Limpieza y análisis de datos

Tipología y ciclo de vida de los datos

Master en Ciencia de Datos

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

Para la realización de la práctica se ha decidido optar por uno de los *datasets* propuestos, concretamente el del Titanic, debido principalmente a que el *dataset* generado en la primera práctica de la asignatura no permitía realizar un estudio en profundidad, puesto que dispone de pocas variables, y alguna de ellas eran candidatas a ser eliminadas del análisis al iniciar el procedimiento de limpieza de datos. Por todo ello, se ha optado por trabajar con un *dataset* de calidad reconocida, que permita alcanzar los objetivos propuestos en la práctica.

El *dataset* escogido se denomina train.csv, y proporciona una serie de atributos de los pasajeros del Titanic. Estos atributos se utilizan para realizar predicciones sobre otros, utilizando modelos de aprendizaje supervisado. El listado de atributos es el siguiente:

- **PassengerId**: Número que identifica al pasajero.
- **Survived**: Identifica si el pasajero sobrevivió (valor 1) o no (valor 0).
- **Pclass**: Idenfifica la clase del ticket como primera clase (valor 1), segunda clase (2) y tercera clase (3).
- **Name**: Nombre del pasajero.
- **Sex**: Sexo del pasajero.
- **Age**: Edad del pasajero.
- **SibSp**: Número de hermanos/cónyuges a bordo.
- **Parch**: Número de padres/hijos a bordo.
- **Ticket**: Número del ticket.
- **Fare**: Tarífa pagada.
- **Cabin**: Número de cabina.
- **Embarked**: Puerto donde embarcó.

Con este *dataset* se pretende estudiar la relación entre los atributos que influyen en mayor y menor medida en la probabilidad de supervivencia del pasajero. De este modo, podrán extraerse diversas conclusiones, que nos permitirán inferir el índice de supervivencia dependiendo de la clase, la edad o el sexo de los pasajeros, entre otros.

2. Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes *datasets* o una subselección útil de datos originales, en base al objetivo que se quiera conseguir.

Tal y como indicábamos en el apartado previo, la elección ha venido condicionada por las características del *dataset* de la primera práctica que, a pesar de tener un potencial interesante, era arriesgado utilizar por disponer de una cantidad limitada de variables.

Tras iniciar el trabajo con el conjunto de datos original, se ha decidido sustituirlo por una de las propuestas alternativas que se ofrecen en el enunciado; concretamente, el *dataset* del Titanic.

Tras cargar el conjunto de datos en R, el primer paso a realizar una subselección útil de los datos originales. En este caso, se ha optado por descartar los atributos siguientes:

- PassengerId
- Name
- Ticket
- Fare
- Cabin

Quedando nuestra selección con los atributos que se listan a continuación:

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

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

El siguiente paso que encontramos en el código en R es la gestión de los datos nulos. En particular, se estudian las posibilidades siguientes:

```
colSums(is.na(data))
```

```
colSums(data=="")
colSums(data=="?")
```

Detectándose un total de 177 valores nulos en la variable "Age", y 2 en la variable "Embarked". Para tratar estos valores, se han seguido las estrategias siguientes:

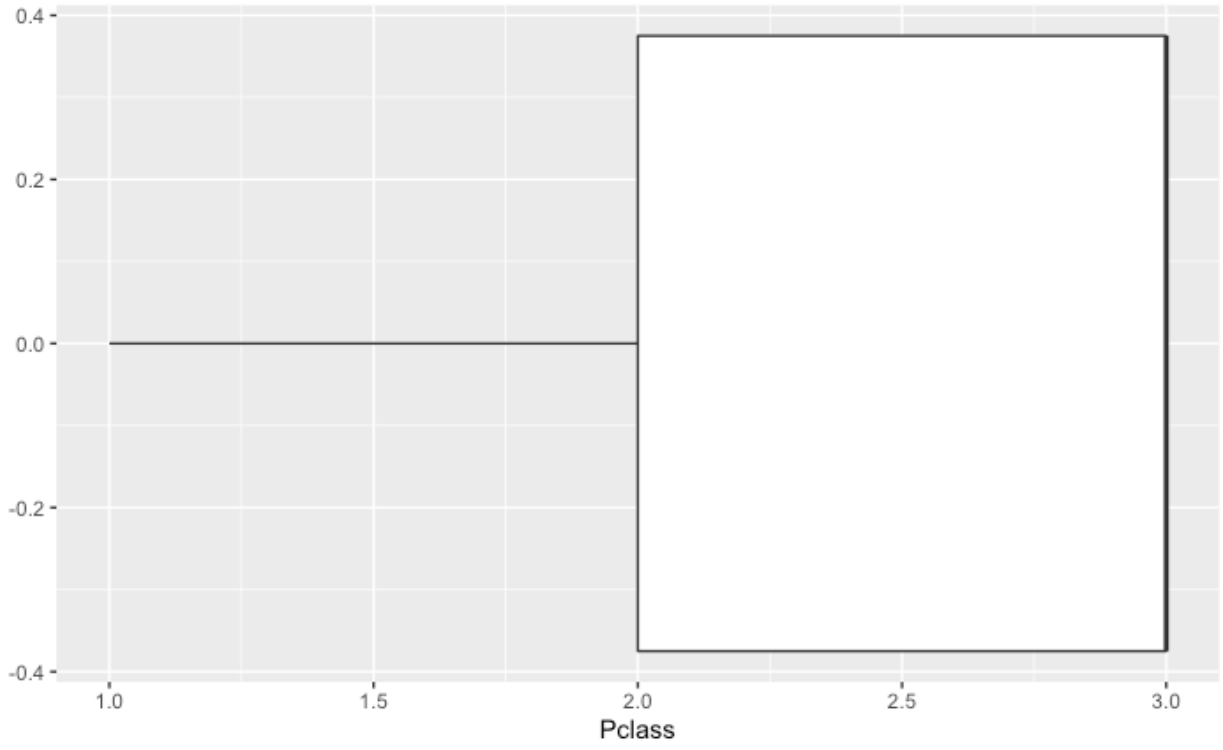
- **Embarked**: Se han eliminado las dos filas.
- **Age**: Se completan todos los valores que faltan utilizando el método de imputación de vecinos.

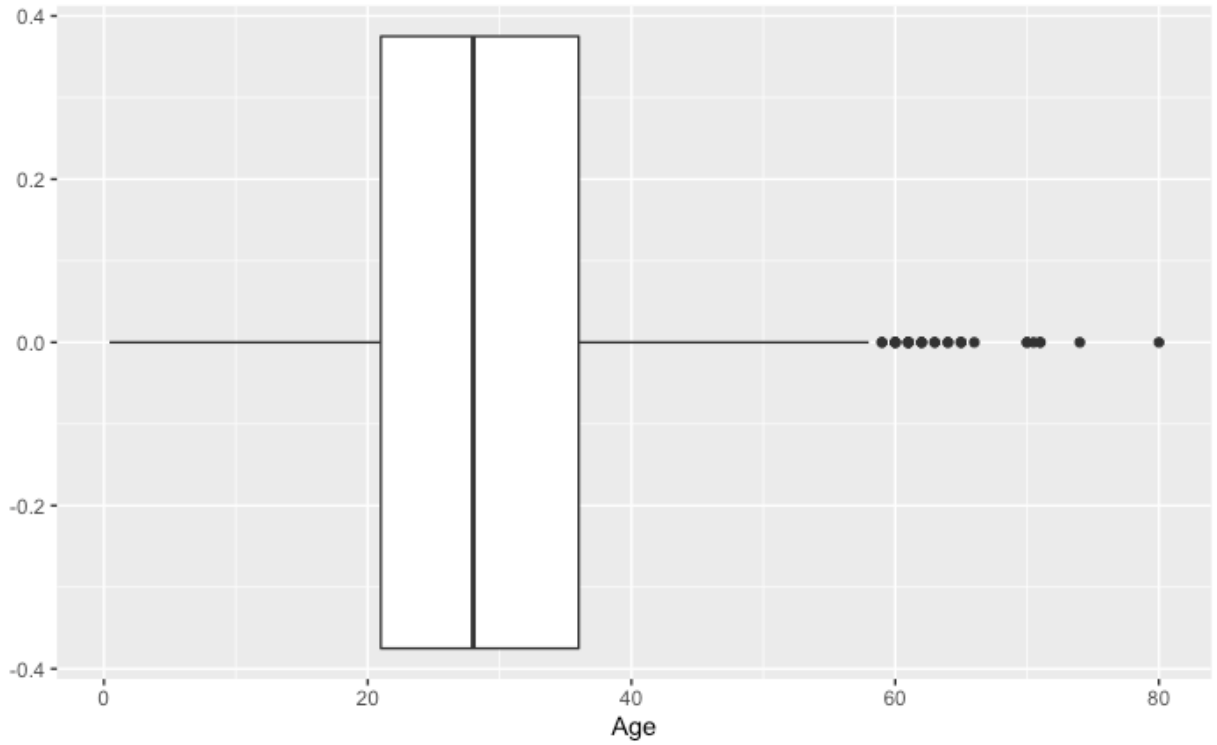
De este modo, el *dataset* resultante queda libre de valores nulos.

