***Sprint #1:***

***Puesta en Marcha del Proyecto y***

***Trabajo con Datos***

1. **Introducción**

En el dinámico entorno urbano de la ciudad de Nueva York, los servicios de taxis y viajes compartidos han revolucionado la forma en que las personas se movilizan. Estos servicios no solo ofrecen una alternativa conveniente al transporte público y al alquiler de automóviles, sino que también generan vastas cantidades de datos que ofrecen una ventana única hacia los patrones de movilidad y la demanda de transporte en la ciudad.

Sin embargo, este avance en la movilidad no está exento de preocupaciones ambientales. El cambio climático, acelerado por la dependencia de combustibles fósiles, ha planteado la urgente necesidad de reducir las emisiones de carbono y mejorar la calidad del aire en las áreas urbanas. En este contexto, una empresa de servicios de transporte de pasajeros, con una visión orientada hacia un futuro más sostenible, busca incursionar en el sector de taxis y viajes compartidos en automóviles, considerando la posibilidad de integrar vehículos eléctricos a su flota.

Nuestro equipo ha sido contratado para acompañar este proceso de toma de decisiones, utilizando un enfoque basado en datos para analizar la relación entre los medios de transporte particulares y la calidad del aire en la ciudad de Nueva York. A través de la recopilación, depuración y análisis de datos provenientes de diversas fuentes, así como el desarrollo de modelos de machine learning, buscamos proporcionar información valiosa que respalde la evaluación de la viabilidad y los impactos potenciales de la transición hacia una flota de vehículos eléctricos.

En este proyecto, nos sumergiremos en el rico conjunto de datos disponibles, explorando las interconexiones entre el movimiento de los taxis, la calidad del aire y la contaminación sonora. A través de un enfoque multidisciplinario que combina análisis de datos, conocimientos en ciencias ambientales y técnicas de machine learning, aspiramos a ofrecer una visión integral que oriente las decisiones estratégicas de nuestro cliente hacia un futuro más limpio y sostenible para la movilidad urbana en Nueva York.

1. **Objetivos**

* **Recopilar, depurar y organizar datos relevantes.**
* Recopilación exhaustiva de datos sobre el transporte de pasajeros en Nueva York.
* Depuración y limpieza de datos para garantizar su calidad y fiabilidad.
* Organización estructurada de la información para facilitar su análisis posterior.
* **Desarrollar un Dashboard Interactivo para una mejor visualización y comprensión de los datos.**
* Creación de un Dashboard Interactivo que permita a los interesados visualizar análisis de valor basados en KPI clave.
* Diseño de una interfaz intuitiva y fácil de usar para facilitar la comprensión de los datos.
* Integración de herramientas de visualización avanzada para representar la información de manera efectiva.
* **Evaluar y comparar los costos y beneficios de los vehículos de combustión y eléctricos.**
* Análisis detallado de los costos asociados a los vehículos de combustión y eléctricos.
* Comparación de los beneficios económicos y ambientales de ambos tipos de vehículos.
* Identificación de oportunidades para promover la adopción de vehículos eléctricos en el sector del transporte de pasajeros.

1. **Alcance**

Este alcance proporciona un marco completo para abordar los objetivos del proyecto y ayudará a la empresa de servicios de transporte en su proceso de toma de decisiones relacionadas con la implementación de vehículos eléctricos en su flota.

1. **Recopilación y depuración de datos:**

* Recopilar datos de diversas fuentes y conjuntos de datos relevantes disponibles públicamente.
* Depurar y limpiar los datos para eliminar duplicados, inconsistencias y datos irrelevantes.
* Integrar los datos en un Data Warehouse para su fácil acceso y análisis.

1. **Análisis exploratorio de datos (EDA):**

* Realizar un análisis exploratorio de datos para comprender su estructura y características.
* Identificar patrones, tendencias y relaciones entre las variables, especialmente en lo que respecta al movimiento de los taxis, la calidad del aire y la contaminación sonora.

1. **Análisis de relaciones y conclusiones:**

* Investigar y analizar las relaciones entre variables.
* Identificar posibles factores que contribuyan a estas relaciones y sacar conclusiones.

1. **Desarrollo de modelo de machine learning:**

* Entrenar y evaluar un modelo de machine learning de regresión de tipo supervisado: RandomForestRegressor, que aborde el problema planteado y conecte con los objetivos del proyecto.
* Utilizar el modelo para analizar y predecir posibles impactos en la calidad del aire y otros factores ambientales debido a cambios en el tipo de vehículos utilizados en el transporte de pasajeros.

