

# Predicción del tiempo esperado de recorrido en la Línea 1 del Metrobús de la CDMX

Raquel Yunoen Badillo Salas y Juan Javier Santos Ochoa

**Resumen**—El objetivo de este trabajo es predecir los tiempo de recorrido en el Metrobús. Ocupamos un modelo de árboles de decisión de XGBoost con una función Poisson. Ocupamos variables de interés como predictores (distancia, estación, tiempo del último metrobus, tráfico, número de estaciones por recorrer y si está lloviendo). Los resultados son mostrados una aplicación interactiva: <https://goo.gl/RjR53s>.

## I. INTRODUCCIÓN

LA Ciudad de México es la quinta ciudad con mayor población en mundo con 21 millones 581,000 habitantes y debido a esto uno de sus mayores problemas es la congestión en los sistemas de transporte. Esto acompañado con poca información de los horarios y tiempos aproximados de espera agravan el problema de congestión.

Uno de los sistemas de transporte más organizados en la Ciudad de México es el Metrobús, sin embargo, no existe ninguna fuente oficial de horarios de Metrobús que ayude al usuario a decidir su ruta. El objetivo de nuestro proyecto es ayudar a los usuarios a saber cuánto tardarán en el Metrobús y en cada estación. Esto contribuirá a que tomen mejores decisiones de transporte. De igual forma, las autoridades encargadas del sistema Metrobús también pueden contar con información oportuna para planear las operaciones.

En este proyecto creamos un modelo estadístico que calcula el tiempo esperado que demorará un viaje entre diferentes estaciones de la Línea 1 de Metrobús. El modelo usa datos históricos y en tiempo real para hacer estimaciones más precisas. Los resultados pueden ser consultados en este enlace: <https://goo.gl/RjR53s>.

El enfoque que aquí desarrollamos usa datos abiertos y puede ser fácilmente extendido a otras líneas del sistema e incluso se puede usar para la planeación de nuevas rutas alternativas ya que permite hacer extrapolaciones a cualquier punto geográfico de la ciudad.

## II. DATOS

En este documento ocupamos los datos abiertos en la página de Hack CDMX para la línea 1 del Metrobús y proxies para mejorar la estimación. Estas bases contienen datos del identificador del viaje, localización, hora, fecha, identificador del vehículo, velocidad y horario de salida.

La base de datos que utilizamos fue la liberada el día 15 de noviembre del presente año en la página de Hack CDMX, que contiene la información de la operación de los buses de la línea 1 del día 14 de noviembre en el horario de las 16:20 hasta las

23:59. Inicialmente la base contaba con aproximadamente 70 mil observaciones, sin embargo, restringimos las observaciones a aquellas que contaban con un identificador del vehículo y del viaje, ya que esto era fundamental para nuestro enfoque. También eliminamos las observaciones que estaban fuera del área geográfica de influencia y las que tenían comportamientos inusuales, como por ejemplo, que con muy pocos segundos de diferencia estaban en lugares muy distantes. Tras el proceso de depuración nos quedamos con casi 35 mil observaciones válidas para el análisis.

Nuestro enfoque consistió en analizar el desplazamiento de los buses a nivel geográfico y en el tiempo para un mismo viaje. Un viaje es un trayecto de ida y regreso de un Metrobús desde su base inicial. La cantidad de observaciones que estaban disponibles para cada viaje eran variables, así como la posición y el tiempo en el que estas medidas se recolectaban. Si representamos cada observación de la siguiente forma, donde:

$$O_{v,t,p}$$

Es la observación para el viaje  $v$ , en el momento  $t$ , en la posición  $p$ . Nuestro análisis consistió principalmente en comparar las observaciones que pertenecían a un mismo viaje con respecto a la observación del periodo anterior ( $O_{v,t_1,p_1}$  vs  $O_{v,t_2,p_2}$ ). De esta forma calculábamos cuánto varió el tiempo ( $t_2 - t_1$ ) y la posición ( $dist(p_1, p_2)$ ) entre dos observaciones consecutivas. Con este procedimiento pudimos obtener dos variables importantes para nuestro modelo: tiempo del recorrido y distancia recorrida.

Utilizamos la información disponible en la base de datos y la combinamos con otras fuentes para crear variables proxies que nos ayudaron a especificar el modelo. Estas variables explicativas son: **1) distancia** Calculamos la distancia lineal entre los dos puntos de las estaciones de Metrobús. **2) Intersecciones** El número de intersecciones viales, es decir, cuantos cruces debía pasar el vehículo de estación a estación. **3) Sentido** Hacia qué sentido va el viaje (norte-sur o sur-norte). **4) Velocidad promedio de los últimos viajes** Es la velocidad promedio de los metrobuses que cubrieron el mismo trayecto durante la última media hora **5) Tráfico** Con ayuda de la API de Google Maps recolectamos datos de cuanto tráfico hay usualmente durante la hora del recorrido. **6) Número de estaciones** el número de estaciones iniciales y finales en el recorrido y finalmente, **8) lluvia** creamos una variable que indica el nivel de precipitaciones en la Ciudad de México a la hora del recorrido con datos de metroblue <sup>1</sup>

