Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

Unidad Iztapalapa

Caracterización de datos de trayectorias individuales

Presentado por:
Jorge Rafael Martínez Buenrostro

Asesora: Dra. Elizabeth Pérez Cortés

México, CDMX, a 31 de julio de 2025

Resumen

Una trayectoria se define como la secuencia de desplazamientos realizada por un individuo en movimiento, compuesta por tres elementos principales: `puntos de recorrido`, `tiempos de pausa` y `longitudes de vuelo`. Los puntos de recorrido corresponden a las ubicaciones o pasos específicos por los que transita el individuo. Los tiempos de pausa representan los intervalos durante los cuales el individuo permanece detenenido en un mismo punto de recorrido. Por último, la longitud de vuelo se refiere a la distancia recorrida entre dos puntos de recorrido consecutivos. En el presente trabajo se describe el proceso de caraterización de datos de movilidad, es decir, el proceso de limpieza y depuración de la información, mediante el cual elimina aquellos campos y registros que no le aportan valor a la trayectoria individual. El objetivo es identificar la mayor cantidad de trayectorias individuales, para así poder crear un modelo que permita simular el movimiento de individuos.

Contenido

Lista de Figuras

Lista de Códigos

Capítulo 1

Introducción del Proyecto

1.1 Descripción general del proyecto

La simulación de una red de comunicaciones con dispositivos personales requiere modelos que representen fielmente los patrones de movimiento de las personas. De lo contrario, las conclusiones derivadas de dicha simulación pueden ser poco útiles. Para avanzar hacia la definición de un modelo de trayectorias individuales, se propone caracterizar los datos de una base existente que permita modelar trayectorias de forma eficaz.

1.2 Objetivos y propósitos

El objetivo principal del proyecto es obtener una caracterización estadística de las trayectorias individuales.

Los propósitos específicos son:

- Caracterizar la base de datos para extraer las trayectorias contenidas.
- Aplicar un modelo de inteligencia artificial para identificar y analizar dichas trayectorias.

1.3 Alcance del sistema

El sistema se enfoca en la identificación de trayectorias peatonales individuales y su análisis mediante herramientas de IA. El alcance incluye:

- Caracterización de la base de datos existente.
- Identificación de trayectorias individuales.
- Generación de reportes y visualizaciones de los resultados.

No se incluye la creación de modelos de IA desde cero; se emplearán herramientas y modelos ya existentes.

Capítulo 2

Requisitos del sistema

2.1 Requisitos

Docker: version 28.2.2, build e6534b4

Docker Compose: version 1.29.2, build unknown

2.1.1 Instrucciones de instalación

1. Clonar el repositorio del proyecto desde el siguiente enlace el proyecto se encuentra dentro de la carpeta `Implementación`.
2. Para levantar y acceder al contenedor, ejecutar el siguiente script:[1]

```
./start_container.sh     # Linux/Mac
.\start_container.bat    # Windows
```

Código 2.1: Iniciar contenedor del proyecto.

3. Para cerrar el contenedor del proyecto, ejecutar el siguiente script:

```
./stop_container.sh   # Linux/Mac
.\stop_container.bat    # Windows
```

Código 2.2: Iniciar contenedor del proyecto.

4. Para ver reflejados los cambios realizados en el código, ejecuta el siguiente script:

```
./restart_container.sh   # Linux/Mac
.\restart_container.bat    # Windows
```

Código 2.3: Reiniciar contenedor del proyecto.

[1]Para más detalles sobre el uso de Docker y Docker Compose, consulte el Apéndice A.

Capítulo 3

Caracterización de datos de trayectorias individuales

3.1 Caracterización de datos de trayectorias individuales

El análisis de datos comienza con una etapa fundamental: la caracterización del conjunto de datos. Esta fase tiene como objetivo examinar y comprender la estructura, el contenido y las principales propiedades de los datos antes de aplicar técnicas analíticas más complejas. En el caso de los datos de trayectorias individuales, la caracterización permite identificar posibles inconsistencias, redundancias y elementos irrelevantes que puedan afectar la calidad del análisis. Las tareas principales llevadas a cabo en esta etapa son las siguientes:

- Explorar las primeras filas del conjunto de datos para obtener una visión general de su estructura.
- Verificar la cantidad total de registros y columnas disponibles.
- Identificar y eliminar columnas que no aportan información relevante para el análisis o inconsistentes.
- Identificar y eliminar las filas que no aportan información relevante para el análisis o inconsistentes.

A continuación, se describen en detalle las acciones específicas realizadas durante el proceso de caracterización.

3.1.1 Exploración inicial del conjunto de datos

Como primer paso en la caracterización, se realizó una exploración preliminar del conjunto de datos con el fin de comprender su estructura general. Para ello, se inspeccionaron las primeras dos filas, lo cual permitió identificar las columnas presentes y observar ejemplos representativos de sus valores. El código utilizado para realizar esta exploración se encuentra en el Apéndice B.1. A continuación, se presenta un resumen de las columnas detectadas junto con una muestra de sus respectivos valores:

1. `id`: Identificador numérico único por registro
   `['34284565','34284566']`
2. `identifier`: UUID del dispositivo
   `['f2640430-7e39-41b7-80bb-3fddaa44779c']`
3. `identifier_type`: Tipo de ID (ej. `'gaid'` para Android)
   `['gaid', 'gaid']`
4. `timestamp`: Fecha-hora del registro
   `['2022-11-07 02:04:21']`
5. `device_lat`/`device_lon`: Coordenadas GPS
   `['21.843149']`, `['-102.196838']`
6. `country_short`/`province_short`: Códigos de ubicación
   `['MX']`, `['MX.01']`
7. `ip_address`: Dirección IPv6
   `['2806:103e:16::']`
8. `device_horizontal_accuracy`: Precisión GPS en metros
   `['8.0']`
9. `source_id`: Hash de la fuente de datos
   `['449d086d...344']`
10. `record_id`: Hash único por registro
    `['77d795df...']`
11. `home_country_code`: País de residencia
    `['MX']`
12. `home_geog_point`/`work_geog_point`: Coordenadas en WKT
    `['POINT(-102.37038 22.20753)']`
13. `home_hex_id`/`work_hex_id`: ID hexagonal (H3)
    `['85498853fffffff']`
14. `data_execute`: Fecha de procesamiento
    `['2023-05-30']`
15. `time_zone_name`: Zona horaria
    `['America/Mexico_City']`

3.1.2 Dimensiones del conjunto de datos

Para verificar las dimensiones del conjunto de datos, se utilizó la bilbioteca Dask, que permite trabajar con grandes volúmenes de datos de manera eficiente. Junto con Python se usó el código en el Apéndice B.2. Como resultado ahora sabemos que el conjunto de datos contiene un total de **69,980,000** registros y **19** campos. Esto indica que hay una cantidad significativa de datos disponibles para el análisis.