1. **Informe y presentación de resultados:**

* Elaborar un informe detallado que incluya los hallazgos del análisis exploratorio de datos, las relaciones identificadas entre variables y los resultados del modelo de machine learning.
* Presentar los resultados a la empresa de servicios de transporte en un Dashboard interactivo, destacando las implicaciones para la toma de decisiones relacionadas con la implementación de vehículos eléctricos en su flota.

1. **Iteración y mejora continua:**

* Recibir retroalimentación de la empresa y otras partes interesadas para posibles mejoras.

1. **Planteamiento de KPI´s**

* Aumentar las ventas promedio en un 10% respecto al mes anterior.
* Incrementar el número de viajes en un 10% respecto al mes anterior.
* Reducir la contaminación acústica en un 40% anual en el servicio de taxis dentro del Distrito de Manhattan.
* Disminuir el consumo de combustible en un 10% anual dentro de los Distritos de Manhattan y Queens.

1. **Solución Propuesta**
2. **Metodología de Trabajo**

Adoptaremos una metodología ágil tipo Scrum, utilizando Trello como herramienta central para la organización y seguimiento de tareas.

Se llevarán a cabo reuniones diarias según la disponibilidad de los integrantes del equipo para revisar el progreso, identificar posibles obstáculos y ajustar el plan de trabajo según sea necesario.

Horario de Daylis: Lunes – Viernes (9:00 – 11:00 / 16:00 – 18:00) HR ARG.

1. **Equipo de Trabajo. Roles y Responsabilidades**

* **Data Anaysts:** Joyce Ruiz, Marco Delgado y Jovanni Escamilla.

Dedicados principalmente a tareas relacionadas con el análisis de datos, son responsables de extraer, limpiar, analizar y presentar datos de una manera significativa para tomar decisiones informadas.

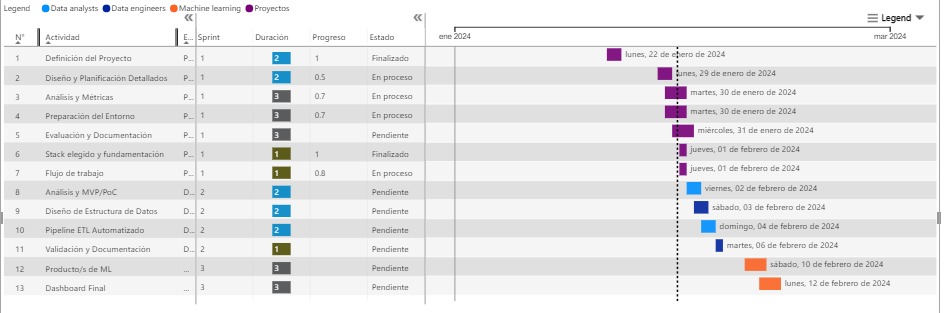
* **Data Engineers:** Joyce Ruiz, Jovanni Escamilla, Lucas Salzotto, Matias Barna.

Centrados principalmente en tareas relacionadas con la ingeniería de datos, se encargan de diseñar, implementar y mantener las infraestructuras de datos y pipelines que permiten la recopilación, procesamiento y almacenamiento eficiente de los datos.

* **Especialistas en Machine Learning:** Matias Barna, Marco Delgado, Lucas Salzotto.

Enfocados en el desarrollo de modelos de Machine Learning y la implementación de soluciones basadas en inteligencia artificial para mejorar la toma de decisiones y automatizar procesos

