UOC – MÁSTER EN CIENCIA DE DATOS
TIPOLOGÍA Y CICLO DE VIDA DE DATOS – PRÁCTICA 2

Titanic: Machine Learning from Disaster

Xavier Ricci y Guillermo Monzón. **Junio 2020**

UOC – MÁSTER EN CIENCIA DE DATOS
TIPOLOGÍA Y CICLO DE VIDA DE DATOS – PRÁCTICA 2

Indice

1. Descripción del dataset.

Hemos escogido estudiar el dataset del Titanic de kaggle.com, porque se trata de un reto en el que participan personas que se dedican al análisis de datos de todo el mundo y nos queríamos participar en este reto. Se trata de una forma de medir nuestras capacidades dentro de este mundo profesional, ya que, aunque esta es una competición para principiantes, en ella participan muchos profesionales y equipos de personas. Nuestro reto inicial, era conseguir estar dentro de la lista del 10% con mejores resultados de los 23 mil equipos participantes. Cabe destacar que, aunque la mayoría de personas en esta lista están participando con códigos propios, también hay muchas personas que utilizan métodos tramposos, ya sea simplemente copiando los resultados finales y consiguiendo un imposible 100% de aciertos o utilizando los códigos de otros con alto rendimiento. Es de lógica entender que en cualquier situación natural estamos hablando de probabilidades a partir de una cantidad limitada de datos. Aunque se tuvieran miles de datos de cada uno de los pasajeros del Titanic una hora antes del accidente, sería imposible predecir todas las circunstancias que llevan a la supervivencia de cada una de las personas. Los que resbalaron, los que decidieron ceder su asiente en un bote salvavidas, los que pisaron a otra persona para ocupar su lugar… son circunstancias impredecibles que no se pueden calcular a partir de una docena de características. Entendemos que a base de mandar resultados se puede ir ajustando las respuestas a las correctas, aunque eso no sería la creación de un modelo de predicción, sino una especie de proceso de descubrimiento de una clave, y este no es nuestro objetivo. A decir verdad, nuestra sensación es que solo con los datos originales, nos parece fantástico poder superar el 80% de aciertos.

Para llevar a cabo este estudio, los responsables de Kaggle, nos ofrecen un listado de 891 personas con 11 características y el resultado de supervivencia o no de cada uno de ellos. A partir de estos datos se nos propone predecir la supervivencia de 418 pasajeros a partir de las mismas 11 caracteristicas. Para decidir los datos de interés a analizar entendemos que primero es necesario conocer los datos en profundidad y ver su estado, así, que empezamos directamente con el limpiado de los datos, para posteriormente decidir cuales son realmente útiles para nuestra predicción.

Los atributos del dataset son los siguientes:

- **PassengerId:**
- **Survived:** indicador de supervivencia (0 - sobrevive / 1 - muere).
- **Pclass:** indicador de la clase del billete (1 - primera / 2 - segunda / 3 - tercera).
- **Name:** nombre del pasajero.
- **Sex:** sex (male - hombre / female - mujer).
- **Age**: edad en años (si es menor de un año se estima de forma fraccionaria).
- **SibSp:**
- **Parch:**
- **Ticket:** número de ticket.
- **Fare:** tarifa.
- **Cabin:** número de camarote.
- **Embarked:** puerto de embarque (C - Cherbourg / Q - Queenstown / S - Southamton).

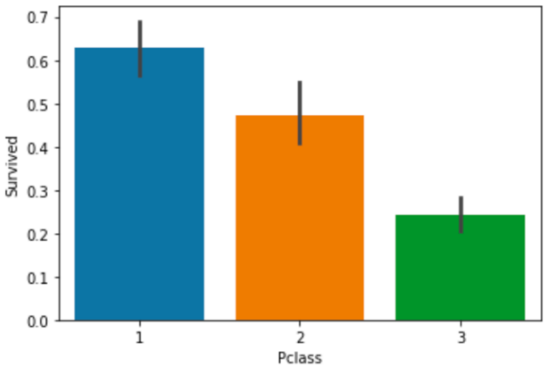
2. Integración y selección de los datos de interés a analizar.

