**Hundimiento del crucero ARA General Belgrano**

Franco Marsico

Universidad de José C Paz

**Objetivo**

Introducirse en el análisis exploratorio de datos para la generación de un modelo de aprendizaje automatizado. Para esto se tomará como modelo una base de datos (ficticia) de individuos que se encontraban a bordo del ARA General Belgrano al momento de su hundimiento. Con un set de datos de “entrenamiento” se buscarán patrones que permitan la generación de un algoritmo predictivo que indique quienes murieron y quiénes no en el buque.

**Contexto histórico del ejemplo**

“El 2 de mayo de 1982 durante la guerra de Malvinas, se produjo el hundimiento del crucero ARA General Belgrano como consecuencia del ataque el submarino nuclear británico HMS Conqueror, en el que perdieron la vida 323 argentinos. El hecho desató una enorme polémica en ambos países, ya que el ataque se produjo fuera del área de exclusión establecida por el gobierno británico alrededor de las islas. Los británicos acusaron a la entonces primera ministra Margaret Thatcher de haber dinamitado toda posibilidad de acuerdo. Del lado argentino, el brutal episodio fue visto como una violación al derecho de guerra y las normas que encuadran este tipo de enfrentamiento. Es el único caso de un barco hundido en guerra por un submarino nuclear.” Diario Perfil

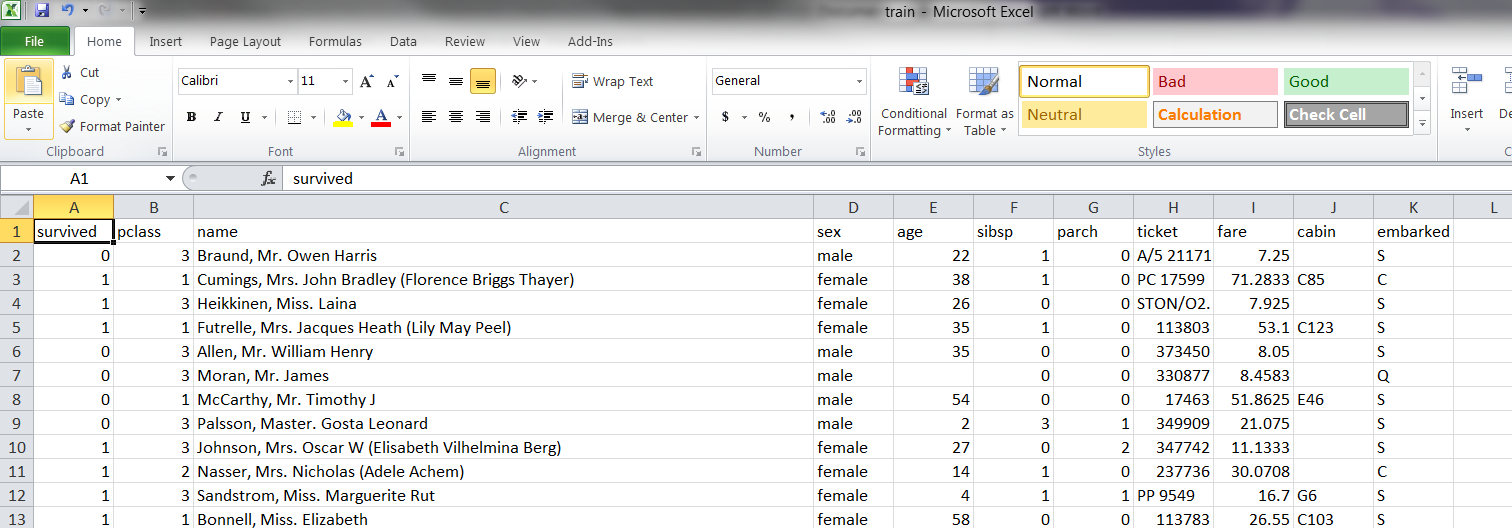
*Aclaración: los datos del ejercicio son ficticios, no así el resumen histórico.*

**El reto**

La competencia es simple: usen los datos de los pasajeros del General Belgrano (nombre, edad, sexo, clase socioeconómica, etc.) para intentar predecir quién sobrevivirá y quién morirá.

