nominal",
                 "Continua",
                 "Nominal",
                 "Nominal"
                ),
  "Tipo de Dato" = c("Entero",
                 "Entero",
                 "Entero",
                 "Cadena de texto",
                 "Cadena de texto",
                 "Número",
                 "Entero",
                 "Entero",
                 "Cadena de texto",
                 "Número",
                 "Cadena de texto",
                 "Cadena de texto"
                )
)
```

knitr::kable(df, col.names = gsub("[.]", " ", names(df)))
```

## Creación de nuevas variables

Observamos que las variables SibSp y Parch informan el número de hermanos/cónyuges y de de padres/hijos, respectivamente; de manera que podemos determinar el tamaño de la familia como una nueva variable cuantitativa discreta.

```{r}
datos$familia <- datos$SibSp + datos$Parch + 1
```

Ordenación de atributos

Dado que tenemos atributos de distinto tipo, los ordenamos de manera que luego podamos operar con mayor facilidad los cuantitativos y los cualitativos.

```{r}
# Ordenamos los datos por tipo
datos <- datos[c(1,2, 10, 6,13,7,8, 3, 5,12,4,9,11)]
```

Cantidad de registros únicos

Ahora nos interesa saber para cada atributo cuál es la cantidad de valores únicos. Esta información nos será de utilidad para definir posteriormente las visualizaciones más adecuadas para cada variable.

```{r}
# Verificamos cantidad de registros únicos
apply(datos, 2, function(x) length(unique(x)))
```

Análisis de valores nulos/vacíos

Para verificar los posibles valores no disponibles en el set de datos usaremos la función is.na()

```{r}
# Estadísticas de valores nulos
colSums(is.na(datos))

# Estadísticas de valores nulos
colSums(datos=="")

# Estadísticas de valores nulos
colSums(datos==" ")
```

> Nuestro conjunto tiene datos ausentes en las variables edad (Age), Número de cabina del pasajero (Cabin) y Puerto de embarque (Embarked).

Tratamiento de valores nulos/vacíos

De las variables que poseen datos nulos, procederemos a tratar particularmente el atributo edad (Age).
El tratamiento de los valores nulos/vacío puede abordarse de diferentes maneras.
En esta prueba, analizo los siguientes tres enfoques:

- Enfoque basado en media: En este caso determinamos la media de la variable Age para todos los datos no nulos, de manera que dicho promedio lo imputamos en cada dato NA. La media es un dato que se ve alterado por los extremos, de manera que nuestro promedio estará afectado por outliers.
- Enfoque basado en mediana: En este caso determinamos la mediana de la variable Age para todos los datos no nulos, de manera que dicho valor lo imputamos en cada dato NA. La mediana es un dato que no se ve afectado por valores extremos, de manera que al imputar por este valor se conservará esta característica.
- Enfoque basado en vecinos más cercanos: En este caso determinamos un valor de la variable Age para todos los datos no nulos usando los k-vecinos próximos, de manera que dicho valor lo imputamos en cada dato NA. Esta estrategia es mucho más completa que las anteriores, pero requiere contar con al menos 1 atributo numérico distinto a la variable que se tratará, y que además, tengan sentido que en la imputación de Age. En este caso, ninguna otra variable cuantitativa tiene sentido para aplicarla en la imputación de Age.

> Por lo anterior, abordaremos la estrategia de imputación por mediana.

```{r}
# Resumen variable Edad
summary(datos$Age)

# Imputación por mediana
datos$Age[is.na(datos$Age)] <- median(datos$Age, na.rm = TRUE)

# Verificamos variable Age
colSums(is.na(datos))
```

Recodificación de variables

Procederemos a recodificar las siguientes variables:

- Embarked (Puerto de embarque), codificicando C = Cherbourg, Q = Queenston y S = Southapmtpon. Además, existen dos registros sin datos que codificaremos como "No especificado"
```

```{r}
# Registros únicos de Embarked
unique(datos$Embarked)

# Recodificamos variable Embarked
datos$Embarked[datos$Embarked=="C"]="Cherbourg"
datos$Embarked[datos$Embarked=="Q"]="Queenston"
datos$Embarked[datos$Embarked=="S"]="Southapmtpon"
datos$Embarked[datos$Embarked==""]="Cherbourg"
```

Valores extremos

Procederemos a revisar los valores extremos o atípicos del conjunto de datos, respecto a sus variables cuantitativas Fare, Age, SibSp y Parch.
Para ello, primero crearemos gráficos de caja y bigotes.

