TA1 - Enzo Cozza - Agustín Fernández

Ejercicio 1

Problema de predicción: se busca predecir la supervivencia de un pasajero al hundimiento del Titanic.

Atributos

Survived: si el pasajero sobrevivió o no al accidente.
Pclass: designa el nivel de clase en la que el pasajero abordó.
Name: nombre del pasajero.
Sex: sexo del pasajero.
Age: edad del pasajero.
SibSp: cantidad de hermanos o cónyuges a bordo.
Parch: cantidad de padres o hijos a bordo.
Ticket: el número de ticket del pasajero.
Fare: precio del ticket abonado por el pasajero.
Cabin: cabina asignada al pasajero.
Embarked: puerto en el que embarcó el pasajero.

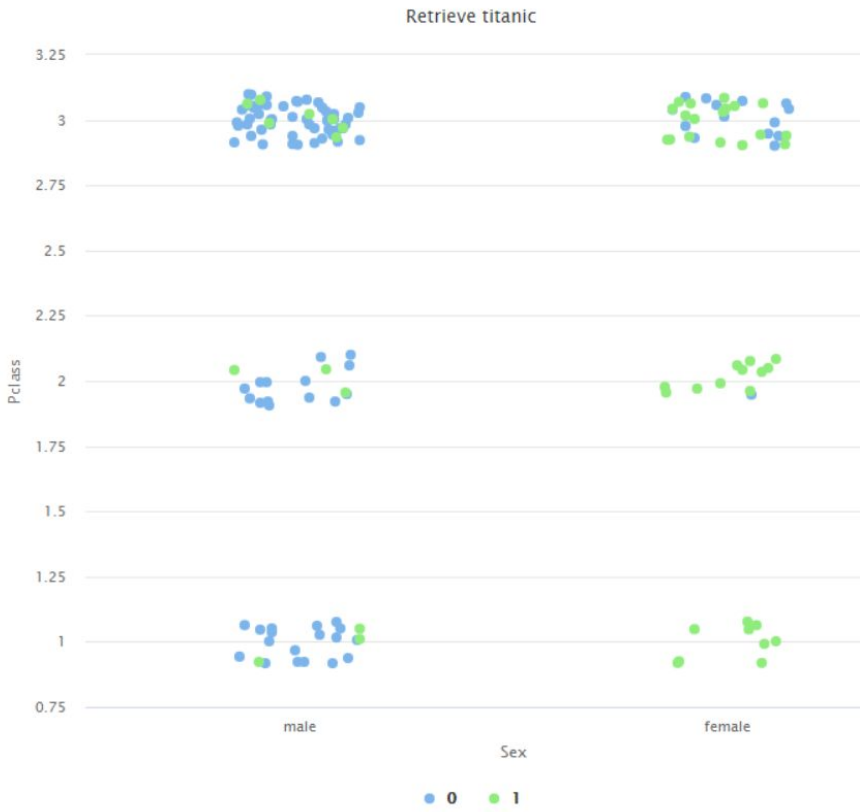
Atributos faltantes:

Age: faltan 28 valores.
Cabin: faltan 120 valores.
Embarked: falta 1 valor.

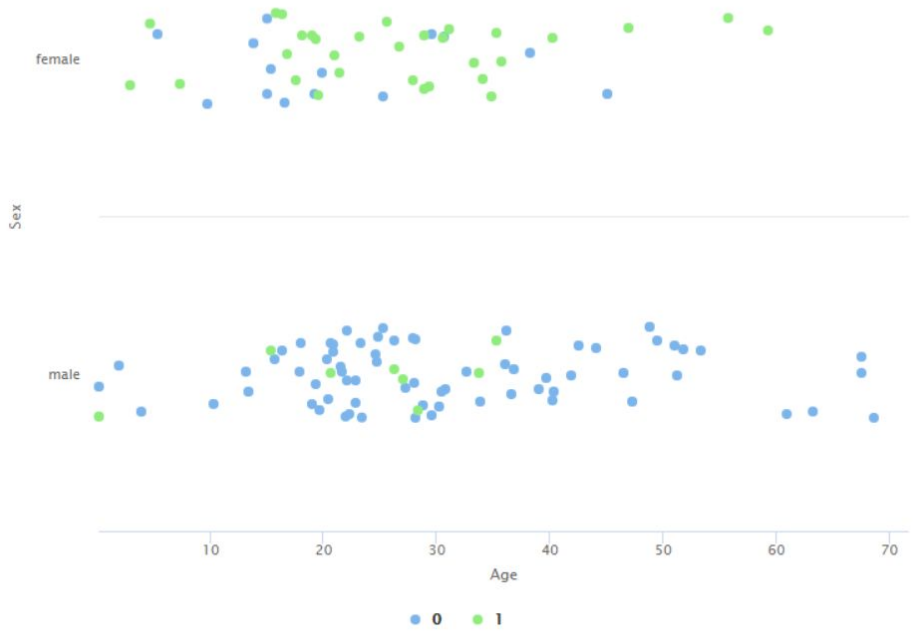
Estadísticas de los atributos:

Survived: 0: 99; 1: 51.
Pclass: 1: 30; 2: 29; 3: 91.
Age: rango 0.83-71; media 28.35; desviación estándar 14.68.
SibSp: rango 0-5; media 0.63; desviación estándar 1.07.
Parch: rango 0-5; media 0.39; desviación estándar 0.87.
Fare: rango 6.75-263; media 28.62; desviación estándar 40.04.
Embarked: S: 106; C: 32; Q: 11.

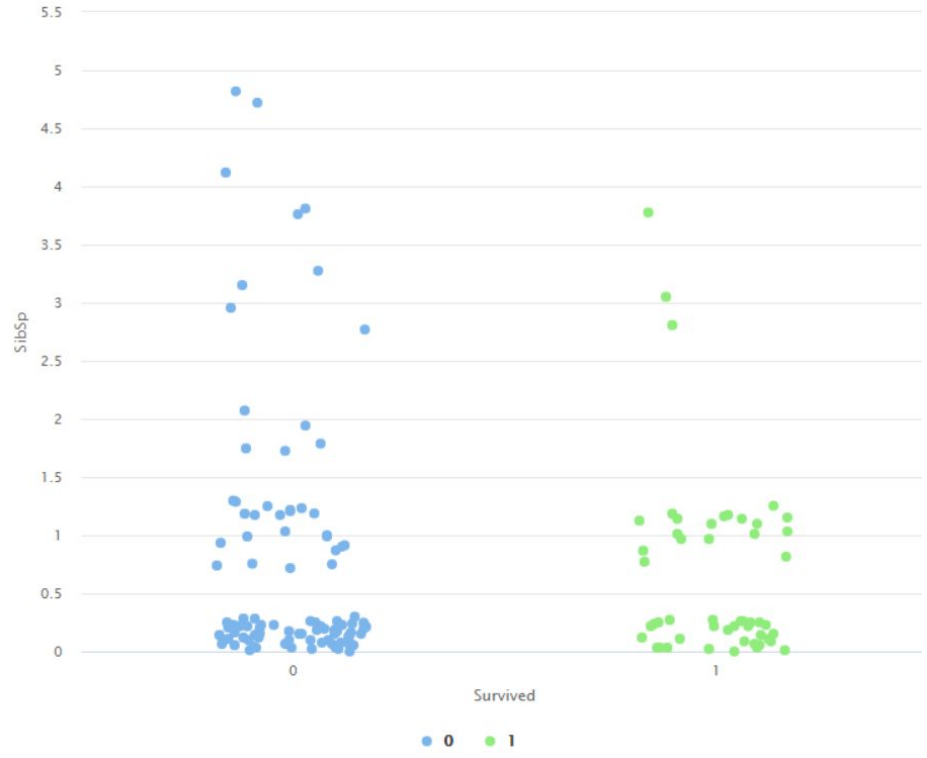
Relaciones con la variable de clasificación:



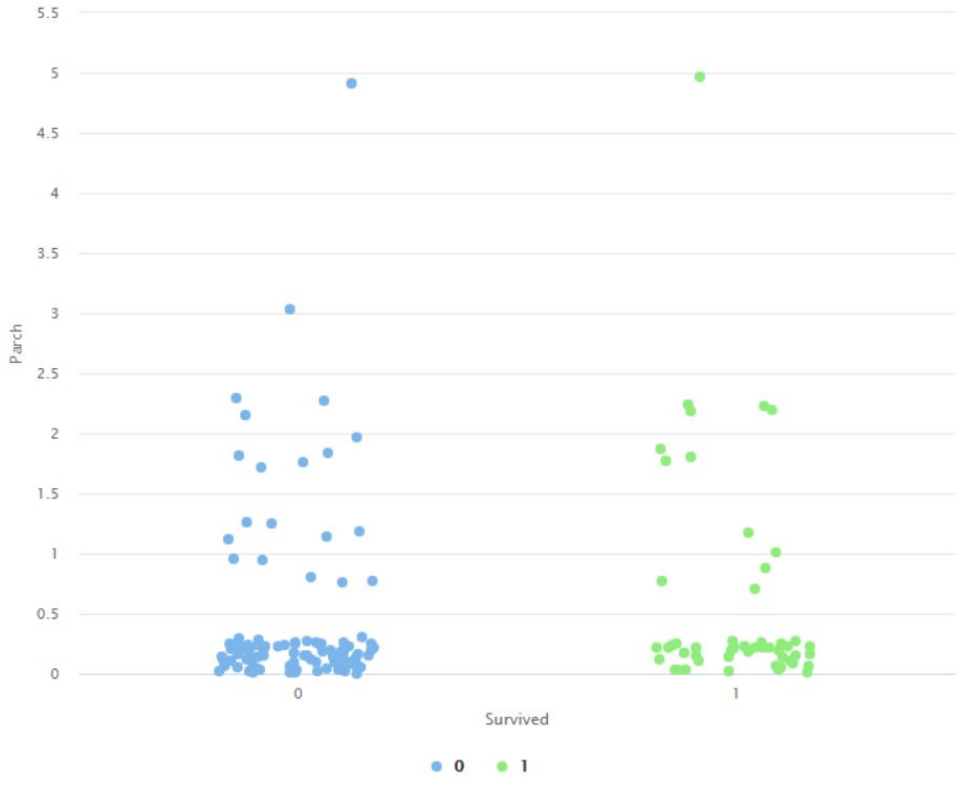
Para este caso, se advierte que el sexo y la clase en la que viajaba la persona influían directamente en la probabilidad de supervivencia de la persona. En este dataset, casi todas las mujeres que viajaban en primera y segunda clase sobrevivieron, mientras que más de la mitad de las que viajaban en tercera también lo hicieron. Sin embargo, en cada una de las clases, menos de la mitad de los hombres sobrevivieron.