**Datos, ya limpios por suerte ;)**

entrenamiento.csv y prueba.csv. El nombre de archivo .csv significa valores separados por comas, donde cada valor (edad, Sexo, etc.) en una fila está separado por una coma. Esto permite que Excel interprete los datos como columnas. Si carga el archivo se vería así:



**Tengo los datos, ¿qué tengo que hacer?**

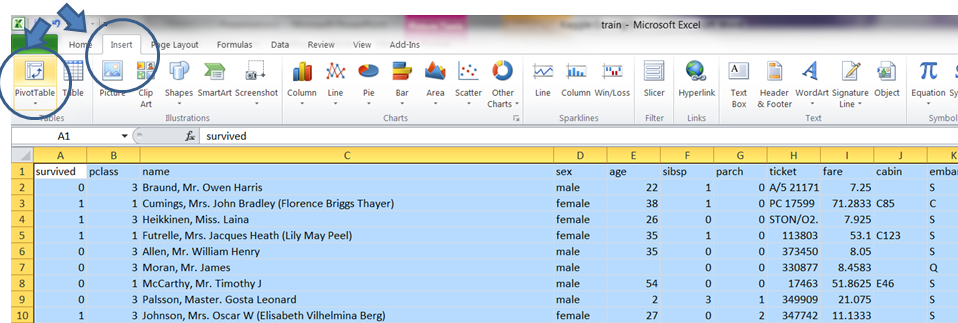
Entrenamiento.csv tiene los detalles de un subconjunto de los pasajeros a bordo (891 para ser exactos) y si sobrevivieron o no. Usando los patrones que encuentren en los datos, tendrán que predecir si los otros 418 pasajeros a bordo (que se encuentran en prueba.csv) sobrevivieron.

**Entonces necesito encontrar patrones en los datos ... ¿Por dónde empiezo?**

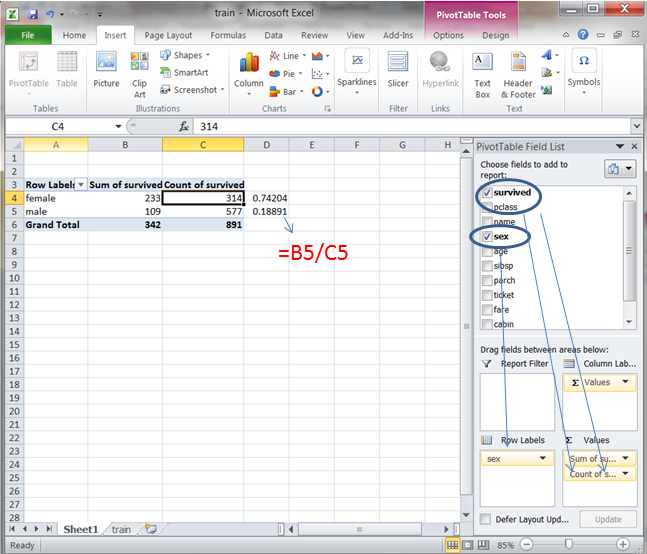
Abran entrenamiento.csv en Excel. Como pueden ver, estan los detalles de una serie de pasajeros a bordo del ARA. Los datos que se muestran son:

Pclass, la clase de su boleto (1 °, 2 ° o 3 °), su nombre, su sexo, edad, Subsp (que es el número de hermanos / cónyuges que tenían a bordo), hijos (que es el número de padres / hijos que tenían a bordo), su tipo de boleto, la tarifa que pagaron por su boleto, su número de cabina y el punto de Embarcación (donde subieron). También si sobrevivieron o no (1 = sí, 0 = lamentablemente, no). Esta es la información que usarán para hacer sus predicciones. Queremos encontrar si existe una relación entre una de las variables y la supervivencia final.

