Analytica Data Solutions

TRANSFORMANDO DATOS EN DECISIONES

Proyecto de Movilidad Sostenible NYC

INTEGRANTES:

María Marcela Balzarelli
Pablo Nahuel Barchiesi Ponce
Michael Williams Martinez Chinchilla
Jorgelina Paola Lujan Ramos

TABLA DE CONTENIDO

1.INTRODUCCIÓN

En este tercer sprint, avanzaremos en la fase crítica de nuestro proyecto, que se centra en la creación de un impactante panel de control (dashboard), la generación de informes significativos y el ajuste necesario de nuestro modelo de machine learning. Este sprint representa un paso esencial para convertir nuestros datos en conocimiento valioso y, en última instancia, en acciones estratégicas para nuestro cliente.

Hasta ahora, hemos establecido una base sólida con la creación y optimización de nuestro pipeline de datos en el segundo sprint. Además, hemos puesto en marcha un datawarehouse automatizado y la carga incremental para garantizar la eficiencia y la actualización constante de nuestros datos.

Nuestro equipo ha demostrado una impresionante capacidad de organización y colaboración para abordar las demandas del cliente. No obstante, también hemos reconocido la necesidad de introducir ajustes en la arquitectura general del proyecto y en la definición de los indicadores clave de rendimiento (KPIs) para garantizar que nuestros resultados sean precisos y satisfagan plenamente las necesidades del cliente.

En este sprint, nos embarcamos en la emocionante tarea de visualizar y comunicar nuestros hallazgos de una manera clara y efectiva. Estamos comprometidos en llevar a cabo un análisis profundo y en ofrecer soluciones informadas para mejorar el rendimiento y la toma de decisiones de nuestro cliente.

2.OBJETIVOS

Objetivos del proyecto

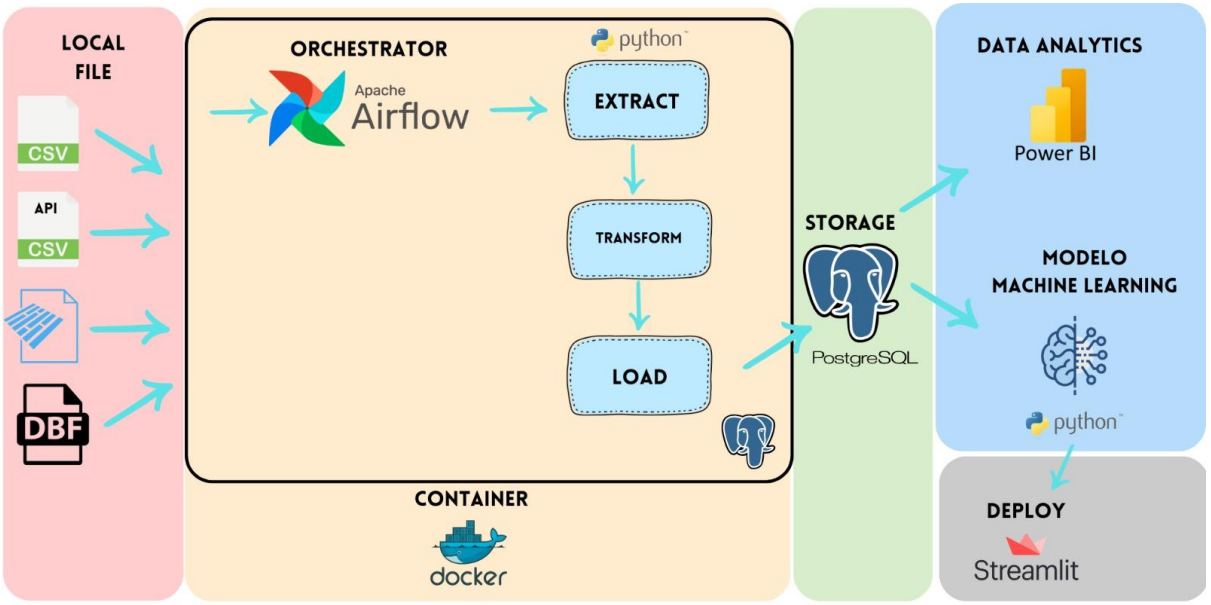
- Investigar y analizar los conjuntos de datos de taxis y viajes compartidos en Nueva York para identificar patrones y tendencias de movimiento.
- Entender la correlación entre los datos de viajes y las métricas ambientales con el fin de proporcionar una base sólida para las decisiones futuras relacionadas con la flota de vehículos.
- Proveer información relevante y confiable a la empresa para respaldar su toma de decisiones sobre la implementación de vehículos eléctricos en su flota.
- Contribuir a la visión de un futuro menos contaminado y ajustarse a las tendencias de mercado actuales.