```{r}
# Caja y bigotes de Fare
boxplot(datos$Fare, main="Box Plot Fare", col="green")
boxplot.stats(datos$Fare)$out

# Caja y bigotes de Age
boxplot(datos$Age, main="Box Plot Age", col="red")
boxplot.stats(datos$Age)$out

# Caja y bigotes de SibSp
boxplot(datos$SibSp, main="Box Plot SibSp", col="blue")
boxplot.stats(datos$SibSp)$out

# Caja y bigotes de Parch
boxplot(datos$Parch, main="Box Plot Parch", col="yellow")
boxplot.stats(datos$Parch)$out

# Caja y bigotes de Familia
boxplot(datos$familia, main="Box Plot Familia", col="yellow")
boxplot.stats(datos$familia)$out
```

> No se recomienda eliminar los datos atípicos de un dataset por el solo hecho de serlo.
Además, existe una importanye cantidad de datos extremos, por lo que no es recomendable analizarlos con miras a quitarlos de nuestro set.

Análisis de los datos

Análisis de frecuencias

Analizaremos las variables cuantitativas mediante histogramas de frecuencia, que nos permiten distinguir la simetrías, los picos y las colas.

Primero determinaremos el número de clases, para ello, usaremos la regla de Sturges que nos permitirá determinar un valor entre 7 y 15 que es el adecuado según Gil (Introducción al análisis de datos, p. 16).

```{r}
# Regla de Sturges
k = nclass.Sturges(datos$PassengerId)
k
```

Variables Cuantitativas

Obviando los atributos 1 y 2 (PassengerId y Survived), procedemos a analizar las variables cuantitativas de nuestro dataset.

```{r}
# Variable Fare
ggplot(datos, aes(Fare)) + geom_histogram(bins = 11) +
  labs(title = 'Histograma de frecuencia',
       x = 'Fare',
       y = 'conteos',
       subtitle = 'Variable Fare',
       caption = 'Número de clases basado en regla de Sturges')

# Variable Age
ggplot(datos, aes(Age)) + geom_histogram(bins = 11) +
  labs(title = 'Histograma de frecuencia',
       x = 'Age',
       y = 'conteos',
       subtitle = 'Variable Age',
       caption = 'Número de clases basado en regla de Sturges')

# Variable SibSp
ggplot(datos, aes(SibSp)) + geom_histogram(bins = 11) +
  labs(title = 'Histograma de frecuencia',
       x = 'SibSp',
       y = 'conteos',
       subtitle = 'Variable SibSp',
       caption = 'Número de clases basado en regla de Sturges')

# Variable Parch
ggplot(datos, aes(Parch)) + geom_histogram(bins = 11) +
  labs(title = 'Histograma de frecuencia',
```

```
    x = 'Parch',
    y = 'conteos',
    subtitle = 'Variable Parch',
    caption = 'Número de clases basado en regla de Sturges')

# Variable familia
ggplot(datos, aes(familia)) + geom_histogram(bins = 11) +
  labs(title = 'Histograma de frecuencia',
    x = 'familia',
    y = 'conteos',
    subtitle = 'Variable familia',
    caption = 'Número de clases basado en regla de Sturges')
```

> Este análisis nos muestra que la variable:
- Fare es unimodal con una asimetría a la derecha y datos extremos en la última clase.
- Age es unimodal con simetría.
- SibSp es unimodal con una asimetría a la derecha y datos extremos en la última clase.
- Parch es unimodal con una asimetría a la derecha.
- Familia es unimodal con una asimetría a la derecha y datos extremos en la última clase.

Normalidad y Homogeneidad

Normalidad

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba shapiro-wilk, dado que tenemos menos de 5.000 registros.

```{r}
# Variable Fare
shapiro.test(datos$Fare)

# Variable Age
shapiro.test(datos$Age)

# Variable SibSp
shapiro.test(datos$SibSp)

# Variable Parch
shapiro.test(datos$Parch)

# Variable familia
shapiro.test(datos$familia)
```

> Dado que nuestro valor p-value para cada variable es menor que nuestro nivel de significancia elegido (0,05), rechazamos la hipótesis nula de que cada variabla se distribuye normalmente.

Homogeneidad

Completado el análisis anterior, interesa comparar las varianzas de los pasajeros de acuerdo a la edad y grupo familiar de cada uno.

Para esto se procederá con un análisis de homogeneidad de varianza usando el test de Fligner-Killeen, dado que la edad no se distribuye normalmente.