1. **Cronograma General de Actividades – Gantt**

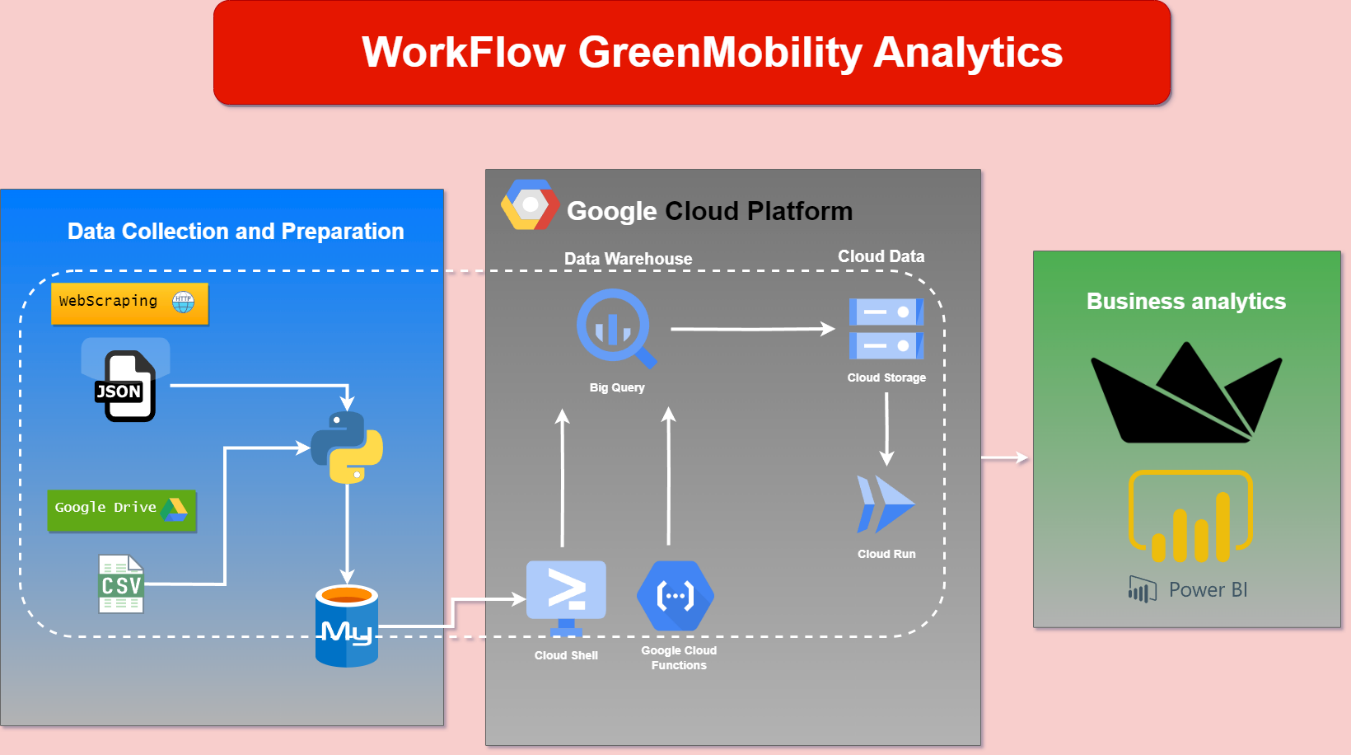
****

1. **Stack Tecnológico**

Las herramientas seleccionadas para el proyecto son las siguientes:

* Python: Para análisis de datos y Machine Learning.
* Pandas, Numpy, Matplotlib: Manipulación y visualización de datos.
* Seaborn, Folium: Visualización de datos geoespaciales.
* Streamlit: Para crear aplicaciones web interactivas y personalizadas para la ciencia de datos y el aprendizaje automático.
* MySQL: permite a los usuarios almacenar, organizar y recuperar datos de manera eficiente.
* Trello: Plataforma de gestión de tareas.
* Google Cloud SQL: Proporciona bases de datos relacionales gestionadas en la nube basadas en MySQL, PostgreSQL y SQL Server.
* GitHub: Para creación y actualización del repositorio.

1. **Flujo de Trabajo (WorkFlow)**

****

1. **Informe de Análisis de los Datos**
2. **Fuentes y confiabilidad de los datos**

Los datos son extraídos de las recolecciones hechas por NYC Taxi and Limousine Commission y otros organismos de NYC.

* Del siguiente enlace, se recolectó la información acerca de los registros de viajes en taxis amarillos en la ciudad de New York:

<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

Después de llevar a cabo la limpieza, se consideró trabajar con las siguientes columnas del dataset:

* **PULocationID:** Zona de taxis en la que se activó el taxímetro
* **Borough:** Distrito
* **Zone:** Zona
* **Service\_Zone:** Zona de servicio
* **Passenger\_Count:** El número de pasajeros en el vehículo.
* **Trip\_Distance:** La distancia de viaje transcurrida en millas informada por el taxímetro.
* **DOLocationID:** Zona de taxis de TLC en la que se desactivó el taxímetro
* **RatecodeID:** El código de tarifa final vigente al final del viaje.
* **Store\_and\_fwd\_flag:** Esta marca indica si el registro de viaje se mantuvo en la memoria del vehículo antes de enviarlo al proveedor, también conocido como "almacenar y reenviar", porque el vehículo no tenía conexión con el servidor.
* **Total\_Amount:** El importe total cobrado a los pasajeros. No incluye propinas en efectivo.
* **Date\_pickup:** Fecha de recogida
* **Time\_pickup:** Hora de recogida
* **Date\_dropoff:** Fecha de finalización
* **Time\_dropoff:** Hora de finalización
* Del siguiente enlace, se obtuvo información para hacer un comparativo entre información de autos eléctricos y autos a gasolina:

<https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64>

Después de llevar a cabo la limpieza, se consideró trabajar con las siguientes columnas del dataset:

* **Model year:** El año al que pertenece el modelo del auto
* **Make:** EL fabricante del auto
* **Model:** El modelo del auto
* **Vehicle class:** El tipo de vehículo
* **City (Gal/mi):** Cantidad de galones que consume por milla recorrida dentro de la ciudad
* **Highway (Gal/mi):** Cantidad de galones que consume por milla recorrida en carreteras
* **CO2 emissions (g/km):** Emisión de CO2 emitido a la atmosfera en una medida de gramos por kilometro recorrido
* **CO2 rating:** Las emisiones de dióxido de carbono del tubo de escape del vehículo se clasifican en una escala del 1 (peor) al 10 (mejor).
* **Smog rating:** Las emisiones del tubo de escape del vehículo de contaminantes que forman smog se clasifican en una escala del 1 (peor) al 10 (mejor).
* **City (kWh/100 mi):** Cantidad de kiloWhats Hora que se consumen por cada 100 millas recorridos dentro de la ciudad
* **Highway (kWh/100 mi):** Cantidad de kiloWhats Hora que se consumen por cada 100 millas recorridos dentro de carreteras
* **Range (mi):** Millas que puede recorrer el auto con una carga completa de su batería
* **Recharge time (h):** Cantidad de horas que se requieren para cargar por completo la batería del auto
* **Type\_Car:** Define si el auto es de gasolina o es eléctrico
* Del siguiente enlace, se recolectó la información acerca de la calidad del aire en New York:

<https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r/about_data>

Después de llevar a cabo la limpieza, se consideró trabajar con las siguientes columnas del dataset:

* **Pollutant:** El tipo de contaminante medido en la observación.
* **Measure Info:** Información adicional sobre la medida del contaminante. En este caso, indica la unidad de medida utilizada para expresar la concentración del contaminante.
* **Geo Place Name:** El nombre del lugar geográfico donde se realizó la medición. Estos podrían ser nombres de vecindarios, distritos o áreas geográficas específicas donde se recopilaron los datos de contaminación.
* **Year**: El año en que se realizó la medición de la contaminación.
* **Data Value:** El valor numérico de la concentración del contaminante medido, expresado en la unidad especificada en la columna "Measure Info".
* Del siguiente enlace, se recolectó la información de registros acerca de sonidos en New York, utilizando el sistema SONYC Urban Sound Tagging (SONYC-UST), que es un conjunto de datos para el desarrollo y evaluación de sistemas de escucha automática para una monitorización realista del ruido urbano:

<https://zenodo.org/records/3966543>

Después de llevar a cabo la limpieza, se consideró trabajar con las siguientes columnas del dataset:

* **Borough:** El código del distrito o municipio al que pertenece la ubicación registrada: 1 (Manhattan), 2 (Brooklyn), 3 (Queens), 4 (Bronx) y 5 (Staten Island)
* **Latitude:** La latitud geográfica de la ubicación.
* **Longitude:** La longitud geográfica de la ubicación.
* **Year:** El año en que se registró la observación.
* **Day:** El día del mes en que se registró la observación.
* **Hour:** La hora del día en que se registró la observación.
* **Engine Sound:** Representa el nivel de sonido del motor, categorizado como "Low" (bajo), "Medium" (medio) o "High" (alto).
* Del dataset “Vehicle Fuel Economy Data.csv”, re pudo recolectar información acerca del consumo de combustible de los autos según su marca y modelo:

Después de llevar a cabo la limpieza, se consideró trabajar con las siguientes columnas del dataset:

* **Year:** Año de fabricación del vehículo.
* **Manufacturer:** Marca del vehículo.
* **Model:** Modelo del vehículo.
* **CO2 (p/mile):** Emisiones de CO2 por milla.
* **Miles per gallon (mpg):** Consumo de combustible en millas por galón.
* **ef\_x\_comb:** Puntuación de eficiencia del vehículo por combustible.
* **comb\_x\_year:** Gasto estimado de combustible por año.
* **Fuel:** Tipo de combustible utilizado por el vehículo.
* **score\_ghg:** Puntuación de emisiones de gases de efecto invernadero (GHG) por vehículo (Cuanto mayor sea la puntuación, menos emisiones de gases de efecto invernadero emitirá el vehículo.).
* **ef\_x\_comb\_city:** Eficiencia de combustible en condiciones de conducción en la ciudad.
* **youSaveSpend:** Indica cuánto se puede ahorrar o gastar en combustible en comparación con un vehículo promedio.
* Alternative Fuel: Tipo de combustible alternativo que puede utilizar.
* **createdOn:** Indica la fecha de creación de los registros en el conjunto de datos.
* **modifiedOn:** Indica la fecha de modificación de los registros en el conjunto de datos.