Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos/ Visual Analytics and Big Data

Optimización de redes de transporte urbano en la ciudad de Nueva York mediante tecnologías Big Data y deep learning.

Trabajo fin de estudio presentado por:	Diego Puerta Martín
Tipo de trabajo:	Desarrollo de software
Director/a:	Jesús Cigales
Fecha:	10/02/2025

Resumen

La congestión urbana en Nueva York representa un desafío crítico con repercusiones económicas (pérdidas de 9,1 billones de dólares anuales), ambientales (aumento de emisiones y contaminación acústica) y sociosanitarias (enfermedades respiratorias, cardiovasculares y problemas de bienestar psicológico).

El proyecto analiza datos reales de: *ride-hailing,* de la *Taxi and Limousine Commission*; eventos de la ciudad, de la *Office of Citywide Event Coordination and Management*; meteorológicos, de la librería *Meteostat;* y la demanda de metro, ofrecidos por *Metropolitan Transportation Authority*. Se han procesado los datos mediante tecnología Big Data como *PySpark* y librerías de visualización como *Folium*, *NetworkX*, *Matplotlib* y *Seaborn*, ofreciendo un modelo de análisis de datos escalable y exportable a otras ciudades. Este estudio analiza cómo diversas variables temporales (hora del día, día de la semana), espaciales (ubicación geográfica) y contextuales (eventos, clima) influyen en la congestión y demanda de transporte privado (*ride-hailing*) y público (metro) en Manhattan. El tráfico varía según la zona de la ciudad, siendo sur y centro más afectados, y según la hora del día y día de la semana, siendo los días laborables, al inicio y fin de la jornada laboral, los que presentan mayores niveles de congestión. Los atributos meteorológicos y de eventos tienen correlación en el tráfico y la demanda de transporte urbano, siendo los eventos los que tienen una mayor correlación con el tráfico y la demanda de metro que las condiciones meteorológicas. Finalmente, se utiliza *PyTorch* para implementar una red neuronal gráfica, que modela las relaciones complejas entre nodos de transporte y clasifica el tráfico entre dos puntos en denso o fluido con precisión. Esta red neuronal gráfica ha demostrado ser competitiva frente a modelos tradicionales como *Random Forest,* mejorando en un 3% el rendimiento general y ofreciendo ventajas en términos de capacidad para capturar relaciones espaciales y temporales no lineales.

Palabras clave: *Big Data, Deep Learning, Urban Mobility, Ride-hailing, Public Transport*

Abstract

Urban congestion in New York represents a critical challenge with economic (losses of $9.1 billion annually), environmental (increased emissions and noise pollution) and social-health (respiratory disease, cardiovascular and psychological wellness problems) impacts.

The project analyzes real data on: ride-hailing, from the *Taxi and Limousine Commission*; city events, from the *Office of Citywide Event Coordination and Management*; weather, from the *Meteostat* library; and subway demand, provided by the *Metropolitan Transportation Authority*. The data have been processed using Big Data technology such as *PySpark* and visualization libraries such as *Folium, NetworkX, Matplotlib* and Seaborn, offering a scalable data analysis model that can be exported to other cities. This study analyzes how various temporal (time of day, day of week), spatial (geographic location) and contextual (events, weather) variables influence congestion and demand for private (ride-hailing) and public (subway) transportation in Manhattan. Traffic varies by area of the city, with south and downtown being more affected, and by time of day and day of the week, with weekdays, at the beginning and end of the workday, having the highest levels of congestion. Weather and event attributes correlate with traffic and urban transit demand, with events having a higher correlation with traffic and subway demand than weather conditions. Finally, *PyTorch* is used to implement a graph neural network, which models complex relationships between transportation nodes and classifies traffic between two points into dense or fluid accurately. This graph neural network has proven to be competitive with traditional models such as Random Forest, improving overall performance by 3% and offering advantages in terms of ability to capture nonlinear spatial and temporal relationships.

Keywords: *Big Data, Deep Learning, Urban Mobility, Ride-hailing, Public Transport*

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

La movilidad urbana en las grandes ciudades ha sido siempre un desafío crítico para su desarrollo y buen funcionamiento. Con el crecimiento de la población y la creciente demanda de una movilidad más eficiente, esta problemática se ha intensificado (Newman and Kenworthy, 2015). Nueva York, una de las ciudades más densamente pobladas del mundo, con aproximadamente 10.200 habitantes por kilómetro cuadrado (World Population Review, 2024), enfrenta una congestión creciente que impacta negativamente tanto en su economía como en la calidad de vida de sus habitantes (Adams, 2023; Pishue, 2024). Según el informe *Global Traffic Scorecard* de INRIX (2023), los neoyorquinos perdieron más tiempo en desplazamientos que los ciudadanos de cualquier otra ciudad, registrándose un aumento del 13% en los viajes hacia el centro de la ciudad en 2023 (p. 5). Ver **Tabla 1.**

Tabla 1. *Variación porcentual del número de viajes registrados hacia el centro de la ciudad.*

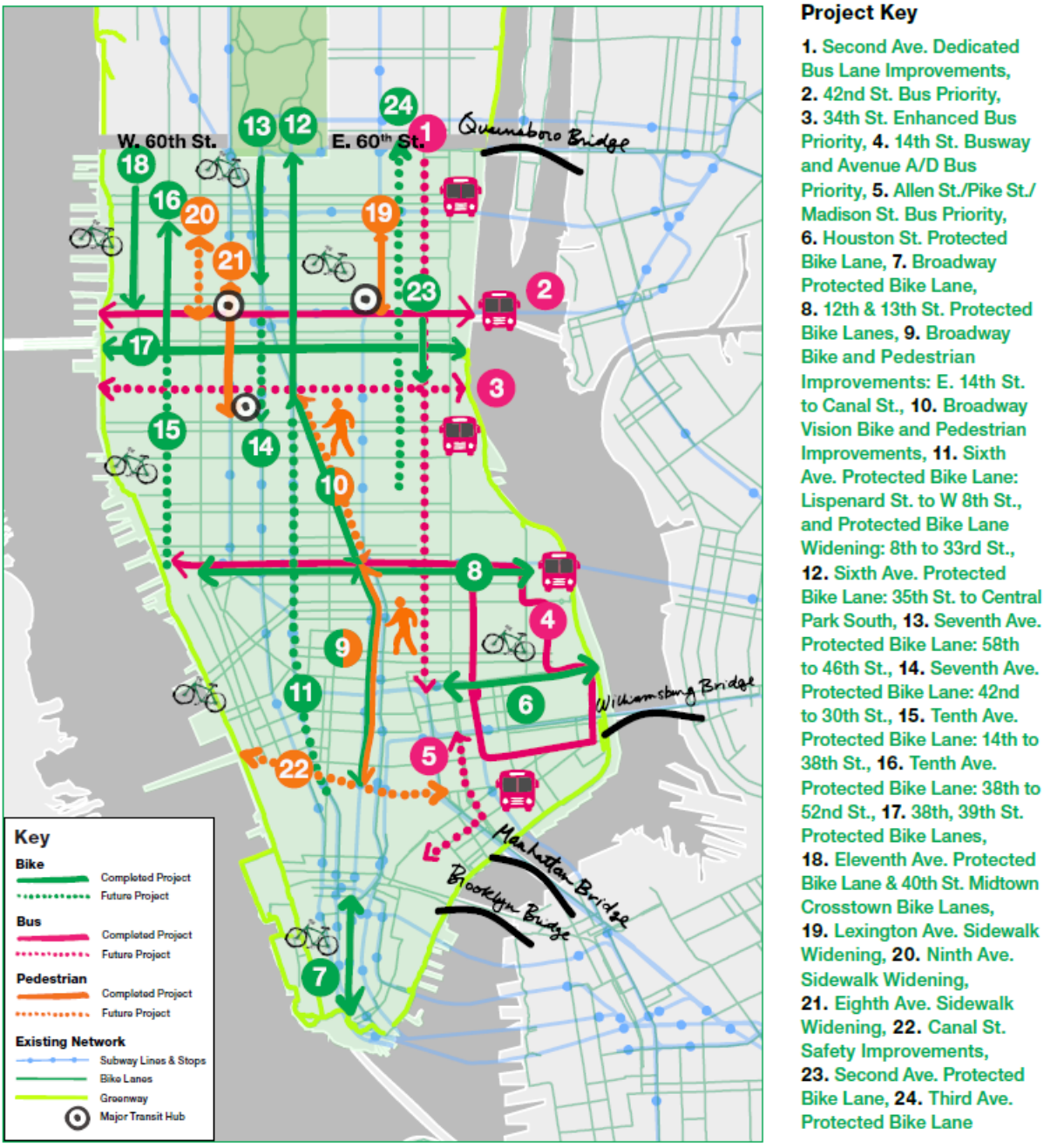
Country	Downtown/City Center	Change
U.S.	Atlanta	7%
	Chicago	-1%
	Dallas	3%
	Houston	4%
	Los Angeles	5%
	Miami	4%
	New York	13%
	Philadelphia	7%
	Phoenix	4%
	Washington D.C.	7%

Fuente: INRIX, 2023a.

Desde 2008, la *Metropolitan Transportation Authority* de Nueva York (MTA) ha implementado diversas medidas para aliviar esta congestión y mejorar la movilidad. Entre ellas destacan: la creación de 20,5 kilómetros de carriles exclusivos para buses, la prioridad en semáforos en más de 300 intersecciones, la habilitación de 224 kilómetros de carriles bici en el *Central Business District* (CBD) de Manhattan y la expansión de más de 93.000 metros cuadrados de

espacio peatonal con el programa *Open Streets* (New York City Department of Transportation, 2023, p. 5). Ver **Figura 1**. No obstante, a pesar de estas medidas, la velocidad promedio de los vehículos en el CBD se redujo a 8 kilómetros por hora en 2023, lo que evidencia que el problema no solo persiste, sino que se ha agravado (Adams, 2023).

Figura 1. *Mejoras y ampliaciones para el Central Business District del programa Open Streets.*



Fuente: Adams, 2023.

Para hacer frente a esta situación, a partir del 7 de enero de 2025, se han implementado peajes dinámicos dentro de la zona sur del área metropolitana de Manhattan, teniendo en cuenta variables como la hora del día, el tipo de vehículo, el método de pago y el cruce de créditos. Dentro de los vehículos afectados por estos peajes están los taxis y los vehículos de alquiler, aplicándose un cargo al pasajero durante determinadas franjas horarias. Aunque los datos aún son preliminares, los resultados apuntan a que, durante la primera semana de aplicación de los peajes dinámicos, se ha reducido el tráfico en un 8% y los tiempos de viaje entre un 30% y un 40% (MTA, 2025).

El aumento de la congestión tiene consecuencias significativas en varios aspectos que afectan a la calidad de vida de los habitantes. A nivel medioambiental, el crecimiento del número de vehículos ha provocado un aumento del 66% en las emisiones de gases contaminantes entre 2010 y 2018, con un notable impacto de los servicios de transporte como Uber y Lyft (Roberton et al., 2020). El tráfico es responsable del 14% de todas las emisiones de PM2.5, nanopartículas responsables de alrededor de 2.000 muertes anuales, según el propio gobierno de la ciudad (NYC Environment and Heatlh Data Portal, 2021). Adicionalmente, a nivel de salud, la Organización Mundial de la Salud (OMS) (2024) asocia la contaminación del aire con múltiples enfermedades, como infecciones respiratorias, asma, enfermedades cardiovasculares, cáncer y diabetes (p. 6, 7). Asimismo, según datos de la propia OMS (2019), la tasa de mortalidad por cada 100.000 habitantes atribuible a la contaminación doméstica y ambiental es de 14.2. Ver **Tabla 2**.

Tabla 2. *Tasa de mortalidad estandarizada por sexo atribuida a la contaminación del aire doméstico y ambiental (por 100.000 habitantes).*

Dimensión de desagregación	Muertes por 100k, 2019
Hembra	11.3
Masculino	17.5
Total	**14.2**

Fuente: OMS, 2019.

A nivel emocional, ésta también se asocia con un impacto en el bienestar psicológico de los ciudadanos (Pelgrims et al., 2021). Por otro lado, el aumento del tráfico convierte a Nueva York en la ciudad más ruidosa de Estados Unidos (Steel Guard Safety Products, 2024). La exposición al ruido se ha relacionado con problemas psiquiátricos, pérdida de audición y enfermedades cardiovasculares (Organización Mundial de la Salud, 2024, p. 7). Por último, durante el año 2023, cada neoyorquino perdió en promedio 101 horas debido a los retrasos en el tráfico, lo que resultó en pérdidas económicas de aproximadamente 9,1 billones de dólares por la reducción de la productividad (Pishue, 2024). Ver **Tabla 3**.

Tabla 3. *Ranking de ciudades con mayores pérdidas temporales y pérdidas económicas por persona.*

2023 US Rank (2022 Rank)	Urban Area	2023 Delay (2022)	Compared to Pre-COVID	2023 Cost per Driver	2023 Cost per City
1 (1)	New York City NY	101 (105)	11%	$1,762	$9.1 B
2 (2)	Chicago IL	96 (87)	18%	$1,672	$6.1 B
3 (3)	Los Angeles CA	89 (78)	-4%	$1,545	$8.3 B
4 (4)	Boston MA	88 (78)	-1%	$1,543	$2.9 B
5 (6)	Miami FL	70 (66)	18%	$1,219	$3.1 B
6 (5)	Philadelphia PA	69 (67)	2%	$1,209	$2.9 B
7 (8)	Washington DC	63 (52)	-9%	$1,095	$2.7 B

Fuente: INRIX, 2023b.

Dada esta situación, se plantea la necesidad de optimizar las redes de transporte urbano en la ciudad de Nueva York, integrando nuevas tecnologías que permitan mejorar la eficiencia de los desplazamientos y reducir las consecuencias negativas que actualmente afronta la ciudad.

1.1. Motivación

El problema que se aborda en este trabajo es la congestión del sistema de transporte urbano en Nueva York, donde miles de personas pierden grandes cantidades de tiempo en sus desplazamientos diarios. La congestión no solo afecta a la productividad, sino que sus consecuencias, como el aumento de las emisiones y de la contaminación acústica, tienen un impacto en el medioambiente y en la salud y el bienestar de los ciudadanos.

La evaluación del valor del tiempo de viaje es importante tanto para la evaluación de políticas de transporte, como los peajes para controlar la congestión. Asimismo, el valor que el

ciudadano le da al tiempo de viaje siempre será positivo, pues el tiempo se considera un bien limitado (ITF, 2019).

