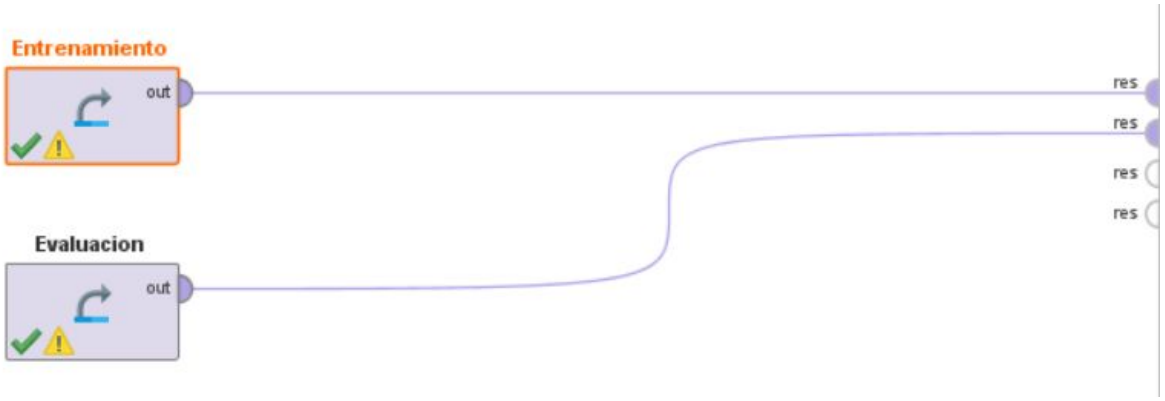
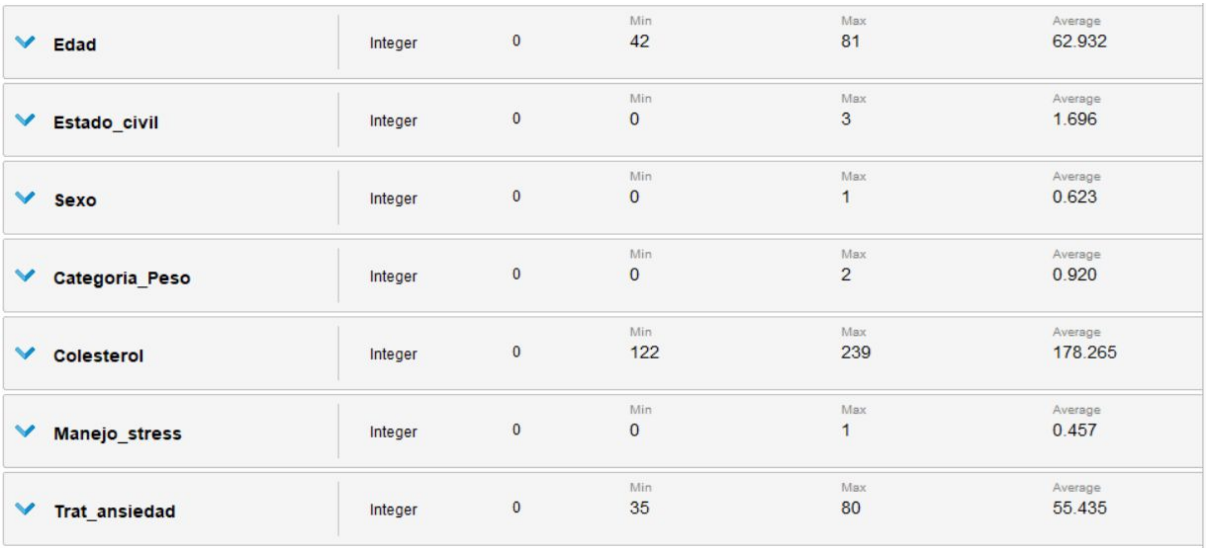
TA4 - Enzo Cozza - Agustín Fernández

Ejercicio 1

1. 

