AWAKELAB

Plan Formativo: Ciencia de Datos	Nivel de Dificultad
Módulo: Fundamentos del Big Data	Bajo / Medio
Tema: SQL Spark	

Intención del aprendizaje o aprendizaje esperado:
Manipular grandes volúmenes de datos utilizando Spark SQL para resolver un problema

Ejercicios planteados
PARTE I **Machine learning para clasificar** Python es el lenguaje preferido para la ciencia de datos debido a NumPy, Pandas y matplotlib, que son herramientas que facilitan el trabajo con matrices y dibujan gráficos y pueden trabajar con grandes matrices de datos de manera eficiente. Pero Spark está diseñado para trabajar con una enorme cantidad de datos, distribuidos en un clúster. **Datos del paciente del corazón** Encontrará adjunto la base de datos. Las columnas son: 1. Años

2. Sexo
3. Tipo de dolor torácico (4 valores)
4. Presión arterial en reposo
5. Colesterol sérico en mg/dl
6. Azúcar en sangre en ayunas > 120 mg/dl
7. Resultados electrocardiográficos en reposo (valores 0,1,2)
8. Frecuencia cardíaca máxima alcanzada
9. Angina inducida por el ejercicio
10. Oldpeak = depresión del ST inducida por el ejercicio en relación con el reposo
11. Pendiente del segmento ST de ejercicio máximo
12. Número de vasos principales (0-3) coloreados por fluoroscopia
13. Thal: 3 = normal; 6 = defecto fijo; 7 = defecto reversible

El campo que indica si el paciente tiene un problema cardíaco. Los números son los siguientes:

Un valor de 3 significa que el paciente está sano (normal). Un valor de 6 significa que se ha solucionado el problema de salud del paciente. Un valor de 7 significa que se puede solucionar el problema de salud del paciente.

Entonces, escriba esta función ENFERMO() para marcar 0 como negativo y 1 como positivo, porque la regresión logística binaria requiere uno de dos resultados.

También debe crear el dataframe de Spark raw_data usando la operación transform() y seleccionando solo la columna de características.

Continuará en AE4 - PARTE II

AWAKELAB

Caso
Preguntas guía
- SQL Spark - Formatos de Archivo
Recursos Bibliográficos:
Referencias [1] ¿Qué es SQL? https://youtu.be/TCam1GMMjTg [2] SQL spark https://cloud.ibm.com/docs/AnalyticsEngine?topic=AnalyticsEngine-working-with-sql&locale=es#:~:text=La%20CLI%20de%20SQL%20Spark%20es%20una%20herramienta%20pr%C3%A1ctica%20para,de%20la%20l%C3%ADnea%20de%20mandatos. [3] Official web site – Apache Spark https://spark.apache.org/sql/