Práctica 2: Limpieza y análisis de datos

Raquel Gómez Pérez y Jorge Serra Planelles

9 de junio de 2020

Resolución

1. Descripción del dataset

El conjunto de datos que vamos a analizar se ha obtenido a partir de este enlace https://www.kaggle.com/c/titanic/.
Este conjunto de datos contiene información de parte de los pasajeros que subieron a bordo del transatlántico Titanic, el 10 de abril de 1912 desde el puerto Southampton y los pasajeros que se incorporaron en los puertos de Cherburgo, Francia, y en Queenstown en Irlanda. Entre estos pasajeros se encuentran pasajeros de muy diferentes clases sociales. En el incidente ocurrido el 14 de abril de 1912, murieron 1514 personas de las 2223 que abordaron. El estricto protocolo de salvamento que se utilizó seguía el principio "Mujeres y niñas primero".
En los dataset cada fila representa a una persona. Hay información de un total de 1309 pasajeros, divididos en dos dataset: 891 en el conjunto de train y 418 en el conjunto de test. Las columnas describen diferentes atributos sobre la persona, tenemos un total de 12 columnas en el dataset de train, y 11 columnas en el dataset de test, ya que no se incluye la columna survived que es la que se desea predecir.

- **PassengerId:** Número de identificación del pasajero.
- **Survived:** Indica si la persona sobrevivió o no al incidente (0 no sobrevivió, 1 sobrevivió).
- **Pclass:** Clase en la que viajaba el pasajero (1^{a}, 2^{a} o 3^{a}).
- **Name:**Nombre del pasajero.
- **Sex:** Sexo del pasajero, femenino o masculino.
- **Age:** Edad del pasajero.
- **SibSp:** Número de hermanos que el pasajero tenía a bordo.
- **Parch:** Número de padres (del pasajero) que estaban a bordo.
- **Ticket:** Número de ticket que el pasajero entregó al abordar.
- **Fare:** Indica el precio que el pasajero pago para obtener su pasaje.

- **Cabin:** Indica la cabina que fue asignada al pasajero.
- **Embarked:** Indica el puerto donde el pasajero abordó (C = cherbourg, Q = Queenstown, S= Southampton).

A partir de este conjunto de datos se pretende responder a la pregunta de qué variables fueron las más determinantes sobre la supervivencia o no de los pasajeros, comprobando por ejemplo cuanto influyó la posición económica en el momento del rescate o si realmente se cumplió el protocolo "mujeres y niños primero.ª la hora de evacuar.

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

Comenzamos con la carga de datos, para realizar la selección de las variables de interés. Vamos a realizar la práctica con el lenguaje de programación Python.

```
import pandas as pd
df_train= pd.read_csv("train.csv")
df_train.head(4)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

```
df_test= pd.read_csv("test.csv")
df_test.head(4)
```

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S

Mostramos también el tipo de datos y un resumen de las variables

```
In [19]: df_train.dtypes
```

```
Out[19]: PassengerId      int64
         Survived         int64
         Pclass           int64
         Name            object
         Sex             object
         Age            float64
         SibSp            int64
         Parch            int64
         Ticket          object
         Fare           float64
         Cabin           object
         Embarked        object
         dtype: object
```

```
In [20]: df_test.dtypes
```

```
Out[20]: PassengerId      int64
         Pclass           int64
         Name            object
         Sex             object
         Age            float64
         SibSp            int64
         Parch            int64
         Ticket          object
         Fare           float64
         Cabin           object
         Embarked        object
         dtype: object
```

```
df_train.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
df_test.describe()
```

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200

Analizadas las variables presentes en el dataset, las variables más representativas son: survived, pcclas, sexo, age y fare. el atributo survival es imprescindible para saber quien sobrevivió al incidente, los atributos pcclas y fare, pueden darnos noción de que tanto influye la posición económica al momento de rescatar a alguien, con el atributo de age y sex, podremos corroborar si realmente al momento de evacuar, mujeres y niños fueron prioridad. Los atributos omitidos: name, sibsp, parch, ticket, cabin, *embarked. El nombre de los pasajeros almacenado en el atributo name puede llegar a influenciar el si sobrevivió o no, dado que algunas personas su nombre es “importante” pero con el fare y clase en la que viajaba es posible saber si tenía buena posición económica. sibsp y parch indican si un pasajero tenía acompañantes durante el viaje o al menos un familiar a bordo, este atributo también podría llegar a influenciar el si se salvo o no, dado que por cuestiones sentimentales podrias dejarle tu lugar en algún bote salvavidas a tu familiar, pero llegar a una conclusión así conlleva a analizar muchos otros factores que son irrelevantes en este caso. Conocer el atributo ticket es algo irrelevante, dado que solo era para control de entrada al barco. El atributo de cabin es usado indica donde estuvo habitando el pasajero durante el viaje, algunos son desconocidos, pero conociendo la clase en que viajaba podrías darte una idea que tipo de cabina se le asignó. El puerto donde embarcó indicado en el atributo embarked, no influye tanto en si sobrevivió o no, dado que en todos los puertos abordaron tanto personas ricas como pobre, niños o adultos, etc.

3. Limpieza de los datos

3.1 Elementos vacíos

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?.

3.2 Valores extremos

Identificación y tratamiento de valores extremos.

4. Análisis de los datos

4.1 Selección del grupo de datos

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

4.2 Normalidad y homogeneidad de la varianza

Comprobación de la normalidad y homogeneidad de la varianza.

4.3 Pruebas estadísticas para comparar los grupos de datos

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.

5. Representación de los resultados

Representación de los resultados a partir de tablas y gráficas.

6. Resolución del problema

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

7. Código

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos a través de este enlace.

Recursos

1. .
2. .

3. .

4. .

5. .

Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	RGP, JSP
Redacción de las respuestas	RGP, JSP
Desarrollo código	RGP, JSP