```{r}

fligner.test(Age ~ familia, data = datos)
```

> Puesto que obtenemos un p-valor menor a 0.05, no aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

Asimetría y Curtosis

A continmuación, analizaremos en particular la variable Age que es la que muestra una distribución con mayor simetría.

```{r}
# Asimetría
skewness(datos$Age, method = "moment")

# Curtosis
kurtosis(datos$Age, method = "moment")
```

> La asimetría es 0.5085. Este valor implica que la distribución de los datos está sesgada a la derecha o positivamente sesgada .
Para la curtosis, tenemos 3.9726 lo que implica que la distribución de los datos es leptocurtica, lo que indica que los datos están muy concentrados en la media, siendo una curva muy apuntada.

El análisis visual es el siguiente:

```{r}
# Gráfico
ggplot(datos, aes(x = Age), binwidth = 2) +
    geom_histogram(aes(y = ..density..), fill = 'red', alpha = 0.5) +
    geom_density(colour = 'blue')
```

Variables Cualitativas

Ahora analizaremos las variables cualitativas mediante gráficos circulares.

```{r}
# Pclass
bar <- ggplot(datos, aes(x = factor(1), fill = factor(Pclass))) + geom_bar(width = 1)
pie <- bar + coord_polar(theta = "y") + theme_void() +
  labs(title = 'Gráfico circular',
       subtitle = 'Variable Pclass')
bar <- ggplot(datos, aes(Pclass)) + geom_bar() + coord_flip() +
  labs(title = 'Gráfico de barra',
       x = 'Pclass',
       y = 'conteos',
       subtitle = 'Variable Pclass')

plot_grid(bar, pie, labels = "AUTO")


plot_grid(bar, pie, labels = "AUTO")

# Name
bar <- ggplot(datos, aes(x = factor(1), fill = factor(Name))) + geom_bar(width = 1)
pie <- bar + coord_polar(theta = "y") + theme_void() +
  labs(title = 'Gráfico circular',
       subtitle = 'Variable Name')
bar <- ggplot(datos, aes(Name)) + geom_bar() +
  labs(title = 'Gráfico de barra',
       x = 'Name',
       y = 'conteos',
       subtitle = 'Variable Name')

# Sex
bar <- ggplot(datos, aes(x = factor(1), fill = factor(Sex))) + geom_bar(width = 1)
pie <- bar + coord_polar(theta = "y") + theme_void() +
  labs(title = 'Gráfico circular',
       subtitle = 'Variable Sex')
bar <- ggplot(datos, aes(Sex)) + geom_bar() + coord_flip() +
  labs(title = 'Gráfico de barra',
       x = 'Sex',
       y = 'conteos',
       subtitle = 'Variable Sex')

plot_grid(bar, pie, labels = "AUTO")


# Embarked
bar <- ggplot(datos, aes(x = factor(1), fill = factor(Embarked))) + geom_bar(width = 1)
pie <- bar + coord_polar(theta = "y") + theme_void() +
  labs(title = 'Gráfico circular',
       subtitle = 'Variable Embarked')
bar <- ggplot(datos, aes(Embarked)) + geom_bar() +
  labs(title = 'Gráfico de barra',
       x = 'Embarked',
       y = 'conteos',
       subtitle = 'Variable Embarked')

plot_grid(bar, pie, labels = "AUTO")
```

Análisis de Correlación

Cuando hablamos de correlación nos referimos al grado de relación que hay entre dos variables, pero que no establece ningún tipo de relación de causa ni efecto de una sobre otra (Liviano y Pujol, 2013).

Para analizar la correlación usaremos distintos procedimientos, siguiendo las recomendaciones de Reguant-Álvarez, Vilà-Baños y Torrado-Fonseca (2018).

Estos autores, proponen que los procedimientos y gráficos a usar para analizar la correlación de variables deben considerar el tipo y su escala asociada.

La siguiente tabla, muestra la secuencia a seguir para las variables de nuestro estudio considerando su tipo y escala.