3.1.3 Depuración de columnas

Dado que el conjunto de datos original contiene 19 campos, es fundamental identificar y eliminar aquellas columnas que no aportan valor al análisis. Para ello, se realizó una revisión de los valores únicos presentes en cada campo, con el objetivo de detectar información redundante o irrelevante. A partir de este análisis, se identificaron las siguientes columnas como innecesarias para los fines del estudio:

- `id`
- `identifier_type`
- `country_short`
- `province_short`
- `ip_address`
- `source_id`
- `home_country_code`
- `home_geog_point`
- `work_geog_point`
- `home_hex_id`
- `work_hex_id`
- `data_execute`

En lugar de eliminar columnas explícitamente, se optó por seleccionar únicamente aquellas que se desean conservar. El código utilizado para esta tarea se encuentra incluido en el Apéndice B.3. Dicho script emplea la biblioteca `dask` para cargar y guardar una nueva versión del conjunto de datos que contiene exclusivamente las siguientes columnas relevantes:

- `identifier`
- `timestamp`
- `device_lat`
- `device_lon`
- `device_horizontal_accuracy`
- `record_id`
- `time_zone_name`

Como resultado, se genera un nuevo archivo `CSV` que conserva únicamente la información útil para el análisis posterior, optimizando así el tamaño y la calidad del conjunto de datos.

3.1.4 Depuración de filas

Una vez obtenida una versión más ligera del conjunto de datos, el siguiente paso consiste en identificar y eliminar aquellas filas que no aportan valor al análisis. Para ello, se generaron representaciones gráficas que permiten observar la distribución de los datos y facilitar la toma de decisiones. Las columnas seleccionadas para este proceso fueron:

- `identifier`: Identificador único del dispositivo.
- `device_horizontal_accuracy`: Precisión del GPS en metros. A menor valor, mayor precisión.

La primera columna a analizar será `device_horizontal_accuracy`, que refleja la precisión del GPS en metros. Este valor depende tanto del sistema de medición como de la fuente de datos, y suele clasificarse según la siguiente escala:

- GPS puro (satelital): 1–20 metros.
- A-GPS (asistido por red): 5–50 metros.
- Triangulación por WiFi o redes móviles: 20–500 metros.
- Geolocalización por IP: 1000–5000 metros.

Con base en esta escala, primero hay que identificar el rango de valores presentes en la columna. Para ello se utilizó el código mostrado en el Apéndice B.4, el cual extrae los valores únicos de `device_horizontal_accuracy` y los guarda en un archivo de texto. El resultado indicó que los valores oscilan entre 0.916 y 199.9, lo que permitió construir un histograma (Apéndice B.5) para analizar la frecuencia de cada valor y así evaluar su relevancia para el análisis. El resultado se muestra en la siguiente figura:

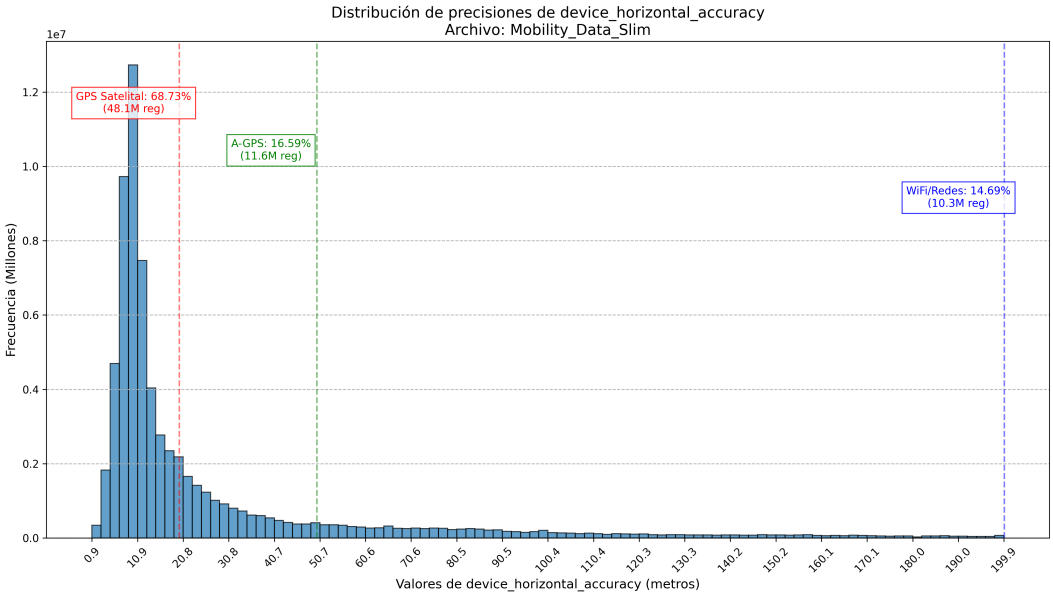


Figura 3.1: Frecuencia de aparación de los valores de 'device_horizontal_accuracy'.

Para el objetivo de este proyecto, se busca que la configuración del GPS sea lo más precisa posible, por lo que aquellos que estén dentro del rango del GPS puro (1-20 metros) son los más relevantes. Como se puede ver en la Figura 3.1, el **68.73 %** de los valores se encuentran dentro de este rango. Sin embargo, el **31.27 %** de registros con están por encima de este rango, precisión A-GPS (5-50 metros) y triangulación por WiFi/red móvil (20-500 metros).

La siguiente columna a evaluar es `identifier`, corresponde al identificador único de cada dispositivo. Para analizar la frecuencia de aparición de estos valores se empleó un script que agrupa las repeticiones por rangos y grafica la cantidad de valores únicos usando escala logarítmica (ver Apéndice B.6).