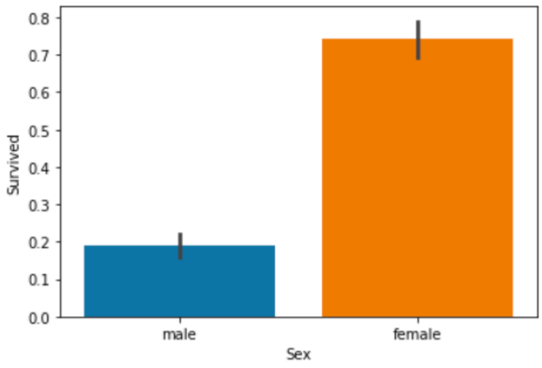
En un primer visionado de los datos, vemos que entre los datos hay muchos datos vacíos, especialmente en *Cabin* y en *Age*, también observamos algunos valores vacíos en *Embarked* y en *Fare*.

Vemos que un 62% de los pasajeros murieron mientras que un 38 % se salvaron.

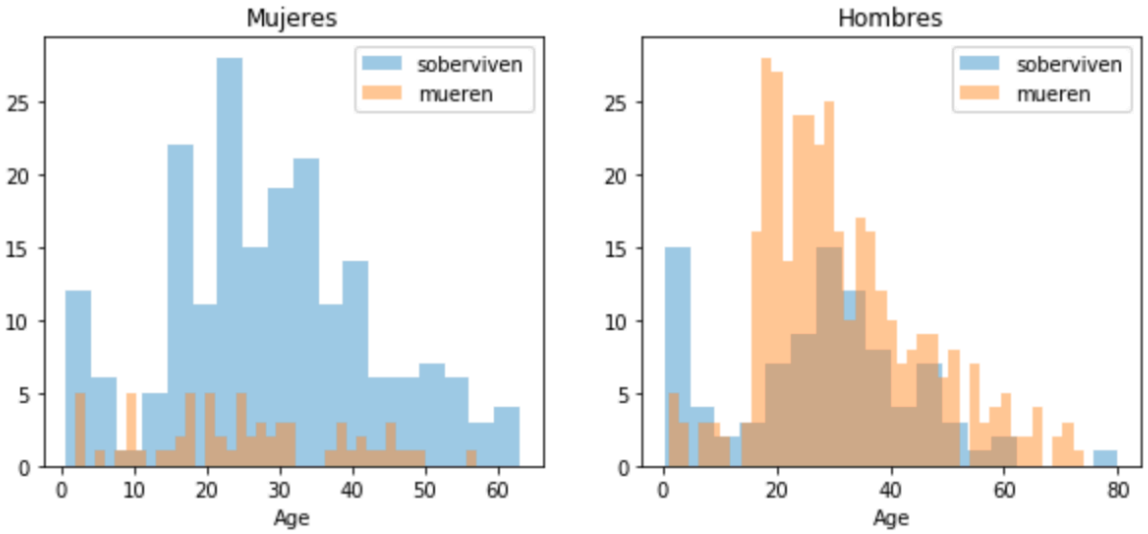
Podemos observar como un 63% de los pasajeros de PRIMERA CLASE sobrevivió al naufragio. Un 47 % para SEGUNDA CLASE y los que viajaban en TERCERA CLASE tan sólo sobrevivió un 24 %.

También se observa que las mujeres sobreviven significativamente más que los hombres.