```{r, echo=FALSE}
df <- data.frame(
  "Tipo de escala" = c("2 nominales",
               "2 ordinales",
               "1 ordinal y 1 escalar",
               "2 escalares"
              ),
  "Procedimiento" = c("Chi cuadrado",
               "Correlación de Spearman",
               "Correlación de Spearman",
               "Correlación de Pearson"
              ),
  "Gráfico" = c("Mosaico",
            "Barra",
            "De cajas y bigotes",
            "Dispersión"
```

```
        )
)
knitr::kable(df, col.names = gsub("[.]", " ", names(df)))
```

Análisis: 2 escalares

Las variables denominadas comúnmente escalares o escalas continuas, corresponden a las variables de escala de intervalo y de razón.

Procederemos a graficarlas para analizar la correlación entre ellas:

```{r}
# Gráfico de correlación
ggpairs(datos[,3:7])
```

Los resultados obtenidos nos muestran correlaciones débiles entre todas las variables analizadas.

Las únicas variables que muestra correlaciones superiores a 0.7 y que se consideran adecuadas son familia, Parch y SibSp, sin embargo, eso se debe a la colinealidad ya que la variable familia la construimos en base a Parch y SibSp.

```{r}
# Aplicamos correlación de pearson | Parch y SibSp
pearson <- cor.test(x = datos$Parch, y = datos$SibSp, exact = FALSE)
pearson
```

Finalmente, los resultados obtenidos nos muestran una correlación moderada entre Parch y SibSp.

Al analizar el p-value, dado que es menor que 0.05, confirmamos que el coeficiente de correlación es estadísticamente significativo, por lo que rechazamos la hipótesis nula.

Por úlitmo, el intervalo de confianza (IC) al 95% para el coeficiente de correlación nos informa el rango de valores que contiene con una confianza del 95% el coeficiente de correlación verdadero.

Dado que el intervalo inferior del IC es menor a 0.4, es posible que nuestra correlación no sea moderada sino baja.

Análisis: ordinal y 1 escalar

Analizaremos la variable ordinal Clase de pasajero con las variables escalares.
Para ello, utilizaremos la correlación de Spearman y nos enfocaremos en los resultados de Clase pasajero con las otras variables.

```{r}
# Gráfico de correlación
mydata <- datos[,3:8]
M <- cor(mydata, method = "spearman")
corrplot.mixed(M)
```

Los resultados muestran una correlación negativa entre Pclass y Fare.

```{r}
# Aplicamos correlación de spearman | Pclass y Fare.
spearman <- cor.test(x = datos$Pclass, y = datos$Fare, method = 'spearman', exact = FALSE)
spearman

# Finalizamos con un gráfico de caja y bigotes para examinar la relacion entre las dos variables analizadas.
boxplot(datos$Fare  ~ datos$Pclass, main="Box Plot Pclass  ~  Fare", col="green")
```

Finalmente, los resultados obtenidos nos muestran una correlación negativa moderada entre Pclass y Fare.

Al analizar el p-value, dado que es menor que 0.05, confirmamos que el coeficiente de correlación es estadísticamente significativo, por lo que rechazamos la hipótesis nula.

Análisis: 2 nominales

Las variables de escala nominal son Name, Sex, Ticket, Cabin y Embarked.
Dado el contexto del problema, analizaremos la correlación de las variables Sex y Embarked de manera que en este caso procederemos con chi-cuadrado y gráfico de mosaico, siguiendo las recomendaciones de Few y Edge (2014).

```{r}
# Creamos una tabla de contingencia.
cuadro <- table(datos$Embarked, datos$Sex)
cuadro

# Analizamos el gráfico que despliega informacion para examinar la relacion entre las dos variables.
mosaicplot(cuadro, color=TRUE, main="Gráfico de mosaico",
        sub = NULL, xlab = "Ciudad de Embarque", ylab = "Sexo")

# Aplicamos chi-cuadrado
chisq.test(cuadro)

# Aplicamos test de fisher
fisher.test(x = cuadro, simulate.p.value = TRUE)

# tamaño de la fuerza de asociación
assocstats(cuadro)
```

```

Al analizar el p-value, dado que es menor que 0.05, confirmamos que el resultado es estadísticamente significativo, por lo que rechazamos la hipótesis nula.

Por otro lado, el tamaño de la fuerza de asociación es de 0.12 que se clasifica como pequeño.