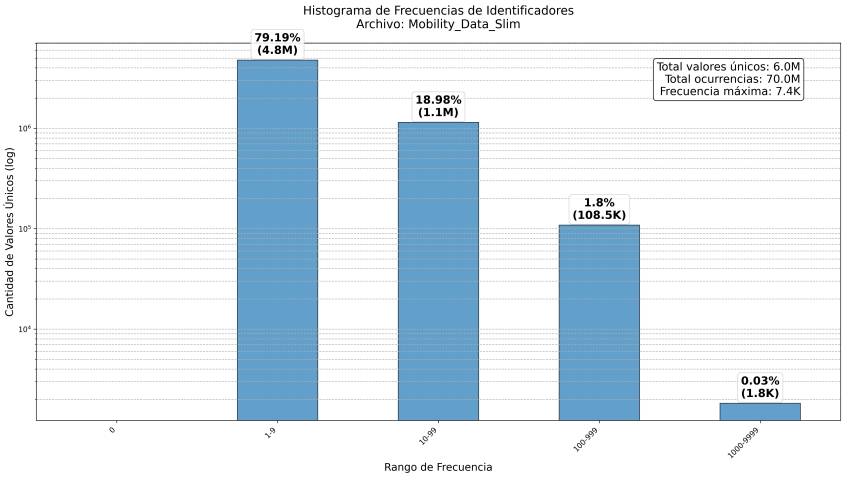
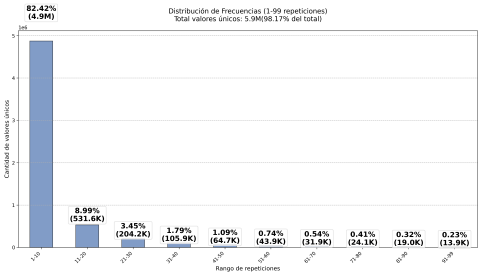
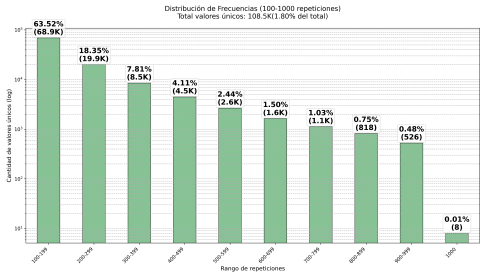


Figura 3.2: Frecuencia de aparición de los identificadores únicos.

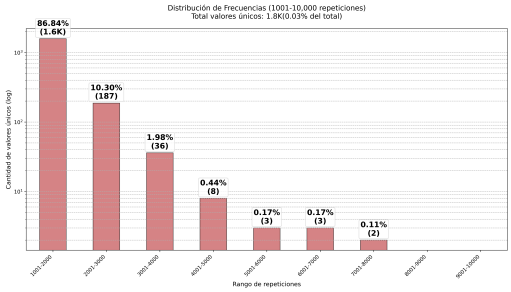
De este script sabemos que el total de individuos es de **6,022,772** de los cuales el **79.19 %** tienen una frecuencia de aparición de una a nueve veces, esto es **4,769,317** de individuos. Así mismo de la Figura 3.2 podemos observar que hay poco más de un **20 %** de individuos con más de 99 repeticiones. Por lo que se necesita hacer un análisis más detallado, para ello se ejecuta el código del Apéndice B.7, el cual segmenta los datos en tres rangos: 1-99, 100-1000 y 1001-10000 repeticiones.



(a) Histograma 1-99 repeticiones



(b) Histograma 100–1000 repeticiones



(c) Histograma 1001–10000 repeticiones

Figura 3.3: Comparación de histogramas por rangos de repeticiones.

Con la información obtenida de los histogramas de la figura anterior, se puede observar que el **98.17 %** de los identificadores únicos tienen entre 1 y 99 repeticiones, lo que equivale a **5,912,437** individuos. Por otro lado, el **1.83 %** restante tiene entre 100 y 10,000 repeticiones, lo que equivale a **110,335** individuos. Con base en esta información aún no se puede determinar que registros eliminar. Por lo que ahora se eliminarán aquellos registros que sean duplicados, es decir, aquellos que tengan el mismo `timestamp`, `device_lat` y `device_lon`. Para ello se utilizó el código del Apéndice B.8, que elimina los duplicados y genera un nuevo archivo CSV con los registros únicos.

Apéndice A

Uso de Docker y Docker Compose

En la sección 2.1 se establece como requisito el uso de Docker y Docker Compose para la ejecución del proyecto. A continuación, se detallan las instrucciones necesarias para su instalación, ya que ambas herramientas son fundamentales para la implementación. Además, se describe el archivo `docker-compose.yml`, el cual permite crear un contenedor que incluye todas las dependencias requeridas para el correcto funcionamiento del sistema.

A.1 ¿Qué son Docker y Docker Compose?

Docker es una plataforma de virtualización ligera que permite desarrollar, empaquetar y ejecutar aplicaciones en contenedores aislados. Un contenedor incluye el código, las dependencias y configuraciones necesarias para que la aplicación se ejecute de manera consistente en cualquier entorno. Esto facilita la portabilidad, escalabilidad y despliegue de software.

Docker Compose es una herramienta que permite definir y ejecutar aplicaciones multicontenedor mediante archivos de configuración YAML. A través de un solo archivo `docker-compose.yml`, es posible especificar los servicios, redes y volúmenes que componen una aplicación, simplificando así su orquestación.

Estas herramientas son fundamentales en este proyecto para garantizar que el entorno de ejecución sea replicable y controlado, independientemente del sistema operativo o configuración local del usuario.

A.2 Instalación en Linux (Ubuntu/Debian)

Para instalar Docker y Docker Compose en un sistema Linux basado en Debian o Ubuntu, siga los siguientes pasos:

1. Actualizar los paquetes del sistema:

```
1 sudo apt update
2 sudo apt upgrade
```

Código A.1: Actualizar el sistema.

2. Instalar Docker:

```
sudo apt install docker.io
sudo systemctl enable docker
sudo systemctl start docker
```

Código A.2: Instalar Docker.

3. Verificar que Docker está instalado correctamente:

```
docker --version
```

Código A.3: Verificar instalación de Docker.

4. Instalar Docker Compose:

```
sudo apt install docker-compose
```

Código A.4: Instalar Docker Compose.

5. Verificar la instalación:

```
docker-compose --version
```

Código A.5: Verificar instalación de Docker Compose.

A.3 Instalación en Windows

Para instalar Docker y Docker Compose en Windows, se recomienda utilizar Docker Desktop, que incluye ambas herramientas de forma integrada.

