Equipo Jerry

Reporte Etapa 2: Data Understanding

Abstract

En esta etapa 2 “Data Understanding” de CRISP-DM, analizaremos los datos que se nos han entregado, realizando los procesos de obtención, descripción, exploración y verificación.

Obtención de datos

Reporte inicial de recolección de datos

En el proceso de obtención se nos compartió un dataset preparado para trabajar con el proyecto, sin necesidad de hacer cambios o modificaciones, es decir, previamente revisado, dichos datos se cargaron a un servidor en el que se realizará todo el procesamiento.

Legalmente no hay problemas con el tratamiento de los datos ya que han sido preprocesados con el fin de obtener el anonimato de los usuarios, con valores únicos, sin embargo fue necesario obtener información adicional a la proporcionada, de las universidades a analizar (latitudes y longitudes).

En cuanto a información relevante, el socio formador nos especificó que los datos pertenecen a alrededor del 30% de la población de Santiago de Chile, debido a que la empresa Movistar, la cual proporciona estos datos, tiene ese porcentaje de usuarios de todas las demás operadoras móviles.

Para cargar los datos utilizamos las herramientas Pandas y Spark..

Descripción de datos

Reporte de descripción de los datos

Dentro de los datos contamos con diferentes columnas con diferentes tipos de datos:

- Phone_ID: Esta columna contiene el ID del usuario en un String
- Timestamp: Contiene la fecha y hora en la que se conectó el usuario a la antena. El tipo de dato es string.
- bts_id: Contiene los ID de las antenas a la cual se hizo la conexión. El tipo de dato es String.
- Lat: contiene la latitud de la antena en cuestión. El tipo de dato es float.
- Lon: contiene la longitud de la antena en cuestión. El tipo de dato es float.

Se cuenta con 49,618,132 registros cada uno con 5 campos.

Valores únicos:

count(DISTINCT PHONE_ID)	count(DISTINCT timestamp)	count(DISTINCT bts_id)	count(DISTINCT lat)	count(DISTINCT lon)
1353435	86400	1871	1198	1264

PHONE_ID = 1,353,435 bts_id = 1871 lon = 1264
timestamp = 86,400 lat = 1198

Valores nulos:

```
+--------+---------+------+---+---+
|PHONE_ID|timestamp|bts_id|lat|lon|
+--------+---------+------+---+---+
|       0|        0|     0|  0|  0|
+--------+---------+------+---+---+
```

PHONE_ID = 0 bts_id = 0 lon = 0
timestamp = 0 lat = 0

Exploración de los datos

Reporte de Exploración

Ya que se obtuvieron los datos, decidimos obtener recomendaciones de hardware para usar este dataset, ya que es muy grande (5.91 GB).

Para ello lo que planeamos fue usar un recurso externo a nuestras laptops ya que cuentan con recursos algo limitados, considerando que Windows y Linux con entorno gráfico por sí solos ya ocupan mucha RAM y parte de la CPU. Las máquinas virtuales no son una buena opción.

El recurso que decidimos utilizar fue una instancia de Oracle Cloud, este dispone de un procesador de servidor ARM de 4 núcleos y 24 GB de memoria RAM. El sistema operativo que instalamos fue Ubuntu Server. Este por sí solo sólo ocupa alrededor de 270 MB de RAM y 0.01% de CPU en estado inactivo. Ya cuando tenemos el entorno de Jupyter iniciado, el uso de RAM aumenta a 326 MB y el CPU está en 0.1%, esto nos deja casi todo el hardware disponible para trabajar.

Para conocer a profundidad nuestras variables desplegamos una descripción estadística de los datos, la cual nos indica el conteo total de valores, la media, la desviación estándar, su valor mínimo y máximo. El resultado de dicho análisis se encuentra en la siguiente figura:

```
+-------+--------------------+--------------------+--------+-------------------+-------------------+
|summary|            PHONE_ID|           timestamp|  bts_id|                lat|                lon|
+-------+--------------------+--------------------+--------+-------------------+-------------------+
|  count|            49618132|            49618132|49618132|           49618132|           49618132|
|   mean|                null|                null|    null|-33.491992948062865|-70.68544975733191|
| stddev|                null|                null|    null|0.12192806227645299|0.1482996866680921|
|    min|00000000e3fc09803...|2021-01-01T00:00:...|   11SEP|           -32.9255|           -70.0588|
|    max|ffffea064367ea293...|2021-01-01T23:59:...|   ZPTG1|           -34.0268|           -71.4888|
+-------+--------------------+--------------------+--------+-------------------+-------------------+
```

Verificación de calidad de datos

Reporte de Calidad de datos

De acuerdo con el reporte de descripción de los datos, podemos decir que los datos se encuentran completos y que son de calidad ya que no existen datos nulos, no hay datos faltantes y tampoco hay varios tipos de datos mezclados en una sola columna.

El formato y valores de cada columna es consistente y la cantidad de datos es inmensa para realizar un buen análisis desde nuestro punto de vista.