Además, si bien hay muchos estudios relacionados con la predicción del tráfico a corto plazo (Fang et al., 2021; Ranjan et al., 2020; Shahriari et al., 2020; Shi et al., 2021; Zhao et al., 2020), no son tantos los que lo abordan a largo plazo (Li et al., 2021; Wang et al., 2021), por lo que se precisa más investigación en este ámbito.

Ante esto, el trabajo pretende ser una contribución para mitigar la problemática enunciada, optimizando las redes de transporte urbano en Nueva York mediante la tecnología Big Data y la inteligencia artificial (IA). Por tanto, se pretende guiar, tanto a administraciones como organismos privados, en la manera de organizar la movilidad, contribuyendo así mejorar las tres grandes problemáticas enunciadas: el bienestar físico y emocional de la ciudadanía, reduciendo el ruido; medio ambiente, reduciendo las emisiones; y la economía de la ciudad de Nueva York, reduciendo las pérdidas de tiempo en los desplazamientos de los neoyorquinos.

1.2. Planteamiento del trabajo

1.2.1. Medios técnicos

Este trabajo se plantea como una solución para la optimización de las redes de transporte urbano en Nueva York, utilizando técnicas de Big Data y algoritmos de inteligencia artificial. El enfoque se basa en el análisis de datos reales proporcionados por la *Taxi and Limousine Commission* (TLC), los datos de la demanda de recorridos de metro de la *Metropolitan Transportation Authority*, datos de eventos permitidos en Nueva York, proporcionado por la *Office of Citywide Event Coordination and Management* (CECM) y los datos meteorológicos de *Meteostat*. El objetivo es utilizar las tecnologías Big Data para realizar un exhaustivo análisis de datos del transporte urbano en Nueva York, desde un enfoque multimodal. Asimismo, se explorará utilizar algoritmos para la optimización de rutas e inteligencia artificial para la predicción del tráfico de la ciudad.

Para alcanzar los objetivos planteados, se implementará un clúster en *Azure Databricks Community Edition*, aprovechando su escalabilidad y capacidad de procesamiento en la nube. Se creará un sistema de ficheros en *Databricks File System* (DBFS) para almacenar los datos

históricos del sistema de transporte de la TLC en formato *parquet*, que serán fundamentales para los análisis posteriores. Además, se integrarán datos climatológicos a través de la librería *Meteostat*, con el fin de considerar el impacto de las condiciones meteorológicas en los tiempos de viaje. Por otro lado, también se conectará a la API de la *NYC OpenData* para obtener los datos de eventos y la demanda de recorridos de metro.

El procesamiento masivo de datos se llevará a cabo utilizando *PySpark* en *Databricks*, que permite manejar grandes volúmenes de información de manera eficiente. Para las transformaciones de datos y la construcción de *pipelines* de procesamiento, se emplearán librerías como *SparkSQL*, que facilita la manipulación de datos estructurados.

Para el análisis y la visualización de datos en este trabajo, se emplearán diversas librerías de *Python* que permiten realizar un análisis exploratorio de datos (EDA) efectivo y generar representaciones visuales tanto estáticas como interactivas. La librería *pandas* se utilizará para la manipulación y análisis de datos tabulares, ya que ofrece estructuras de datos flexibles y potentes herramientas para el procesamiento de grandes conjuntos de datos. Para la visualización estática, se utilizarán *matplotlib* y *seaborn*, que permiten crear gráficos informativos y estilizados: *matplotlib* es una librería versátil para generar gráficos básicos y personalizados, mientras que *seaborn* facilita la creación de gráficos estadísticos más avanzados con un diseño mejorado. Además, se empleará *networkx* para la visualización y análisis de redes, permitiendo representar rutas de transporte como grafos y visualizar su estructura en mapas interactivos. Por último, se usará *folium* para generar mapas interactivos, donde se mostrarán rutas y mapas de calor superpuestos sobre el área metropolitana de Nueva York, proporcionando una comprensión espacial clara de los patrones de transporte. Estas herramientas combinadas permiten un análisis completo e intuitivo de los datos.

En cuanto a la modelización mediante inteligencia artificial, se hará uso de *scikit-learn* para implementar algoritmos de aprendizaje automático como *Random Forest*. Adicionalmente, se explorarán librerías avanzadas de *deep learning*, tales como *Torch* o *Keras*, dependiendo de la necesidad de modelos más complejos para la predicción de tráfico y la optimización de rutas. Esta combinación de herramientas asegura un enfoque robusto y escalable para el análisis y modelado de los datos.

La implementación en *Azure Databricks Community Edition* ofrece una infraestructura para manejar y procesar grandes volúmenes de datos. Con el almacenamiento en *DBFS*, los datos de *ride-hailing*, la demanda de recorridos de metro y los conjuntos de datos de eventos y meteorológicos pueden centralizarse para una rápida accesibilidad, mientras que *PySpark* en *Databricks* facilitará el procesamiento de datos a gran escala (Zaharia et al., 2010).

Se valorará el uso de *SparkSQL* y *SparkMLib*, que también permiten realizar operaciones eficientes sobre los datos y aplicar algoritmos de *machine learning* (ML), incluyendo regresiones y modelos supervisados que puedan alimentar el sistema de predicción de rutas. Por ejemplo, *SparkMLib* permite construir modelos de clasificación y regresión a gran escala, ajustándose así al volumen y la velocidad de los datos de transporte (Meng et al., 2016).

Para modelos de mayor complejidad, se explorará el uso de *TensorFlow*, *Keras* y *PyTorch* pueden ofrecer un marco sólido para la construcción de redes neuronales avanzadas, aprovechando la flexibilidad de *TensorFlow* para implementar algoritmos personalizados de *deep learning* (DL) (Abadi et al., 2016).

1.2.2. Marco teórico

Para abordar de manera completa los fundamentos teóricos y técnicos de este trabajo, es esencial profundizar en los conceptos clave que sustentan la optimización de redes de transporte mediante el uso de tecnologías avanzadas como la *inteligencia artificial*, el *machine learning* y el *deep learning*, así como en los algoritmos específicos que se implementarán, como el algoritmo de *Dijkstra* y las redes neuronales de grafos.

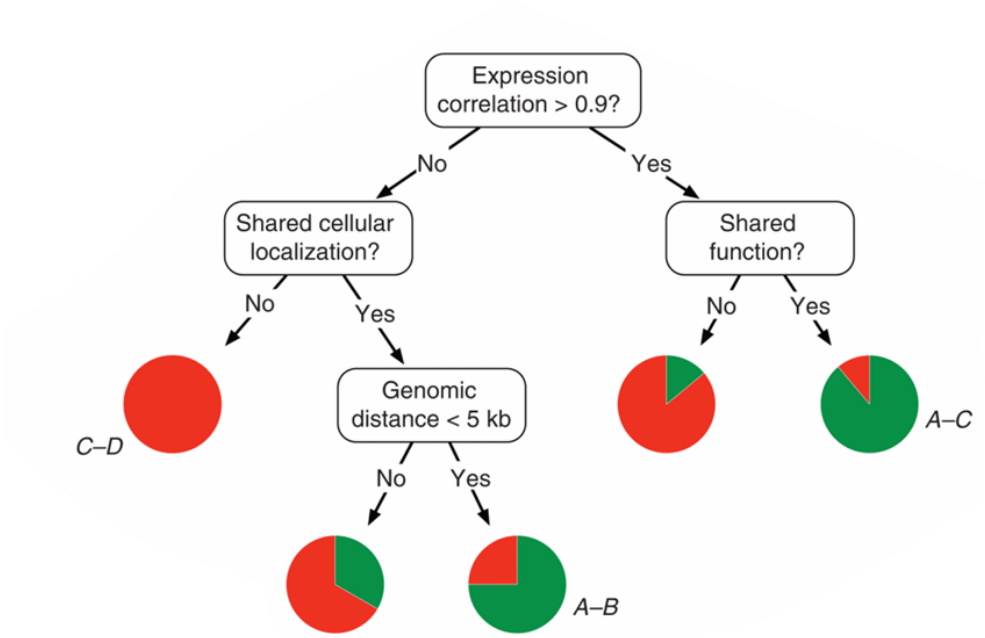
Inteligencia artificial se refiere a la capacidad de una máquina para imitar funciones cognitivas humanas como el aprendizaje y la resolución de problemas (Russell y Norvig, 2020). Dentro del ámbito de la optimización de redes de transporte, la IA permite desarrollar sistemas que analizan grandes volúmenes de datos y toman decisiones o realizan predicciones para mejorar la eficiencia de los sistemas de transporte. Este enfoque resulta especialmente útil en contextos de movilidad urbana, donde la IA puede evaluar los atributos que le afectan, siendo capaz de adelantarse a eventos futuros y contribuyendo así a una posible mejora de la organización del transporte.

Machine learning, como subcampo de la IA, incluye técnicas que permiten a los sistemas aprender de los datos sin ser programados explícitamente (Murphy, 2012). En este proyecto, el ML se empleará para desarrollar un modelo que prediga el tráfico entre dos rutas y evalúen la eficiencia de rutas en función de múltiples factores, como las horas pico, el clima y los eventos urbanos. Estos modelos son esenciales para manejar la variabilidad del estado del tráfico en una red urbana compleja y multimodal como la de Nueva York.

Dentro del ML, las técnicas se dividen principalmente en dos categorías: aprendizaje supervisado y aprendizaje no supervisado, cada una de ellas con aplicaciones específicas según el tipo de problema y datos disponibles.

- **Aprendizaje supervisado**. En esta categoría, los algoritmos se entrenan con datos etiquetados, es decir, datos de entrada acompañados de su correspondiente salida esperada. Esto permite al modelo aprender relaciones específicas entre las variables y realizar predicciones precisas. Por ejemplo, en este trabajo, un modelo supervisado podría predecir los tiempos de viaje entre dos puntos del área metropolitana de Manhattan, en función de características como la hora del día, el día de la semana, el clima, los patrones históricos de demanda y los eventos de la ciudad. Algoritmos comunes en esta categoría incluyen, entre otros, la regresión lineal, bosques aleatorios y árboles de decisión. Estos últimos son explicados en detalle por Kigsford et. al (2008), que indican que un árbol de decisión es un tipo de clasificador que predice etiquetas de clase para datos mediante una serie de preguntas jerárquicas sobre las características de los elementos. Una vez construida la estructura en forma de árbol, se entrena con ejemplos etiquetados para luego clasificar nuevos datos de manera rápida y precisa. Ver **Figura 2**.

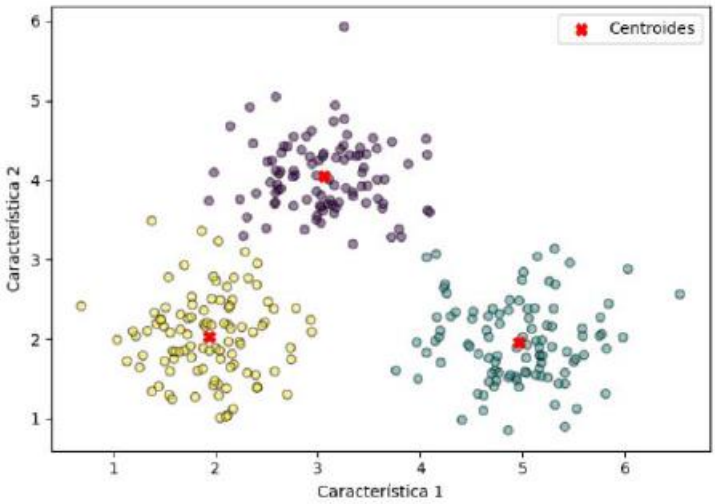
Figura 2. *Ejemplo de árbol de decisión.*



Fuente: Kingsford et. al, 2008

- **Aprendizaje no supervisado**. Aquí, los algoritmos trabajan con datos no etiquetados, identificando patrones ocultos o relaciones dentro de los datos. Este enfoque es útil para descubrir agrupaciones de recorridos con comportamientos similares o identificar patrones de congestión en redes de transporte. Las técnicas de reducción de dimensionalidad (ej.: PCA) o algoritmos como el *clustering* (ej.: K-Means), como muestran Rincón et al. (2024) en su estudio, son ejemplos comunes de este enfoque. Ver **Figura 3**.

Figura 3. *Ejemplo de clustering con K-Means.*

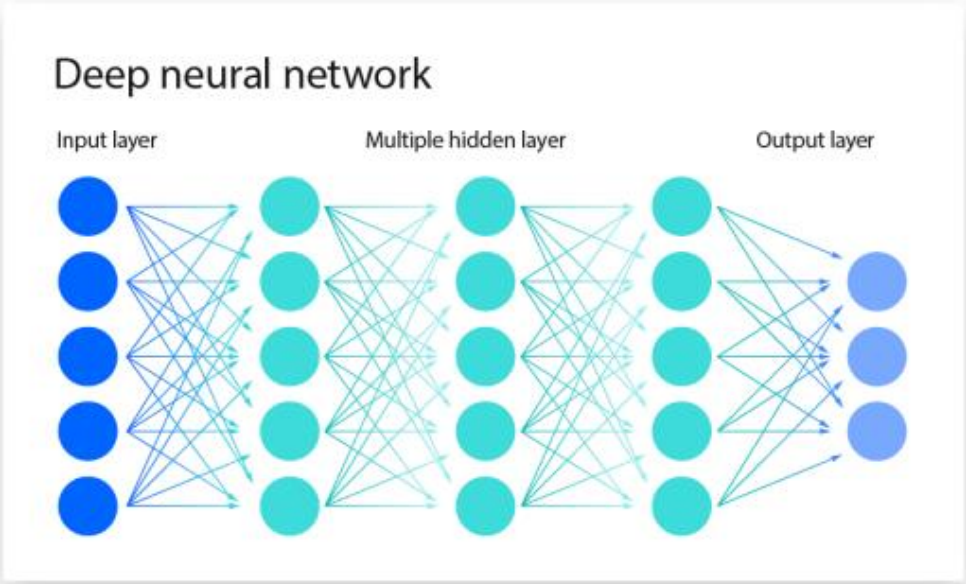


Fuente: Rincón et. al, 2024

Deep Learning, una especialización dentro del ML, utiliza redes neuronales profundas para analizar grandes volúmenes de datos no estructurados, como imágenes y audio, y también ha demostrado ser altamente efectivo en el procesamiento de secuencias temporales y relaciones espaciales (LeCun et al., 2015). En el contexto de este trabajo, el DL se emplea para desarrollar modelos complejos que incluyan los datos históricos de los viajes de *ride-hailing*, de las condiciones meteorológicas y de los eventos sociales, cuyo objetivo sea la categorización del estado del tráfico entre dos puntos de la ciudad.

