Caso 1 – Dataset Hospital con PySpark

Hans Jorge Contreras Robledo

Fecha: 15/09/2025

Introducción

En este ejercicio trabajamos con un dataset hospitalario utilizando PySpark. El objetivo fue simular un pipeline batch en donde primero se cargan los datos, se limpian, se realizan consultas con SQL y finalmente se entrena un modelo de Machine Learning para predecir resultados en base a la información disponible.

1. Inicialización de Spark

Instalamos PySpark en entorno de Google Collab y creamos una sesión de Spark con `SparkSession`, necesaria para trabajar con datos distribuidos, ejecutar consultas SQL y modelos de Machine Learning.

2. Carga de datos

Cargamos el archivo CSV hospitalario con `spark.read.csv`, activando `header=True` e `inferSchema=True`. Mostramos número de filas, esquema y registros de ejemplo para revisar la estructura.

3. Limpieza de datos

Se reemplazaron valores nulos (cero en numéricas, 'desconocido' en categóricas) y se eliminaron duplicados. Esto deja los datos listos para análisis.

4. Consultas SQL

Con `createOrReplaceTempView` registramos la tabla temporal `hospital`. Ejecutamos consultas SQL como el conteo por diagnóstico o servicio clínico. También aplicamos `ROUND(AVG(...),2)` para redondear promedios.

5. Preparación para Machine Learning

Creamos una columna `label` en base a `dias_estancia` (1 si mayor al percentil 75, 0 en caso contrario). Indexamos variables categóricas, aplicamos OneHotEncoder y ensamblamos las variables en un vector de características.

6. Entrenamiento del modelo

Entrenamos dos modelos: Regresión Logística y Random Forest. Se dividieron los datos en entrenamiento (80%) y prueba (20%).

7. Evaluación del modelo

Se evaluaron con AUC (capacidad de separación) y Accuracy (porcentaje de aciertos). Esto permitió comparar cuál modelo tenía mejor desempeño.

8. Guardado de modelos y métricas

Se guardaron los modelos entrenados en la carpeta `models/` y las métricas en `artifacts/` para su reutilización sin necesidad de reentrenar.

Conclusiones

Este ejercicio mostró el flujo completo en PySpark: carga, limpieza, consultas SQL y modelos ML. Con ello se simuló un caso real de análisis hospitalario en Big Data, enfatizando la preparación y evaluación de datos.