Se puede empezar si hay alguna diferencia en cantidad de muertes por sexo. Excel tiene una herramienta útil para este tipo de exploración llamada tabla dinámica. Resalten todo el conjunto de datos y vaya a 'Insertar' -> 'Tablas dinámicas'.



Esto debería crear una nueva hoja de cálculo en su documento. A la derecha deben aparecer todas las variables que seleccionó y cuatro cuadros en la parte inferior. Estamos interesados ​​en quién sobrevivió, por lo que desean ver cómo varía este valor a medida que seleccionan otras variables. En el lado derecho, arrastren la palabra 'sobrevivió', que tiene una casilla de verificación al lado, hasta el cuadro inferior derecho. Esto mostrará la suma de supervivientes. Ahora, para ver cuántas mujeres y hombres sobrevivieron, arrastren la variable 'Sexo' a las 'Etiquetas de fila'. Dado que Supervivencia es 1 o 0, la suma de esto es el número total de sobrevivientes. Si desean encontrar la proporción, arrastren la variable supervivencia nuevamente al cuadro de valores (por lo que ahora hay dos allí), hagan clic en la flecha hacia abajo, seleccionen la configuración del campo de valor y cambien 'Sumar' a 'contar'. Esto le indicará el número total de filas (o pasajeros en este caso). Ahora, en la celda al lado de la tabla, simplemente pueden escribir = B4 / C4, lo que mostrará la proporción de mujeres que sobrevivieron.

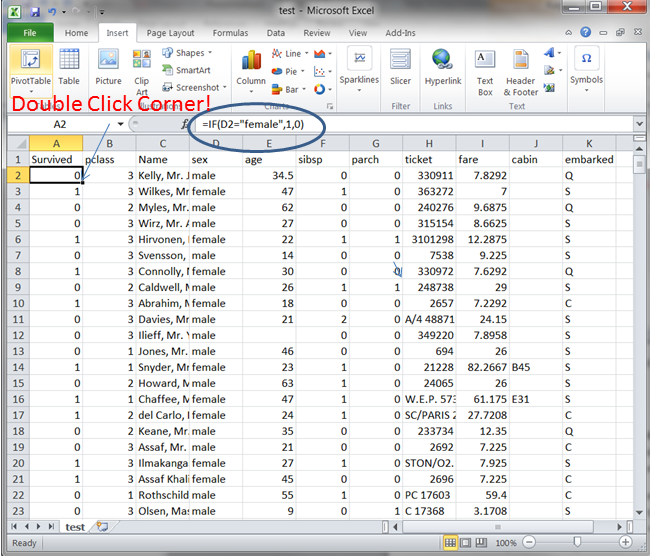


¡De esto se puede ver que casi el 75% de las mujeres sobrevivieron! Sin embargo, solo el 19%.

**Haciendo predicciones**

El Sexo parece ser un fuerte indicador de supervivencia. Para hacer las primeras predicciones basadas en esto abran prueba.csv en Excel e inserten una nueva columna en la primera columna con el nombre de columna “supervivencia”.

Para hacer un modelo que establezca que si la pasajera es mujer, ella sobrevive en la primera celda, escriban una declaración lógica 'si'. Escriba = if (E2 = "femenino", 1,0), lo que significa que si (Sexo = femenino, entonces haga esta celda = 1, si no, entonces haga esta celda = 0). Luego hagan doble click en la esquina inferior derecha de este cuadro para que se arrastre a través de todas las filas. Ahora ha creado una fórmula calculada para cada fila en Excel, pero se desea un valor definitivo, 0 o 1. Así que vuelvan a pegar la misma columna sobre sí mismo con Pegado especial > Valores. Guarden el archivo como ModeloSexo.csv.

