Ejercicio 1

1. Importa en RM el dataset de entrenamiento (“cardiac-training.csv”).
a) Verifica que la primera fila se configura como nombres de los atributos.
b) Setea el atributo “2do_Ataque_Corazon” de acuerdo al problema

i. ¿Por qué hacemos esto?
La variable 2do ataque al corazón es la variable que queremos predecir en base al resto de los datos, por ello hay que indicar que es la variable objetivo.

ii. Verificar que los valores posibles son efectivamente de 2 clases (si/no)
Efectivamente lo son

c) Verifica que el atributo “2do_Ataque_Corazon” esté configurado como variable de predicción, o agregar un “set role” posterior

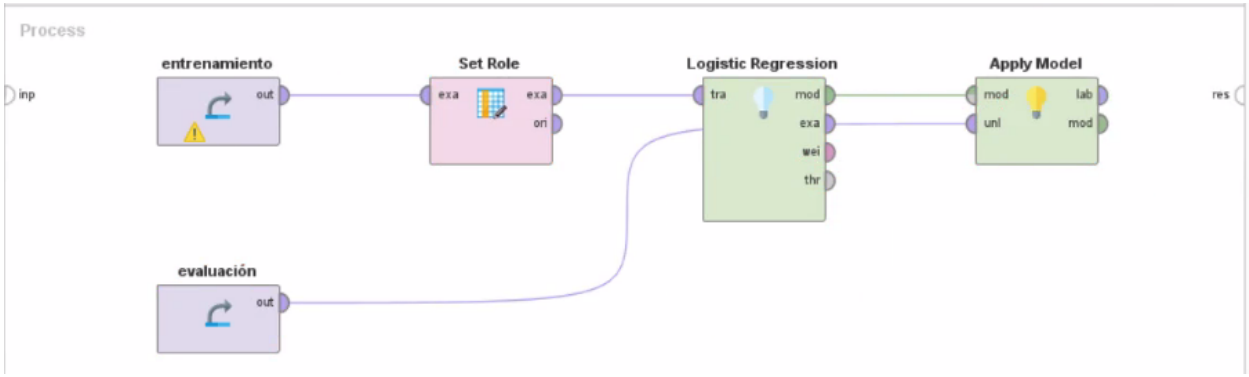
Seteamos la variable como label.

d) Completa el proceso de importación de datos y agregar el dataset a un nuevo proceso principal en blanco. Renombrar el operador “retrieve” del dataset a “entrenamiento”.

2. Importa el archivo de test / evaluación / predicción (“cardiac-scoring.csv”).
a. Verifica que el tipo de datos de los atributos es “integer”.

i. ¿por qué hacemos esto?
Porque para aplicar un modelo de regresión lógistica necesitamos trabajar con números.

b. Completa el proceso de importación y renombrar el operador “retrieve” a “evaluación”



3. Corre el modelo y comparar los rangos de los atributos entre los datasets de entrenamiento y evaluación.
a. ¿Cómo son, comparativamente, estos rangos?

Los rangos de los atributos entre los dos conjuntos. Son idénticos, puede suceder que como los rangos son parecidos luego cuando usemos otros datos que estén fuera de rango nuestro modelo no sea capaz de predecir con exactitud.

Verificar que no va a existir en el dataset scoring ningún valor que esté fuera de los rangos del dataset de training. Puede suceder que si en producción las unidades de medida son distintas a las de los datasets de training se el modelo falle.
Una buena política es normalizar los datos antes de aplicar el modelo para evitar este tipo de problemas, de esta forma tendrán todos el mismo rango y la misma magnitud. Una regla de sensatez es identificar aquellos valores que se van de rango para evitar que el sistema sea inválido, si metemos basura va a salir basura cuando lo pongamos en producción.

b. ¿Están todos los atributos de los ejemplos de evaluación / predicción en los rangos de los atributos del dataset de entrenamiento?

i. ¿Por qué tenemos que verificar esto?