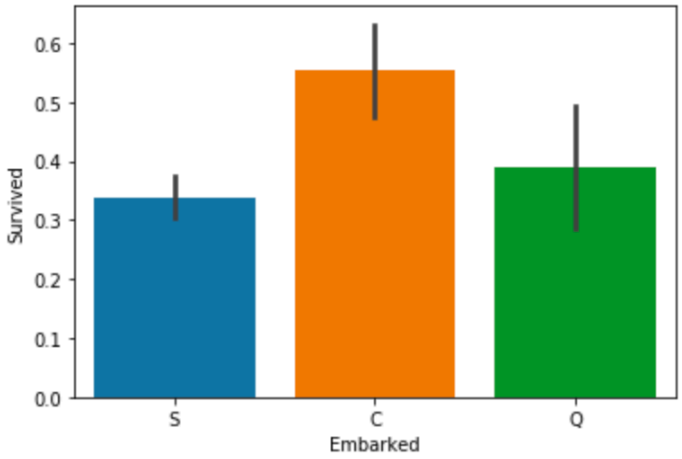




Observamos que los niños también tienen tasas de supervivencia muy superiores comparados con los adultos.



Otra característica que podemos observar de un análisis previo de los datos es que lo pasajeros que embarcan en el puerto francés de Chebourg sobreviven de forma notoriamente superior a los pasajeros del resto de puertos



Aunque en los primeros análisis de los datos ya se puede ver que el sexo, la clase y la edad son elementos determinantes en la supervivencia, para mejorar el resultado del análisis final es necesario limpiar primero los datos para ver mas claramente la importancia de otras características que no parecen tan relevantes a primera vista.

3. Limpieza de los datos.

Cada caso ha sido tratado específicamente. En los casos en que solo faltaba algún dato puntual, se ha recogido la tendencia de la mayoría, o la tendencia de grupos de personas con características comunes, cuando esto se ha considerado relevante (como en el caso de la edad). Siempre teniendo en cuenta la posible influencia de los valores extremos en estos resultados y modificándolos previamente cuando se ha considerado pertinente (como en el caso del precio de los billetes).En esta línea, en algunos casos incluso hemos aplicado regresión lineal para para estimar los resultados.

A continuación detallamos estos procesos:

Respecto al nombre:

La variable name, que inicialmente parecía inútil se convierte en muy relevante gracias a la aparición de títulos descriptivos que nos pueden ayudar a determinar la edad, la clase o incluso la profesión de las personas. Esta información nos servirá para completar los datos vacios.

Para comprender mejor el significado de cada uno de los títulos y así tener una mayor comprensión de los datos, hemos creado un pequeño glosario con los títulos:

Agunas notas sobre los títulos:

Nobles: the Countess, lady is what you use to address someone of Nobility.

Mlle es equivalente a Miss

Madame es equivalente a Mrs. Usually, a servant (in Britain) addersses her Mistress as Madame. But only if the mistress is married.

Master: title for an underage male. If a person is under 18. En el caso del titanic todos los masters son de menos de 14.5 años.

Colonel is a honorary title of conferred by several states in the US and certain military units of the Commonwealth of Nations.

"Ms" is a recent term for those ladies who don't think anyone needs to know whether they are married or not, like the generic

Jonkheer is an honorific in the Low Countries denoting the lowest rank within the nobility

Hemos visto que en muchos casos este está relacionado con la supervivencia.

de 517, Mr sobreviven: 81, un 0.15667311411992263
de 125, Mrs sobreviven: 99, un 0.792
de 182, Miss sobreviven: 127, un 0.6978021978021978
de 40, Master sobreviven: 23, un 0.575
de 1, Don sobreviven: 0, un 0.0
de 6, Rev sobreviven: 0, un 0.0
de 7, Dr sobreviven: 3, un 0.42857142857142855
de 1, Mme sobreviven: 1, un 1.0
de 1, Ms sobreviven: 1, un 1.0
de 2, Major sobreviven: 1, un 0.5
de 1, Lady sobreviven: 1, un 1.0
de 1, Sir sobreviven: 1, un 1.0
de 2, Mlle sobreviven: 2, un 1.0
de 2, Col sobreviven: 1, un 0.5
de 1, Capt sobreviven: 0, un 0.0
de 1, the Countess sobreviven: 1, un 1.0
de 1, Jonkheer sobreviven: 0, un 0.0

En los hombres:

Sr (Adultos en general): Con muy poco indice de supervivencia. Apenas el 15% sobreviven.