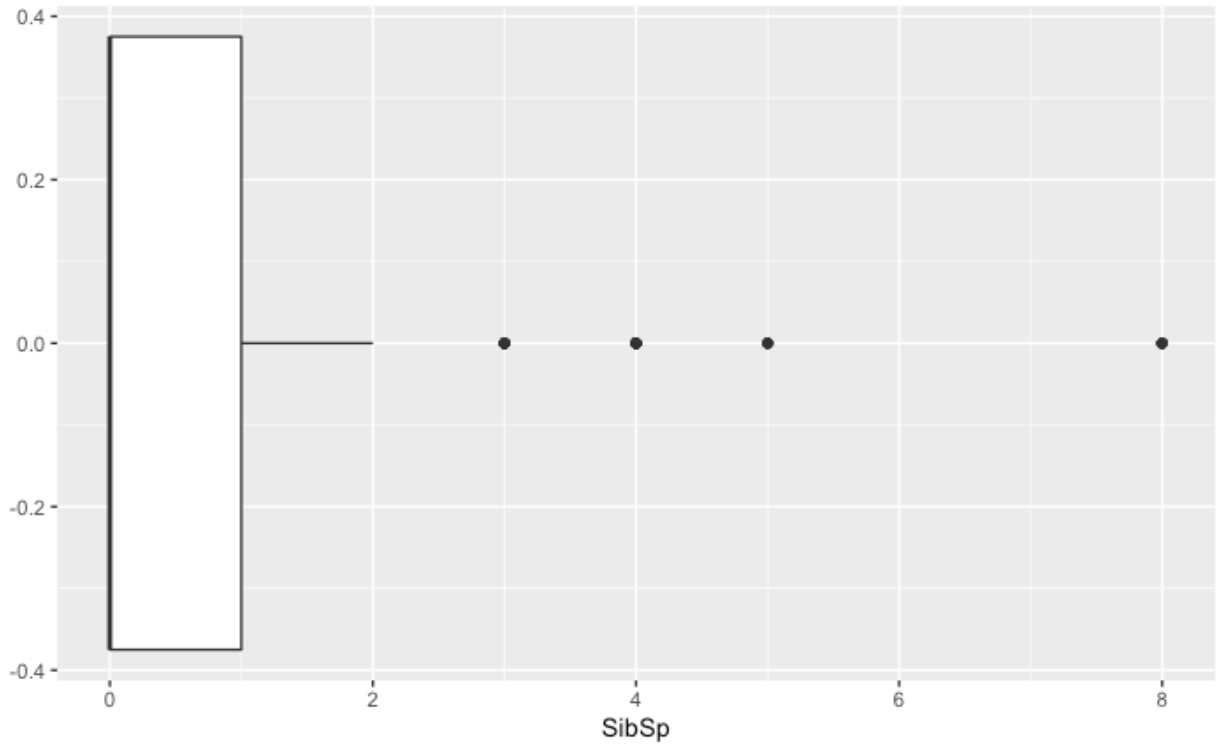
3.2. Identifica y gestiona los valores extremos.