1. Acceder al sitio oficial: https://www.docker.com/products/docker-desktop/
2. Descargar el instalador correspondiente para Windows.
3. Ejecutar el instalador y seguir el asistente de instalación.
4. Reiniciar el sistema si es necesario.
5. Verificar que Docker y Docker Compose estén correctamente instalados desde la terminal de Windows (PowerShell o CMD):

```
docker --version
docker-compose --version
```

Código A.6: Verificar instalación de Docker y Docker Compose.

Nota: Docker Desktop requiere que la virtualización esté habilitada en la BIOS del sistema. También es necesario contar con Windows 10 o superior.

A.4 Descripción del archivo `docker-compose.yml`

El archivo `docker-compose.yml` permite definir y configurar el entorno de ejecución del proyecto utilizando un contenedor de Docker. A continuación, se presenta su contenido y una explicación de cada uno de sus elementos:

```
1  version: "3.8"
2
3  services:
4    data-analysis:
5      image: python:3.13-bookworm
6      container_name: data-analysis
7      runtime: nvidia
8      tty: true
9      stdin_open: true
10     volumes:
11       - ./:/app
12       - python-packages:/usr/local/lib/python3.13/site-packages
13     command: sh -c "pip␣install␣-r␣requirements.txt␣&&␣pip␣
           install␣-e␣.␣&&␣python3␣/app/src/main.py"
14     working_dir: /app
15     environment:
16       - PYTHONPATH=/app
17
18 volumes:
19   python-packages:
```

Código A.7: Archivo docker-compose.yml

A continuación se explica el propósito de cada sección:

- `version: "3.8"`
  Define la versión del esquema de Docker Compose utilizado. La versión 3.8 es compatible con la mayoría de las características modernas de Docker.
- `services` → `data-analysis`
  Se define un servicio llamado `data-analysis`, que representa el contenedor principal del proyecto.
- `image: python:3.13-bookworm`
  Utiliza una imagen oficial de Python 3.13 basada en Debian Bookworm como entorno base.
- `container_name: data-analysis`
  Asigna un nombre personalizado al contenedor para facilitar su identificación.
- `runtime: nvidia`
  Indica que el contenedor utilizará el runtime de NVIDIA para permitir acceso a la GPU. Requiere tener instalado `nvidia-docker`.
- `tty: true` y `stdin_open: true`
  Habilitan la interacción con el terminal del contenedor, lo que es útil para

ejecutar comandos manuales si es necesario.

- `volumes`
  - `./:/app`: Monta el directorio actual del proyecto como `/app` dentro del contenedor.
  - `python-packages:/usr/local/lib/python3.13/site-packages`: Crea un volumen persistente para las bibliotecas de Python instaladas.
- `command`
  Ejecuta una serie de comandos cuando el contenedor inicia: instala las dependencias del archivo `requirements.txt`, instala el proyecto en modo editable (`pip install -e .`) y ejecuta el archivo `main.py`.
- `working_dir: /app`
  Establece el directorio de trabajo dentro del contenedor como `/app`.
- `environment`
  Define la variable de entorno `PYTHONPATH` para que Python pueda encontrar correctamente los módulos dentro del proyecto.
- `volumes → python-packages`
  Declara un volumen persistente llamado `python-packages`, que se utiliza para almacenar los paquetes instalados sin perderlos entre reinicios del contenedor.

Este archivo permite que el entorno de desarrollo sea fácilmente replicable y ejecutable, sin necesidad de instalar manualmente dependencias o configurar rutas en el sistema anfitrión.

A.5 Scripts de control del contenedor

Para facilitar el manejo del contenedor durante el desarrollo del proyecto, se han creado tres scripts auxiliares en Bash que automatizan las operaciones más comunes: iniciar, reiniciar y detener el contenedor.

A.5.1 `start_container.sh`

Este script verifica si el contenedor `data-analysis` ya se encuentra en ejecución. En caso de que no esté activo, lo inicia utilizando `docker-compose up -d`. Posteriormente, ejecuta el archivo `main.py` dentro del contenedor.

```
#!/bin/bash

if ! docker ps --filter "name=^/data-analysis$" --filter "
    status=running" | grep -q data-analysis; then
    echo "Contenedor no esta corriendo. Levantando con docker-
        compose..."
    docker-compose up -d
    echo "Esperando que se instalen las dependencias..."
    while ! docker exec data-analysis pip show colorama &> /dev
        /null; do
        sleep 2
    done
    echo "Dependencias instaladas correctamente."
else
    echo "Contenedor ya esta corriendo. Usando instancia
        existente."
fi

echo "Ejecutando script..."
docker exec -it data-analysis python3 /app/src/main.py
```

Código A.8: Script para iniciar el contenedor.

A.5.2 `restart_container.sh`

Este script reinicia completamente el contenedor (equivalente a detenerlo y volverlo a levantar), lo cual resulta útil cuando se han modificado archivos como `requirements.txt` o `setup.py`. Tras reiniciar, vuelve a ejecutar el archivo principal del proyecto.

```
#!/bin/bash
docker restart data-analysis
sleep 2
docker exec -it data-analysis python3 /app/src/main.py
```

Código A.9: Script para reiniciar el contenedor.

A.5.3 `stop_container.sh`

Este script detiene y elimina el contenedor junto con los volúmenes asociados. Debe utilizarse con precaución, ya que elimina todas las dependencias instaladas en el entorno del contenedor. Solo es necesario en casos donde se requiere limpiar completamente el entorno.

```
1 #!/bin/bash
2 docker-compose down --volumes
```

Código A.10: Script para detener y eliminar el contenedor y sus volúmenes.

A.6 Proceso de uso y desarrollo del contenedor

A continuación se describe el flujo recomendado para desarrollar y ejecutar el sistema dentro del contenedor de Docker:

1. Verifique que Docker y Docker Compose están instalados (Apéndice A.6).
   Nota para usuarios de Windows: Si se utiliza Windows como sistema operativo, se deben usar los archivos con extensión `.bat` en lugar de `.sh`, y deben ser ejecutados desde la terminal de Windows (por ejemplo, CMD o PowerShell).
2. Asigne permisos de ejecución a los scripts:

```
1 chmod +x start_container.sh restart_container.sh
     stop_container.sh
```

Código A.11: Dar permisos de ejecución a los scripts.