Master (Jovenes): Aunque a primera vista parecia que Master podia referirse a la clase social aparte de a la edad comprobamos a continuación que simplemente se refiere a los menores de 15 anyos. De los jovenes el 57% sobreviven

Reverendos: Vemos que todos mueren (6 de 6 ya empieza a ser un valor interesante) 0% sobreviven

Doctores: 2 de 6 el 33% sobreviven

Major, Col y Capt (Militares) vemos que los rangos militares sobreviven en un 3/5, un 60% sobreviven

Los nobles: Sir, Don y Jonkheer 1/3 baja al 33.3% sobreviven

En las mujeres:

Mrs y Mme (Casadas): Con muy alto indice de supervivencia del 79% sobreviven

Miss y Mille (Solteras 0 y 63 años): 69% sobreviven

El resto son muy pocos casos para generalizar pero vemos que una doctora sobrevive y que la Lady y la Countess sobreviven.

Con toda esta información, hemos reagrupado los títulos en conjuntos que nos han parecido mas coherentes: 'Mr' 'Mrs' 'Miss' 'Master' 'Arist' 'Rev' 'Dr' 'Army'

En algunos casos incluso hemos modificado los datos originales para mantener esta coherencia, como por ejemplo hemos converntido en Master a todos los pasajeros menores de 15 años.

Respecto a la edad:

La problemática de la variable *Age*, como hemos visto en el apartado de anterior, estriba en que tiene un número elevado de valores perdidos, 263 concretamente. A continuación procedemos a resolver esta problemática asignando un valor para cada uno de los registros desconocidos de la variable edad.

Para rellenar los datos vacios con la máxima precisión posible, los agrupamos según el titulo par ver si existe alguna relación entre este y la edad, como hemos visto que sucedia con los *Master*.

```
La mediana de edad de los 576 Mr con Age es de 29.0. Faltan: 176
La mediana de edad de los 171 Mrs con Age es de 35.0. Faltan: 27
La mediana de edad de los 213 Miss con Age es de 22.0. Faltan: 51
La mediana de edad de los 58 Master con Age es de 6.0. Faltan: 8
La mediana de edad de los 6 Arist con Age es de 39.5. Faltan: 0
La mediana de edad de los 8 Rev con Age es de 41.5. Faltan: 0
La mediana de edad de los 7 Dr con Age es de 49.0. Faltan: 1
La mediana de edad de los 7 Army con Age es de 53.0. Faltan: 0
```

Vemos que efecticamente los titulos estan relacionados con la edad. Así que utilizamos esta referencia para rellenar los datos. Primero corregimos los que hay menos casos: *Dr* y *Master*, asignándoles la mediana.

Para obtener las edades no registradas de pasajeros con los títulos de Mr, Mrs y Miss (los mas abundantes) aplicaremos una regresion lineal para cada grupo, a partir de las variables *Pclass*, *Fare* y *Family*. Esta última la crearemos a partir de sumar las variables de *SibSP* y *Parch*.

Para poder hacer la regresion nos tuvimos que asegurar primero de que no había elementos vacios en las columnas que utilizadas para el modelo.

Para completar el valor del único vacío en el precio del billete, utilizamos la mediana del precio de pasajeros con las mismas características que el vacio.

Una vez completados todos los vacios en las edades, realizamos un nuevo ajuste sobre la clasificacin de datos generando la equivalencia femenina del valor Master entre las mujeres de menos de 15 años. A estos valores los clasificaremos como Girl

De igual modo vamos a crear una separación para los pasajeros mayores de 60, clasificándolos como MrSenior para los hombres mayores de 60 años y MrSSenior para las mujeres.

Entendemos que estas agrupaciones tienen un sentido lógico, aunque subjetivo. En este ultimo caso, estamos generando agrupaciones que consideramos que comparten características comunes que influyen en la supervivencia.

Respecto al precio:

Hemos podido observar que hay un grupo de personas que no pagan el billete y que, en algunos casos, en la variable Ticket aparece la anotación LINE. En general son hombres que viajan solos y embarcaron en el puerto de Southamton y que en su mayoria murieron (al margen de la clase en que viajaran). Se podría deducir de ello que se trata de personas con alguna relación especial con la compañía propietaria del barco. Se opta por identficarles como Empleados en la variable Title.