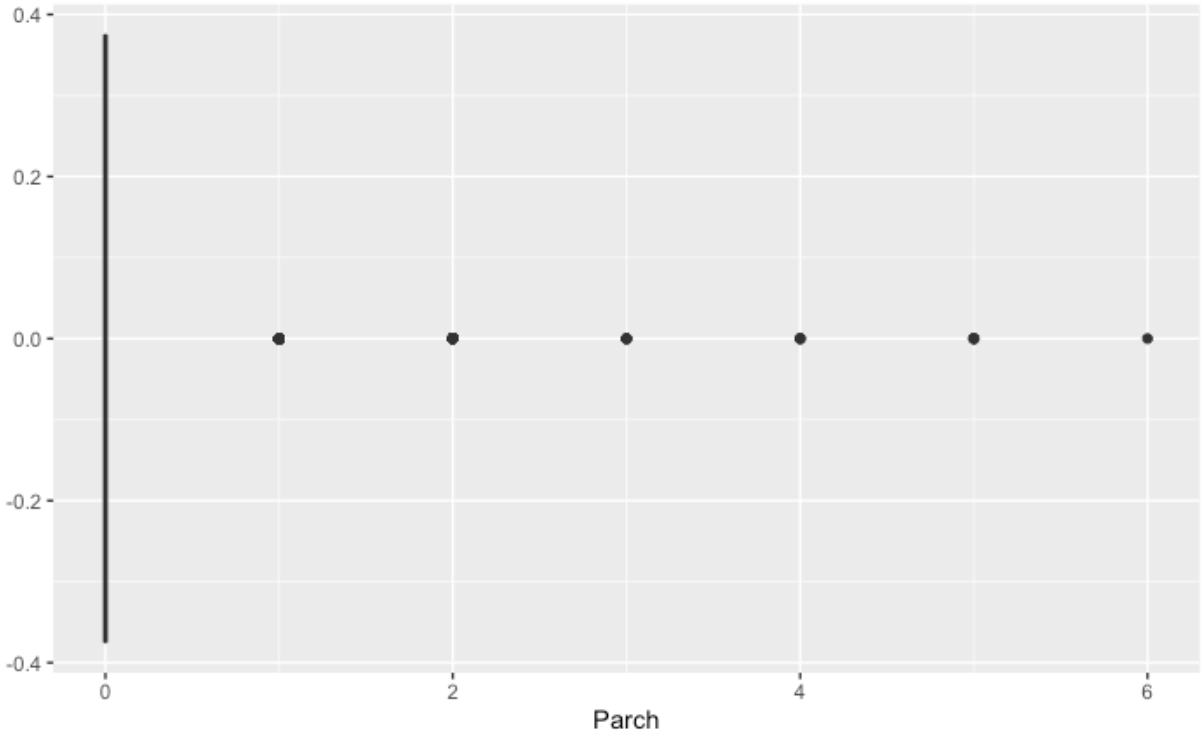
Para detectar los valores extremos, se ha procedido a generar una representación gráfica de todas las variables numéricas. De ese modo, hemos podido confirmar que las variables "SibSp", "Parch" y "Age" tienen *outliers*. En cualquier caso, se ha decidido no eliminarlos, puesto que se trata de valores reales que nos serán útiles al construir nuestro modelo.

A continuación, se adjuntan las representaciones generadas en R:




0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
80
Age

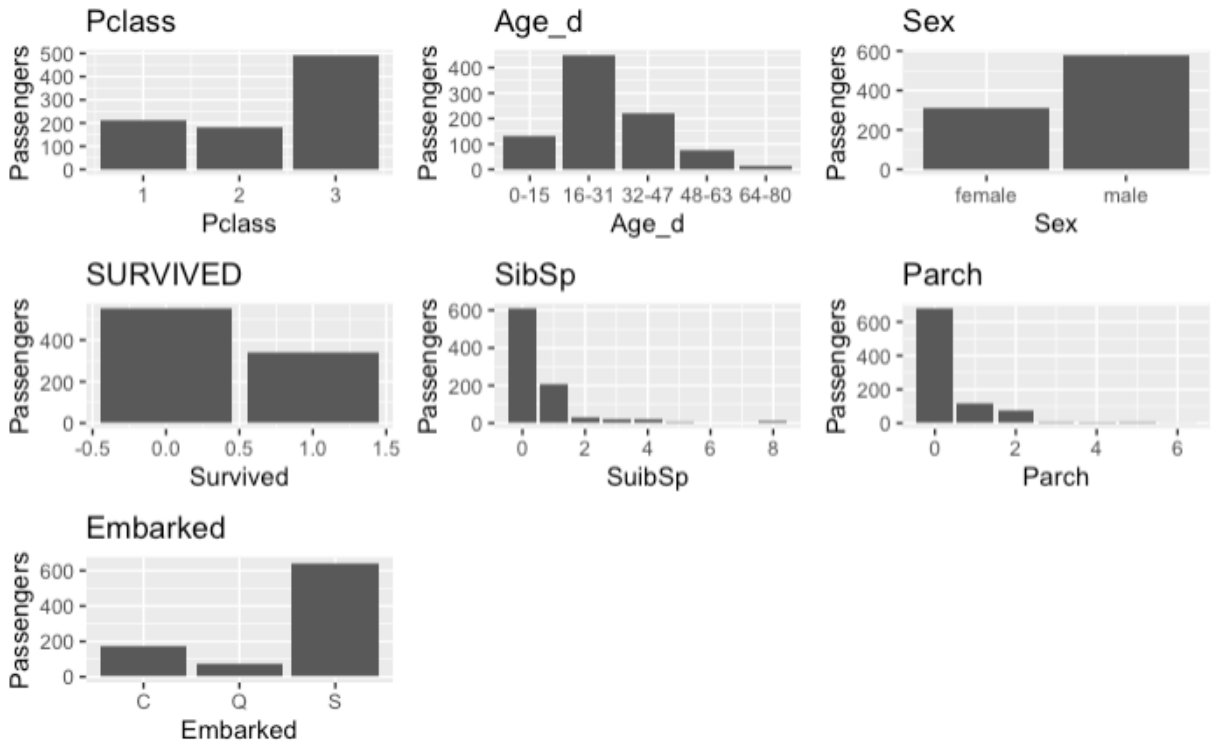

0.4
0.2
0.0
-0.2
-0.4
0
2
4
6
8
SibSp



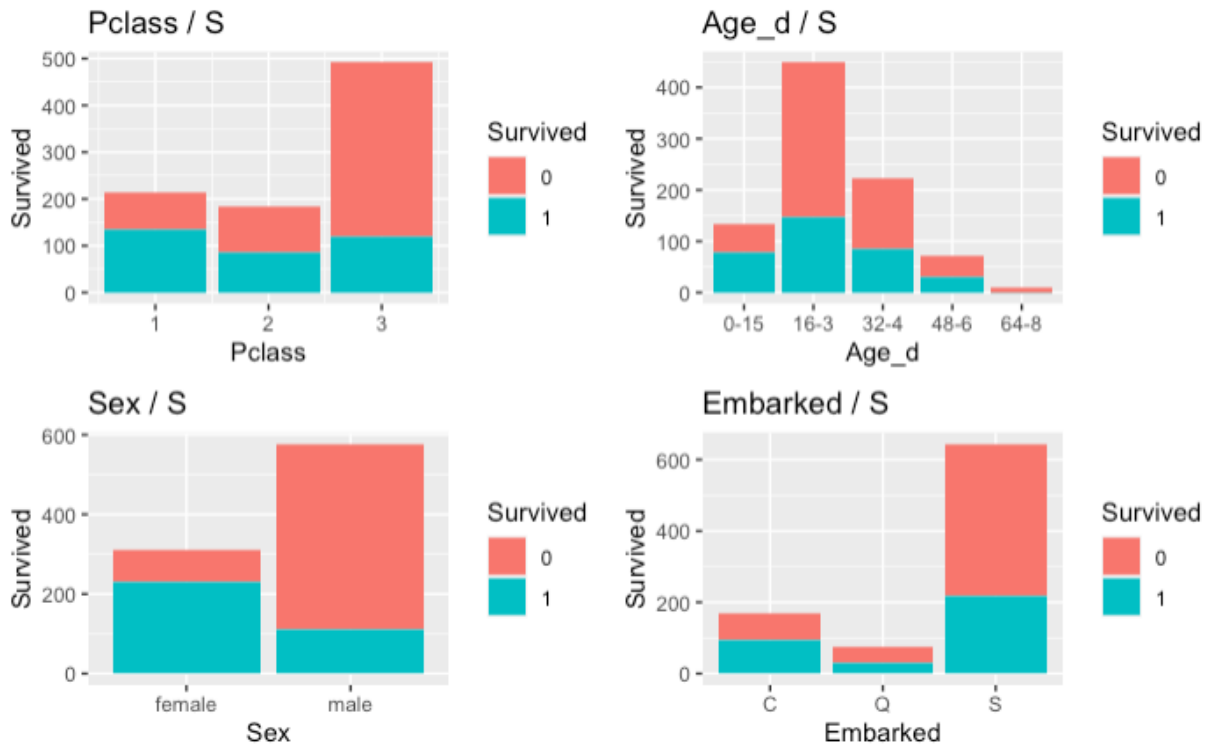
4. Análisis de los datos

4.1. Selección de grupos de datos que se quieren analizar / comparar (p.e. si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se va a aplicar?

Como primer paso, se ha realizado un análisis exploratorio de los datos, que nos ha permitido extraer unas primeras conclusiones a partir de las representaciones gráficas de las diferentes variables:



A partir de estas conclusiones, se ha estudiado la relación de las variables con la variable que se quiere predecir, que es la relativa a la supervivencia de los pasajeros. Para ello, ha sido necesario transformar las variables a factores. Adicionalmente, se ha decidido prescindir de las variables "Parch" y "SibSp", ya que la gran mayoría de los casos se concentran en 0.



Para comprobar qué atributos tienen una mayor influencia en la variable a predecir, se ha procedido a realizar un test de significancia. De ese modo, aumentará la precisión de la predicción, y reduciremos el *overfitting* al contar con menos datos engañosos. Los test de significancia escogidos han sido el cálculo de la V de Cramer y el coeficiente de Phi.

Puesto que todos los resultados son superiores a 0.1, conservaremos todas las variables que tenemos en el *dataset*.

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

Para comprobar si las variables siguen una distribución normal, se ha procedido a aplicar la prueba de Anderson-Darling. Con esta prueba se obtiene un p-valor, que si es superior al valor alfa prefijado, se considera que la variable sigue una distribución normal.

En este caso, se confirma que ninguna de las variables del *dataset* sigue una distribución normal.

Para estudiar la homegeneidad de la varianza, se ha aplicado el test de Fligner-Killeen, que se trata de la alternativa no paramétrica del test de Leneve, utilizada cuando los datos no cumplen con la condición de normalidad.