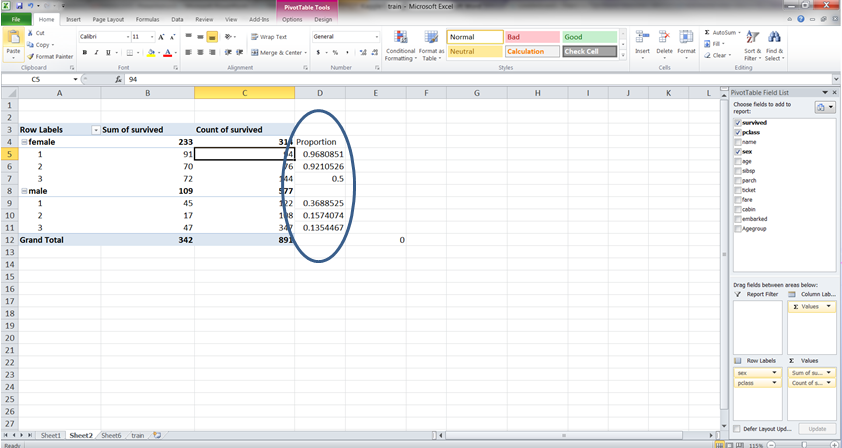


**Otro Modelo**

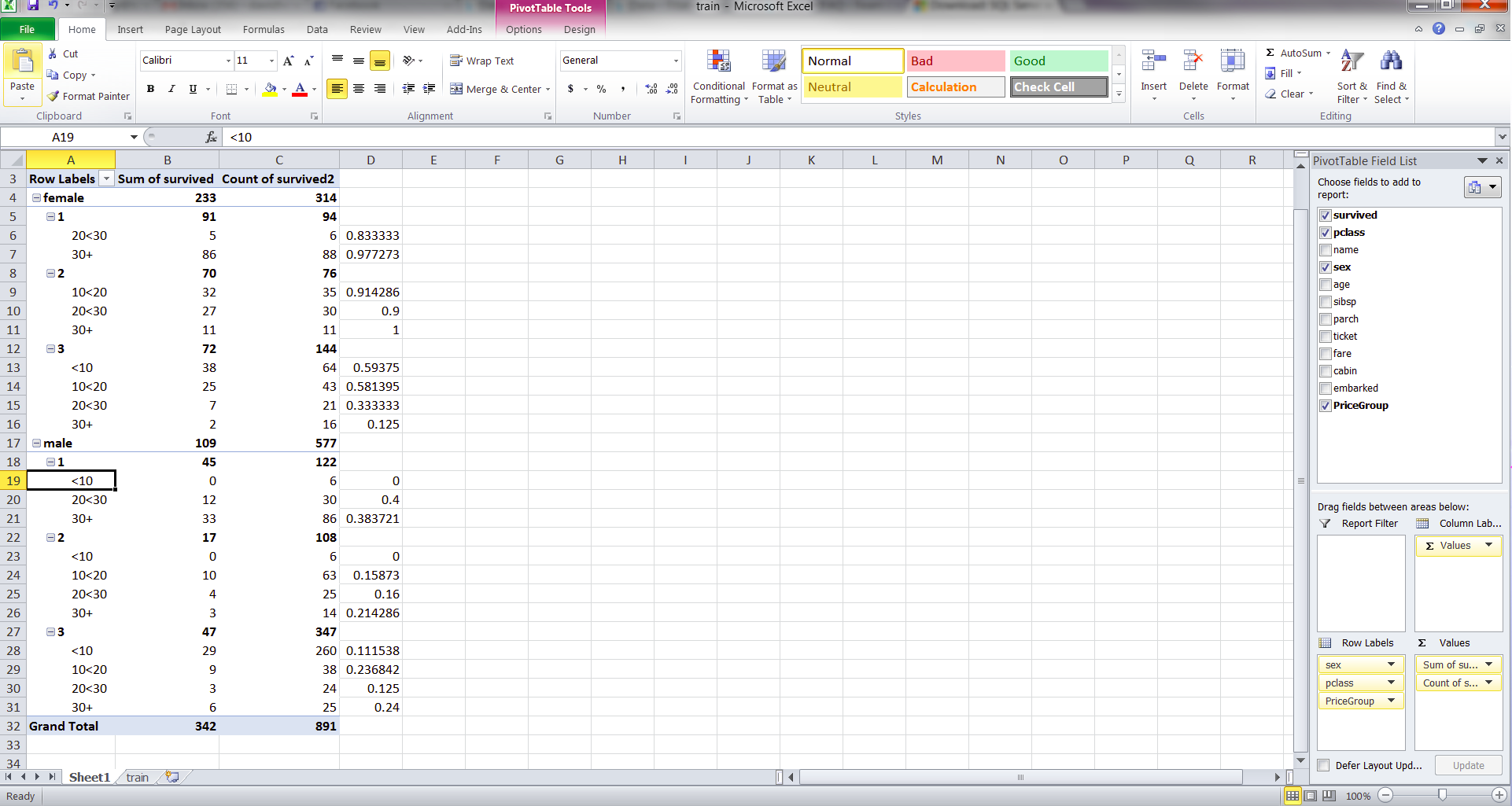
Es probable que el primer modelo no sea el mejor. A medida que analizamos más aprendemos sobre los datos, aparecen publicaciones que discuten técnicas que brindan información y pensaremos en nuevas formas de mejorar su modelo. Intentemos mejorar el nuestro.