Según IBM, una red neuronal es un programa, o modelo, de *machine learning* que toma decisiones de forma similar al cerebro humano, utilizando procesos que imitan la forma en que las neuronas biológicas trabajan juntas para identificar fenómenos, sopesar opciones y llegar a conclusiones (IBM, 2024). Ver **Figura 4.**

Figura 4. *Estructura básica de una red neuronal.*



Fuente: IBM, 2024.

Las redes neuronales profundas (DNN), caracterizadas por múltiples capas ocultas, son capaces de capturar patrones complejos y representaciones jerárquicas de datos mediante un proceso de aprendizaje jerárquico (LeCun et al., 2015). En una DNN, las capas iniciales suelen aprender características de bajo nivel, mientras que las capas posteriores se centran en representaciones de mayor nivel de abstracción.

El mecanismo central de aprendizaje en estas redes se basa en el algoritmo de retropropagación, que ajusta los pesos de las conexiones entre neuronas utilizando el descenso por gradiente para minimizar una función de pérdida (Rumelhart et al., 1986). Además, la introducción de funciones de activación no lineales, como ReLU (*Rectified Linear Unit*), permite a las redes aprender representaciones más complejas, evitando problemas como la saturación en las funciones lineales (Nair & Hinton, 2010).

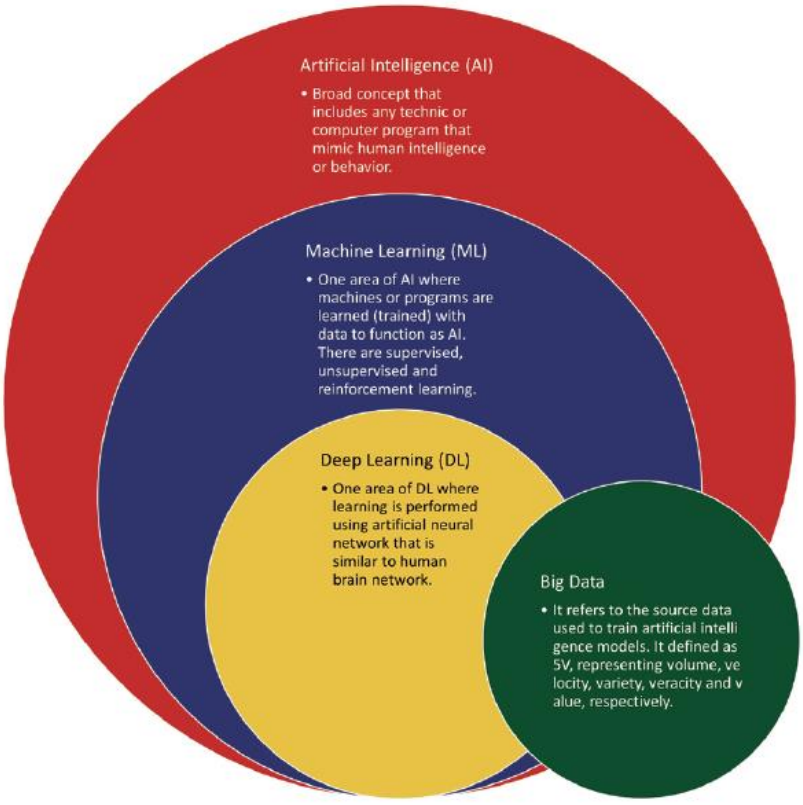
Entre las arquitecturas más destacadas, las redes convolucionales (CNN) son ampliamente utilizadas para datos espaciales como imágenes, debido a su capacidad de aprender patrones locales mediante la operación de convolución. Por otro lado, las redes recurrentes (RNN) son esenciales para datos secuenciales, ya que mantienen un estado interno que permite capturar

dependencias temporales. Sin embargo, problemas como el desvanecimiento de gradientes han llevado al desarrollo de variantes avanzadas como LSTM (*Long Short-Term Memory*) y GRU (*Gated Recurrent Unit*), que superan estas limitaciones al implementar mecanismos de memoria más robustos (Cho et al., 2014; Hochreiter & Schmidhuber, 1997).

El DL también se beneficia del incremento de capacidades computacionales y la disponibilidad de grandes volúmenes de datos. Herramientas, como GPUs, y *frameworks,* como *TensorFlow* y *PyTorch,* han facilitado el entrenamiento de modelos cada vez más complejos. En el contexto del transporte, estas arquitecturas permiten modelar la interacción entre múltiples factores, como tráfico, clima y eventos sociales, haciendo posible realizar predicciones y optimizaciones en redes urbanas de alta complejidad.

Según Chung et al. (2022) estas áreas de conocimiento pueden visualizarse en un diagrama que muestra la relación jerárquica entre IA, ML y DL. Ver **Figura 5**.

Figura 5. *Diagrama de relación entre inteligencia artificial, machine learning y deep learning*.



Fuente: Chung et al, 2022.

El algoritmo de *Dijkstra* es un método de búsqueda de caminos que encuentra la ruta más corta entre dos nodos en un grafo, lo cual es especialmente útil en redes de transporte donde se requiere minimizar el tiempo de viaje o la distancia (Dijkstra, 1959). Este algoritmo se explorará en nuestro trabajo para calcular rutas óptimas dentro del sistema de transporte de Nueva York, considerando las restricciones de la red de metro y de las rutas de taxis.

Según *Graph Everywhere* (2024), un grafo es una composición de un conjunto de objetos, conocidos como nodos, que se relacionan con otros nodos a través de un conjunto de conexiones llamadas aristas. Ver **Figura 6**.

Figura 6. *Estructura básica de un grafo.*

Fuente: Graph Everywhere, 2024.

Las redes neuronales de grafos (GNN) son una aplicación avanzada de redes neuronales que permite procesar y extraer información de datos en forma de grafos, como los que representan las redes de transporte (Scarselli et al., 2009). Este enfoque es adecuado para este trabajo, ya que permite modelar la interdependencia entre diferentes puntos de la red y predecir tiempos de viaje con base en factores como la congestión o las transferencias entre rutas. A través de la GNN, podremos mejorar la precisión de las predicciones en un sistema de transporte multimodal al tener en cuenta la compleja estructura de las conexiones entre estaciones. Una imagen conceptual del funcionamiento de una GNN puede ilustrar cómo los

nodos (estaciones o puntos de transferencia) se actualizan iterativamente en función de sus conexiones, lo cual ayuda a comprender cómo se capturan las relaciones espaciales y temporales en la red de transporte.

Este enfoque integrador de IA, ML, DL y algoritmos de optimización en una plataforma de Big Data en la nube proporcionará una visión detallada y precisa del sistema de transporte de Nueva York.

1.3. Estructura del trabajo

Este trabajo se organiza en los siguientes capítulos, cada uno diseñado para desarrollar y profundizar en los aspectos clave de la optimización de las redes de transporte urbano mediante el uso de **Big Data** e **Inteligencia Artificial**:

- **Capítulo 2: contexto y estado del arte**

  En este capítulo se analiza cómo el Big Data ha transformado la planificación y optimización del transporte urbano, con un enfoque en el uso de modelos predictivos, sistemas de transporte inteligentes y nuevas tecnologías, tales como los vehículos conectados y autónomos. Se abordan los avances en predicción de tráfico, la integración de diversas fuentes de datos y los desafíos asociados, incluyendo la privacidad y la calidad de los datos. También se destacan las oportunidades y retos que estas herramientas representan para mejorar la eficiencia y sostenibilidad en ciudades como Nueva York.

- **Capítulo 3: objetivos concretos y metodología de trabajo**

  En este capítulo se detallan los objetivos generales y específicos del trabajo, que incluyen el análisis exploratorio de los datos del transporte urbano en Nueva York (*ride-hailing* y metro) y la creación de una red neuronal basada en grafos para predecir el estado del tráfico entre dos puntos de la ciudad. Se presentará la metodología utilizada para la recopilación y análisis de los datos de la TLC, los datos de la CECM, los datos de la demanda de metro y los datos meteorológicos. La metodología seguirá un enfoque iterativo similar a la técnica CRISP-DM (*Cross Industry Standard Process for Data Mining*) para el análisis de datos, que permitirá extraer *insights* valiosos para la toma de decisiones en la optimización del transporte. Además, se describirán las

herramientas y tecnologías empleadas: tecnología *cloud* para el almacenamiento masivo de datos, *PySpark* para el procesamiento masivo de datos y librerías asociadas para lograr los objetivos del trabajo.

- **Capítulo 4: marco normativo**

  Este capítulo se centrará en la revisión de las normativas actuales relacionadas con la privacidad y protección de datos en Estados Unidos.

- **Capítulo 5: desarrollo específico de la contribución**

  En este capítulo se describe de forma detallada el desarrollo de un modelo para la optimización del transporte urbano en Manhattan, basado en datos reales. El proceso comienza con la integración y análisis de múltiples fuentes de datos, como registros de transporte privado proporcionados por la TLC de Nueva York, datos de la demanda de metro de la CECM, información meteorológica y datos de eventos relevantes en la ciudad. El objetivo principal es identificar patrones, evaluar el impacto de factores externos y desarrollar herramientas predictivas y de optimización.

  El primer paso consiste en la identificación de los principales cuellos de botella y la caracterización de la demanda de transporte privado de *ride-hailing.* Mediante técnicas avanzadas de visualización, se analizarán zonas y rutas con mayor congestión y su distribución temporal. Los mapas de calor se utilizarán para plasmar gráficamente los tiempos de viaje promedio, velocidades de circulación y variaciones a lo largo del día y la semana. Esto permitirá, no solo identificar áreas críticas en horarios pico, sino también proporcionar un entendimiento más claro de cómo la demanda varía en función del tiempo y del contexto urbano.

  Otro aspecto crucial es el análisis del impacto de las condiciones externas, como el clima y los eventos en la ciudad. Se integrarán datos meteorológicos, incluyendo variables como precipitaciones, temperatura y velocidad del viento, para estudiar cómo éstas afectan a la demanda y a la velocidad de circulación. Por ejemplo, se evaluará la influencia del tiempo en la movilidad de la ciudad. De igual manera, los eventos que tengan lugar en la ciudad, como conciertos o eventos deportivos, serán objeto de análisis para determinar su influencia en los flujos de movilidad y en las rutas de congestión.

En paralelo, se llevará a cabo un estudio sobre la demanda de transporte público, en concreto de metro, y estudiar esta alineación temporal con la demanda de *ride-hailing*. Asimismo, se realizará un estudio de la influencia de las variables previamente comentadas en la demanda de transporte.

Finalmente, se abordará la creación de un modelo predictivo que combina diversas técnicas avanzadas. Se diseñará una red neuronal basada en grafos, que permitirá clasificar el tráfico como denso o fluido. Se representará la red de transporte como un grafo, en el cual las localizaciones de recogida y de destino son los nodos. Los pesos o aristas de este grafo estarán determinados por factores como número de eventos en la ciudad, condiciones meteorológicas, hora del día y día de la semana. Además, se aplicará un modelo tradicional de *Random Forest* para realizar una comparativa de rendimiento.

- **Capítulo 6: código fuente y datos analizados**

  En este capítulo se explicará donde se encuentra alojado el código fuente utilizado para el desarrollo del proyecto, proporcionando un enlace al repositorio donde se encuentra todo el desarrollo. Asimismo, se indicará un repositorio propio dónde se encuentran los datos analizados.

- **Capítulo 7: conclusiones**

  En este capítulo se ofrecerán las conclusiones del trabajo. Se discutirá cómo los resultados obtenidos ayudan a mejorar la movilidad en Nueva York, destacando los avances realizados en el análisis de datos y predicción del tráfico.

- **Capítulo 8: limitaciones y trabajo futuro**

  Este capítulo abordará las limitaciones del trabajo, como la falta de acceso a ciertos datos o la capacidad computacional. Además, se discutirá el trabajo futuro que podría llevarse a cabo para mejorar los modelos desarrollados y su potencial aplicación en otras ciudades o entornos. También se explorarán líneas de investigación adicionales para seguir optimizando las redes de transporte mediante el uso de tecnologías avanzadas.

2. Contexto y estado del arte

2.1. Contexto del problema

El Big Data ha transformado profundamente la manera en que se aborda la planificación y optimización de redes de transporte urbano. Este cambio ha sido impulsado por la proliferación de fuentes de datos como registros de teléfonos móviles, tarjetas de transporte inteligentes, datos GPS y sensores en vehículos y carreteras. Las ciudades modernas generan enormes cantidades de datos que permiten una observación detallada de los patrones de movilidad de los ciudadanos (Chen et al., 2016). Estos datos ofrecen oportunidades para mejorar la eficiencia del transporte público, la gestión del tráfico y la planificación urbana, lo que es esencial para ciudades densamente pobladas como Nueva York (Lv et al., 2015).

2.2. Estado del arte

Uno de los avances más destacados en la optimización del transporte es el uso de modelos predictivos basados en aprendizaje profundo. Lv et al. (2015) aplicaron un enfoque de *deep learning* para la predicción de flujo de tráfico, obteniendo resultados significativos que mejoraron la capacidad de prever patrones de tráfico en tiempo real. Estos modelos permiten anticipar la demanda de transporte y ajustarse dinámicamente para evitar embotellamientos, algo que es crucial en un entorno urbano complejo como Nueva York. Adicionalmente, estudios como el de Ke et al. (2018) han explorado el uso de redes neuronales recurrentes y convolucionales para la predicción de la demanda de taxis, un enfoque que puede aplicarse para gestionar la movilidad en tiempo real. Estos modelos predicen no solo el tráfico de vehículos, sino también la demanda de transporte público, lo cual permite mejorar la asignación de recursos en horas pico y reducir el tiempo de espera.

Los Sistemas Inteligentes de Transporte (ITS) integran datos de múltiples fuentes, como sensores de tráfico, cámaras de vigilancia y redes de telecomunicaciones para ofrecer soluciones en tiempo real a problemas de tráfico (Zhu et al., 2019). En este sentido, el Estado de Nueva York ha sido pionero en la implementación de ITS, los cuales permiten un monitoreo en tiempo real del tráfico y la automatización de la toma de decisiones, como el ajuste de los tiempos de apertura y cierre de los semáforos y la redirección del tráfico en casos de

accidentes o eventos inesperados (Department of Transportation, 2024). El concepto de vehículos conectados es también un área de gran relevancia en los ITS. Alanazi (2023) resalta que los vehículos conectados y autónomos (CAVs) pueden interactuar tanto con otros vehículos como con la infraestructura circundante, compartiendo datos sobre el tráfico y el estado de las vías. Este tipo de comunicación no solo optimiza el flujo vehicular, sino que también mejora la seguridad y la eficiencia, particularmente en intersecciones y zonas críticas. Estas tecnologías representan un enfoque prometedor para abordar los retos de la congestión y los accidentes en entornos urbanos.