Objetivos Específicos

- Recopilar y depurar datos de diferentes fuentes para crear una base de datos (DataWarehouse).

- Realizar un análisis exploratorio de los datos para encontrar relaciones
- Crear un dashboard interactivo y visualmente atractivo que integre los resultados del análisis exploratorio de datos
- Entrenar y poner en producción un modelo de machine learning para resolver el problema de inversión en el sector.

3.SOLUCIÓN DATA PIPELINE



Ingesta de Datos:
Los archivos de datos en los formatos CSV, Parquet y DBF están almacenados localmente . Estos archivos contienen información esencial sobre la movilidad urbana y son la base de nuestro análisis.

Orquestación con Apache Airflow en Docker:
Utilizamos Apache Airflow para orquestar y programar el flujo de trabajo de ingesta, procesamiento y carga de los datos. Airflow asegura que las tareas se ejecuten automáticamente y en el orden correcto, optimizando la eficiencia del sistema.

Proceso ETL Automatizado:
El proceso de Extracción, Transformación y Carga (ETL) se realiza a través del lenguaje de programación Python, además, hemos implementado con airflow el mecanismo de carga incremental para actualizar los datos de manera eficiente.

Almacenamiento en PostgreSQL:
Los datos procesados se almacenan en una base de datos PostgreSQL.

Análisis de Datos en Power BI:

Una vez que los datos se encuentran en la base de datos, se pueden analizar y visualizar utilizando Power BI. Esto permite identificar tendencias, patrones y obtener información valiosa para la toma de decisiones informadas.

Modelo de Machine Learning y Streamlit:
Se creará un modelo de Machine Learning en Python. Este modelo se implementará en una aplicación interactiva utilizando Streamlit, lo que permitirá a los usuarios utilizar el modelo y obtener predicciones en tiempo real.

4.ETL

De los 8 dataset entregados por la empresa se escogieron tres que son Electric and Alternative Fuel Charging Stations.csv ,en el cual se realizaron transformaciones como limpieza de nulos y se redujo a utilizar el 30% de los datos ya que son los prodigios para nuestros análisis y se le cambió el nombre a Station_NY.csv.

