UNIDAD TEMÁTICA 3: Algoritmos Lineales

Trabajo de Aplicación 4 - Aplicación de Regresión logística utilizando Rapid Miner, con predicción y análisis

ESCENARIO

El Dr. García es un cardiólogo que atiende en un importante centro de salud. Más allá de su extensa experiencia, entiende que sería sumamente conveniente contar con la asistencia de un sistema que, aprovechando la información histórica de muchísimos pacientes, le ayude a determinar mejor los riesgos de problemas coronarios. Claramente, el sobrepeso, sexo y niveles de colesterol tienen una relación importante con la enfermedad coronaria.

El Dr. García está convencido de las enormes ventajas de la medicina preventiva, desea tener herramientas que le ayuden a determinar, en forma eficiente, a cuáles de sus pacientes proponer distintos tratamientos para reducir sus riesgos, y, en particular, desea enfocarse en la prevención de un ***segundo ataque cardíaco,*** es decir, proponer a los pacientes con riesgo elevado cambios en el estilo de vida. Como los costos resultantes de la enfermedad cardíaca, para el sanatorio, son muy elevados, el Dr. sabe que podrá fácilmente instaurar un programa de ayuda para bajar de peso, mejorar la dieta y reducir el colesterol, y para manejar el estrés.

Ahora bien, la parte difícil está al tratar de determinar qué pacientes podrían beneficiarse de estos programas.

Datos

El Dr. García ha obtenido una base de datos que contiene información de historias clínicas de pacientes. La base contiene dos datasets:

- El primero contiene datos históricos de pacientes que ya han tenido un ataque al corazón, con un atributo que indica si han tenido o no más de uno ("***cardiac-training.csv***"). Lo utilizaremos para entrenamiento.
- El segundo contiene datos de pacientes actuales que han tenido un ataque pero **no un segundo.** Utilizaremos este dataset para predicción **("*cardiac-scoring.csv*").** El Dr. García pretende ayudar a las personas de este segundo grupo a evitar un segundo ataque.

Los atributos de estos datasets son:

Edad: la edad en años redondeada al entero más cercano.

Estado_civil: codificado medinate un número: 0 = soltero; 1 = casado, 2 = divorciado y 3 = viudo.
Sexo: 0 = femenino; 1= masculino.
Categoria_Peso: el peso de la persona, categorizado en uno de tres posibles niveles: 0 = normal; 1 = sobrepeso; 2 = obeso.
Colesterol: nivel de colesterol de la persona, tal como se ha registrado en el momento del tratamiento indicado cuando su más reciente ataque al corazón.

Manejo_stress: un atributo binario que indica si el paciente ha participado previamente de cursos de manejo del estrés: 0 = no; 1 = si.

Trat_ansiedad: valor entre 0 y 100 indicativo del nivel natural de estrés de cada persona y de su habilidad para manejar este estrés. Poco tiempo después de que la persona se recuperara de su primer ataque, se le administró un test de ansiedad natural estándar. Los valores están tabulados en incrementos de 5. Un valor de 0 indicaría que la persona nunca siente ansiedad, presión o estrés en ninguna situación, mientras que un valor de 100 indicaría que la persona vive en un estado continuo de sobrecarga e incapaz de lidiar con su situación.

2do_Ataque_Corazon: Este atributo existe solamente en el dataset de entrenamiento. Es la variable objetivo o de predicción ("label" en RM). En el dataset de entrenamiento, este atributo contiene "SI" para aquellos individuos que han sufrido un segundo ataque al corazón, y "no" en caso contrario.

RESPONDER PREGUNTAS PROYECTADAS EN PANTALLA

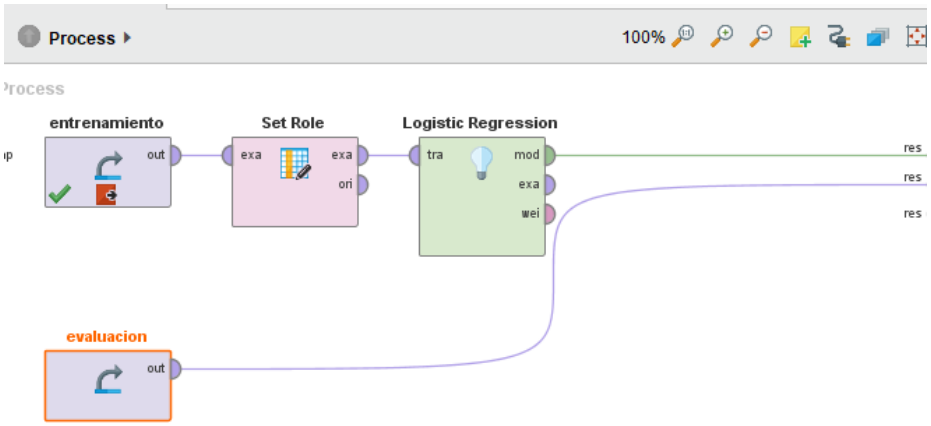
Ejercicio 1

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?
      ii. Verificar que los valores posibles son efectivamente de 2 clases (si/no)
   c) Verifica que el atributo "2do_Ataque_Corazon" esté configurado como variable de predicción, o agregar un "set role" posterior
   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".