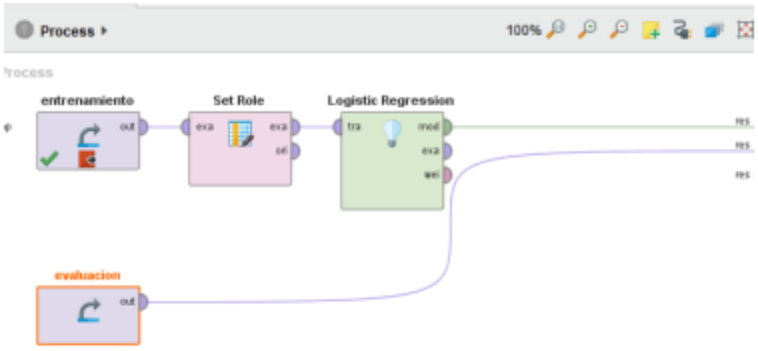
Por la razón mencionada en la respuesta anterior, el hecho de que el conjunto no sea lo suficientemente disperso puede generar problemas de rendimiento.

c. ¿Hay más tareas de preparación previa de los datos para hacer?

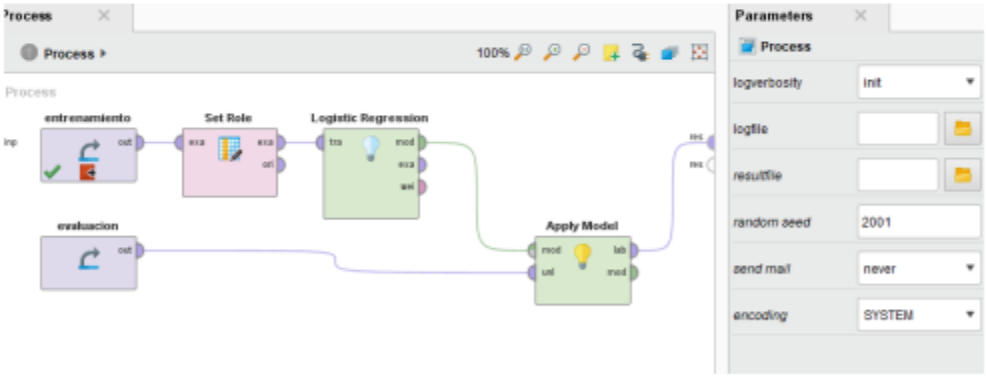
Hay que controlar los outliers cuando se utiliza regresión logística porque pueden afectar negativamente a los coeficientes de los atributos. También se deberían filtrar variables que estén altamente correlacionadas entre ellas. En general la regresión lineal no funciona bien cuando existen valores muy dispersos o demasiado dispersos.

Ejercicio 2

1. Agregar un “set role “ luego del “retrieve” del dataset de entrenamiento
2. Agregar un operador “logistic regression”
3. Analizar los parámetros del operador
4. Ejecutar para aprender el modelo, y observar los coeficientes resultantes del modelo



5. **agrega** un operador “apply model”, y conectar a sus entradas:
   a. a “mod” conectar la salida del operador “logistic regression”
   b. a “unl” el dataset “scoring”
6. **verifica** que los puertos “lab” y “mod” estén conectados a las salidas “res” del proceso



7. **Ejecuta el proceso**, y en los resultados, observa los coeficientes obtenidos para los diferentes atributos, éstos conforman el modelo de regression logística.
8. **Observa** las estadísticas
   a. nuevos atributos generados
   b. prediction
   c. confidence (si)
   d. confidence (no)

Observar los coeficientes que aplican acá, son los coeficientes a los cuales se le aplica el log para llegar a la curva logística. En general cuando el valor absoluto del coeficiente se aleja de cero el p-value es menor y el z-value mayor. Todo esto tiene un análisis matemático como fundamento

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Edad	-0.119	-0.937	0.078	-1.528	0.127
Estado_civil	-1.278	-1.055	0.531	-2.409	0.016
Sexo	-0.215	-0.104	0.851	-0.253	0.801
Categoria_Peso	-4.056	-3.102	0.976	-4.154	0.000
Colesterol	-0.009	-0.279	0.015	-0.587	0.557
Manejo_stress	0.071	0.035	0.949	0.075	0.941
Trat_ansiedad	0.054	0.663	0.065	0.818	0.413
Intercept	12.419	0.342	4.690	2.648	0.008