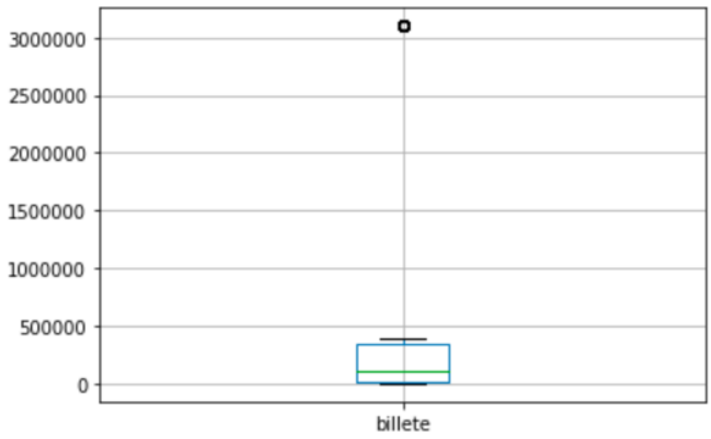
Adicionalmente, para no distorsionar la variable *Fare* se decide asignarles,en este caso, el valor de la media de la variable *Fare* según la clase a la que pertenezcan.

Una vez eliminado el precio ilógico de 0, vemos que hay un outlayer entre los precios altos, los analizamos y los justamos a los siguientes precios más elevados para evitar que se distorsione el analisis general de este dato

Respecto al billete:

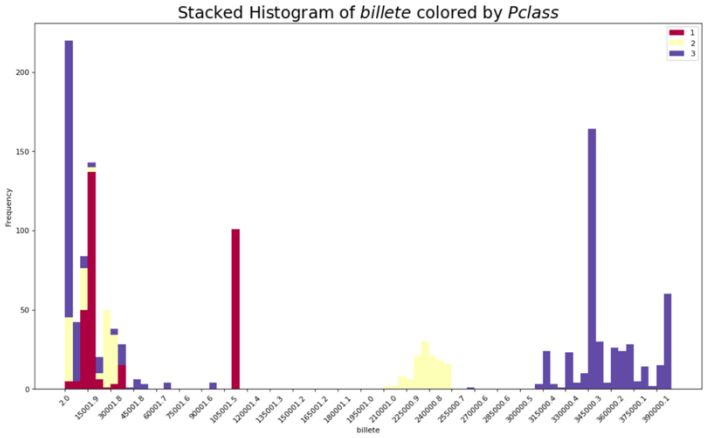
Al intentar analizar los Tickets en detalle vemos que se trata de un dato muy confuso, que en algunos casos es numérico y en otros es alfanumérico, para unificar el criterio de forma simple, recogemos solo los números finales, ignorando las letras iniciales.

También vemos que entre los números de los billetes hay números extremos que va a 3 millones, frente al siguiente que no llega a 400 mil.

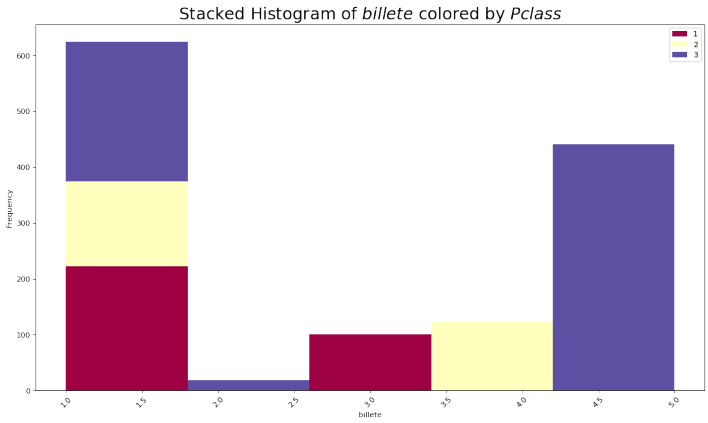


Modificamos el número extremo a 400.000

A partir de esta representación visual, vemos que el número de billlete tiene claramente alguna relación con la clase, especialmente a partir del billete número 40000 aproximadamente.