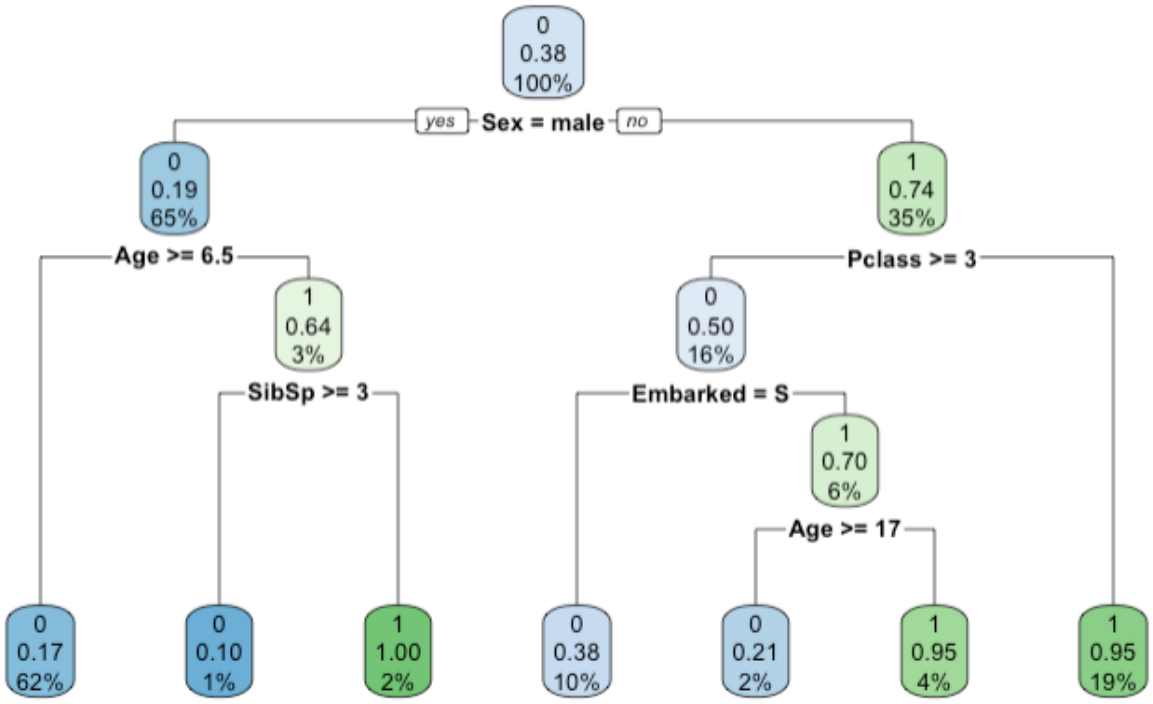
De aquellas variables susceptibles a poder aplicar el test, se ha confirmado que tanto "Pclass" como "Parch" presentan varianzas estadísticamente diferentes para distintos grupos de la variable "Survived", ocurriendo lo contrario con la variable "SibSp".

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.

En primer lugar, se ha aplicado regresión lineal, un modelo matemático que tiene como objetivo aproximar la relación de dependencia lineal entre una variable dependiente y una serie de variables independientes. Siendo el coeficiente de determinación (R-squared) una medida de calidad del modelo, que toma valores entre 0 y 1. Si este valor es alto, indica una fuerte correlación entre ambas variables, que no se da en ninguna de las comparaciones realizadas en el análisis.

Otro método es el coeficiente de correlación, que mide la asociación entre dos variables. Puede tomar valores entre -1 y 1, donde los extremos indican una correlación perfecta, y el 0 indica ausencia de correlación. El signo es negativo cuando los valores elevados de una variable se asocian con valores pequeños de la otra, y es positivo cuando ambas variables tienden a incrementar o disminuir simultáneamente. El coeficiente de correlación más utilizado es el de Pearson, pero requiere que la distribución de las dos variables comparadas sea normal, y además que se cumpla el criterio de homocedasticidad. Por tanto, en nuestro caso tenemos que optar por la correlación de Spearman, que aparece como alternativa no paramétrica que mide el grado de dependencia entre dos variables, siendo la correlación más elevada entre las variables "Parch" y "SibSp", y nula en resto de comparaciones.

Por último, se ha procedido a la generación de un árbol de decisión, que nos permita extraer conclusiones que nos ayuden a inferir cuales son las características base de los pasajeros en relación con su índice de supervivencia, siendo el resultado el siguiente:



5. Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

Las diversas representaciones gráficas, así como comentarios variados sobre la evolución del análisis, pueden consultarse directamente en el código en R, así como en los diversos apartados de este documento.

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

En base a nuestro planteamiento inicial:

> *"Con este dataset se pretende estudiar la relación entre los atributos que influyen en mayor y menor medida en la probabilidad de supervivencia del pasajero. De este modo, podrán extraerse diversas conclusiones, que nos permitirán inferir el índice de supervivencia dependiendo de la clase, la edad o el sexo de los pasajeros, entre otros."*

Podemos concluir que los métodos aplicados han dado respuesta a las preguntas que realizamos inicialmente, puesto que sí disponemos de una relación clara entre los atributos y el índice de supervivencia de los pasajeros del Titanic.

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.

Se adjunta el código en R en el fichero TCVD_PRA2.Rmd.

Contribuciones	Firma
Investigación previa	JP GA
Redacción de las respuestas	JP GA
Desarrollo del código	JP GA