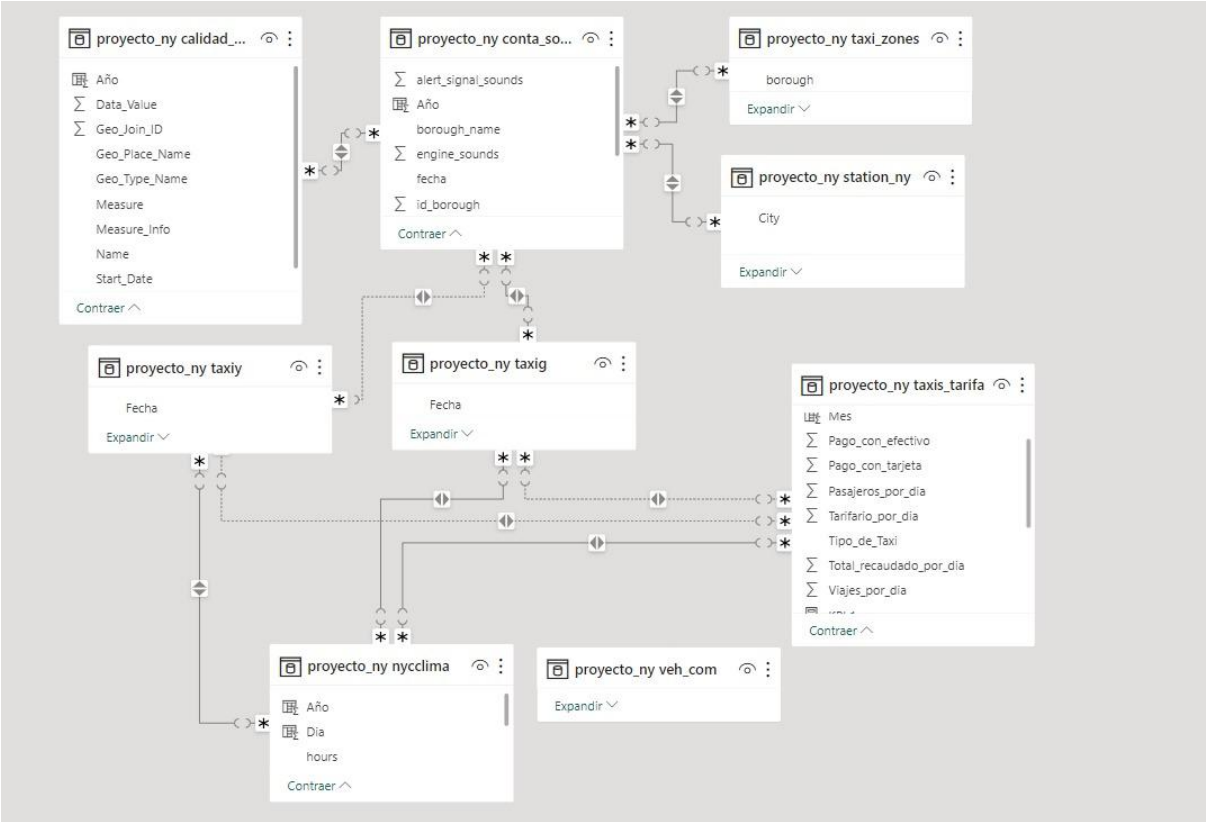
taxi_zones.dbf y taxi+_zone_lookup.csv se realizó un merge entre los dos dataset para formar un dataset llamado taxis zonas.csv y el cual se eliminaron algunas columnas irrelevantes para nuestro estudio.

Los otros 5 dataset se excluyeron porque tenían muchos valores nulos como es el caso del “ alternative fuel vehicles us.csv” “ light duty vehicles.csv ” y los demás no van con nuestro caso de estudio por lo que encontramos y se extrajeron dataset de fuentes externas.

Los dataset extraidos fueron calidad_del aire.csv , conta_sonora.csv, NYCclima.csv y vehiculos_combustion_co2_2023 y un conjunto de datos taxis.parquet_2022_2023 que representan las rutas de los taxis amarillos y verdes que transitan en nueva york en el cual este ultimo se realizaron transformaciones como cambiar tipo de datos, filtro por dia, creacion de columnas nuevas, cambio de nombre a español, eliminacion de valores nulos por ultimo se escogieron las columnas a utilizar y se concatenaron los dataframe de taxis_amarillos y taxis_verdes para posteriormente concatenar estos dos ultimos en el cual se llego al taxis_tarifa.csv.

Los dataset dados por la empresa y extraídos se encuentran el la carpeta llamada Datasets de nuestro repositorio y los datasets ya limpios en el cual se utilizaran para nuestro análisis y modelo de m.l se encuentran dentro de sprint_2 llamada la carpeta dataset_limpios.

5.DIAGRAMA ENTIDAD RELACIÓN



6. ANÁLISIS EXPLORATORIO DE DATOS

En el proceso de realizar el Análisis Exploratorio de Datos (EDA), se ha evidenciado un notorio incremento en los niveles de contaminación sonora durante los últimos años. Específicamente, este fenómeno se muestra más acentuado en el área urbana de Nueva York. Este hallazgo sugiere una correlación significativa entre la contaminación sonora y los ruidos generados por los motores de los vehículos.