La predicción precisa de la demanda de transporte y los flujos de tráfico es esencial para una planificación eficiente. Según Zhang et al. (2017), el uso de redes neuronales profundas para modelar datos espaciales y temporales ha permitido predicciones más precisas sobre el flujo de tráfico en tiempo real. Esto se aplica directamente a la gestión de redes de transporte urbano, ya que permite a los operadores prever picos de demanda y ajustar la oferta de transporte en consecuencia. González et al. (2008) también estudiaron cómo los datos de telefonía móvil pueden usarse para entender los patrones de movilidad humana, lo que proporciona una fuente adicional de datos valiosos para mejorar la planificación urbana. En este contexto, Nueva York podría beneficiarse de la integración de estos métodos, especialmente en áreas altamente transitadas como Manhattan, donde los patrones de tráfico y movilidad son complejos y cambiantes.

Uno de los mayores desafíos en la optimización de redes de transporte mediante Big Data es la integración de datos de diversas fuentes. Según Chen et al. (2016), la variabilidad y heterogeneidad de los datos, como los provenientes de sensores de tráfico, GPS, redes sociales y registros de uso de transporte, complican su integración y análisis. Las técnicas de minería de datos y el uso de algoritmos avanzados de aprendizaje automático permiten superar estos obstáculos, pero aún existen limitaciones, especialmente en términos de calidad y representatividad de los datos. Otro desafío importante que hay que considerar es el de la privacidad de los usuarios. El uso de datos personales, como los registros de teléfonos móviles y tarjetas de transporte, plantea preocupaciones éticas (Zheng et al., 2014). La anonimización de los datos y la implementación de políticas de privacidad son esenciales para garantizar que los usuarios no se vean afectados negativamente por el uso de sus datos personales.

Las aplicaciones de Big Data no solo se limitan a la optimización del tráfico en tiempo real, sino también a la planificación urbana a largo plazo. Colak et al. (2015) demostraron cómo los datos de movilidad a gran escala, como los registros de telefonía móvil, pueden utilizarse para la planificación de infraestructuras de transporte más eficientes. En ciudades como Nueva York, donde el crecimiento poblacional y los cambios en los patrones de trabajo y movilidad están en constante evolución, estos enfoques son clave para planificar futuras expansiones del sistema de transporte. El estudio de Borgi et al. (2017) también resalta el papel de Big Data en la mejora de la logística urbana, una aplicación crucial en Nueva York debido al elevado volumen de entregas de mercancías que se realizan diariamente en la ciudad. La optimización de rutas de vehículos de carga a través de Big Data puede mejorar significativamente la eficiencia del transporte y reducir las emisiones contaminantes.

En el futuro cercano, tecnologías emergentes como los vehículos autónomos y los vehículos eléctricos tendrán un impacto significativo en la optimización del transporte urbano. Zhu et al. (2019) argumentan que la combinación de Big Data con estas tecnologías puede mejorar la gestión del tráfico, reducir la congestión y mejorar la sostenibilidad del transporte en áreas urbanas. Otra tecnología emergente son los vehículos conectados, que pueden comunicarse con infraestructuras de transporte inteligentes para optimizar el flujo de tráfico en tiempo real. Este tipo de tecnología no solo mejorará la seguridad vial, sino que también permitirá la optimización dinámica del transporte público, ajustándose automáticamente a la demanda.

El uso de Big Data en la optimización de redes de transporte urbano representa una oportunidad transformadora para ciudades como Nueva York. A través de modelos predictivos avanzados, sistemas de transporte inteligentes y la integración de diversas fuentes de datos, es posible mejorar la eficiencia del transporte, reducir los tiempos de viaje y minimizar las emisiones contaminantes. Sin embargo, aún existen desafíos significativos en la integración de datos, la privacidad y la implementación de nuevas tecnologías, lo que subraya la importancia de seguir investigando en este campo.

La movilidad urbana ha sido un campo de estudio ampliamente enriquecido por avances como el análisis de correlaciones espaciales y temporales mediante redes neuronales de grafos. Kim et al. (2022) han demostrado cómo los mapas de flujo de trayectorias permiten identificar patrones en el tráfico a gran escala, mientras que Zhu et al. (2022) han optimizado la

predicción del flujo vehicular incorporando información espaciotemporal dinámica. Sin embargo, estas investigaciones no abordan de manera exhaustiva cómo factores externos, como la meteorología o los eventos urbanos, pueden alterar significativamente la movilidad en una ciudad compleja y dinámica como Nueva York.

Este proyecto, basado en datos reales y en abierto, avanza en esta dirección al integrar estos factores externos en el análisis. La incorporación de eventos y variables climáticas, combinada con modelos de redes neuronales de grafos, no solo permite predecir patrones de tráfico con precisión, sino también entender su impacto en la demanda multimodal. Este enfoque innovador proporciona a las administraciones locales herramientas concretas para mejorar la planificación y optimización del transporte urbano, algo que no se ha explorado profundamente en trabajos previos.

Por último, varios estudios han utilizado datos de registros de viajes de Taxi, Uber y Lyft, aunque siempre con objetivos diferentes a los de este proyecto (Liu et al., 2021; Poongodi et al., 2022; Yao et al., 2023). Estos estudios se han centrado en comparar la eficacia de distintos modelos de predicción, desarrollar modelos que ayuden a los conductores de transporte bajo demanda a encontrar pasajeros de manera más eficiente, predecir la demanda o analizar los factores que influyen en el precio del taxi. Sin embargo, ninguno aborda un enfoque que considera los datos climáticos y de eventos en la ciudad, como se pretende en este trabajo.

2.3. Conclusiones

El uso de Big Data en la optimización de redes de transporte urbano representa una oportunidad transformadora para ciudades como Nueva York. A través de modelos predictivos avanzados, sistemas de transporte inteligentes y la integración de diversas fuentes de datos, es posible mejorar la eficiencia del transporte, reducir los tiempos de viaje y minimizar las emisiones contaminantes. Sin embargo, aún existen desafíos significativos en la integración de datos, la privacidad y la implementación de nuevas tecnologías, lo que subraya la importancia de seguir investigando en este campo.

Este estado del arte resalta la necesidad de continuar desarrollando métodos avanzados de análisis de datos y algoritmos predictivos que sean capaces de manejar las complejidades de

las redes de transporte urbano y de adaptarse a las demandas cambiantes de las ciudades modernas.

3. Objetivos concretos y metodología de trabajo

3.1. Objetivo general

El objetivo general del proyecto es incrementar el conocimiento y la comprensión en el ámbito de la movilidad urbana en la ciudad de Nueva York, mediante el análisis exhaustivo de los datos de transporte privado, así como de su relación con los datos de demanda de metro de Nueva York. Este conocimiento permitirá proporcionar a las administraciones locales herramientas y datos clave para la mejora de la planificación urbana, con un enfoque en la optimización de las redes de transporte, la reducción de tiempos de desplazamiento y la mejora de la accesibilidad en toda la ciudad.

Para ello, el proyecto se enfocará en crear un modelo de análisis de datos escalable y exportable a otras ciudades, junto con modelos predictivos basados en redes neuronales de grafos, que permitan estimar con el tráfico e identificar los puntos de congestión.

Este trabajo pretende agregar valor añadido a la bibliografía existente por medio del desarrollo de un análisis acerca de cómo factores externos, como la meteorología y los eventos urbanos, influyen en la movilidad dentro del área metropolitana. Fenómenos como lluvias intensas, la velocidad del viento o grandes eventos culturales y deportivos pueden alterar significativamente los patrones de tráfico, la demanda de transporte privado y el uso del transporte público, generando desafíos únicos para la planificación y gestión de la movilidad urbana. Asimismo, este trabajo también pretende analizar la demanda de transporte en la ciudad desde un enfoque multimodal, estudiando los patrones de demanda del metro y de transporte privado.

Este enfoque proporcionará datos valiosos que ayudarán a las autoridades urbanísticas a tomar decisiones informadas sobre nuevas infraestructuras, mejoras en la conectividad y el diseño de nuevas rutas, considerando las necesidades de los usuarios y los impactos medioambientales. Integrar estas variables en el análisis permitirá evaluar de manera más precisa las opciones más eficientes y sostenibles para gestionar el transporte en una ciudad dinámica y en constante evolución.

3.2. Objetivos específicos

De forma más concreta, en el proyecto se pretende:

1. Identificar los principales cuellos de botella del distrito de Manhattan en relación con el tráfico: áreas de mayor demanda, las rutas más y menos demandadas y la distribución temporal de la demanda, la velocidad del tráfico y los tiempos de viaje.
2. Identificar los momentos de mayor congestión de tráfico en la ciudad: mapas de calor con la distribución los retrasos en la recogida y de la demanda.
3. Estudiar y analizar la influencia de las condiciones meteorológicas en el tráfico.
4. Estudiar y analizar la influencia del número de eventos en el área metropolitana de Manhattan en el tráfico de la ciudad.
5. Explorar la eficiencia del metro y su alineación con la demanda de *ride-hailing*.
6. Estudiar y analizar la influencia del número de eventos y las condiciones meteorológicas en el área metropolitana de Manhattan en la demanda de metro.
7. Crear un modelo escalable de procesamiento de datos distribuido para el análisis de la movilidad en Nueva York.
8. Crear una red neuronal basada en grafos para predecir el nivel de tráfico de una ruta determinada en base a las variables previamente comentadas y demostrar su competitividad frente a otros algoritmos clásicos de *machine learning*, como los *Random Forest*.

3.3. Metodología del trabajo

La metodología CRISP-DM es un enfoque estructurado y ampliamente utilizado para la realización de proyectos de ciencia de datos. Aunque fue diseñado específicamente para minería de datos, sus fases son aplicables a proyectos de analítica, optimización y predicción en redes de transporte. En este contexto, CRISP-DM proporciona un marco claro para estructurar las fases del análisis de datos en la optimización de redes de transporte urbano. A continuación, se detallan cada una de las fases aplicadas al análisis y optimización de redes de transporte urbano:

1. **Comprensión del negocio.** Esta fase inicial se centra en entender los objetivos del proyecto y su contexto en el ámbito de la movilidad urbana. Para optimizar una red de transporte es fundamental definir qué se quiere mejorar, por ejemplo, tiempos de viaje, cobertura de rutas, frecuencia o eficiencia multimodal. En esta fase, se debe colaborar con expertos en transporte y con las administraciones responsables para identificar los problemas críticos del sistema actual y plantear metas concretas, como

la reducción del tiempo promedio de desplazamiento entre puntos clave de la ciudad o la mejora en la conectividad entre diferentes medios de transporte.

2. **Comprensión de los datos.** En esta fase, se procede a una exploración exhaustiva de las fuentes de datos disponibles. Para el caso de este proyecto, se utilizarán los datos de los viajes de *ride-hailing* del año 2024 de la TLC, los datos de eventos de la CECM, los datos meteorológicos proporcionados por la librería *Meteostat* y los datos de la demanda de metro proporcionados por la *Metropolitan Transportation Authority*, que son de dominio público. El análisis inicial de los datos se centra en evaluar la calidad, la completitud y la fiabilidad de las distintas fuentes, identificando posibles problemas como datos faltantes, registros atípicos o discrepancias entre conjuntos de datos. La comprensión de los datos permite una visión general sobre su distribución, sus limitaciones y su potencial valor para el modelo.
3. **Preparación de los datos.** Una vez que los datos han sido comprendidos, se inicia una fase crítica, la de preparación, que incluye una serie de pasos fundamentales para garantizar que la información esté lista para su análisis y modelado. Esta fase involucra la limpieza de los datos, la selección de atributos relevantes, la integración de las distintas fuentes y el procesamiento necesario para que todos los datos tengan un formato y una estructura compatibles. En el contexto de un proyecto de optimización de redes de transporte, la integración de datos de múltiples fuentes es esencial para desarrollar modelos robustos y precisos.

   Una de las claves de esta fase es la inclusión de datos meteorológicos en los registros de transporte privado, como los de Uber y Lyft. Las condiciones climáticas, tales como la temperatura, las precipitaciones o la dirección del viento, podrían tener un impacto directo en la demanda del servicio y en los tiempos de viaje. Incorporar estos datos meteorológicos permite, por un lado, analizar cómo las variaciones de estos atributos afectan los patrones de uso del transporte privado y, por otro, ajustar las predicciones del tráfico en condiciones cambiantes. La integración de estos datos debe realizarse con una alineación temporal y espacial precisa. Para ello, se aplican transformaciones complejas, como la normalización temporal, que asegura que los datos climáticos coincidan con los momentos exactos de los viajes, y la normalización espacial, para garantizar que las condiciones meteorológicas se correspondan con el área de Nueva York.

Dentro de la fase de preparación, el análisis exploratorio de los datos se convierte en una etapa esencial para comprender los patrones subyacentes en el sistema de transporte. Durante el EDA, se analizarán métricas clave como el tiempo de viaje, los retrasos de recogida, la velocidad de circulación y demanda promedio, y sus correlaciones con el resto de los atributos. En consecuencia, este análisis se realizará con un enfoque temporal, buscando patrones estacionales según la hora del día o el día de la semana. De esta manera, por ejemplo, se podrá estudiar si un clima adverso incrementa la demanda de *ride-hailing* o si ralentiza la velocidad de circulación.

Para facilitar la interpretación y extracción de conocimientos, se utilizarán técnicas avanzadas de visualización de datos, como gráficos de dispersión y mapas de calor. Estas visualizaciones permitirán ilustrar la relación entre las variables y ayudarán a identificar patrones tanto espaciales como temporales en el comportamiento de los viajes. Los mapas de calor, por ejemplo, se utilizarán para visualizar las zonas de alta y baja demanda según los patrones temporales de la movilidad, mientras que los gráficos de series temporales proporcionarán información valiosa sobre las fluctuaciones horarias en la velocidad de circulación y la demanda. Este enfoque de análisis visual será crucial para optimizar los modelos y ajustar los recursos disponibles a los patrones observados en los datos.

4. **Modelado.** En la fase de modelado, el objetivo principal es aplicar algoritmos y técnicas de análisis de datos para crear modelos predictivos u optimizar soluciones. Para el proyecto de optimización de redes de transporte urbano, se utilizarán modelos de aprendizaje automático que permitan comprender la problemática (clusterización) y de aprendizaje profundo para predecir el tráfico de las rutas. En este contexto, la construcción de un grafo de transporte es una parte crucial. Un grafo de transporte modela las rutas y conexiones entre diferentes paradas o estaciones de transporte, donde los nodos representan las paradas y las aristas los caminos entre ellas, con un peso asociado a cada conexión que podría representar el tiempo de viaje, el coste o incluso la distancia.

