Modelo de clasificación para predecir sobrevivencia después de un año de un ataque cardíaco

Luis Florian & René Tarot

Introducción

Este trabajo tiene como objeto mostrar los pasos que se llevaron a cabo durante la creación de un modelo de clasificación capaz de predecir lo más acertado posible si un paciente luego de subrir un ataque cardiacó sobrevivirá por lo menos un año.
El set de datos utilizado fue obtenido de una publicación de UCI Machine Learning Repository, la data fue recolectada por Dr. Evlin Kinney, The Reed Institute de Miami.
En este set de datos, todos los pacientes registrados sufrieron ataques cardíacos en algún momento del pasado. Algunos todavía están vivos y otros desafortunadamente no. Las variables survivel y still-alive, cuando se toman juntas, indican si un paciente sobrevivió durante al menos un año después del ataque cardíaco.
Utilizamos redes neurales para realizar dicho trabajo, y se obtuvo un resultado bastante aceptable.

Explorando el set de datos

survival	still-alive	age-at-heart-attack	...	name	group	Alive-at-1
11.0	0.0	71.0	...	name	1.0	0.0
19.0	0.0	72.0	...	name	1.0	0.0
16.0	0.0	55.0	...	name	1.0	0.0
57.0	0.0	60.0	...	name	1.0	0.0
19.0	1.0	57.0	...	name	1.0	0.0
...	...	...	...	...	...	...
7.5	1.0	64.0	...	name	NaN	NaN
41.0	0.0	64.0	...	name	NaN	NaN
36.0	0.0	69.0	...	name	NaN	NaN
22.0	0.0	57.0	...	name	NaN	NaN
20.0	0.0	62.0	...	name	NaN	NaN

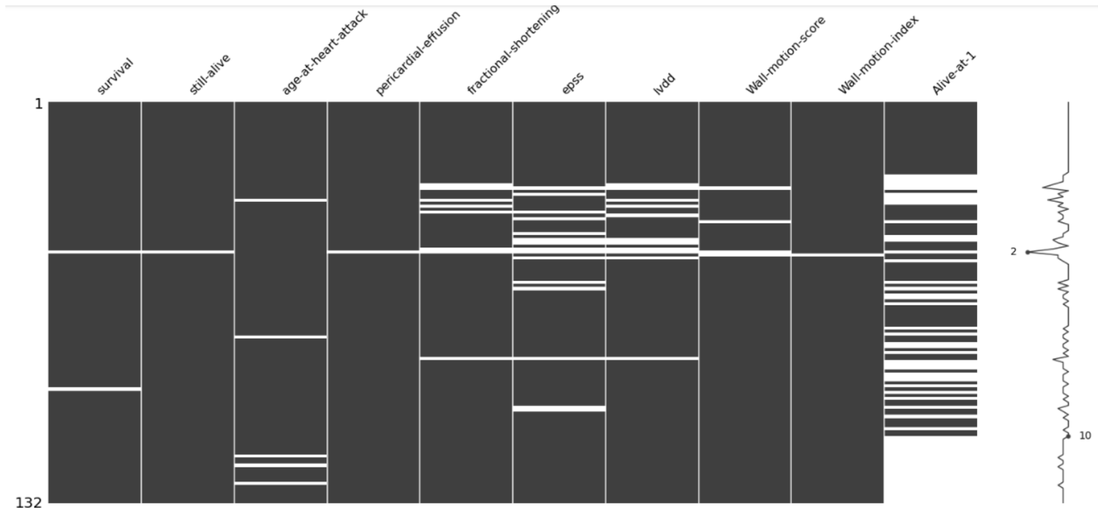
El set de datos consta de 132 instancias de pacientes y 12 variables categóricas y numéricas. En la publicación del Dataset se indica que las variables mult, name y group no aportan ningún valor. Por lo que obviarán en nuestro estudio.

Hipótesis

Aún siendo un dataset tan pequeño se podrá obtener un accurcy mayor a 80.

