UNIDAD DE INGENIERÍA EN COMPUTACIÓN
Sede Central Cartago
Bases de Datos 2
Profesor: Kenneth Obando Rodríguez

TEC | Tecnológico de Costa Rica

Trabajo Corto 3 (%) Fecha de Entrega: 15 de mayo, 2024

Spark

1 Objetivos

- Familiarizarse con el manejo básico de Apache Spark en Google Colab.
- Explorar y visualizar el set de datos ”California Housing” usando PySpark.

2 Requisitos Previos

- Acceso a Google Colab.
- Conocimientos básicos de Python y PySpark.

3 Parte 1: Configuración del Entorno

3.1 Instalación de Spark en Colab

Instalar todas las dependencias necesarias para ejecutar Apache Spark en Google Colab, incluyendo `findspark` y `pyspark`.

3.2 Inicialización de Spark Session

Importar y configurar `SparkSession` para comenzar a trabajar con Spark.

4 Parte 2: Carga y Exploración de Datos

4.1 Carga de Datos

Cargar el dataset ”California Housing” disponible en el directorio `sample_data` de Google Colab.

4.2 Exploración Inicial

Mostrar las primeras filas del conjunto de datos para verificar su correcta carga y utilizar `describe()` para obtener un resumen estadístico de las variables.

4.3 Parte 3: Manipulación de Datos

4.3.1 Selección y Renombramiento de Columnas:

Seleccionar las columnas relevantes y renombrarlas si es necesario para facilitar su manejo, y convertirlas al tipo de datos correcto.

4.3.2 Limpieza de Datos:

Verificar y manejar valores faltantes en las columnas total_bedrooms y cualquier otra columna relevante.

5 Parte 4: Análisis Estadístico de los Datos

5.1 Cálculo de Medidas de Tendencia Central y Dispersión

Calcular la media, mediana y desviación estándar para las variables total_rooms, total_bedrooms, population, y median_income.

5.2 Análisis de Percentiles

Determinar los percentiles 25, 50 y 75 para la variable median_house_value para entender mejor la distribución de los valores de las viviendas.

5.3 Correlaciones entre Variables

Utilizar el método .corr() para calcular la correlación entre las variables median_income, median_house_value, y total_rooms. Este análisis ayudará a entender qué tan relacionadas están estas variables entre sí.

6 Parte 5: Visualización Avanzada de Datos

6.1 Gráficos de Barras para Categorías de Edad

Generar un gráfico de líneas para visualizar cómo varía la median_income a lo largo de diferentes categorías de edad de las casas.

6.2 Gráfico de Líneas para Evolución de la Mediana de Ingresos

Crear gráficos de barras que muestren el número promedio de habitaciones por categoría de edad de la vivienda (housing_median_age).

6.3 Gráficos de Densidad para Variables Clave

Utilizar gráficos de densidad para visualizar la distribución de median_house_value y median_income, comparando la densidad de estas dos variables para identificar patrones o anomalías.

6.4 Visualización de Correlaciones

Crear un mapa de calor para visualizar las correlaciones entre las variables, usando bibliotecas como Seaborn o Matplotlib.

7 Parte 6: Conclusiones y Recomendaciones

Elaboren conclusiones basadas en los patrones estadísticos y las tendencias visuales observadas.

8 Cierre

Recordar guardar y documentar adecuadamente el trabajo realizado en Google Colab.