3. Para iniciar el contenedor y ejecutar el proyecto con los cambios más recientes del código fuente:

```
1 ./start_container.sh
```

Código A.12: Iniciar contenedor y ejecutar el proyecto.

4. Si se realizan cambios en las dependencias o archivos de configuración del entorno (como `requirements.txt`), utilice:

```
1 ./restart_container.sh
```

Código A.13: Reiniciar el contenedor completamente.

5. Para detener el contenedor y eliminar todos los volúmenes asociados:

```
1 ./stop_container.sh
```

Código A.14: Eliminar el contenedor y limpiar el entorno.

Este conjunto de scripts permite un desarrollo ágil dentro del contenedor, ya que los cambios realizados en el código fuente local se reflejan de inmediato gracias al uso de `volumes`. Además, se reduce la necesidad de ejecutar manualmente comandos repetitivos, facilitando el trabajo del usuario final y asegurando la correcta ejecución del proyecto.

Apéndice B

Scripts para la caracterización de datos

En la sección 3.1 se describen los pasos del proceso de caracterización de datos de trayectorias individuales. En este anexo se presentan los scripts utilizados para llevar a cabo dicho proceso.

```
1  import dask.dataframe as dd
2  import sys
3
4  print("Exploracion␣inicial␣de␣datos␣con␣Dask\n")
5
6  if len(sys.argv) < 2:
7      print("Error:␣Debe␣especificar␣un␣archivo␣CSV")
8      sys.exit(1)
9
10 ruta_archivo = sys.argv[1]
11
12 ddf = dd.read_csv(
13     ruta_archivo,
14     encoding="utf-8",
15     sep=",",
16     dtype="object",
17 )
18
19 columnas = ddf.columns.tolist()
20
21 print("Columnas␣y␣2␣ejemplos␣por␣cada␣una:\n")
22 for col in columnas:
23     ejemplos = ddf[col].head(2).values.tolist()
24     print(f"-␣{col}:␣{ejemplos}")
25
26 input("Presiona␣Enter␣para␣continuar...")
```

Código B.1: csv_glance.py, exploración inicial del conjunto de datos.

```
import dask.dataframe as dd
import sys
import os

def contar_registros(ruta_archivo):

    columnas_usar = ["record_id"]
    try:
        print(f"\nCargando archivo {ruta_archivo}...")
        ddf = dd.read_csv(
            ruta_archivo,
            usecols=columnas_usar,
            sep=",",
            dtype={"record_id": "str"},
            blocksize="256MB",
        )

        print("Contando registros (paciencia para archivos grandes)...")
        total_registros = ddf.shape[0].compute()

        print(f"\nAnalisis completado:")
        print(f"Archivo analizado: {ruta_archivo}")
        print(f"Total de registros: {total_registros:,}")

    except Exception as e:
        print(f"\nOcurrio un error inesperado: {str(e)}")

if __name__ == "__main__":
    print("=== Contador de registros en archivos CSV grandes ===")

    if len(sys.argv) < 2:
        print("Uso: python csv_count_registers.py <nombre_del_archivo.csv>")
        sys.exit(1)

    archivo = sys.argv[1]
    contar_registros(archivo)
```

Código B.2: csv_count_registers.py, conteo de registros en el conjunto de datos.

```
import dask.dataframe as dd

columnas_deseadas = [
  'identifier',
  'timestamp',
  'device_lat',
  'device_lon',
  'device_horizontal_accuracy',
  'record_id',
  'time_zone_name'
]

df = dd.read_csv('Mobility_Data.csv', usecols=
    columnas_deseadas)

df.to_csv('Mobility_Data_Slim.csv', index=False, single_file=
    True, encoding='utf-8-sig')
```

Código B.3: remove_columns.py, eliminación de campos innecesarios en el conjunto de datos.

```
import pandas as pd
from tqdm import tqdm
import os
import sys
from src.menus.menu import MainMenu
def main():
    print("\n" + "="*50)
    print("␣EXTRACTOR␣DE␣VALORES␣UNICOS␣DE␣COLUMNAS␣CSV")
    print("="*50 + "\n")

    if len(sys.argv) < 2:
        print("Uso:␣python␣extract_unique.py␣<archivo.csv>")
        sys.exit(1)

    csv_file = sys.argv[1]

    if not os.path.exists(csv_file):
        print(f"Error:␣El␣archivo␣'{csv_file}'␣no␣existe.")
        sys.exit(1)

    chunk_size = 1_000_000

    try:
        available_columns = pd.read_csv(csv_file, nrows=0).
            columns.tolist()
    except Exception as e:
        print(f"Error␣leyendo␣el␣archivo:␣{e}")
        sys.exit(1)

    try:
        selected_index = MainMenu.display_available_columns(
            available_columns)
        target_column = available_columns[selected_index]
    except (ValueError, IndexError):
        print("Seleccion␣invalida.")
        sys.exit(1)
    except Exception as e:
        print(f"Error␣inesperado␣al␣seleccionar␣columna:␣{e}"
            )
        sys.exit(1)

    safe_column_name = target_column.replace("␣", "_").
        replace("/", "_")
    output_file = f"valores_unicos_{safe_column_name}.txt"

    unique_values = set()
    print(f"\nProcesando␣columna:␣{target_column}\n")

    try:
        for chunk in tqdm(pd.read_csv(csv_file, usecols=[
```

```
            target_column], chunksize=chunk_size)):
47              unique_values.update(chunk[target_column].dropna
                    ().astype(str))
48      except Exception as e:
49          print(f"Error durante el procesamiento: {e}")
50          sys.exit(1)
51
52      try:
53          numeric_values = sorted([float(v) for v in
                unique_values])
54          is_numeric = True
55      except ValueError:
56          is_numeric = False
57
58      try:
59          with open(output_file, "w", encoding="utf-8") as f:
60              if is_numeric:
61                  min_val = numeric_values[0]
62                  max_val = numeric_values[-1]
63                  f.write(f"# Rango de valores: {min_val} - {
                        max_val}\n")
64                  f.write("\n".join(str(v) for v in
                        numeric_values))
65              else:
66                  sorted_values = sorted(unique_values)
67                  f.write("# Rango de valores: No numerico\n")
68                  f.write("\n".join(sorted_values))
69      except Exception as e:
70          print(f"Error guardando los resultados: {e}")
71          sys.exit(1)
72
73      print(f"\nSe encontraron {len(unique_values):,} valores
            unicos.")
74      print(f"Resultados guardados en: {output_file}")
75
76      print("\nMuestra de valores unicos (primeros 10):")
77      print("\n".join(sorted(unique_values)[:10]))
78
79  if __name__ == "__main__":
80      main()
```