Siguiendo esta apreciación visual categorizamos la variable *billete* en 5 grupos:



Respecto a la cabina:

Suponiendo que las letras hacen referencia a las distintas cubiertas del barco y considerando este hecho un factor primordial un la supervivencia, recogemos estos datos y rellenamos los desconocidos con la letra U de Unknown.

Llamamos Planta a la transformación en categoría numérica de este atributo.

Respecto al embarque:

La variable *Embarked* tan sólo tiene dos valores perdidos. Vamos a tratar darle un valor a estos dos registros a partir de los de la variable *Fare*.

4. Análisis de los datos.

Antes de probar modelos, para evitar el sobre entrenamiento, ahora que ya conocemos los datos seleccionados, seleccionamos los apartados que nos parecen más relevantes y eliminamos los que aportan menos informacion.

```
DF2=DF2.drop('PassengerId', axis=1) # Esta información es irrelevante
DF2=DF2.drop('Name', axis=1) # La información mas relevante ya esta en Title
DF2=DF2.drop('SibSp', axis=1) # Esta información ya esta en Family
DF2=DF2.drop('Parch', axis=1) # Esta información ya esta en Family
DF2=DF2.drop('Ticket', axis=1) # Esta información ya esta en billete
DF2=DF2.drop('Cabin', axis=1) # Esta información ya esta en Planta
DF2=DF2.drop('Fare', axis=1) # Esta información ya esta en precio
```

Y nos quedamos con:

Pclass, Sex, Age, Embarked, Title, Family, precio, billete, Planta.

Preparando los datos mediante dumies para los atributos categóricos:

Dando un data set de 23 atributos numerales (24 en el training con survived)

```
Survived           891 non-null float64
Pclass            1309 non-null int64
Age               1309 non-null int64
Family            1309 non-null int64
precio            1309 non-null int64
billete           1309 non-null int64
Planta            1309 non-null int64
Titulo_Arist       1309 non-null uint8
Titulo_Army        1309 non-null uint8
Titulo_Dr          1309 non-null uint8
Titulo_Empleados   1309 non-null uint8
Titulo_Girl        1309 non-null uint8
Titulo_Master      1309 non-null uint8
Titulo_Miss        1309 non-null uint8
Titulo_Mr          1309 non-null uint8
Titulo_MrSSenior   1309 non-null uint8
Titulo_MrSenior    1309 non-null uint8
Titulo_Mrs         1309 non-null uint8
Titulo_Rev         1309 non-null uint8
Puerto_C           1309 non-null uint8
Puerto_Q           1309 non-null uint8
Puerto_S           1309 non-null uint8
Sexo_female        1309 non-null uint8
Sexo_male          1309 non-null uint8
```

Que finalmente normalizados entre 0 y 1

	Pclass	Age	Family	precio	billete	Planta	Titulo_Arist	Titulo_Army	Titulo_Dr	Titulo_Empleados	...	Titulo_Mr	T
891	1.0	0.4250	0.0	0.111111	1.00	1.0	0.0	0.0	0.0	0.0	...	1.0	(
892	1.0	0.5875	0.1	0.000000	1.00	1.0	0.0	0.0	0.0	0.0	...	0.0	(
893	0.5	0.7750	0.0	0.333333	0.75	1.0	0.0	0.0	0.0	0.0	...	0.0	(
894	1.0	0.3375	0.0	0.333333	1.00	1.0	0.0	0.0	0.0	0.0	...	1.0	(
895	1.0	0.2750	0.2	0.444444	1.00	1.0	0.0	0.0	0.0	0.0	...	0.0	(

5. Creación de modelos

A continuación crearemos diversos modelos y comprobaremos su efectividad con la finanalidad de participar en la competición activa en Kaggle:

Modelos de regresión

Aplicamos cross validation con el 20% para test sobre los siuiente 4 modelos (adjuntamos el resultado medio de las cross validation)

LogisticRegression: **80% de aceirtos**

LogisticRegressionCV: **81% de aceirtos**

RandomForestClassifier: **83% de aceirtos**

GradientBoostingClassifier: **81% de aceirtos**

En un analisi mas profundo de cada cross validation vemos que en general los resultados son bastante inestables. Por ejemplo, vemos que los resultados del método de Random Forest Classifier oscila entre el 69 y el 94% de aciertos según donde se haya hecho el corte de la cross validation.

