**AIR TICKET FORECAST**

**Adquisición de Datos:**

En la fase inicial del proyecto, la adquisición de datos es crucial ya que servirán como base para el entrenamiento del modelo predictivo, en este caso, una red neuronal. Después de realizar un análisis exhaustivo del mercado aeronáutico colombiano, se decidió enfocar la adquisición de datos en dos aerolíneas: LATAM y WINGO. Estas aerolíneas ofrecen una amplia variedad de conexiones entre las distintas ciudades del país.

En este proyecto, se centrará en cuatro rutas específicas: Bogotá-Cali, Bogotá-Medellín, Bogotá-San Andrés y Bogotá-Santa Marta. Se tomó esta decisión después de observar que estas rutas se encuentran en los itinerarios de ambas aerolíneas, además de ser las ciudades más importantes y visitadas del país, especialmente durante las épocas de vacaciones. Por lo tanto, se considera que estas rutas son las más relevantes para el propósito del proyecto en su fase inicial. Esta fase inicial de Adquisición de datos cuenta con tres pasos:

* **Recopilación de datos:** la recopilación se hace a través de un programa desarrollado en Python que su función principal es hacer web Scraping que se entiende como el proceso de recopilación de información de sitios web mediante el uso de software especializado que "raspa" o extrae los datos relevantes de las páginas web y los almacena en una base de datos o archivo. Sin embargo, es importante tener en cuenta que algunas técnicas de web scripting pueden ser ilegales o violar los términos de servicio de los sitios web. Pero Bajo leyes colombianas y teniendo en cuenta nuestro modelo de negocio no infringimos la ley de competencia desleal ya que solo mostraremos información de sus precios a los consumidores, y no suministraremos a ningún competidor del sector aeronáutico a si a su vez tendremos en cuenta todas las reglamentaciones necesarias en cuanto a tratamiento de datos para nuestros clientes “usuarios”.

Esta requirió bastante trabajo ya que lograr identificas como realizar el código y además identificar los datos de valor de este pudiera dar.

Para ejemplificar este paso se adjunta un video y un repositorio en github.

Video: Scraping Aero.mp4

Github:

**Limpieza de datos:**

Debido a que el proceso de adquisición de los datos, se realizaron varios procesamientos que permitieran que los datos de ambas aerolíneas tuviesen un mismo tipo y formato. Para las variables tipo time, la limpieza se realizo bajo funciones que permitieron trasformar estos datos a formatos entendibles pro Python, al precio del le debieron quitar algunos indicadores como “$, comas, COP” esto no permitían que se reconocieran como valores numéricos.

Limpie los datos para eliminar errores, duplicados, datos faltantes o inconsistentes que puedan afectar la precisión del modelo.

* **Preparación de datos:** Prepare los datos para el análisis, transformándolos en un formato adecuado para el modelo. Puede incluir la selección de variables relevantes, la normalización de datos, la codificación de variables categóricas, etc.