   Se emplearán redes neuronales de grafos, que son técnicas más sofisticadas y eficientes para manejar relaciones complejas entre nodos de un grafo, ideales para predecir patrones de tráfico considerando las interdependencias entre distintas rutas.

El proceso de prueba y validación será fundamental en esta fase. Los modelos se entrenarán con conjuntos de datos históricos y se evaluarán en función de su capacidad para predecir con precisión el tráfico. Los resultados de estas pruebas servirán para ajustar y mejorar los modelos antes de su implementación en el sistema final.

5. **Evaluación.** En la última fase se llevará a cabo la evaluación de los modelos de inteligencia artificial en cuanto a su capacidad para predecir el tráfico. Para ello, se emplearán matrices de confusión y diversas métricas de rendimiento, como *precision*, *recall*, *F1-score* y *acuraccy*. Asimismo, se introducirán visualizaciones comparativas que permitirán analizar el desempeño de diferentes modelos y enfoques, proporcionando una visión clara sobre sus fortalezas, debilidades y aplicabilidad en escenarios concretos. Estas herramientas facilitarán una evaluación integral y fundamentada para determinar el modelo más adecuado según las características del problema.

Cabe mencionar que la última fase relativa al despliegue de la metodología CRISP-DM no se aborda en el presente proyecto, por lo que no se sigue con exactitud dicha metodología, solamente toma muchas de sus fases.

4. Marco normativo

Para el desarrollo de este Trabajo de Fin de Máster sobre la optimización de redes de transporte urbano en la ciudad de Nueva York, mediante tecnologías de Big Data y *deep learning,* es fundamental considerar las normativas de protección de datos aplicables, especialmente al manejar información geolocalizada.

La ciudad de Nueva York promueve la transparencia y el acceso a la información pública a través de iniciativas como *NYC Open Data*, que proporciona datos abiertos de diversas agencias municipales para su uso por parte de ciudadanos, investigadores y desarrolladores (NYC OpenData, 2024b).

Estos conjuntos de datos están diseñados para proteger la privacidad de los individuos, asegurando que la información publicada no permita la identificación de personas específicas. La política de privacidad de la ciudad establece que no se recopilan datos con fines comerciales

ni se intercambia, vende o distribuye la información recopilada a través de NYC.gov para dichos fines (NYC.gov, 2024a).

Aunque Estados Unidos no cuenta con una ley federal única de protección de datos, existen regulaciones sectoriales y estatales que pueden ser relevantes según la naturaleza de los datos y su uso, tal y como dicta la Ley de Privacidad de 1974 (U.S. Department of Justice, 2022). En el caso de Nueva York, la Ley de Libertad de Información (FOIL) regula el acceso a los registros públicos y establece excepciones para proteger la privacidad de los individuos (The Official Website of New York State, 2024a).

La *Office of Technology & Innovation* establece un marco integral para la protección de la información identificativa en las agencias de la ciudad, fundamentado en nueve principios clave de privacidad: transparencia, confianza pública, responsabilidad, minimización de datos, limitación de uso, gobernanza responsable, calidad de datos, seguridad y equidad. Estos principios promueven la recolección legal y justa de datos, su uso restringido a fines específicos y su protección mediante prácticas seguras y actualizadas, garantizando la precisión y la no discriminación. Además, se identifican las tipologías de datos a proteger, como la información identificativa y sensible, y se destaca la necesidad de acuerdos claros para el intercambio de datos, especificando propósitos, usuarios autorizados y medidas de seguridad. También se subraya la importancia de protocolos internos que aseguren el cumplimiento de leyes y regulaciones, junto con la responsabilidad de las agencias en implementar prácticas de privacidad, capacitar a sus oficiales y actualizar sus políticas ante nuevas amenazas. Este enfoque estructurado y ético fomenta la confianza pública y el cumplimiento normativo (Fitzpatrick, 2023).

En el caso de querer exportar esta idea al territorio nacional, dado que esta investigación utiliza datos geolocalizados anonimizados, es decir, sin información que permita identificar a individuos, estos datos no se consideran datos personales y, por lo tanto, no estarían sujetos a las normativas de protección de datos personales (AEDP, 2022). Sin embargo, es esencial garantizar que los datos permanezcan anonimizados durante todo el proceso de investigación para evitar cualquier riesgo de reidentificación.

5. Desarrollo específico de la contribución

Con el fin de alcanzar los objetivos descritos en el presente proyecto, el primer paso consistió en la exploración de las distintas fuentes *Open Data* de la ciudad de Nueva York, buscando aquellos conjuntos de datos de los que se esperaba un valor académico potencial. Tras esta fase de investigación, se seleccionaron los *datasets* de viajes de *Ride-Hailing* en la página de la TLC, los datos de metro de la *Metropolitan Transportation Authority*, los datos de eventos de la CECM y finalmente se obtuvieron los datos meteorológicos de la librería de *Meteostat*. El criterio de selección es la completitud (que se abordará posteriormente) y la fiabilidad de las fuentes.

Cabe mencionar que la exportación de los datos fue muy sencilla:

1. Los datos de *Ride-Hailing* se publican de forma mensual en formato .parquet. Se disponen datos desde 2024. Se han descargado los datos correspondientes a los primeros seis meses del año 2024 (de enero a junio).
2. Desde la página de *NYC OpenData* se pueden realizar consultas para exportar en formato .csv los eventos en el área de Nueva York o bien, conectarse directamente a su api.
3. Los datos de la demanda de recorridos de metro de la *Metropolitan Transportation Authority* se pueden obtener de idéntica manera a la referida al punto 2.
4. Los datos meteorológicos de *Meteostat* son fácilmente manipulables y convertibles a un *dataframe* de *Pandas* en *Python* directamente desde la librería *Meteostat*, seleccionando el rango de fechas y horas que se necesita, así como el punto geográfico objeto de estudio.

Todos los conjuntos de datos están muy bien documentados: incluyen un *Data Dictionary* en formato tabla/.csv o, en su defecto, disponen de documentación en línea con las características de cada atributo. Todos estos aspectos se abordarán en el apartado de compresión de los datos.

Todos estos conjuntos de datos se han subido a *Azure Databricks File System* en Azure *Databricks Community Edition,* y se ha creado un *clúster* con dos nodos.

5.1. Comprensión de los datos

Como se anticipó en la parte introductoria, el primer paso es entender los datos que se disponen e identificar los atributos más interesantes para alcanzar los objetivos definidos.

A continuación, se presenta una tabla con todos los atributos, sus descripciones y el tipo de dato correspondientes al conjunto de datos proporcionado por la *Taxi and Limosine Comission*: Ver **Tabla 4.**

Tabla 4. *Data Dictionary de la Taxi and Limosine Comission.*

Atributo	Descripción	Tipo de dato
Hvfhs_license_num	El número de licencia TLC de la base o empresa HVFHS	String
Dispatching_base_num	El número de licencia TLC Base de la base que despachó el viaje.	String
Pickup_datetime	La fecha y hora en que comenzó el viaje.	Timestamp
DropOff_datetime	La fecha y hora en que terminó el viaje.	Timestamp
PULocationID	Zona de taxi TLC donde comenzó el viaje.	String
DOLocationID	Zona de taxi TLC donde terminó el viaje.	String
originating_base_num	Número de la base que recibió la solicitud inicial del viaje.	String
request_datetime	Fecha/hora en que el pasajero solicitó ser recogido.	Timestamp
on_scene_datetime	Fecha/hora en que el conductor llegó al lugar de recogida (solo para vehículos accesibles).	Timestamp
trip_miles	Millas totales del viaje del pasajero.	Double
trip_time	Tiempo total en segundos del viaje del pasajero.	Double
base_passenger_fare	Tarifa base del pasajero antes de peajes, propinas, impuestos y tasas.	Long
tolls	Coste total de todos los peajes pagados en el viaje.	Double
bcf	Coste total recaudado en el viaje para el Black Car Fund.	Double
sales_tax	Coste total recaudado en el viaje para el impuesto sobre ventas del estado de NY.	Double
congestion_surcharge	Coste total recaudado en el viaje para el recargo por congestión del estado de NY.	Double
airport_fee	Tarifa de $2.50 por recogida y entrega en los aeropuertos de LaGuardia, Newark y JFK.	Double
tips	Coste total de propinas recibidas del pasajero.	Double
driver_pay	Pago total al conductor (sin incluir peajes ni propinas y neto de comisión, recargos o impuestos).	Double

shared_request_flag	¿El pasajero aceptó un viaje compartido, independientemente de si fue emparejado? (S/N).	String
shared_match_flag	¿El pasajero compartió el vehículo con otro pasajero que reservó por separado en algún momento del viaje? (S/N).	String
access_a_ride_flag	¿El viaje fue administrado en nombre de la Autoridad Metropolitana de Transporte (MTA)? (S/N).	String
wav_request_flag	¿El pasajero solicitó un vehículo accesible para sillas de ruedas (WAV)? (S/N).	String
wav_match_flag	¿El viaje ocurrió en un vehículo accesible para sillas de ruedas (WAV)? (S/N).	String

Fuente: NYC.gov, 2024b.

De todos los atributos presentes, los más atractivos para analizar la movilidad son los relativos al tiempo de viaje, las distancias y las ubicaciones de recogida y de llegada. Con estos atributos es posible extraer información sobre las rutas más demandadas, realizar un análisis temporal de la demanda, analizar la velocidad promedio e identificar las zonas con mayores retrasos.

Adicionalmente, la TLC proporciona la relación geográfica de los id que se indican en las columnas *PULocationID* y *DOLocationID*, mediante un fichero .shp, un .dbf y un .shx. Un fichero .shp (*shapefile*) almacena geometrías vectoriales como puntos, líneas o polígonos que representan elementos geográficos. El fichero .dbf contiene los datos tabulares asociados a esas geometrías (atributos como nombres, valores, etc.). El fichero .shx actúa como un índice que conecta las geometrías del .shp con los datos del .dbf, mejorando la eficiencia en la búsqueda. Estos tres archivos son necesarios juntos para trabajar correctamente con *shapefiles*. Para el procesamiento de estos ficheros se utiliza *GeoPandas.* El integrar esta información espacial es crucial para la visualización de mapas de calor sobre la ciudad de Nueva York.

El enorme tamaño del conjunto de datos es el principal motivo por el que se decide utilizar un entorno distribuido y *Spark* para su procesamiento. En este caso, los archivos .parquet tienen un tamaño promedio de 0.4GB, conteniendo millones de registros. Se maneja, por tanto, un *dataframe* de 120 millones de filas. Se han filtrado los registros correspondientes a los viajes desde y hacia el área metropolitana de Manhattan para poder abordar el proyecto con la capacidad computacional disponible (15GB de memoria) y adicionalmente se ha decidido reducir el tamaño de la muestra a un 5% de los datos, es decir, se trabaja con un *dataframe* de en torno a 6 millones de registros.

Seguidamente, se presenta una tabla con todos los atributos, sus descripciones y el tipo de dato correspondientes al conjunto de datos proporcionado por la Oficina de Coordinación de eventos de la ciudad de Nueva York. Ver **Tabla 5**:

Tabla 5. *Data Dictionary de los eventos en la ciudad de Nueva York.*

Atributo	Descripción	Tipo de dato
Event ID	Este es el ID del evento.	Integer
Event Name	Este es el nombre del evento.	String
Start Date/Time	Esta es la fecha/hora de inicio de este evento. Para la mayoría de los eventos, será el tiempo de configuración del evento	Floating Timestamp
End Date/Time	Esta es la fecha de finalización de este evento. Para la mayoría, será el tiempo de desmontaje del evento	Floating Timestamp
Event Agency	Esta es la agencia de NYC considerada la agencia principal que otorga permisos para este evento.	String
Event Type	Este es el tipo de evento.	String
Event Borough	Este es el distrito donde se llevará a cabo el evento.	String
Event Location	Esta es la ubicación del evento.	String
Event Street Side	Este es el lado de la calle donde se llevará a cabo el evento si la ubicación es una calle.	String
Street Closure Type	Este es el tipo de cierre de calle si la ubicación está en una calle.	String
Community Board	Este es el Community Board donde está ubicado el evento.	String
Police Precinct	Este es el recinto policial donde está ubicado el evento.	String

Fuente: NYC OpenData, 2025

Con los atributos descritos es posible calcular el número de eventos por tipo, fecha y hora. Los resultados de estos cálculos se podrán cruzar con los datos de la TLC, para su ulterior análisis.

La tabla con los datos de eventos de la ciudad contiene unos 350.000 registros, correspondientes a los eventos de la ciudad durante el año 2024.

Seguidamente, en las tablas inferiores se describen los atributos que pone a disposición del usuario la librería *Meteostat*, las unidades de cada variable y el significado de la variable *coco*, indicativa de la condición meteorológica particular. Ver **Tabla 6; Tabla 7; y Tabla 8**.

Tabla 6. *Data Dictionary de Meteostat.*

Código	Significado
TEMP	Temperatura
TAVG	Temperatura promedio
TMIN	Temperatura máxima
TMAX	Temperatura mínima
DWPT	Punto de condensación
PRCP	Precipitación total
WDIR	Dirección del viento
WSPD	Velocidad del viento promedio
WPGT	Velocidad máxima del viento
RHUM	Humedad relativa
PRES	Presión
SNOW	Profundidad de la nieve
TSUN	Duración de la luz diurna
COCO	Código de condición de tiempo

Fuente: Meteostat, 2024.

Tabla 7. *Unidades de los atributos de los datos de Meteostat.*

Parámetro	Unidades
Temperatura	°C
Precipitación	mm
Luz Diurna	Minutes
Presión del aire	hPa
Velocidad del viento	km/h
Dirección del viento	Grados
Visibilidad, altura de las nubes	m
Humedad relativa	%

Fuente: Meteostat, 2024.

Tabla 8. *Significado código COCO de Meteostat.*

Código	Condición climatológica
1	Despejado
2	Agradable
3	Nublado
4	Cubierto
5	Niebla
6	Niebla helada
7	Lluvia ligera
8	Lluvia
9	Lluvia intensa
10	Lluvia helada
11	Lluvia helada intensa

12	Aguanieve
13	Aguanieve intensa
14	Nevisca ligera
15	Nevada
16	Nieve intensa
17	Chaparrón
18	Chaparrón fuerte
19	Chaparrón de aguanieve
20	Chaparrón de aguanieve intenso
21	Chaparrón de nieve
22	Chaparrón de nieve intenso
23	Relámpagos
24	Granizo
25	Tormenta de relámpagos
26	Tormenta de relámpagos fuerte
27	Tormenta

Fuente: Meteostat, 2024.