Dentro de las estadísticas descriptivas del volumen total de contaminación sonora en la ciudad de Nueva York, se destaca que el 50% de esta proviene directamente de los ruidos emitidos por los motores de los vehículos. Esta proporción resalta que exactamente la mitad de la contaminación sonora tiene su origen en los sonidos generados por los motores de los vehículos en circulación.

Además, se ha identificado una relación estrecha entre el tamaño de los motores de los vehículos y las emisiones de dióxido de carbono (CO_2). Este patrón muestra una tendencia clara: a medida que el tamaño del motor aumenta, también lo hacen las emisiones de CO_2, lo que resalta la importancia de abordar no solo la contaminación sonora, sino también las implicaciones ambientales asociadas con los vehículos de mayor envergadura.

Otro aspecto destacable del análisis es la concentración de actividad de taxis en las áreas residenciales. Se ha observado que las zonas con mayor densidad de movimiento de taxis coinciden con las áreas de los diferentes barrios de la ciudad.

Esto sugiere un patrón de demanda de transporte en las zonas urbanas más concurridas, lo que podría contribuir de manera significativa a los niveles de contaminación sonora y a la complejidad del entorno acústico en estas áreas.

Este análisis exploratorio proporciona una visión inicial pero valiosa de la relación entre la contaminación sonora, los vehículos y los patrones de actividad en Manhattan. Estos resultados resaltan la importancia de abordar de manera integral la gestión del ruido ambiental y sus múltiples interacciones con el entorno urbano y los medios de transporte.

En cuanto al análisis de la calidad del aire, se centró en componentes clave como el Ozono (O3), Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2) y Partículas Finas (PM2.5). Estos componentes son fundamentales para evaluar el impacto en la salud y el medio ambiente debido a la contaminación del aire. Se realizó un análisis de los últimos 3 años recopilados, observando la evolución de los valores de Ozono (O3) que se mantienen relativamente constantes en el tiempo. También se notaron picos de Dióxido de Nitrógeno (NO2) en estaciones específicas del año y una leve reducción en los valores de Partículas Finas (PM2.5) a lo largo del tiempo.

En relación con el tráfico vehicular, se analizaron componentes clave, destacando la diferencia en la densidad de tráfico entre vehículos de pasajeros y camiones. La circulación de vehículos de pasajeros es notablemente mayor en comparación con la de camiones, lo cual refleja una dinámica importante en la movilidad urbana.

Se procedió a realizar visualizaciones por diferentes tipos de áreas geográficas en la ciudad de Nueva York, como Distritos Comunitarios (CD), Boroughs, Uniformed Hospital Fund Area (UHF) y Citywide. Estas visualizaciones indicaron que las áreas más pequeñas (CD) concentran más los componentes clave de la calidad del aire.

En cuanto al transporte, se observó que los autos de combustión son los más utilizados, aunque se notó una incipiente implementación de autos verdes. Esta transición hacia vehículos más sostenibles es reciente y se refleja en los datos.

Se exploró la correlación entre la cantidad de viajes de taxis y las precipitaciones. En el caso de taxis verdes, no se observaron aumentos notables durante precipitaciones elevadas, mientras que en taxis amarillos hubo un pequeño incremento durante periodos de precipitaciones más altas. Además, se planteó la posibilidad de que la temperatura tenga una influencia más fuerte en la cantidad de viajes de taxis que las precipitaciones, basado en las observaciones de los gráficos.

En resumen, el Análisis Exploratorio de Datos proporcionó una comprensión más profunda de cómo los componentes clave, la densidad de tráfico y las condiciones climáticas influyen en diversos aspectos de la calidad del aire y el transporte en la ciudad de Nueva York. Aunque algunas comparaciones y análisis específicos no

pudieron realizarse debido a la falta de datos, se extrajeron valiosas conclusiones sobre la dinámica urbana y la interacción entre estos factores.

