

## Ejercicio 1

- 1) Se setea el atributo como label para que sea considerado en el modelo como variable dependiente, es decir la variable que se va a estimar.

La variable label efectivamente asume 2 valores, sí y no.

- 2) Esto se realiza porque la regresión precisa de valores de tipo integer para poder adaptar el modelo, si las mismas son de tipo binomial debe reinterpretarlas asumiendo algún criterio que les asigne números.
- 3) a). Estos rangos son iguales para ambos data sets.  
  
b) Los rangos de los atributos del data set son iguales por tanto es posible realizar correctamente el entrenamiento y la posterior prueba. Esto debe ser verificado porque si no entrenamos situaciones que luego van a aparecer en el test no vamos a tener un comportamiento ajustado para dichos atributos.  
  
c) Otras tareas previas son eliminar outliers, que en este caso no se identifican. Por otra parte es posible también normalizar los atributos y eliminar atributos correlacionados.

## Ejercicio 2

El operador logistic regression tiene los parámetros:

**solver** es opcional, IRLSM es más rápido en problemas con poca cantidad de predictores, L\_BFGS es mejor para datasets con muchas columnas, COORDINATE\_DESCENT es el IRLSM con actualización de covarianzas, COORDINATE\_DESCENT\_NAIVE es IRLSM con actualización ingenua de descenso de gradientes.

**reproducible** : hace reproducible al modelo, si se setea entonces el máximo número de hilos (maximum\_number\_of\_threads) controla el nivel de paralelismo del modelo, si no se utiliza el por defecto.

**maximum number of threads** (optional) : Controla el paralelismo.

**use regularization** (optional) : al checkear este campo se puede especificar lambda, Alpha, y la búsqueda de parámetros relacionados con lambda.

**standardize** : Estandariza las variables numéricas para que tengan media 0 y varianza 1

**non-negative coefficients** : restringe los coeficientes para que sean no negativos.

4)

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Estado_civil.1	161.972	161.972	136.986	1.182	0.237
Estado_civil.3	82.721	82.721	140.403	0.589	0.556
Estado_civil.0	104.135	104.135	393.366	0.265	0.791
Categoria_Peso.0	46.831	46.831	152.953	0.306	0.759
Categoria_Peso.2	-233.182	-233.182	198.303	-1.176	0.240
Sexo.1	-88.350	-88.350	85.230	-1.037	0.300
Manejo_stress.0	64.966	64.966	144.549	0.449	0.653
Edad	-0.944	-7.415	0.815	-1.158	0.247
Colesterol	-2.134	-68.857	1.013	-2.106	0.035
Trat_ansiedad	4.037	49.945	7.690	0.525	0.600
Intercept	154.709	-59.587	-451.018	-0.343	0.732

7) Luego de ejecutados los pasos 5) y 6)

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Estado_civil.1	458.455	458.455	187.906	2.440	0.015
Estado_civil.3	696.648	696.648	249.894	2.788	0.005
Estado_civil.0	69.379	69.379	415.555	0.167	0.867
Categoria_Peso.0	-74.789	-74.789	211.487	-0.354	0.724
Categoria_Peso.2	-141.274	-141.274	202.300	-0.698	0.485
Sexo.1	-29.988	-29.988	106.734	-0.281	0.779
Manejo_stress.0	145.502	145.502	152.332	0.955	0.339
Edad	86.083	86.083	85.524	1.007	0.314
Colesterol	-254.934	-254.934	93.093	-2.739	0.006
Trat_ansiedad	-132.334	-132.334	84.833	-1.560	0.119
Intercept	-373.073	-373.073	179.626	-2.077	0.038

8) Atributos generados, los atributos polinominales fueron regeneradas en tantos atributos como categorías tenía el atributo. Cada nuevo atributo toma el valor 1 cuando el atributo representa a su valor original y 0 de lo contrario.

Row No.	2do_Ataque...	prediction(2do_Ataque_Corazon)	confidence(Si)	confidence(No)
1	No	No	0	1
2	No	No	0.000	1.000
3	No	Si	1	0
4	Si	No	0	1
5	Si	Si	1	0
6	No	No	0.000	1.000
7	Si	Si	1.000	0.000
8	No	No	0	1
9	Si	Si	1	0
10	No	No	0	1
11	Si	Si	1	0
12	Si	Si	1	0
13	Si	Si	1	0

Se observan las predicciones de acuerdo a los valores reales, y el nivel de Confidence con respecto a asignar al atributo el valor Sí y el valor No (1-valor sí).

accuracy: 82.93%

	true Si	true No	class precision
pred. Si	18	5	78.26%
pred. No	2	16	88.89%
class recall	90.00%	76.19%	

También se observa la precisión del modelo 82,9%, es decir las veces que se predijo acertadamente el sí en el primer cuadrante (18), las veces que se predijo sí y la respuesta era No (5), las veces que se predijo No siendo la respuesta sí (2) y por último las veces que acertadamente se predijo No (16).

### Ejercicio 3

La decisión del Dr. García depende bastante del área específica en la que está actuando, al ser el área de la salud un área sensible seguramente dependerá de los recursos disponibles a la hora de priorizar a quién atender y/o destinar recursos.

De acuerdo al modelo analizado, esta persona no está entre aquellas vulnerables a sufrir un ataque cardíaco dados los atributos disponibles, por lo que podría priorizarse personas que como las de la tupla 11 que tiene un 99,3% de confianza de sí tener uno. Podría optar por elaborarse una lista priorizando los más susceptibles a sufrir un ataque.

358 personas tienen predicción de ataque cardíaco según este modelo. En este caso el umbral a considerar depende la sensibilidad del tema a tratar, se podría exigir un nivel de confianza muy alto para poder decidir no tratar a un paciente, para asegurarse de no cometer errores.

Se puede tener una estimación de la performance global del modelo con el porcentaje de acierto y error de las predicciones en el caso de los datos con etiqueta sumado a la confianza con la que se predicen los casos Sí y No.