IA 1

UT 02 - TA 01

Participantes:
Federico Beconha
Bruno Cattaneo
Gerardo Fernandez
Felipe Mestre

El problema del titanic

Definición

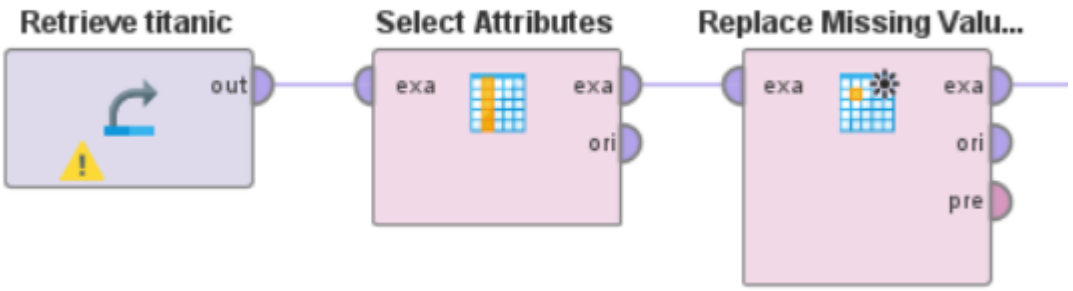
El problema consta de calcular la posibilidad de sobrevivir que tiene una persona en el titanic en base a ciertos atributos.

Atributos

Nombre	**Tipo**	**Descripción**
Pclass	Ordinal (entero)	La clase del pasajero
Survived	Ordinal (entero)	Si sobrevivió o no
Name	Polynomial	Su nombre
Sexo	Nominal	Su sexo
Edad	Ordinal (real)	Su edad
Sibsp	Ordinal (entero)	El número de hermanos que tenía a bordo
Parch	Ordinal (entero)	El número de padres o hijos que tenía a bordo
Ticket	Nominal	El número de pasaje que tenía a bordo
Fare	Ordinal (real)	La tarifa que pagó
Cabin	Nominal	El número de cabina
Puerto de embarcación	Nominal	Puerto donde embarcó
Boat	Nominal	Bote salvavidas, lo tienen solo aquellos que sobrevivieron
Cuerpo	Ordinal (entero)	Número de cuerpo, solo lo tienen aquellos que murieron y su cuerpo fue encontrado.

Sacamos

- Boat: Sacamos este atributo porque estaba altamente correlacionado con la variable que queremos predecir.
- Body: Igual que el anterior.
- Cabin: Tenía 1015 missing values
- Embarked: Es un dato que no tiene relación con el hecho de la muerte de una persona
- home.dest: Es un dato que no tiene relación con el hecho de la muerte de una persona
- name: Es un dato que no tiene relación con el hecho de la muerte de una persona
- parch: Es un dato que no tiene relación con el hecho de la muerte de una persona
- sibsp: Es un dato que no tiene relación con el hecho de la muerte de una persona
- ticket: (alta correlación por el precio)



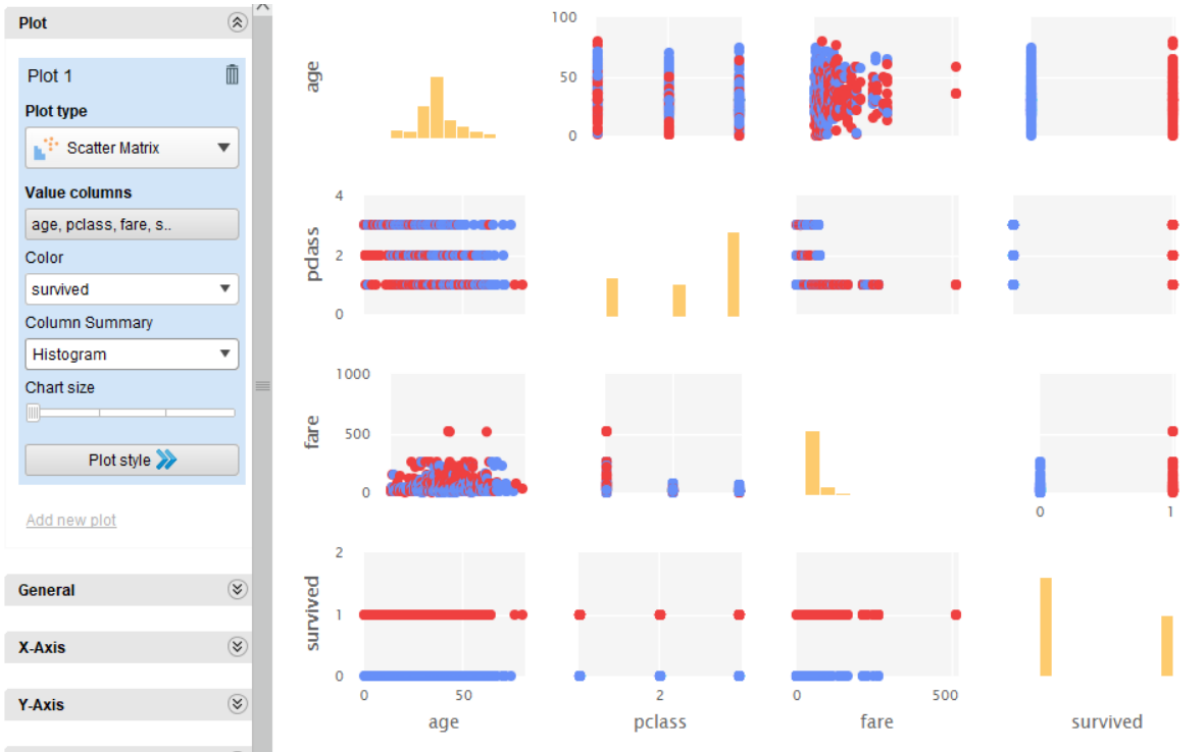
Para quitar los atributos utilizamos el bloque select attributes de Rapid Miner. Este bloque nos permite seleccionar un subconjunto de los atributos del dataset original y lo retorna en su salida. Los atributos que seleccionamos en este operador son:

- Edad: En los hechos es bastante probable que esta variable esté relacionada con la muerte de personas, los más viejos son más susceptibles.
- Tarifa: En el caso de la tarifa lo seleccionamos porque la tarifa tenía incidencia del área del barco donde se situaba la gente
- Clase: La clase del pasaje también tenía incidencia en la ubicación en el barco y en la disponibilidad de botes.
- Sexo: El sexo es determinante porque en general se aplica que los niños y las mujeres van primero.
- Sobrevivió

Luego sustituimos los valores faltantes de los datos con el bloque replace missing values. Seleccionamos los atributos edad, tarifa, clase y sexo y cambiamos los valores faltantes por el valor promedio. Edad tenía 264 valores faltantes en 1310 filas, tarifa 2, sexo y clase 1. Osea que faltaban los datos de la edad de casi el 20% de las personas.

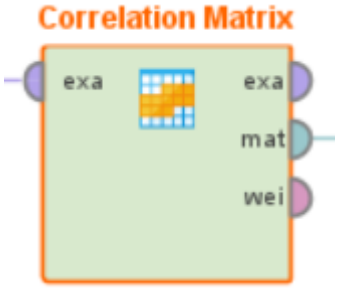
A continuación procedimos a visualizar los datos

Scatter Plot Matrix



En esta matriz vemos cómo se relacionan los atributos entre ellos de a pares. Es bastante útil para interpretar las relaciones que tienen los datos. Podemos ver que hay personas de todas las edades que sobrevivieron, incluso las personas más ancianas aparecen como sobrevivientes. En cuanto a la tarifa podemos ver en el extremo que representa a las personas que pagaron más hay valores de supervivencia. Relaciones de este tipo son a las cuales nos referimos más arriba.

Matriz de correlación



Para continuar con nuestro análisis utilizamos el bloque de matriz de correlación.

Estos son los atributos que permite configurar este operador. El primero permite seleccionar qué atributos van a estar en la matriz. El segundo invierte la selección de atributos hecha, osea que la matriz se va a hacer con los atributos que no estén seleccionados. El tercer atributo incluye categorías que tengan roles excepcionales como id, prediction, label, etc. El cuarto atributo es normalize weights que indica que los resultados finales van a estar normalizados. El último atributo hace que la matriz de correlaciones muestre correlaciones cuadradas en lugar de las simples.

En general las variables están débilmente relacionadas. En este caso vemos una correlación débil entre el sexo de la persona y si sobrevivió o no, en este caso eso es bastante razonable porque sabemos que la mayoría de las personas que murieron fueron hombres. La relación entre la tarifa paga y la clase del pasajero es una relación bastante obvia.

Attribut...	pclass	sex	age	fare	survived
pclass	1	0.124	-0.366	-0.558	-0.312
sex	0.124	1	0.057	-0.185	-0.529
age	-0.366	0.057	1	0.172	-0.050
fare	-0.558	-0.185	0.172	1	0.244
survived	-0.312	-0.529	-0.050	0.244	1