Código B.4: unique_values.py, obtención de valores únicos de la columna 'device_horizontal_accuracy'.

```
1   import os
2   import pandas as pd
3   import matplotlib.pyplot as plt
4   import numpy as np
5   import sys
6   from tqdm import tqdm
7
8   def classify_tech(valor):
9       if 1 <= valor <= 20:
10          return 'GPS Satelital'
11      elif 5 <= valor <= 50:
12          return 'A-GPS (Asistido por red)'
13      elif 20 <= valor <= 500:
14          return 'Triangulacion WiFi/Redes Moviles'
15      else:
16          return 'Fuera de rango'
17
18  def format_count(count):
19      if count >= 1_000_000:
20          return f"{count/1_000_000:.1f}M"
21      elif count >= 1_000:
22          return f"{count/1_000:.1f}K"
23      return str(count)
24
25  def main():
26      if len(sys.argv) < 2:
27          print("Error: Debe especificar un archivo CSV como
                argumento")
28          sys.exit(1)
29
30      csv_file = sys.argv[1]
31      filename = os.path.splitext(os.path.basename(csv_file))
            [0]
32      column = "device_horizontal_accuracy"
33      bins = 100
34
35      print(f"\nIniciando procesamiento del archivo: {
            csv_file}")
36      print(f"Columna analizada: {column}")
37
38      os.makedirs("img", exist_ok=True)
39      print("Directorio 'img' verificado/creado")
40
41      print("\nProcesando datos y clasificando tecnologias...
            ")
42      frequency = pd.Series(dtype=float)
43      tech_counts = {
44          'GPS Satelital': 0,
45          'A-GPS (Asistido por red)': 0,
46          'Triangulacion WiFi/Redes Moviles': 0,
```

```
47              'Fuera de rango': 0
48          }
49
50          total_row = sum(1 for _ in pd.read_csv(csv_file,
                usecols=[column], chunksize=1_000_000))
51
52          with tqdm(total=total_row, unit='M rows') as pbar:
53              for chunk in pd.read_csv(csv_file, usecols=[column
                    ], chunksize=1_000_000):
54                  chunk_clean = chunk[column].dropna()
55
56                  for valor in chunk_clean:
57                      tech = classify_tech(valor)
58                      tech_counts[tech] += 1
59
60                  counts = chunk_clean.value_counts()
61                  if not frequency.empty or not counts.empty:
62                      frequency = pd.concat([frequency, counts],
                            axis=0).groupby(level=0).sum()
63                  pbar.update(1)
64
65          total = sum(tech_counts.values())
66          percentage = {k: (v/total)*100 for k, v in tech_counts.
                items()}
67
68          print("\nGenerando histograma con estadIsticas...")
69          counts, edges = np.histogram(frequency.index, bins=bins
                , weights=frequency.values)
70
71          plt.figure(figsize=(14, 8))
72          plt.bar(edges[:-1], counts, width=np.diff(edges), align
                ='edge', edgecolor='black', alpha=0.7)
73
74          plt.axvline(x=20, color='r', linestyle='--', alpha=0.5)
75          plt.axvline(x=50, color='g', linestyle='--', alpha=0.5)
76          plt.axvline(x=200, color='b', linestyle='--', alpha
                =0.5)
77
78          gps_str = f"GPS Satelital: {percentage['GPS Satelital
                ']:.2f}%\n({format_count(tech_counts['GPS Satelital
                '])} reg)"
79          agps_str = f"A-GPS: {percentage['A-GPS (Asistido por
                red)']:.2f}%\n({format_count(tech_counts['A-GPS (
                Asistido por red)'])} reg)"
80          wifi_str = f"WiFi/Redes: {percentage['Triangulacion
                WiFi/Redes Moviles']:.2f}%\n({format_count(
                tech_counts['Triangulacion WiFi/Redes Moviles'])}
                reg)"
81
82          plt.text(10, max(counts)*0.9, gps_str, ha='center',
                color='r', fontsize=10,
```

```
83                  bbox=dict(facecolor='white', alpha=0.8,
                        edgecolor='r'))
84          plt.text(40, max(counts)*0.8, agps_str, ha='center',
                color='g', fontsize=10,
85                  bbox=dict(facecolor='white', alpha=0.8,
                        edgecolor='g'))
86          plt.text(190, max(counts)*0.7, wifi_str, ha='center',
                color='b', fontsize=10,
87                  bbox=dict(facecolor='white', alpha=0.8,
                        edgecolor='b'))
88
89
90          plt.title(f"DistribuciOn de precisiones de {column}\
                nArchivo: {filename}", fontsize=14)
91          plt.xlabel(f"Valores de {column} (metros)", fontsize
                =12)
92          plt.ylabel("Frecuencia (Millones)", fontsize=12)
93          plt.xticks(edges[::5], rotation=45)
94          plt.grid(axis='y', linestyle='--')
95          plt.tight_layout()
96
97          output_path = os.path.join("img", f"histograma_{column}
                _{filename}.png")
98          plt.savefig(output_path, dpi=300, bbox_inches='tight')
99          plt.close()
100
101         print("\n=== DISTRIBUCION DE TECNOLOGIAS DE
                GEOLOCALIZACION ===")
102         for tech, count in tech_counts.items():
103             print(f"{tech}: {count:,} registros ({percentage[
                    tech]:.2f}%)")
104
105         print(f"\nHistograma generado exitosamente")
106         print(f"Archivo guardado en: {output_path}")
107         print(f"Total registros analizados: {total:,}\n")
108
109         if __name__ == "__main__":
110         main()
```