Label			Least	Most	Values
2do_Ataque_Corazon	Binominal	0	Si (68)	No (70)	No (70), Si (68)
Edad	Integer	0	Min 42	Max 81	Average 62.978
Estado_civil	Integer	0	Min 0	Max 3	Average 1.696
Sexo	Integer	0	Min 0	Max 1	Average 0.623
Categoria_Peso	Integer	0	Min 0	Max 2	Average 0.920
Colesterol	Integer	0	Min 122	Max 239	Average 177.391
Manejo_stress	Integer	0	Min 0	Max 1	Average 0.442
Trat_ansiedad	Integer	0	Min 35	Max 80	Average 55.435

Se debe utilizar el atributo como binomial, debido a que la regresión logística exige que la variable objetivo sea de este tipo para efectuar sus predicciones de clasificación.

2. 

			Min	Max	Average
Edad	Integer	0	42	81	62.932
Estado_civil	Integer	0	0	3	1.696
Sexo	Integer	0	0	1	0.623
Categoria_Peso	Integer	0	0	2	0.920
Colesterol	Integer	0	122	239	178.265
Manejo_stress	Integer	0	0	1	0.457
Trat_ansiedad	Integer	0	35	80	55.435

   El método de modelado de datos regresión logística exige que todos los atributos utilizados para realizar el modelo sean del tipo ‘integer’.

3. a. Los rangos son iguales para cada atributo.

   b. Sí, ya que los rangos de los atributos de evaluación son iguales a los de entrenamiento. Esto debe ser verificado antes de proceder, para asegurar que el modelo se comporte de manera correcta con los datos de evaluación, ya que si fue diseñado con datos dentro de un rango específico, y se evalúa con datos fuera de ese rango, no sabrá cómo comportarse correctamente.

   c. Normalización de los datos, eliminar variables correlacionadas, detección y tratamiento de outliers.

Ejercicio 2

3. Parámetros:

Solver: IRLSM para problemas con pocos predictores y búsqueda lambda con penalización L1; L_BFGS escala bien para datasets con muchas columnas; COORDINATE_DESCENT es IRLSM con actualizaciones en la covarianza; COORDINATE_DESCENT_NAIVE.
Reproducible: hace que el modelo sea reproducible. Dependiendo de lo seleccionado, queda definido el nivel de paralelismo en base a la cantidad de hilos.
Use regularization: marcar este parámetro si se debe regularizar.
Standarize: estandarizar las columnas para que cada una tenga media 0 y varianza 1.
Non-negative coefficients: restringir los coeficientes a ser no negativos.
Add intercept: incluir una variable constante al modelo.
Compute p-values: solicitar los p-values.
Remove collinear columns: en caso de dependencia linear entre columnas, remueve algunas de estas.
Missing values handling: cómo tratar los valores faltantes: ‘Skip’ los saltea, ‘MeanImputation’ les asigna la media.
Max iterations: número máximo de iteraciones.

Max runtime seconds: máximo número de segundos permitidos para el tiempo de ejecución del modelo.

7. Coeficientes:

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. Nuevos atributos generados:
Estado_civil.1, Estado_civil.3, Estado_civil.0, Categoria_Peso.0, Categoria_Peso.2, Sexo.1, Manejo_stress.0.

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%	

Ejercicio 3

1. La decisión del Dr. García para este paciente sería que **no** tendrá un segundo ataque cardíaco, ya que por los datos recolectados a partir del modelado de regresión logística, tenemos un confidence por “No” de 1.000 y un confidence por “Si” de 0, por lo que es casi garantido que no tendrá otro ataque. Sin embargo, el Dr. debe tener en cuenta que la precisión obtenida para el modelo es aproximadamente del 82%.
2. En el caso de la tupla 11, se obtuvo un nivel de confianza de predicción de 1 para el “Sí”, por lo que debería ser priorizado para el tratamiento.

3. La predicción podría ser utilizada para realizar un primer filtro de los pacientes, para así priorizar los que más confianza por ‘Si’ se tenga y poder darles el tratamiento adecuado. 329 pacientes en el dataset tienen una predicción de un probable segundo ataque cardíaco.
   Se podría analizar la performance global del modelo mediante la matriz de confusión, que aporta información de los falsos y verdaderos positivos, y falsos y verdaderos negativos.