INTEGRANTES:

GABRIELA STEPHANIE DONAIRES VELASQUE

**ANGELA GUADALUPE FERNANDEZ GALLO**

BRAYHAN ANDERSON CRUZADO CANCHO

**Trabajo de Estadística con Spark**

**Introducción**

Introducción a Spark y su Aplicación en Estadística

Apache Spark es una poderosa herramienta para el procesamiento y análisis de datos a gran escala. Permite realizar operaciones de análisis estadístico sobre grandes conjuntos de datos de manera eficiente y en tiempo real. Gracias a su capacidad de procesamiento en paralelo y en memoria, Spark es ideal para cálculos estadísticos complejos que requieren gran velocidad y capacidad.

En este trabajo exploraremos el uso de Spark para realizar análisis estadísticos básicos y avanzados, abordando funciones de tendencia central, dispersión y correlación con ejemplos prácticos.

1. Preparación del Entorno de Trabajo

Para utilizar Spark en nuestro análisis estadístico, primero se configura el entorno de trabajo en PySpark, la interfaz de Python para Spark.

python

from pyspark.sql import SparkSession

from pyspark.sql.functions import mean, stddev, variance, corr, count

Creamos una \*SparkSession\*, que es el punto de entrada para el uso de Spark.

python

spark = SparkSession.builder \

.appName("Trabajo de Estadística con Spark") \

.getOrCreate()

Con esta sesión activa, estamos listos para cargar datos y comenzar nuestro análisis.

2. Carga y Exploración de Datos

Utilizaremos un archivo de ejemplo, que contiene registros de ventas, incluyendo las columnas producto\_id, ventas, precio, y cantidad.

python

data = [("P1", 150, 20, 5),

("P2", 300, 15, 10),

("P3", 450, 25, 15),

("P4", 200, 10, 10),

("P5", 400, 18, 22)]

columnas = ["producto\_id", "ventas", "precio", "cantidad"]

df = spark.createDataFrame(data, schema=columnas)

df.show()

El conjunto de datos cargado nos proporciona la base para realizar cálculos estadísticos.

3. Análisis Estadístico con Spark

3.1 Media (Promedio)

La \*media\* o promedio es el valor típico de una serie de datos. Calculamos la media de la columna ventas para obtener un indicador de las ventas promedio.

python

df.select(mean("ventas")).show()

- \*Ejemplo\*: Si el resultado es 300, esto indica que en promedio, cada producto genera ventas de 300 unidades.

3.2 Mediana

La \*mediana\* es el valor central en un conjunto de datos ordenado. Spark no tiene una función de mediana directa, pero podemos calcularla manualmente o usando un enfoque basado en percentiles.

python

# Calculamos el percentil 0.5 (equivalente a la mediana)

df.approxQuantile("ventas", [0.5], 0.01)

- \*Ejemplo\*: Si la mediana es 300, esto indica que el punto medio de las ventas es de 300 unidades, lo cual puede ayudar a entender la distribución.

3.3 Moda

La \*moda\* es el valor que ocurre con mayor frecuencia en un conjunto de datos. En Spark, la moda no tiene una función directa, por lo que se calcula mediante agregación y orden.

python

df.groupBy("ventas").count().orderBy("count", ascending=False).show(1)

- \*Ejemplo\*: Si la moda en ventas es 150, significa que 150 es el valor más común en nuestras ventas, indicando una tendencia.

3.4 Varianza y Desviación Estándar

La \*varianza\* y la \*desviación estándar\* miden la dispersión de los datos respecto a la media. Valores altos indican mayor dispersión en el conjunto de datos.

python

df.select(variance("ventas"), stddev("ventas")).show()

- \*Ejemplo\*: Si la varianza es 1000 y la desviación estándar es aproximadamente 31.6, esto muestra una dispersión significativa en las ventas.

3.5 Correlación

La \*correlación\* evalúa la relación entre dos variables. En este caso, analizaremos la correlación entre ventas y precio para determinar si existe una relación entre estos valores.

python

df.select(corr("ventas", "precio")).show()

- \*Ejemplo\*: Si el resultado es 0.9, existe una fuerte correlación positiva entre ventas y precio, lo cual indica que a medida que aumenta el precio, también lo hacen las ventas.

4. Conclusión

El uso de Spark para el análisis estadístico permite procesar y obtener resultados en grandes volúmenes de datos de manera rápida y eficiente. En este trabajo, hemos calculado medidas de tendencia central y dispersión, además de analizar la relación entre diferentes variables. Estas técnicas proporcionan insights valiosos y aplicables en contextos empresariales.