Volviendo a entrenamiento.csv extendemos la hipótesis anterior de que tal vez la edad también puede ser predictiva, pero pueden agregar la edad a la tabla dinámica. Arrastren la edad a las etiquetas de Fila (donde estaba el sexo) y mostrará la cantidad de pasajeros que coincidieron tanto en sexo como en edad. Sin embargo, como pueden observar, la edad no está dividida de ninguna manera. Es posible que deseen agrupar, tal vez comenzando solo con adultos y niños. Al igual que antes, hagan una columna adicional con una declaración 'si'. En este caso, se leerá = SI (F2> 18, "adulto", "niño"), donde F2 era la columna de edad. Ahora vuelvan a crear la tabla dinámica (tenga en cuenta que debe volver a crearla) y agreguen las mismas variables antes, ¡puede ver si hay más patrones!

Según la tabla, las mujeres adultas (mayores de 18 años) tenían un 78% de posibilidades de supervivencia y los hombres adultos solo tenían un 18% de posibilidades de supervivencia. Puede ver que esto no ha cambiado mucho de las proporciones originales. Esto nos dice que no hay mucha información adicional en la variable edad. Veamos una variable más: la clase de pasajero. Agregando esto a la tabla dinámica podemos ver los resultados:



Como puede ver, las proporciones ahora cambiaron drásticamente, lo que significa que hay algún algo de predictivo en esta variable. Sin embargo, esto todavía no supera la división entre hombres y mujeres. Ahora vamos a agrupar lo que la gente pagó por su boleto, de modo que cada clase se divida en pagos de: (i) menos de $ 10, (ii) entre $ 10 y $ 20, (iii) entre $ 20 y $ 30, y (iv) más de $ 30:



Ahora tenemos las proporciones para muchas variables diferentes. Decido asumir que cualquier grupo con más de la mitad de sobrevivientes sobrevivirá siempre en la predicción (todos igual a 1), y todos aquellos con menos de la mitad nunca sobreviviran (todos igual a 0). Entonces, yendo a prueba.csv, hagan una instrucción SI anidada en cada celda de una nueva columna llamada “Supervivencia”:

**= SI (E2 = "masculino", 0, SI (C2 = 3, SI (J2> 20,0,1), 1))**

Vuelvan a pegar la columna de fórmula con Pegar como valores, luego guárdenlo en un archivo csv.