> Los resultados muestran que existe relación entre las dos variables, Sexo (Sex) y Ciudad de Embarque (Embarked), la que tiene una fuerza de asociación pequeña.

# Reducción de la dimensionalidad

Analizaremos a continuación cuáles son las variables que pueden explicar el modelo de manera más apropiada.
Aplicaremos PCA y Random Forest para determinar la importancia de las variables.

## PCA

Examinando los pesos, vemos que los dos primeros componentes principales miden el equilibrio principalmente entre:

a. (1) Familia (pesos positivos grandes) versus (2) edad (pesos negativos grandes).
Los puntajes altos en el componente principal 1 significan que los pasajeros con alta cantidad de miembros de familia, tienen baja edad.

b. (1) Clase de pasajero (pesos positivos grandes) versus (2) precio del pasaje (pesos negativos grandes).
Los puntajes altos en el componente principal 2 significan que los pasajeros con número alto de clase (3), tienen bajo precio de boleto.

```{r}
# Componentes Principales
pca <- prcomp(datos[,3:8], scale = TRUE)
pca$rotation

# Varianza
prop_varianza <- pca$sdev^2 / sum(pca$sdev^2)

# Gráfica de varianza
ggplot(data = data.frame(prop_varianza, pc = 1:6),
       aes(x = pc, y = prop_varianza)) +
  geom_col(width = 0.3) +
  scale_y_continuous(limits = c(0,1)) +
  theme_bw() +
  labs(x = "Componente principal",
       y = "Prop. de varianza explicada")

# Varianza acumulada
prop_varianza_acum <- cumsum(prop_varianza)
prop_varianza_acum

# Gráfica varianza acumulada
ggplot(data = data.frame(prop_varianza_acum, pc = 1:6),
       aes(x = pc, y = prop_varianza_acum, group = 1)) +
  geom_point() +
  geom_line() +
  theme_bw() +
  labs(x = "Componente principal",
       y = "Prop. varianza explicada acumulada")
```

> En este caso, la primera componente explica el 42.8% de la varianza observada en los datos y la segunda el 28.2%.
Si se empleasen únicamente las dos primeras componentes se conseguiría explicar el 71.1% de la varianza observada y se se emplean las tres primeras se explicaría el 83.9%.

> Las conclusiones que se extraen, son esperables dado que una familia con un alto número de integrantes supone solo dos adultos y se esperaría que el resto sean niños, de ahí la baja edad.
Lo mismo puede decirse de la clase y el ticket, tener un ticket de tercera clase será mucho más económico que un ticket de primera clase que será mucho más costoso.

Random Forest

```{r, include=FALSE}
# Recodificamos la variable Sex
datos$Sex[datos$Sex=="female"]= 1
datos$Sex[datos$Sex=="male"]= 0

# Discretizamos las variables con pocas clases
cols<-c("Survived","Pclass","Sex","Embarked", "familia")
for (i in cols){
  datos[,i] <- as.factor(datos[,i])
}
```

Utilizaremos este enfoque para determinar la importancia (MeanDecreaseGini) que nos proporciona una medida de la contribución de una variable en la clasificación, de manera que números altos indican que una variable es más importante como predictor.

En este caso, incorporamos además la varibles ordinales para crear un bosque de clasificación.

```{r}
set.seed(123)
# Ajustar modelo
```
```

```
modelorf <- randomForest(datos$Survived~., data=datos[,3:10])

# Resumen del ajuste del modelo
modelorf

# Importancia de las variables
modelorf$importance

# Gráfico del modelo
varImpPlot(modelorf)
```

> A partir de los resultados, observamos que las variables Sex, Fare y Age, son buenos predictores.
Luego le siguen clase de boleto y tamaño de familia.

Análisis respecto a la variable objetivo

Dado que el problema consiste en constuir un modelo que nos permita predecir qué pasajeros sobrevivieron al naufragio del Titanic, procederemos a analizar las variables respecto de nuestra variable objetivo.

Relación con la variable Survived

Análisis respecto a variable Sex

```{r}
# Visualizamos la relación entre las variables Sex y Survived
ggplot(data = datos,aes(x=Sex, fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```

> Hay una mayor proporción de mujeres que se salvan que hombres.