Se observa que la mayor cantidad de sobrevivientes pertenece al sexo femenino, pero no se puede observar un patrón definido para la edad.



Puede observarse que la cantidad de hermanos o cónyuges a bordo incide en la supervivencia de manera tal que, cuantos menos de estos seres queridos había a bordo, mayor era su probabilidad de sobrevivir.



Este caso es similar al de arriba, ya que se puede ver que la mayor concentración de supervivientes se encuentra en las personas que no tenían hijos o padres a bordo.

Ejercicio 2

Para el caso de los valores faltantes, se procedió de la siguiente manera:

Age: faltan 28 valores.
Cabin: faltan 120 valores.
Embarked: falta 1 valor.

Para ‘Age’ los valores faltantes fueron reemplazados por la media del conjunto total de datos del dataset. Se consideró el valor más oportuno, ya que la edad de los supervivientes se encuentra en forma distribuida.

Para ‘Cabin’ se tomó la decisión de eliminarlo del dataset, ya que el % de valores faltantes excedía el 75% del total del dataset.

Para ‘Embarked’ se reemplazó el valor faltante por la moda. No se quiso descartar el valor de esa instancia ya que aporta la información de los demás atributos y además no produciría ninguna anomalía por su nuevo valor.

Ejercicio 3

Attribute filter type: es un filtro para seleccionar los atributos del dataset de los cuales se quiere verificar su correlación.

Invert selection: si este parámetro es marcado, entonces se revertirá el filtro seleccionado en el parámetro anterior.

Include special attributes: de marcar este parámetro, los atributos con roles especiales (label, id, etc.) serán incluidos en la matriz.

Normalize weights: si se marca este parámetro, entonces los valores resultantes de los atributos serán normalizados.

Squared correlation: este parámetro indica si se deberá calcular el coeficiente de determinación (R^2).

Attribut...	Age	Embark...	Passen...	Pclass	Name	Sex	SibSp	Parch	Ticket	Fare
Age	1	?	0.073	-0.323	?	-0.158	-0.378	-0.209	?	0.021
Embarked	?	1	?	?	?	?	?	?	?	?
Passeng...	0.073	?	1	0.012	?	-0.157	-0.151	-0.069	?	-0.030
Pclass	-0.323	?	0.012	1	?	0.025	0.092	0.023	?	-0.606
Name	?	?	?	?	1	?	?	?	?	?
Sex	-0.158	?	-0.157	0.025	?	1	0.181	0.105	?	0.009
SibSp	-0.378	?	-0.151	0.092	?	0.181	1	0.398	?	0.267
Parch	-0.209	?	-0.069	0.023	?	0.105	0.398	1	?	0.255
Ticket	?	?	?	?	?	?	?	?	1	?
Fare	0.021	?	-0.030	-0.606	?	0.009	0.267	0.255	?	1

Variables correlacionadas: Fare-Pclass, SibSp-Age, SibSp-Parch

Se entiende que la primera correlación se produce debido a que el precio disminuye a medida que aumenta el número de clase, siendo primera clase la que se debe abonar más.

SibSp-Age están correlacionadas bajo la suposición de que es más frecuente viajar con hermanos a temprana edad y más probable viajar en pareja a mayor edad.

Se podría pensar que los casos de la última correlación comprenden a familias completas en las que viajan padres (al ser más de uno ya se habla de viajar en pareja) e hijos (al ser más de uno ya se habla de hermanos).