Estos datos se proporcionan por fecha y hora del día, por lo que es posible nutrir y enriquecer el *dataframe* de la TLC con datos meteorológicos, para estudiar y correlacionar eventos de tráfico con condiciones meteorológicas adversas. Asimismo, la tabla que contiene los datos meteorológicos de Nueva York tiene 7.846 filas, que resultan de multiplicar las 24 horas del día por el número de días contenidos en el rango de estudio.

Los registros de *ride-hailing*, proporcionados por la TLC, se caracterizan por una granularidad temporal a nivel de minutos, lo que permite capturar detalles precisos sobre cada viaje. Por otro lado, los datos meteorológicos y de eventos están disponibles a nivel de hora, lo que introduce una diferencia en la granularidad temporal. Aunque esta disparidad puede reducir ligeramente la precisión al integrar las fuentes, se buscará mitigar este impacto mediante estrategias de agregación temporal y una unificación cuidadosa de las tablas.

En resumen, después de aplicar los filtros necesarios y los *join,* tiene 1.329.724 filas y 76 columnas.

Finalmente, en la tabla inferior se describen los datos de la demanda de recorridos en la ciudad de Nueva York. Ver **Tabla 9**.

Tabla 9. *Data Dictionary de metro de Nueva York.*

Atributo	Tipo de dato	Descripción
Year	NUMERIC	El año en el que ocurrieron los viajes en metro.
Month	NUMERIC	El mes en el que ocurrieron los viajes en metro.
Day of Week	TEXT	El día de la semana en que ocurrieron los viajes en metro (lunes, martes, etc.).
Hour of Day	NUMERIC	La hora del día en que ocurrieron los viajes en metro. Todas las horas se redondean hacia abajo a la hora más cercana.
Timestamp	DATE	Fecha representativa para el año, mes, día de la semana y hora del día en que ocurrieron los viajes en metro. Esta fecha proviene de la primera semana completa del mes.
Origin Station Complex ID	TEXT	El identificador único del complejo de estaciones de metro donde comenzaron los viajes.
Origin Station Complex Name	TEXT	El nombre del complejo de estaciones de metro donde comenzaron los viajes.
Origin Latitude	NUMERIC	La latitud del complejo de estaciones de metro donde comenzaron los viajes.
Origin Longitude	NUMERIC	La longitud del complejo de estaciones de metro donde comenzaron los viajes.
Destination Station Complex ID	TEXT	El identificador único del complejo de estaciones de metro donde se infiere que terminaron los viajes.
Destination Station Complex Name	TEXT	El nombre del complejo de estaciones de metro donde se infiere que terminaron los viajes.
Destination Latitude	NUMERIC	La latitud del complejo de estaciones de metro donde se infiere que terminaron los viajes.
Destination Longitude	NUMERIC	La longitud del complejo de estaciones de metro donde se infiere que terminaron los viajes.
Estimated Average Ridership	NUMERIC	La estimación del número de pasajeros para un par origen-destino y hora del día, promediada.

Fuente: The Official Website of New York State, 2024b.

Estos datos se proporcionan por fecha y hora del día, por lo que es posible nutrir y enriquecer este *dataframe* con los datos meteorológicos y de eventos de la ciudad, de manera similar a

lo que se realiza en el *dataframe* de la TLC. El volumen de datos que se importan desde el *api* es variable, pero en este caso particular se ha trabajado con un conjunto de datos de 3 millones de registros de pares ubicación-destino.

5.2. Preparación de los datos

El Análisis Exploratorio de Datos es una etapa fundamental en cualquier proyecto de análisis de datos, ya que permite comprender la estructura, las características principales y las relaciones entre las variables de un conjunto de datos. A través de técnicas estadísticas y visualizaciones, el EDA facilita la detección de patrones, tendencias, valores atípicos y posibles inconsistencias en los datos, lo que es clave para orientar las decisiones de modelado y preparación de los datos. En este trabajo, el EDA se realiza para identificar cómo las variables de ubicación, tiempo, condiciones climáticas y los eventos de la ciudad influyen en los tiempos de traslado, velocidad de circulación y demanda, proporcionando un punto de partida sólido para el desarrollo de modelos predictivos.

5.2.1. Limpieza de los datos

Datos de transporte privado del área de Manhattan

El proceso de limpieza de datos engloba la gestión de los datos faltantes y los valores atípicos. Como se comentó previamente, la completitud de los atributos objeto de estudio en todos los conjuntos de datos es máxima, por lo que, en ese sentido, no ha habido que aplicar ninguna técnica de eliminación/relleno de datos, que incluyen:

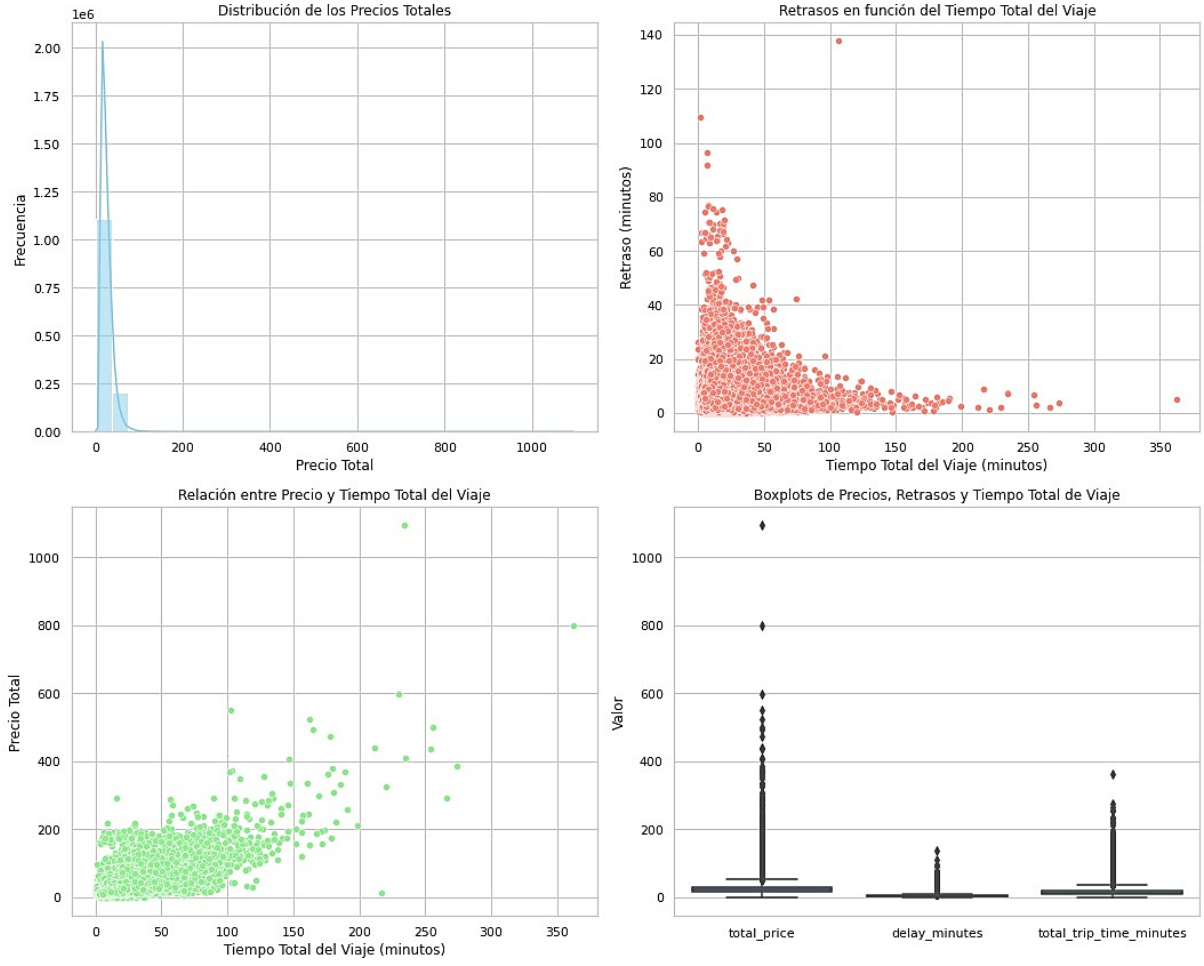
- Descarte de los atributos en el caso de que superen cierto porcentaje de valores faltantes.
- Relleno con la media de los valores, la moda y la mediana.
- Relleno por proximidad.

Los valores negativos de los retrasos, tiempos de viaje y precio se descartaron por entender que no es posible abonar dinero por viajar o que los tiempos sean negativos. Lo más adecuado hubiera sido consultar a la fuente sobre esos datos en concreto, por si se tratase únicamente de un error en los signos.

Por el contrario, sí se ha realizado un análisis de los valores atípicos. Se han analizado los valores anormales de las columnas clave como el tiempo de viaje, la distancia recorrida, el

precio pagado y los retrasos de recogida. Se ha estudiado la distribución de la variable precio, la relación entre el retraso y el tiempo total de viaje, los valores extremos de los kilómetros recorridos y las variables mencionadas, así como la relación entre precio y el tiempo total de viaje. Asimismo, se muestran los diagramas de cajas para las tres variables mencionadas. Todas ellas muestran una foto completa de los valores atípicos del conjunto de datos. Ver **Figura 7.**

Figura 7. *Histogramas, diagramas de burbujas y diagramas de caja para búsqueda de valores atípicos.*



Fuente: Elaboración propia.

Como se puede observar, los valores máximos de todas las variables están muy por encima de lo que se entiende como real en un área metropolitana como la de Manhattan, con unos 21 kilómetros de largo por 4 kilómetros de ancho en su zona más estrecha. Ver **Tabla 10**.

Tabla 10. *Resumen estadístico de las variables de tiempo de viaje, distancia recorrida, precio y retraso de recogida.*

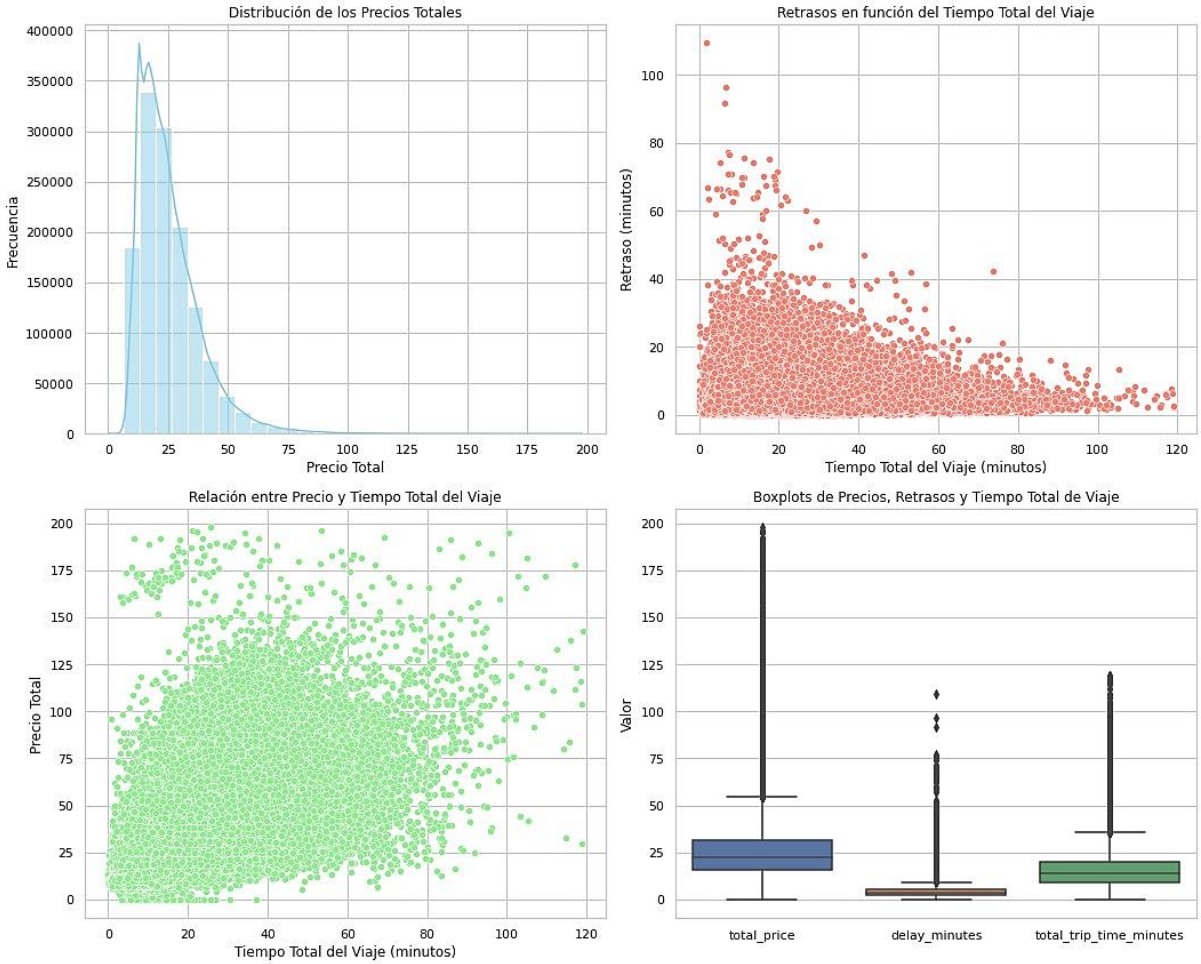
summary	total_trip_time_minutes	trip_kilometers	total_price	delay_minutes
count	1323505	1323505	1323505	1323505
mean	15.614587553503767	4.237957303528551	25.503821300259627	4.2536452701979455
stddev	8.720713317993805	3.343052539734923	13.499226231815133	2.730872361070354
min	0.0	0.0	0.0	0.0
max	362.3	301.8960906	1094.6100000000001	137.88333333333333

Fuente: Elaboración propia.

En cuanto a la estrategia seguida con los valores atípicos encontrados en la distancia de viaje, se deciden eliminar los registros que están por encima de 30 kilómetros, no siendo lógico que se realicen viajes de mayor distancia dentro del área metropolitana de Manhattan. Asimismo, se decide eliminar los valores atípicos de tiempo de viaje, precio y retrasos, estableciendo unos límites máximos de tiempo, pues se entiende que es muy difícil que trayectos de menos de 25 kilómetros tarden en completarse más de 120 minutos (2 horas) o cuyo precio sea mayor de 200 dólares. Siguiendo esta estrategia se pretende mantener representatividad de hechos que pueden considerarse atípicos pero explicables, descartando las que se presuponen extremadamente improbables.