Aun así, este metodo es el que consigue de media un mayor numero de aciertos, llegando al 83%

A partir de los resultados predichos generados por el metodo Random Forest Classifier creamos un csv para comprobar su grado de acierto en kaggle
Esta predicción (adjuntada como "Submision_rf.csv") ha conseguido un **0.79425 en Kaggle**

UOC – MÁSTER EN CIENCIA DE DATOS
TIPOLOGÍA Y CICLO DE VIDA DE DATOS – PRÁCTICA 2

Combinación de modelos regresión

Aprovechamos para hacer también una combinación de las predicciones de los 3 métodos con los mejores resultados (haciendo la media de cada prediccción y redondeando el resultado)

ESta predicción (adjuntada como "Submision_suma_mejores1.csv") se ha reducido hasta **0.78947 en Kaggle**

Modelo de redes neuronal

A continuación preparamos los datos para utilizar una red neuronal completamente conectada con las siguientes características:

```
#Capa de entrada
model.add(Dense(23, input_shape=(23,)))
model.add(BatchNormalization())
model.add(Activation("relu"))
model.add(Dropout(0.4))

#Capas ocultas
model.add(Dense(50))
model.add(BatchNormalization())
model.add(Activation("relu"))
model.add(Dropout(0.4))

model.add(Dense(50, activation="relu"))

# Capa de salida con función sigmoide que da un numero entre 0 y 1
model.add(Dense(1, activation="sigmoid"))
"Submision_NN.csv", index=False)
```

ESta predicción (adjuntada como "Submision_NN.csv" ha conseguido:

0.80382 en Kaggle

Creación de una función para intentar mejorar los resultados

A partir de la estructura anterior intentamos ajustar el modelo para mejorar los resulados. Para ello creamos una función para añadir capas(1 o 2), modular el numero de neuronas por capa entre 25 y 100 y alternar la activación entre sigmoid y relu.

De todas las combinaciones la que obtuvo mejor reultado sobre el test fue 'sigmoid, 100, 2' (ligeramente superior a los optenidos por la combinación de activación relu, 50 neuronas por capa y 1 capa escondida, que es la capa con la que se había conseguido el 0.8 de acierto en kaggle.

A pesar de este resultado esperanzador la predicción de esta red neuronal (adjuntada como "submission_Best_NN_Original2.csv" apenas fue de se ha reducido hasta 0.78947 en Kaggle

Combinación de modelos de red neuronal

Finalmente, optamos por combinar todos los resultados de las predicciones de las 6 redes neuronales generados por la función.

Esta predicción (adjuntada como "submission_sum_NN42.csv") ha conseguido el mismo resultado que la anterior **0.79425 en Kaggle**

6. Conclusiones

Después de muchos otros intentos que no están reflejados en este documento (por falta de capacidad de calculo de mi ordenador), no parece que ninguna estructura en la red neuronal pueda mejorar estos resultados.

Entendemos que para seguir mejorando, deberíamos volver a modificar aspectos de la limpieza original de datos o crear nuevas combinaciones de elementos a partir de los ya existentes.

Aun así, una prediccin del 0.80382, que nos situa entre los 9% mejores de los 23000 participantes en este concurso de Kaggle nos parece bastante aceptable, de momento.

7. Código

https://github.com/xavierricci/Practica2_titanic/wiki

Contribuciones	Firma
Investigación previa	XRV, GMD
Redacción de las respuestas	XRV, GMD
Desarrollo código	XRV, GMD