7. INDICADORES CLAVE DE DESEMPEÑO (KPIs)

En este proyecto, es esencial medir el desempeño y los resultados obtenidos a través de indicadores clave. Los siguientes KPIs se utilizarán para evaluar el éxito de nuestras soluciones y el impacto de la implementación de vehículos eléctricos en la flota de transporte:

KPI: Crecimiento Porcentual de Tarifas de Taxi Verdes dos años

Objetivo: Evaluar si hubo un aumento de al menos el 5% en las tarifas entre los dos últimos años.

KPI: Control de Incremento de Contaminación Sonora en Barrios con Mayor Contaminación

Objetivo: Limitar el incremento anual de los niveles de contaminación sonora en los barrios identificados con mayor contaminación a un máximo del 5% durante los próximos dos años.

KPI: Crecimiento Anual de la Cantidad de Viajes Realizados

Objetivo: Aumentar la cantidad de viajes realizados en un 10% anual durante el último año.

8. MODELO DE MACHINE LEARNING

Se ha desarrollado un modelo de Machine Learning que permite predecir la cantidad de viajes en función de varios parámetros clave. Estos parámetros incluyen el número de días, la temperatura promedio y el tipo de taxi (verde o amarillo).

El modelo se ha desarrollado utilizando un enfoque de aprendizaje supervisado. Se han recopilado datos históricos de viajes y condiciones climáticas, que se han utilizado para entrenar el modelo. El modelo utiliza un algoritmo de regresión lineal para aprender patrones y relaciones entre los factores de entrada y la variable de salida (la cantidad de viajes).

El modelo utiliza un algoritmo de bosque aleatorio para aprender patrones y relaciones entre los factores de entrada y la variable de salida (la cantidad de viajes).

El modelo ha demostrado ser capaz de predecir la cantidad de viajes con una precisión del 95%. Esta precisión es suficiente para tomar decisiones informadas en cuanto a la asignación de recursos y la planificación de servicios de taxi.

El modelo de Machine Learning desarrollado proporciona una valiosa herramienta para la toma de decisiones estratégicas en el negocio de transporte. El modelo permite anticipar cuántos viajes se pueden esperar en un momento dado, lo que ayuda a mejorar la eficiencia y la efectividad de la gestión de la flota y las operaciones.

Para mejorar aún más la precisión del modelo, se recomienda seguir recopilando datos históricos de viajes y condiciones climáticas. Esto permitirá al modelo aprender patrones y relaciones más complejas entre los factores de entrada y la variable de salida.

9.STACK TECNOLÓGICO

Para llevar a cabo el proyecto se han seleccionado las siguientes tecnologías:

- Trabajo diario: Python, Google meet, Github.
- Ingeniería de datos: Python, PostgreSQL, Docker, Apache Airflow.
- Análisis y visualización de datos: Power Bi, Python.
- Modelo de machine learning: Python, Streamlit.
- Gestión de proyectos: Jira.

10. API

Se ha desarrollado una API que permita a los usuarios obtener información sobre la contaminación sonora y la calidad del aire en función de la ubicación geográfica que ingresen y la función específica que elijan.

La API se ha desarrollado utilizando un conjunto de datos de contaminación sonora y calidad del aire recopilados de diversas fuentes. Los datos se han procesado y almacenado en un formato que facilita su consulta a través de la API. Se ha desplegado en Streamlit junto con el modelo de Machine Learning, lo que facilita su uso por parte de los usuarios.

La API permite a los usuarios obtener información sobre los siguientes parámetros:

- Niveles de ruido
- Índice de calidad del aire
- Contaminantes atmosféricos

La API proporciona una herramienta valiosa para la toma de decisiones en materia de medio ambiente. Los usuarios pueden utilizar la API para obtener información actualizada sobre el entorno ambiental en una ubicación determinada, lo que les ayuda a tomar decisiones informadas sobre su salud y seguridad.

Para mejorar la utilidad de la API, se recomienda seguir recopilando datos sobre la contaminación sonora y la calidad del aire. Esto permitirá a la API proporcionar información más completa y precisa.