Después de aplicar la comentada estrategia, las visualizaciones previamente indicadas cambian su forma. Ver **Figura 8.**

Figura 8. *Histogramas, diagramas de burbujas y diagramas de caja después de eliminar valores anómalos.*



Fuente: Elaboración propia.

Con este enfoque, se pretende dar cabida y mantener situaciones que pueden ser anómalas precisamente por estos eventos y condiciones meteorológicas adversas.

Datos de transporte público (metro) del área de Manhattan

En este conjunto de datos no se ha llevado a cabo un proceso de limpieza de datos en el sentido tradicional, dado que los datos utilizados para analizar la demanda de metro provienen de fuentes oficiales que garantizan una alta calidad y completitud. Estas bases de datos son generadas y mantenidas por organismos públicos y sistemas de transporte con protocolos estrictos para asegurar la integridad de la información. La ausencia de valores nulos o inconsistencias estructurales en los datos proporcionados respalda esta decisión,

permitiendo centrar los esfuerzos en el análisis y modelado de los patrones de movilidad. Ver **Tabla 11**.

Tabla 11. *Valores faltantes de los datos de metro.*

date	hour	timestamp	ridership	origin_station_complex_id	destination_station_complex_id	origin_latitude	origin_longitude	day_of_week	destination_longitude	destination_latitude
0	0	0	0	0	0	0	0	0	0	0

Fuente: Elaboración propia.

No obstante, se realizaron ciertas transformaciones necesarias para adecuar los datos al formato requerido para el análisis. En primer lugar, se llevó a cabo una conversión de formatos para integrar los datos en una estructura compatible con las herramientas analíticas empleadas. Adicionalmente, se aplicó un filtro directamente en la *api* de origen para extraer exclusivamente los datos de recorridos desde, hacia y dentro de Manhattan, así como aquellos correspondientes al rango temporal objeto de estudio (01/01/2024 al 01/07/2024). Por último, se generó la columna *date* (por medio de la función de *spark to_date*) para poder realizar los ulteriores *join.* Además, se asegura que las coordenadas geográficas de las estaciones de origen y destino (*origin_latitude, origin_longitude, destination_latitude, destination_longitude)* estén en formato numérico *double* mediante el uso de *cast*. Este enfoque no solo optimizó el proceso de ingesta de datos, sino que también redujo la necesidad de realizar posteriores manipulaciones complejas.

5.2.1.1. Conversión de formatos

En esta etapa, se garantiza que todas las columnas de fecha sean tipo *timestamp* en el mismo formato, ya que este paso es crucial para que los *join* se produzcan eficientemente.

En este proyecto, el manejo de fechas es crucial para garantizar que se integran todas las fuentes de datos correctamente. Esto es común a todos los conjuntos de datos.

5.2.1.2. Generación de atributos derivados

Datos de transporte privado del área de Manhattan

La preparación de atributos a partir de los datos de *ride-hailing* se enfoca en la generación de nuevos atributos clave para el análisis de los patrones de tráfico:

- **Precio total del viaje.** Suma varias columnas de tarifas y tasas (*base_passenger_fare, tolls, bcf, sales_tax, congestion_surcharge, y airport_fee*), manejando valores nulos con *coalesce* (sustituyéndolos por 0).
- **Retraso en minutos**. Calcula la diferencia en minutos entre la hora de recogida y la solicitud.
- **Tiempo total del viaje**. Calcula la diferencia en minutos entre la hora de llegada al destino y la hora de recogida.
- **Conversión de millas a kilómetros**. Transforma la distancia del viaje (*trip_miles*) de millas a kilómetros.
- **Información temporal.** Se obtiene la hora del día y el día de la semana.

Datos de eventos de la ciudad de Nueva York

La preparación de atributos a partir de los datos de eventos se enfoca en convertir información de tipo temporal y categórica en un formato estructurado y procesable. A continuación, se detalla el proceso llevado a cabo:

- **Estandarización de fechas.** Las columnas de inicio y fin del evento *("Start Date/Time" y "End Date/Time"*) se convierten al tipo *timestamp*, gestionando distintos formatos de fecha y asegurando la coherencia temporal. Se eliminan casos donde la fecha de fin es anterior a la de inicio para mantener la integridad de los datos.
- **Generación de secuencias horarias**. Se crea una secuencia de horas entre las fechas de inicio y fin de cada evento, generando una columna para cada hora en la que el evento está activo. Esto permite analizar la ocurrencia de eventos en intervalos horarios específicos.
- **Extracción de atributos temporales**. A partir de la secuencia generada, se extraen atributos clave como la fecha (*"Date"*) y la hora (*"Hour"*) para facilitar el análisis temporal.
- **Conteo de eventos por hora**. Se asigna un valor binario a cada evento, indicando si ocurre en una hora específica, y se agrupan los datos por fecha, hora y tipo de evento para calcular el conteo total de cada tipo.
- **Pivotación y agregación**. Los tipos de evento se convierten en columnas individuales mediante pivotación, permitiendo analizar la distribución e influencia de cada tipo de evento. Los valores nulos se reemplazan por 0.

- **Creación de un atributo general.** Se genera una nueva columna, "*nº events*", que suma el total de eventos activos por fecha y hora. Este atributo resume la intensidad de eventos en cada intervalo de tiempo.

La estructura del *dataframe* es la que se muestra a continuación, donde solo se muestran 6 columnas del total. Ver **Tabla 12**.

Tabla 12. *Estructura resultante del dataframe de eventos en el área metropolitana de Manhattan.*

Date	Hour	Athletic Race / Tour	BID Multi-Block	Bike the Block	Block Party
2024-06-07	23:00	0	0	0	0
2024-08-09	12:00	0	0	0	2
2024-08-28	14:00	0	0	0	0
2024-07-29	00:00	0	0	0	0
2024-04-20	10:00	0	0	0	0
2024-05-10	22:00	0	0	0	0
2024-04-16	18:00	0	0	0	0
2024-10-25	11:00	0	0	0	0
2024-08-21	23:00	0	0	0	0
2024-06-10	19:00	0	0	0	0

only showing top 10 rows

Fuente: Elaboración propia.

Este procesamiento convierte datos complejos de eventos en una estructura apta para análisis y modelado predictivo, facilitando la integración con otras fuentes de datos y la generación de *insights*.

Datos meteorológicos de la ciudad de Nueva York

La generación de atributos en el *dataframe* de datos meteorológicos se basa en transformar la columna *coco* (que representa las condiciones meteorológicas) en un conjunto de columnas binarias. Cada valor único presente en *coco* se convierte en un atributo individual. Para cada fila, la columna correspondiente al valor de *coco* se establece en 1 si coincide con la condición meteorológica registrada, y en 0 en caso contrario. Esta técnica, conocida como codificación *one-hot*, permite representar de forma explícita y procesable las distintas condiciones meteorológicas en el modelo.

La estructura del *dataframe* es la que se muestra a continuación, donde solo se muestran 17 columnas del total. Ver **Tabla 13**.

Tabla 13. *Estructura resultante del dataframe de eventos en el área metropolitana de Manhattan.*

time	temp	dwpt	rhum	prcp	snow	wdir	wspd	wpgt	pres	tsun	coco	coco_1.0	coco_20.0	coco_15.0	coco_17.0	coco_0.0
2024-01-01 00:00:00	6.0	-4.5	47.0	0.0	0.0	250.0	11.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 01:00:00	6.0	-2.9	53.0	0.0	0.0	243.0	7.9	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 02:00:00	6.0	-2.9	53.0	0.0	0.0	261.0	6.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 03:00:00	6.0	-2.9	53.0	0.0	0.0	250.0	9.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 04:00:00	6.0	-2.9	53.0	0.0	0.0	260.0	9.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 05:00:00	6.0	-2.4	55.0	0.1	0.0	260.0	11.0	0.0	1017.0	0.0	12.0	0	0	0	0	0
2024-01-01 06:00:00	6.0	-1.9	57.0	0.0	0.0	232.0	7.0	0.0	1016.0	0.0	3.0	0	0	0	0	0
2024-01-01 07:00:00	5.0	-2.1	60.0	0.0	0.0	235.0	6.8	0.0	1016.0	0.0	3.0	0	0	0	0	0
2024-01-01 08:00:00	5.0	-1.7	62.0	0.0	0.0	250.0	6.0	0.0	1016.0	0.0	3.0	0	0	0	0	0
2024-01-01 09:00:00	4.0	-1.4	68.0	0.0	0.0	250.0	6.8	0.0	1016.0	0.0	3.0	0	0	0	0	0
2024-01-01 10:00:00	5.0	-1.0	65.0	0.0	0.0	290.0	7.0	0.0	1016.0	0.0	3.0	0	0	0	0	0
2024-01-01 11:00:00	4.0	0.7	79.0	0.0	0.0	290.0	6.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 12:00:00	4.0	1.2	82.0	0.0	0.0	240.0	6.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 13:00:00	4.0	0.7	79.0	0.0	0.0	341.0	6.0	0.0	1018.0	0.0	3.0	0	0	0	0	0
2024-01-01 14:00:00	5.0	0.6	73.0	0.0	0.0	350.0	6.0	0.0	1018.0	0.0	3.0	0	0	0	0	0
2024-01-01 15:00:00	6.0	0.5	68.0	0.0	0.0	80.0	6.0	0.0	1018.0	0.0	3.0	0	0	0	0	0
2024-01-01 16:00:00	7.0	-0.2	60.0	0.0	0.0	17.0	6.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 17:00:00	8.0	0.2	58.0	0.0	0.0	25.0	6.0	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 18:00:00	7.0	-1.2	56.0	0.0	0.0	22.0	6.8	0.0	1017.0	0.0	3.0	0	0	0	0	0
2024-01-01 19:00:00	8.0	-1.0	53.0	0.0	0.0	50.0	6.0	0.0	1017.0	0.0	3.0	0	0	0	0	0

only showing top 20 rows

Fuente: Elaboración propia.

5.2.2. Estructuración de los datos

La estructuración de los datos constituye un paso esencial en el procesamiento de grandes volúmenes de información en proyectos de análisis y optimización de transporte urbano. En este proyecto, se ha empleado *Apache Spark*, una herramienta distribuida y escalable, para llevar a cabo las operaciones de limpieza, integración y transformación de los datos. A continuación, se describen detalladamente los pasos seguidos para construir la base de datos final, combinando distintas fuentes mediante operaciones de *join*.

5.2.2.1. Relación entre IDs y coordenadas geográficas

En primer lugar, se ha cruzado la información de las localizaciones de recogida (pickup_*locationid*) y destino (*dropoff*_locationid) con una tabla de referencia (proporcionada por la propia TLC) que contiene la relación entre los identificadores de ubicación y sus respectivas coordenadas geográficas (ID, latitud y longitud). Este cruce se realizó mediante un *join* de tipo *left* para preservar todas las observaciones del conjunto principal y añadir las coordenadas asociadas. Esta operación genera dos nuevas columnas para cada tipo de localización: *pickup_latitude, pickup_longitude, dropoff_latitude y dropoff_longitude*.

5.2.2.2. Integración de eventos urbanos

Los datos de transporte público y privado se complementaron con un conjunto adicional que recoge los eventos urbanos registrados en la ciudad durante el periodo de análisis. Estos eventos incluyen actividades como ferias, festivales, desfiles y competiciones deportivas, categorizados según su tipo. La integración se realizó mediante un *join* de tipo *left*, utilizando como claves de unión la fecha y la hora del evento. Este proceso incorpora una columna por cada tipo de evento, en la cual se registra la suma del número de eventos de cada evento en el momento del viaje. Además, se añadió una columna consolidada (*nº events*) previamente creada que refleja el número total de eventos simultáneos.

5.2.2.3. Integración de datos meteorológicos

Posteriormente, los datos de transporte público y privado se enriquecieron con información meteorológica, asociando cada viaje con las condiciones climáticas predominantes al momento del desplazamiento. Este cruce se llevó a cabo mediante un *join* de tipo *left*, utilizando las columnas de fecha y hora como claves para establecer la relación entre ambos conjuntos de datos. Este paso añade variables como temperatura (*temp*), precipitaciones (*prcp*), humedad relativa (*rhum*), velocidad del viento (*wspd*) y la condición meteorológica específica (ej.: lluvia intensa), entre otras, las cuales son fundamentales para analizar el impacto de las condiciones climáticas en los tiempos de viaje.

5.2.2.4. Resumen

El resultado final es un *dataframe* que integra información geográfica, meteorológica y de eventos urbanos con los datos de transporte, en un formato listo para el análisis avanzado y la construcción de modelos de predicción. Cada paso en el proceso de estructuración ha sido implementado con *Apache Spark*, aprovechando su capacidad para procesar datos en paralelo y manejar eficientemente conjuntos de datos de gran tamaño. La estructura columnar del *dataframe* final es la siguiente. Ver **Tabla 14**.

Tabla 14. *Estructura columnar del dataframe.*

pickup_location_id	**pickup_latitude**	**dropoff_latitude**	**request_datetime**
droppff_location_id	**pickup_longitude**	**dropoff_longitude**	**total_price**
delay_minutes	**total_trip_time_minutes**	**license**	**trip_kilometers**
hour_of_day	**day_of_week**	**date**	**hour_of_day_formatted**
Athletic Race / Tour	**BID Multi-Block**	**Bike the Block**	**Block Party**
Clean-Up	**Farmers Market**	**Grid Request**	**Health Fair**
Open Culture	**Open Street Patner Event**	**Parade**	**Plaza Event**
Plaza Patner Event	**Press Conference**	**Production Event**	**Religious Event**
Sidewalk Sale	**Single Block Festival**	**Special Event**	**Sport - Adult**
Sport - Youth	**Stationary Demonstration**	**Stickball**	**Street Event**
Street Festival	**Theater Load and Load Outs**	**nº events**	**Hour_int**
time	**temp**	**dwpt**	**rhum**
prcp	**snow**	**wdir**	**wspd**
pres	**tsun**	**coco**	**coco_1.0**
coco_20.0	**coco_15.0**	**coco_17.0**	**coco_0.0**
coco_9.0	**coco_12.0**	**coco_18.0**	**coco_14.0**
coco_19.0	**coco_5.0**	**coco_4.0**	**coco_7.0**
coco_2.0	**coco_8.0**	**coco_3.0**	**coco_16.0**
coco_13.0	**total_trip_time_hours**	**speed_kmh**	**Date**

Fuente: Elaboración propia.

5.2.3. Análisis estadístico

5.2.3.1. Ride-hailing

El análisis exploratorio de datos es una etapa clave para comprender los patrones y relaciones subyacentes en el conjunto de datos de transporte privado en Nueva York, que incluye información de Uber, Lyft, etc. A continuación, se detallan las principales observaciones y visualizaciones realizadas, estructuradas en función de los objetivos exploratorios planteados.