Código B.5: accuracy_histogram.py, creación de un histograma de frecuencias de la columna 'device_horizontal_accuracy'.

```
1   import os
2   import pandas as pd
3   import matplotlib.pyplot as plt
4   import numpy as np
5   from collections import Counter
6   import sys
7   from tqdm import tqdm
8   import math
9
10  def format_count(count):
11      if count >= 1_000_000:
12          return f"{count/1_000_000:.1f}M"
13      elif count >= 1_000:
14          return f"{count/1_000:.1f}K"
15      return str(count)
16
17  def main():
18      if len(sys.argv) < 2:
19          print("Error: Debe especificar un archivo CSV como
                argumento")
20          sys.exit(1)
21
22      csv_file = sys.argv[1]
23      filename = os.path.splitext(os.path.basename(csv_file))
            [0]
24      column = "identifier"
25      chunksize = 1_000_000
26
27      print(f"\nIniciando procesamiento del archivo: {
            csv_file}")
28      print(f"Columna analizada: {column}")
29      os.makedirs("img", exist_ok=True)
30      print("Directorio 'img' verificado/creado")
31
32      print("\nProcesando datos y contando frecuencias...")
33      counter = Counter()
34
35      total_chunks = sum(1 for _ in pd.read_csv(csv_file,
            usecols=[column], chunksize=chunksize))
36
37      with tqdm(total=total_chunks, unit=' chunk') as pbar:
38          for chunk in pd.read_csv(csv_file, usecols=[column
                ], chunksize=chunksize):
39              counter.update(chunk[column].dropna().astype(
                    str))
40              pbar.update(1)
41
42      frecuency = pd.Series(counter)
43      total_unique_values = len(frecuency)
44      max_freq = frecuency.max()
```

```
45
46          print(f"Datos␣procesados␣correctamente")
47          print(f"Total␣de␣valores␣Unicos:␣{total_unique_values
                :,}")
48          print(f"Frecuencia␣maxima:␣{max_freq:,}")
49
50          bins = [0] + [10**i for i in range(0, int(np.log10(
                max_freq)) + 2)]
51          group_freq = pd.cut(frecuency, bins=bins, right=False).
                value_counts().sort_index()
52
53          total_ocurrence = frecuency.sum()
54          percentage_per_range = (group_freq /
                total_unique_values * 100).round(2)
55
56          print("\nGenerando␣histograma␣con␣estadisticas...")
57          plt.figure(figsize=(16, 9))
58          ax = group_freq.plot(kind='bar', logy=True, alpha=0.7,
                edgecolor='black')
59
60          formatted_labels = []
61          for interval in group_freq.index.categories:
62              left = int(interval.left)
63              right = int(interval.right - 1)
64              formatted_labels.append(f"{left}-{right}" if left
                    != right else f"{left}")
65
66          plt.xticks(range(len(formatted_labels)),
                formatted_labels, rotation=45, ha='right')
67
68          plt.title(f"Histograma␣de␣Frecuencias␣de␣
                Identificadores\nArchivo:␣{filename}", fontsize=16,
                pad=20)
69          plt.xlabel("Rango␣de␣Frecuencia", fontsize=14)
70          plt.ylabel("Cantidad␣de␣Valores␣Unicos␣(log)", fontsize
                =14)
71          plt.grid(True, which="both", ls="--", axis='y')
72
73          stats_text = (
74              f"Total␣valores␣unicos:␣{format_count(
                    total_unique_values)}\n"
75              f"Total␣ocurrencias:␣{format_count(total_ocurrence)
                    }\n"
76              f"Frecuencia␣maxima:␣{format_count(max_freq)}"
77          )
78          plt.annotate(stats_text,
79                       xy=(0.95, 0.95),
80                       xycoords='axes␣fraction',
81                       fontsize=15,
82                       ha='right',
83                       va='top',
```

```
84                      bbox=dict(boxstyle='round', facecolor='
                            white', alpha=0.9))
85
86          max_val = group_freq.max()
87          min_y = 0.9
88
89          for i, (count, porcent) in enumerate(zip(group_freq.
                values, percentage_per_range.values)):
90              if count > 0:
91                  y_pos = count * 1.1 if count * 1.1 > min_y else
                        min_y * 1.2
92
93                  text = f"{porcent}%\n({format_count(count)})"
94
95                  ax.text(
96                      i, y_pos, text,
97                      ha='center', va='bottom',
98                      fontsize=15,
99                      fontweight='bold',
100                     bbox=dict(
101                         facecolor='white',
102                         alpha=0.85,
103                         edgecolor='lightgray',
104                         boxstyle='round,pad=0.3'
105                     )
106                 )
107
108         output_path = os.path.join("img", f"histograma_{column}
                _{filename}.png")
109         plt.tight_layout()
110         plt.savefig(output_path, dpi=300, bbox_inches='tight')
111         plt.close()
112
113         print("\n=== DISTRIBUCION DE FRECUENCIAS ===")
114         for i, (intervalo, count) in enumerate(group_freq.items
                ()):
115             print(f"Rango {formatted_labels[i]}: {count:,}")
116
117         print(f"\nHistograma generado exitosamente")
118         print(f"Archivo guardado en: {output_path}")
119         print(f"Total ocurrencias analizadas: {total_ocurrence
                :,}\n")
120
121     if __name__ == "__main__":
122         main()
```