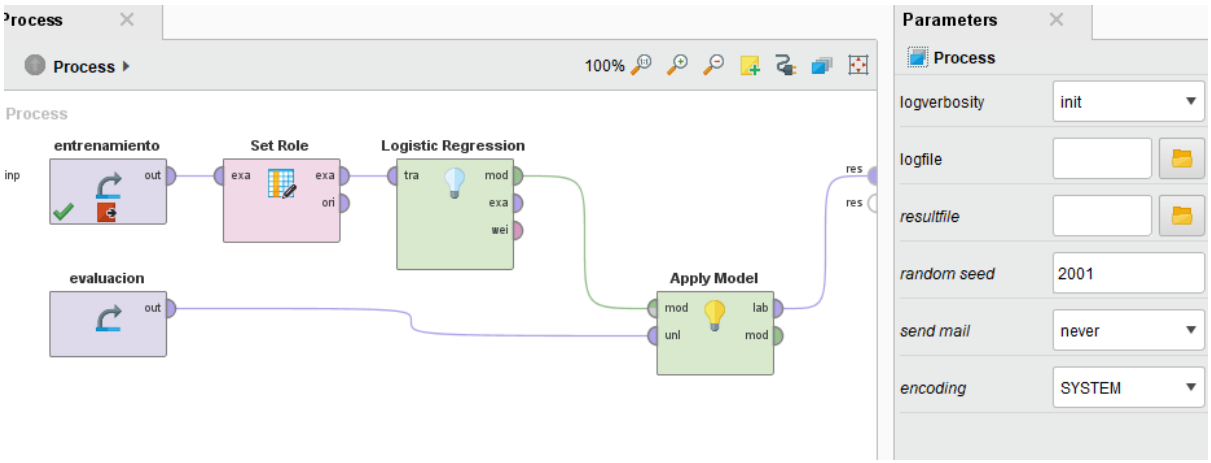
2. Importa el archivo de test / evaluación / predicción (***"cardiac-scoring.csv"***).
   a. Verifica que el tipo de datos de los atributos es "integer".
      i. ¿por qué hacemos esto?
   b. Completa el proceso de importación y renombrar el operador "retrieve" a "evaluación"
3. Corre el modelo y comparar los rangos de los atributos entre los datasets de entrenamiento y evaluación.
   a. ¿cómo son, comparativamente, estos rangos?
   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?
   c. ¿hay más tareas de preparación previa de los datos para hacer?

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)

Ejercicio 3 – EVALUACION e INTERPRETACIÓN DE LOS RESULTADOS

En el resultado podemos observar que para cada persona se ha pronosticado si tendrá o no un 2do ataque cardíaco

En las estadísticas vemos que, de 690 personas representadas, la predicción da que 365 sufrirán un 2do ataque, mientras 325 no…

El Dr. García espera poder tratar a las 365 del primer grupo, y quizás también a algunas del 2do grupo para las cuales el nivel de confianza de la predicción "no" sea bajo…

- …. si se trata de personas reales....
- ¿qué tanta confianza podemos tener en esas predicciones?

1. Veamos la primera tupla:

Hombre, soltero, 61 años, con sobrepeso pero el colesterol es bajo (139 y la media es 178).

- está en el medio de la clase para tratamiento de ansiedad (50) y ha participado en manejo del estrés
- el modelo nos da xx % de confianza en que la predicción "No" es correcta, lo que nos deja un (1-xx %) de duda.

¿Cuál sería la decisión del Dr. García para este paciente?

Para cada persona en el dataset, sus atributos se aplican al modelo de regresión logística y se calcula una predicción con valores de confianza

2. veamos la tupla 11:

Hombre, divorciado, 66 años, está por encima de la media en todos los predictores (analizar)

- tenemos un 99.3% de confianza en la predicción de un 2do ataque cardíaco

3. ¿Cuál sería la decisión del Dr. García para este paciente?

- ¿Cómo usar la predicción en un caso de consultoría o desarrollo de un sistema real?
- ¿cuántos pacientes tienen predicción de ataque cardíaco?
- tener en cuenta los niveles de confianza
- ¿cómo podríamos en RM analizar la performance global del modelo?