ACTIVIDAD SESIÓN INTRODUCIÓN A MACHINE LEARNING ESCALABLE

Una tienda de cosmética quiere desarrollar un sistema inteligente que clasifique productos de **skin care** en diferentes categorías según sus características. La clasificación ayudará a recomendar productos adecuados a los clientes según su tipo de piel.

La tienda proporciona un dataset con información de productos, incluyendo:

- **Ingredientes clave** (como ácido hialurónico, retinol, vitamina C, etc.)
- **Nivel de hidratación**
- **Nivel de absorción**
- **Factor de protección solar (SPF)**
- **Tipo de piel recomendado** (seco, graso, mixto o sensible)

El objetivo es entrenar un modelo de clasificación con MLlib que prediga el tipo de piel recomendado para cada producto.

Importante:

Se debe convertir **Tipo de Piel** en valores numéricos:

- **Seco** → 0
- **Graso** → 1
- **Mixto** → 2
- **Sensible** → 3

INSTRUCCIONES

1. Carga y exploración de datos (2 puntos)

- Cargar los datos desde **skincare_products.csv** en un DataFrame de PySpark.
- Mostrar las primeras filas del dataset.
- Realizar un resumen estadístico de las variables numéricas.

2. Preprocesamiento de datos (2 puntos)

- Convertir la columna "Tipo de Piel" en valores numéricos (0, 1, 2, 3).
- Transformar las variables categóricas (Hidratación y Absorción) en valores numéricos.
    - Hidratación: Bajo (0), Medio (1), Alto (2)
    - Absorción: Bajo (0), Medio (1), Alto (2)
- Unir todas las características en un vector usando **VectorAssembler**.

3. División de datos y entrenamiento del modelo (3 puntos)

- Dividir los datos en 80% entrenamiento y 20% prueba.
- Entrenar un modelo de Árboles de Decisión con MLlib usando las características del **dataset**.

4. Predicción y evaluación (2 puntos)

- Aplicar el modelo al conjunto de prueba y mostrar las predicciones.
- Calcular la precisión del modelo usando **MulticlassClassificationEvaluator**.

5. Análisis de resultados y mejoras (1 punto)

- Explicar en breve (3-5 líneas) qué tan preciso fue el modelo y cómo se podría mejorar (por ejemplo, usando otro algoritmo o ajustando parámetros).

INSTRUCCIONES ADICIONALES:

- Incluye un documento con el análisis de tus resultados y mejoras.
- Puntos totales = 10 puntos.
- Comprimir el archivo en formato .zip o .rar.
- Subir el archivo a la plataforma.