EVALUACIÓN FINAL: ANÁLISIS DE MOVIMIENTOS MIGRATORIOS CON SPARK

Eres parte de un equipo de analistas de datos encargado de estudiar las tendencias de migración humana en el siglo XXI utilizando Big Data. Para ello, trabajarás con un conjunto de datos que contiene información sobre migraciones entre distintos países, sus causas y el impacto socioeconómico en las regiones de origen y destino.

Objetivos de la actividad

1. Aplicar conceptos de Big Data utilizando Apache Spark y PySpark.
2. Explorar y transformar datos con RDDs y DataFrames.
3. Realizar consultas con Spark SQL.
4. Implementar modelos de aprendizaje automático con MLlib.

INSTRUCCIONES

1. Carga y exploración de datos (2 puntos)

- Carga el dataset proporcionado en Spark.
- Convierte los datos en un RDD y un DataFrame.
- Explora los datos: muestra las primeras filas, el esquema y genera estadísticas descriptivas.

2. Procesamiento de datos con RDDs y DataFrames (3 puntos)

- Aplica transformaciones sobre los RDDs (filter, map, flatMap).
- Aplica acciones sobre los RDDs (collect, take, count).
- Realiza operaciones con DataFrames: filtrado, agregaciones y ordenamiento.
- Escribe los resultados en formato Parquet.

3. Consultas con Spark SQL (2 puntos)

- Registra el DataFrame como una tabla temporal.
- Realiza consultas sobre los principales países de origen y destino.

- Analiza las principales razones de migración por región.

4. Aplicación de MLlib para predicción de flujos migratorios (3 puntos)

- Convierte los datos en un formato adecuado para MLlib.
- Aplica un modelo de regresión logística para predecir la probabilidad de migración basada en factores socioeconómicos.
- Evalúa el modelo y analiza su precisión.

INSTRUCCIONES ADICIONALES:

- Puntos totales: 10 puntos.
- Descarga el dataset proporcionado.
- Comprime tu proyecto en un archivo .zip o .rar.
- Adjunta un documento con las reflexiones analíticas.
- Sube tu entrega a la plataforma.