

BunkerDB - Data Scientist - Examen de Ingreso

Este ejercicio deberá ser resuelto en Python, utilizando las mejores prácticas de código. El entregable deberá ser una carpeta de archivos .py en un repositorio de Github privado, con el pipeline armado y un README.md explicando el contenido de cada archivo y cómo se ejecuta el análisis. Por favor darle acceso de lectura a @NicolasMatrices-v2 y @demianmatarazzobunker.

Enunciado:

Las Ecobicis en la Ciudad de Buenos Aires, Argentina son bicicletas públicas que los residentes pueden alquilar por períodos de una hora como máximo. Cada bicicleta se ancla en una estación. La base de datos con los viajes realizados por los usuarios de una estación a otra durante el 2022 será adjuntada en el correo.

Queremos armar un modelo de predicción de demanda de bicicletas para cada cuadrante de la Ciudad de Buenos Aires. Habrá cuatro cuadrantes, identificados como Noreste, Noroeste, Sudeste y Sudoeste y las estaciones deberán ser agregadas a nivel cuadrante. Cada cuadrante se define por los siguientes puntos cardinales (longitud, latitud):

- Punto A: (-58.43, -34.68);
- Punto B: (-58.43, -34.54);
- Punto C: (-58.55, -34.6);
- Punto D: (-58.35, -34.6);

Consignas

1. Asociar cada estación a su cuadrante. Realizar un análisis exploratorio acotado.
2. Preprocesar la base de datos. Definir criterio de outliers y excluirlas.
3. Selección de features e ingeniería de nuevas features que serán incorporadas dentro del modelo para poder predecir la demanda en cada uno de los cuadrantes.
4. Armar al menos dos modelos de predicción de demanda.
5. Definir y evaluar las métricas de error para el modelo en cuestión.
6. Tunear modelo.
7. Visualizar distintas corridas para definir cuál fue la mejor.

A continuación, describimos las variables que están contenidas dentro de la base de datos provista:

Column Name	Description	Variable type
id_recorrido	Unique identifier of trip.	String
duracion_recorrido	Duration of trip in seconds	Integer
Fecha_recorrido (origen/destino)	Timestamp of when the trip occurred. Records when the trip started (origen) and when the trip ended (destino).	Timestamp
Id_estacion (origen/destino)	Unique identifier of bike-sharing station. Records where the trip started (origen) and where the trip ended (destino).	String
Nombre_estacion (origen/recorrido)	Name of the bike-sharing station. Records where the trip started (origen) and where the trip ended (destino).	String
Direccion_estacion (origen/recorrido)	Address of the bike-sharing station. Records where the trip started (origen) and where the trip ended (destino).	String
Long_estacion (origen/destino)	Longitude of the bike-sharing station. Records where the trip started (origen) and where the trip ended (destino).	Integer
Lat_estacion (origen/destino)	Latitude of the bike-sharing station. Records where the trip started (origen) and where the trip ended (destino).	Integer
id_usuario	Unique identifier of user that did the bicycle trip.	String