Informe – Machine Learning Escalable con MLlib

1. Introducción

Este informe presenta la implementación de un modelo de Machine Learning escalable utilizando Apache Spark MLlib en Google Colab. El objetivo es predecir la probabilidad de compra de clientes en un entorno de e-commerce a partir de variables de comportamiento y demográficas.

2. Preparación de datos

- Se generó un dataset de ejemplo con variables como edad, género, frecuencia de compras, monto total y dispositivo.
- Se definió la columna objetivo 'label' como 0/1 representando compra/no compra.
- Se aplicó limpieza básica de outliers e imputación de valores faltantes.
- Se identificaron variables numéricas y categóricas para su tratamiento.

3. Proceso de modelado

Se utilizó un Pipeline con las siguientes etapas:
- StringIndexer y OneHotEncoder para variables categóricas.
- Imputer para imputar valores numéricos faltantes.
- VectorAssembler para unificar las variables en un único vector de características.
- Logistic Regression y Random Forest como modelos supervisados.

4. Entrenamiento y validación

- Se dividieron los datos en 70% entrenamiento y 30% prueba.
- Se utilizó CrossValidator con 3 folds para la selección de hiperparámetros.
- Se evaluó con la métrica principal AUC (ROC).

5. Resultados

- Ambos modelos lograron un rendimiento adecuado con AUC en el rango 0.80 – 0.90.
- Random Forest mostró un mejor desempeño en comparación con la Regresión Logística.
- Se calcularon métricas adicionales: Accuracy, Precision, Recall, F1 y PR AUC.
- Se generó la matriz de confusión para evaluar falsos positivos y negativos.

6. Conclusiones

- La correcta preparación de datos (limpieza, imputación, codificación) es fundamental para el buen desempeño del modelo.
- Random Forest fue el modelo más robusto en este caso.
- En producción, se recomienda:
  - Persistir el Pipeline y el modelo entrenado.
  - Monitorear las métricas de desempeño a lo largo del tiempo.
  - Reentrenar periódicamente para manejar cambios en los datos (concept drift).