Inteligencia artificial avanzada para la ciencia de datos

Reto

Etapa 2. Entendimiento de los datos

Diego Arturo Padilla Domínguez - A01552594

Keyuan Zhao - A01366831

Carolina Herrera Martínez - A01411547

Cutberto Arizabalo Nava - A01411431

Jose Pablo Cobos Austria - A01274631

Campus Querétaro

7 de octubre de 2022

Etapa 2: “Data Understanding”

Como segunda etapa de la metodología CRISP-DM para nuestro proyecto, en este documento desarrollaremos la fase de: "Data Understanding" con el objetivo principal de realizar un análisis completo de nuestros datos, iniciando desde su proceso de obtención, descripción, exploración y finalizando con la verificación de estos mismos.

Colección inicial de datos

Iniciando con la parte de la recolección de datos, debido a la naturaleza del proyecto no fue complicado porque el socio formador nos apoyó dándonos el dataset ya formateado (por cuestiones de legalidad) y un repositorio con información logística. Tras esto podemos decir que todos los datos recopilados representan el 30% de la población en la ciudad de Santiago de Chile, lo cual muestran la conexión por torres en un día normal.

Además, para poder acceder a ellos y realizar su lectura, se hizo uso primeramente de la herramienta de excel, que nos permite visualizarlos de

Descripción de los datos

Diccionario de datos

Nombre	**Descripción**	**Tipo de dato**	**Comentarios**
PHONE_ID	Id único de cada celular.	String	Este es la seudomizacion del número telefónico. El mismo celular siempre tendrá el mismo phone id.
timestamp	Hora de conexión a una torre.	String	Tiempo exacto del día y la hora en el que se conectó el dispositivo a la torre.
bts_id	Id único de cada torre	String	Corresponde al Id único de cada torre.
lat	Latitud de la torre	Float	Coordenada de la torre.
lon	Longitud de la torre	Float	

Cantidad de datos: 49618132

PHONE_ID

Valores únicos	1353435
Valores nulos	0

timestamp

Valores únicos	86400
Valores nulos	0

bts_id

Valores únicos	1871
Valores nulos	0
Más común	MORRF

lat

Valores únicos	1198
Valores nulos	0

lon

Valores únicos	1264
Valores nulos	0

Calidad de datos

Tras haber realizado una exploración de los datos que tenemos, encontramos que los datos son de calidad.

Los valores son de conexiones reales a cada una de las antenas, proporcionadas por 'Telefónica', los valores son consistentes y se encuentran estandarizados. No se encuentran valores faltantes en ninguno de los registros.

El formato de los datos es consistente, ya que todos los registros cuentan con los mismos campos llenados y los valores cuentan con el mismo formato en todos los registros.

La cantidad de datos proporcionados es suficiente para modelar los viajes de la población de Santiago de Chile, porque contamos con más de 5 gb de datos de conexiones a antenas.

Todas las coordenadas de las torres están dentro de Santiago de Chile y todas las torres tienen por lo menos una conexión.