Código B.6: identifier_histogram.py, creación de un histograma de frecuencias de la columna 'identifier'.

```
import os
 import pandas as pd
 import matplotlib.pyplot as plt
 import numpy as np
 from collections import Counter
 import sys
 from tqdm import tqdm

 def format_count(count):
     if count >= 1_000_000:
         return f"{count/1_000_000:.1f}M"
     elif count >= 1_000:
         return f"{count/1_000:.1f}K"
     return str(count)

 def create_histogram(data, bins, title, filename, color='
    skyblue', log_scale=False):
     grouped = pd.cut(data, bins=bins, right=False).
        value_counts().sort_index()
     total_values = len(data)
     max_count = grouped.max()

     plt.figure(figsize=(14, 8))
     ax = grouped.plot(kind='bar', color=color, edgecolor='
        black', alpha=0.7, logy=log_scale)

     bin_labels = []
     for interval in grouped.index.categories:
         left = int(interval.left)
         right = int(interval.right)
         bin_labels.append(f"{left}-{right-1}" if right-left
             > 1 else str(left))

     plt.xticks(range(len(bin_labels)), bin_labels, rotation
        =45, ha='right')
     plt.title(f"{title}\nTotal␣valores␣unicos:␣{
        format_count(total_values)}", fontsize=14, pad=20)
     plt.xlabel("Rango␣de␣repeticiones", fontsize=12)
     plt.ylabel("Cantidad␣de␣valores␣unicos" + ("␣(log)" if
        log_scale else ""), fontsize=12)
     plt.grid(True, which="both", ls="--", axis='y')

     min_y = 0.9
     for i, (count, interval) in enumerate(zip(grouped.
        values, grouped.index)):
         if count > 0:
             percentage = (count / total_values) * 100
             y_pos = count * 1.1 if count * 1.1 > min_y else
                 min_y * 1.2
             text = f"{percentage:.2f}%\n({format_count(
```

```
                    count)})"
42
43                 ax.text(i, y_pos, text,
44                     ha='center', va='bottom',
45                     fontsize=15, fontweight='bold',
46                     bbox=dict(facecolor='white', alpha=0.8,
                           edgecolor='lightgray', boxstyle='round,
                           pad=0.2'))
47
48         output_path = os.path.join("img", filename)
49         plt.tight_layout()
50         plt.savefig(output_path, dpi=300, bbox_inches='tight')
51         plt.close()
52         return output_path
53
54     def main():
55         if len(sys.argv) < 2:
56             print("Error: Debe especificar un archivo CSV")
57             sys.exit(1)
58
59         csv_file = sys.argv[1]
60         filename_base = os.path.splitext(os.path.basename(
               csv_file))[0]
61         column = "identifier"
62         chunksize = 1_000_000
63         os.makedirs("img", exist_ok=True)
64
65         print(f"\nIniciando analisis de: {csv_file}")
66         print(f"Columna analizada: {column}")
67
68         print("\nContando frecuencias...")
69         counter = Counter()
70
71         with tqdm(desc="  Contando filas totales", unit=' filas
               ') as pbar:
72             total_rows = 0
73             for chunk in pd.read_csv(csv_file, usecols=[column
                   ], chunksize=chunksize):
74                 total_rows += len(chunk)
75                 pbar.update(len(chunk))
76
77         with tqdm(total=total_rows, desc="  Procesando datos",
               unit=' filas') as pbar:
78             for chunk in pd.read_csv(csv_file, usecols=[column
                   ], chunksize=chunksize):
79                 counter.update(chunk[column].dropna().astype(
                       str))
80                 pbar.update(len(chunk))
81
82         frequencies = pd.Series(counter)
83         total_unique = len(frequencies)
```

```
84          print(f"\nDatos␣procesados␣-␣Total␣valores␣unicos:␣{
               format_count(total_unique)}")
85
86          print("\nClasificando␣frecuencias...")
87          with tqdm(total=4, desc="␣␣Progreso") as pbar:
88              low_freq = frequencies[(frequencies >= 1) & (
                   frequencies <= 99)]
89              pbar.update(1)
90              mid_freq = frequencies[(frequencies >= 100) & (
                   frequencies <= 1000)]
91              pbar.update(1)
92              high_freq = frequencies[(frequencies >= 1001) & (
                   frequencies <= 10000)]
93              pbar.update(1)
94
95          low_bin = list(range(1, 100, 10)) + [100]
96          mid_bin = list(range(100, 1001, 100)) + [1001]
97          high_bin = list(range(1001, 10001, 1000)) + [10001]
98
99          print("\n===Resumen␣de␣frecuencias␣===")
100         print(f"\nRango␣1-99␣repeticiones:")
101         print(f"␣␣␣-␣Valores␣unicos:␣{format_count(len(low_freq
               ))}␣({len(low_freq)/total_unique:.1%})")
102
103         print(f"\nRango␣100-1000␣repeticiones:")
104         print(f"␣␣␣-␣Valores␣unicos:␣{format_count(len(mid_freq
               ))}␣({len(mid_freq)/total_unique:.1%})")
105
106         print(f"\nRango␣1001-10000␣repeticiones:")
107         print(f"␣␣␣-␣Valores␣unicos:␣{format_count(len(
               high_freq))}␣({len(high_freq)/total_unique:.1%})")
108
109
110         print("\nGenerando␣graficos...")
111         with tqdm(total=3, desc="␣␣Progreso") as pbar:
112             low_path = create_histogram(
113                 low_freq,
114                 bins=low_bin,
115                 title="Distribucion␣de␣Frecuencias␣(1-99␣
                       repeticiones)",
116                 filename=f"histograma_1-99_{column}_{
                       filename_base}.png",
117                 color='#4C72B0'
118             )
119             pbar.update(1)
120
121             mid_path = create_histogram(
122                 mid_freq,
123                 bins=mid_bin,
124                 title="Distribucion␣de␣Frecuencias␣(100-1000␣
                       repeticiones)",
```

```
125                 filename=f"histograma_100-1k_{column}_{
                        filename_base}.png",
126                 color='#55A868',
127                 log_scale=True
128             )
129             pbar.update(1)
130 
131             high_path = create_histogram(
132                 high_freq,
133                 bins=high_bin,
134                 title="Distribucion de Frecuencias (1001-10,000
                        repeticiones)",
135                 filename=f"histograma_1k-10k_{column}_{
                        filename_base}.png",
136                 color='#C44E52',
137                 log_scale=True
138             )
139             pbar.update(1)
140 
141         print("\nGraficos generados exitosamente:")
142         print(f"{low_path}")
143         print(f"{mid_path}")
144         print(f"{high_path}")
145 
146     if __name__ == "__main__":
147         main()
```

Código B.7: identifier_histogram_detailed.py, análisis de frecuencias de la columna 'identifier'.

```
import dask.dataframe as dd
import sys
import os

def delete_duplicates(input_file, output_file):

    ddf = dd.read_csv(input_file)

    print(f"\nProcesando archivo: {input_file}")
    print(f"Numero inicial de registros: {len(ddf):,}")

    ddf_deduplicate = ddf.drop_duplicates(
        subset=['timestamp', 'device_lon', 'device_lat'],
        keep='first'
    )

    print(f"Numero de registros despues de eliminar
       duplicados: {len(ddf_deduplicate):,}")

    ddf_deduplicate.to_csv(
        output_file,
        index=False,
        single_file=True
    )

    print(f"\nArchivo sin duplicados guardado en: {
       output_file}")

if __name__ == "__main__":
    if len(sys.argv) < 2:
        print("Error: Debe especificar un archivo CSV como
           argumento")
        sys.exit(1)

    input_csv = sys.argv[1]
    base_name = os.path.splitext(input_csv)[0]
    output_csv = f"{base_name}_DeDuplicate.csv"

    delete_duplicates(input_csv, output_csv)
```

Código B.8: csv_deduplicate.py, eliminación de duplicados en el conjunto de datos.

Referencias

[1] Autor referencia 1

[2] Autor referencia 2

[3] Autor referencia 3