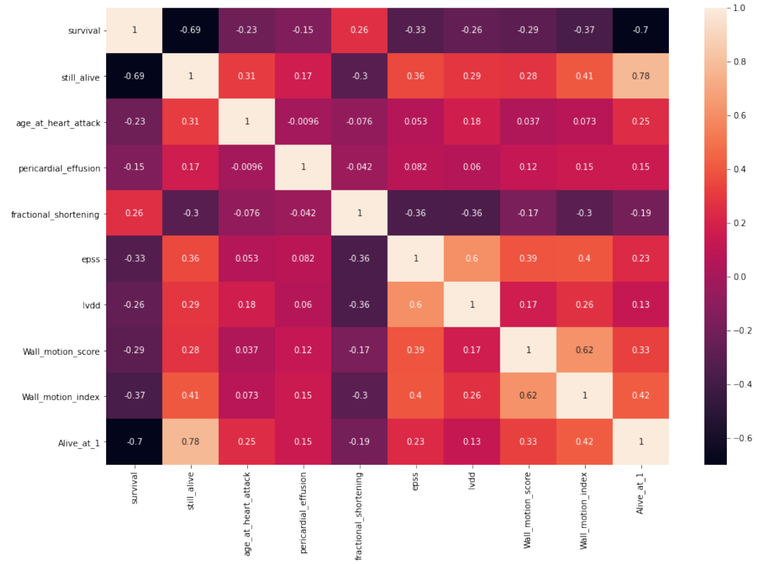
Valores Faltantes en el Dataset

Luego de revisar el dataset se detectaron 103 valores faltantes, los cuales es necesario cubrir debido al pequeño tamaño completo del dataset. Para esta tarea disponemos de varias opciones, sustituir por la media, la moda o utilizar una red neural para predecir los valores faltantes. Lo que después de pruebas nos dió un mejor resultados fue utilizar imputaciones, utilizamos Nearest neighbor imputation, el cual es una herramienta que pondera las muestras utilizando la diferencia de media cuadrática en features que en dos filas tienen data observada. Luego de su aplicación no quedó ningún valor faltante. A continuación se observa una gráfica que muestra la cantidad de valores faltantes representados por espacios en blanco.



Correlación entre variables

Para determinar que variables utilizar en nuestro modelo y cuales descartar utilizamos una correlación de variables. Los colores más claros representan una mayor correlación, al igual que un valor más cercano a 1.



Red totalmente conectada

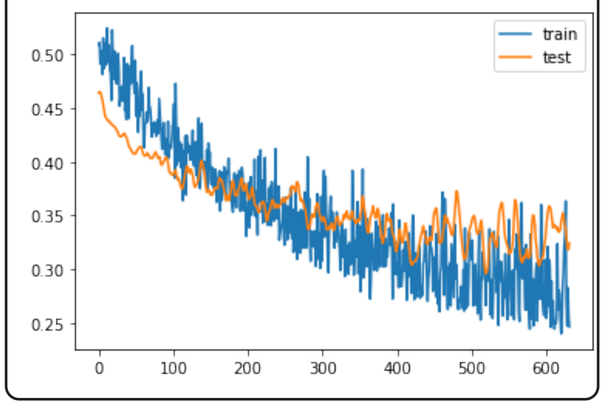
Se utilizó una red totalmente conectada de 5 capas, relu y sigmod como funciones de activación, la función adam de optimizer y binary crossentropy para función loss. Esta red Neural fue entrenada con 4000 epochs.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 8)	80
dense_1 (Dense)	(None, 6)	54
dense_2 (Dense)	(None, 4)	28
dense_3 (Dense)	(None, 2)	10
dense_4 (Dense)	(None, 1)	3

Total params: 175
Trainable params: 175

RESULTADOS: Train: 0.882, Test: 0.900

Gráfica de historial de Loss



Referencia

Galván, F. M. M. (s.f.). *Imputación de datos: teoría y práctica.*
Guttenberg, N. (2015). Learning to generate classifiers.
Kinney, D. E. (2002, month=March). *Echocardiogram data set.* Descargado de `https://archive.ics.uci.edu/ml/datasets/echocardiogram`