Análisis respecto a variable Embarked

```{r}
# Visualizamos la relación entre las variables Embarked y Survived
ggplot(data = datos,aes(x=Embarked, fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```

> Hay una mayor proporción de pasajeros que embarcan en Cherbourg que se salvan respecto a los pasajeros que subieron en otro puerto.

Análisis respecto a variable Pclas

```{r}
# Visualizamos la relación entre las variables Pclass y Survived
ggplot(data = datos,aes(x=Pclass, fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```

> Hay una mayor proporción de pasajeros de primera clase que se salvan respecto de pasajeros de tercera clase.

Análisis respecto a variable Age

```{r}
# Visualizamos la relación entre las variables Age y Survived
ggplot(data = datos,aes(x=Age, fill=Survived))+geom_histogram(bins = 11) +geom_bar(position="fill")+ylab("Frecuencia")

# Ahora, podemos dividir el gráfico de familia por Pclass:
ggplot(data = datos,aes(x=Age, fill=Survived))+geom_histogram(bins = 11) +geom_bar(position="fill")+ylab("Frecuencia")+facet_wrap(~Pclass)
```

> Hay una mayor proporción de pasajeros menores de 40 años que no se salvan respecto de pasajeros mayores a esa edad o incluso niños (menores de 10).
Luego, al aperturar por clase de pasajero se observa que la mayor proporción de personas menores de 40 que mueren las aportan pasajeros que son de tercera clase.

Análisis respecto a variable familia

```{r}
# Visualizamos la relación entre las variables familia y Survived
ggplot(data = datos,aes(x=familia, fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")

# Ahora, podemos dividir el gráfico de familia por Pclass:
ggplot(data = datos,aes(x=familia, fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pclass)
```

> Hay una mayor proporción de familias de 4 o menos integrantes que se salvan, respecto a familias con más de 4 integrantes.
Se observa que en el caso de pasajeros de tercera clase, hay un aumento en la proporción de muerte en caso de familias mononucleares o binucleares.

Construcción de modelos de predicción

Dado que el problema consiste en usar el aprendizaje automático para crear un modelo que prediga qué pasajeros sobrevivieron al naufragio del Titanic, procederemos a crear los siguientes modelos para llevar a cabo el objetivo:

- Modelo de regresión lineal.
- Árbol de decisión.
- Bosque aleatorio.

Para ello cargamos los datos de prueba.

```{r, echo=FALSE}
# Cargamos el juego de datos
datostest <- read.csv('D:/Descargas/UOC/S1/Tipología de los datos/Pruebas/PRAC 2/titanic/test.csv', header=TRUE, sep=";", stringsAsFactors = FALSE)

# Creamos variable
datostest$familia <- datostest$SibSp + datostest$Parch + 1

# Ordenamos los datos por tipo
datostest <- datostest[c(1,2, 10, 6,13,7,8, 3, 5,12,4,9,11)]

# Imputación por mediana
datostest$Age[is.na(datostest$Age)] <- median(datostest$Age, na.rm = TRUE)

# Recodificamos variable Embarked
datostest$Embarked[datostest$Embarked=="C"]="Cherbourg"
datostest$Embarked[datostest$Embarked=="Q"]="Queenston"
datostest$Embarked[datostest$Embarked=="S"]="Southapmtpon"
datostest$Embarked[datostest$Embarked==""]="Cherbourg"

# Recodificamos la variable Sex
datostest$Sex[datostest$Sex=="female"]= 1
datostest$Sex[datostest$Sex=="male"]= 0

# Imputación por mediana
datostest$Fare <- as.numeric(datostest$Fare)
datostest$Fare[is.na(datostest$Fare)] <- median(datostest$Fare, na.rm = TRUE)

# Discretizamos las variables con pocas clases
cols<-c("Survived","Pclass","Sex","Embarked", "familia")
for (i in cols){
  datostest[,i] <- as.factor(datostest[,i])
}
```

Bosque aleatorio

Anteriormente creamos nuestro modelo de árbol aleatorio buscando determinar de manera preliminar cuáles eras las variables predictoras más adecuadas.

Ahora procederemos a evaluar el modelo de clasificación:

```{r}
# Hacer predicciones
predictrf <- predict(modelorf, datostest[,3:10])

# Matriz de confusión
(mc <- with(datostest,table(predictrf, Survived)))

# % correcto
100 * sum(diag(mc)) / sum(mc)
```

> Nuestro modelo basado en Bosque Aleatorio (Random Forest) clasifica correctamente el 91.3% de los casos.

Árbol de decisión

Realizaremos una clasificación basado en árbol de decisión (árbol de clasificación).

```{r}
set.seed(123)
# Ajustar modelo
modeloac <- rpart(datos$Survived~., data=datos[,3:10], method=           "class")

# Resumen del ajuste del modelo
summary(modeloac)

# Extracción de reglas
modeloac

# Gráfico del modelo
rpart.plot(modeloac)

# Hacer predicciones
predictac <- predict(modeloac, datostest[,3:10], type =     "class")

# Matriz de confusión
(mc <- with(datostest,table(predictac, Survived)))

# % correcto
100 * sum(diag(mc)) / sum(mc)
```

> Nuestro modelo basado en Árbol de Clasificación RPART, clasifica correctamente el 91.6% de los casos.

Regresión logística

Realizaremos una clasificación basado en regresión logística.
Para ello seguiremos la siguiente secuencia:

- Crearemos un modelo considerando todas la variables y en base al análisis de los p-value, seleccionaremos las variables predictoras más adecuadas.
- Con estas nuevas variables, crearemos un segundo modelo y analizarmos su capacidad de clasificar a los sobrevivientes.
- Finalmente, usando los datos de test evaluaremos el modelo determinando como clasifica a los pasajeros.

```{r}
set.seed(123)
# Ajustar modelo
modelolog <- glm(datos$Survived~., data=datos[,3:10], family=binomial, maxit = 100)

# Resumen del ajuste del modelo
summary(modelolog)

# Odds ratio
OR=exp(modelolog$coefficients[2:18])
OR
```

Basados en los resultados de p-value crearemos un nuevo modelo que considere las variables Age, Sex y Pclass.

```{r}
# Nuevo modelo
modelolog <- glm(Survived ~ Age + Sex + Pclass, datos, family=binomial, maxit = 100)

# Resumen del ajuste del nuevo modelo
summary(modelolog)
```

Ahora aplicaremos nuestro modelo construido con los datos de test, para comprobar su capacidad predictiva.

```{r}
# Hacer predicciones
modelologpred <- glm(Survived ~ Age + Sex + Pclass, datostest, family=binomial, maxit = 100)

predictrlog <- ifelse(modelologpred$fitted.values > 0.5, 1, 0)

# Matriz de confusión
(mc <- with(datostest,table(predictrlog, Survived)))

# % correcto
100 * sum(diag(mc)) / sum(mc)
```

> Nuestro modelo basado en regresión logística, clasifica correctamente el 100% de los casos.

Conclusiones

En esta práctica se uso el aprendizaje automático para crear modelos con la capacidad de predecir qué pasajeros sobrevivieron al naufragio del Titanic.

Para ello, se inicio haciendo una limpieza de los datos, abordano las siguientes estrategias:

- Tratamiento de valores nulos.
- Análisis de valores atípicos.

Luego, se continuo creando nuevas variables a partir de los datos. En particular se creo la variable Familia que incorporaba al pasajero, más sus conyuges, hijos y padres.

Posteriormente, se hizo un análisis de cada variable de acuerdo a su tipo (nominal, Ordinal, Discreta, Continua).
En el caso de las continuas se analizó la curtosis y simetría.

Posteriormente, se realizaron análisis de correlación de acuerdo al tipo de variable y se usaron las técnicas de:

- R Pearson.
- Spearman.
- chi-cuadrado de contingencia.

Seguimos con un análisis para reducción de dimensionalidad, usando:

- Análisis de Componentes Principales (PCA).
- Análisis de importancia de variables mediante Random Forest.

A continuación, se analizó la variable objetivo Survived, con cada una de las variables consideradas como predictoras.

Por último, se crearon y analizaron tres modelos para predecir a los sobrevivientes:

- Árbol de decisión.
- Bosque aleatorio.
- Regresión logística.

Los primeros dos modelos, tienen resultados similares y logran clasificar correctamente a más del 90% de los casos. En tanto, nuestro modelo de regresión logística es capaz de clasificar correctamente los casos de prueba en un 100%.

8. Contribuciones

Contribuciones	Firma
Investigación previa	Diego Cheuquepán Maldonado
Redacción de las respuestas	Diego Cheuquepán Maldonado
Desarrollo código	Diego Cheuquepán Maldonado