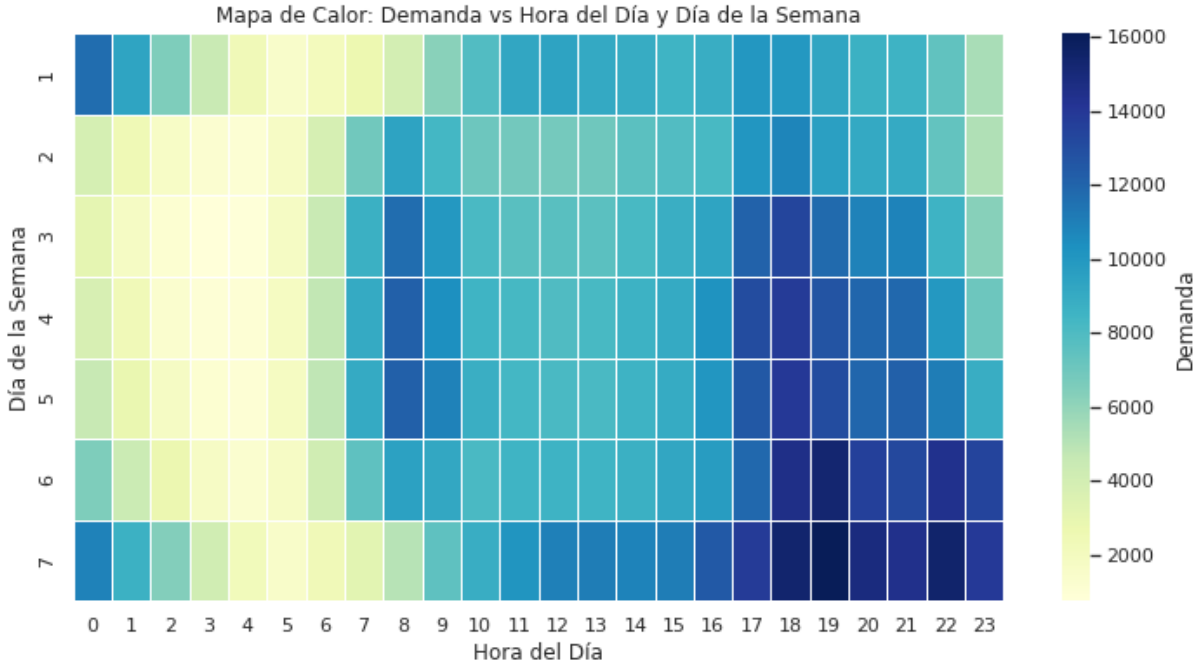
Se crearon mapas de calor para visualizar la variación de la demanda, el tiempo de viaje y la velocidad promedio por día de la semana frente a la hora del día.

Las visualizaciones que se presentan a continuación representan los días de la semana codificados como números del 1 al 7, siguiendo la lógica de la función *dayofweek* utilizada en *PySpark*. Bajo este esquema, los días están mapeados según el estándar heredado de Java y *Spark SQL*, donde el valor 1 corresponde al domingo, 2 al lunes, y así sucesivamente hasta el sábado, que se representa con el número 7. Este enfoque permite analizar tendencias o

patrones temporales relacionados con los días de la semana de forma consistente con la funcionalidad subyacente de *Spark*.

La demanda alcanzó su punto máximo durante las horas pico los días laborables y los fines de semana por la tarde-noche, entendiendo fin de semana desde el viernes por la tarde a domingo por la tarde. Ver **Figura 9.**

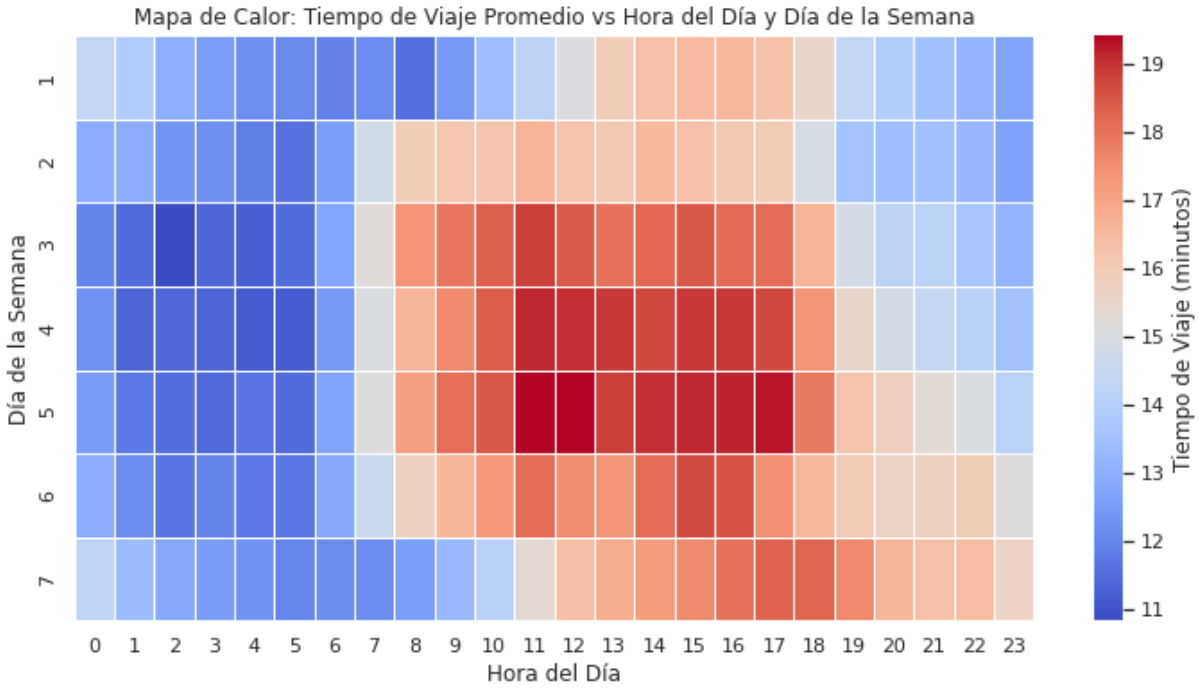
Figura 9. *Mapa de calor de la demanda por hora del día y día de la semana.*



Fuente: Elaboración propia.

El tiempo de viaje se elevó durante las horas pico, mostrando cuellos de botella en las horas de mayor afluencia. Ver **Figura 10**.

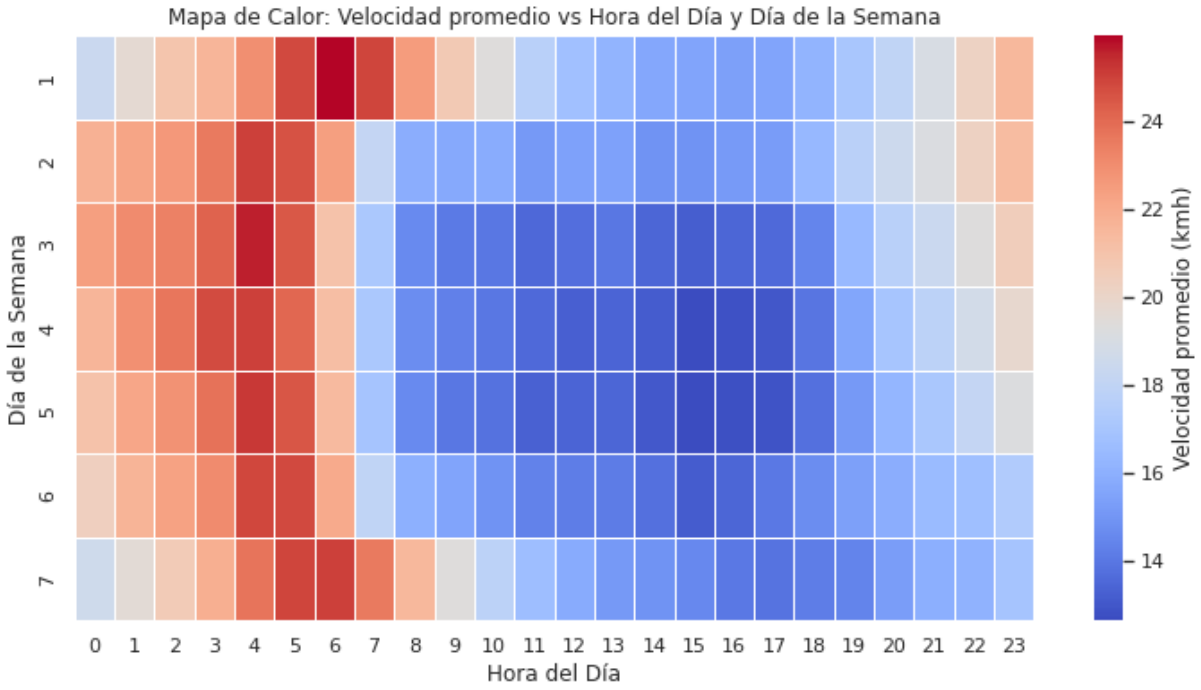
Figura 10. *Mapa de calor del tiempo de viaje por hora del día y día de la semana.*



Fuente: Elaboración propia.

Las velocidades más bajas coincidieron con estos picos de tiempo de viaje, especialmente en días laborables, por la tarde, cuando se entiende que coincide con la salida de los ciudadanos de Nueva York del trabajo. Ver **Figura 11.**

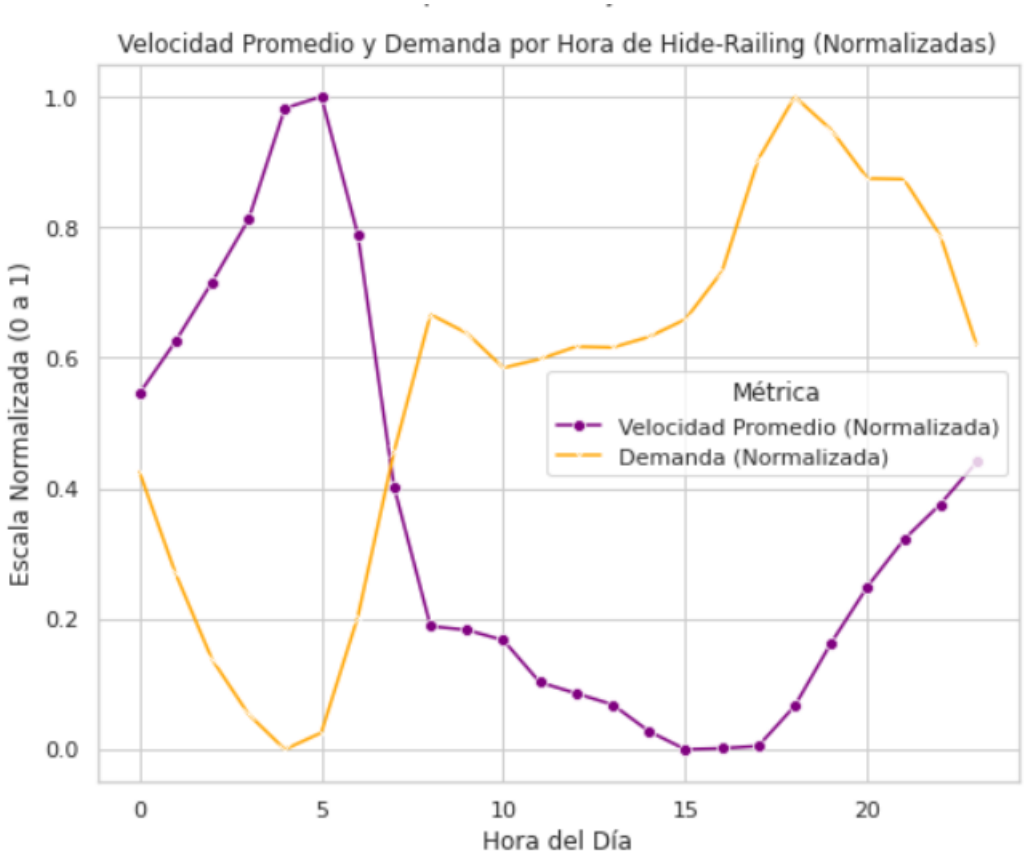
Figura 11. *Mapa de calor de la velocidad promedio por hora del día y día de la semana.*



Fuente: Elaboración propia.

Es interesante el hecho de que el punto de mayor demanda no coincida con los momentos de mayor congestión de la ciudad, que son los momentos de menor velocidad de circulación, por lo que se deduce que podría haber otros factores que podrían estar influyendo en la demanda de *ride-hailing*. Ver **Figura 12.**

Figura 12. *Comparativa de demanda de ride-hailing frente a la velocidad promedio de Manhattan.*



Fuente: Elaboración propia.

El coeficiente de correlación de Pearson ($r_{p)}$ es una medida estadística que evalúa la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. Este coeficiente tiene un rango que va de -1 a 1, donde:

- 1 indica una correlación lineal positiva perfecta, es decir, a medida que una variable aumenta, la otra también aumenta de manera proporcional.
- -1 indica una correlación lineal negativa perfecta, lo que significa que a medida que una variable aumenta, la otra disminuye de manera proporcional.
- 0 indica que no existe una relación lineal entre las dos variables.

El coeficiente de Pearson se calcula utilizando la siguiente fórmula:

$$rp = \frac{\Sigma((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{(\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2}}$$

Donde:

- x_i e y_i son los valores individuales de las variables.
- $\bar{x}$ e $\bar{y}$ son las medias de las variables x e y.

El coeficiente de correlación de Pearson asume que las variables son cuantitativas, tienen una relación lineal, y están distribuidas de manera aproximadamente normal. Aunque mide la relación lineal, no implica causalidad entre las variables (De Winter et al., 2016).

El coeficiente de correlación de Spearman (r_s) es una medida no paramétrica de la relación monótona entre dos variables, calculada a partir de los rangos de los valores de las variables. A diferencia del coeficiente de Pearson, Spearman no evalúa directamente los valores originales de las variables, sino sus rangos, lo que lo hace más robusto frente a distribuciones no normales y valores atípicos. El rango del coeficiente de Spearman también va de -1 a 1, donde:

- 1 indica una relación monótona positiva perfecta (a medida que una variable aumenta, la otra también lo hace consistentemente).
- -1 indica una relación monótona negativa perfecta (a medida que una variable aumenta, la otra disminuye consistentemente).
- 0 indica que no hay una relación monótona entre las dos variables.

El coeficiente de Spearman se calcula con la fórmula:

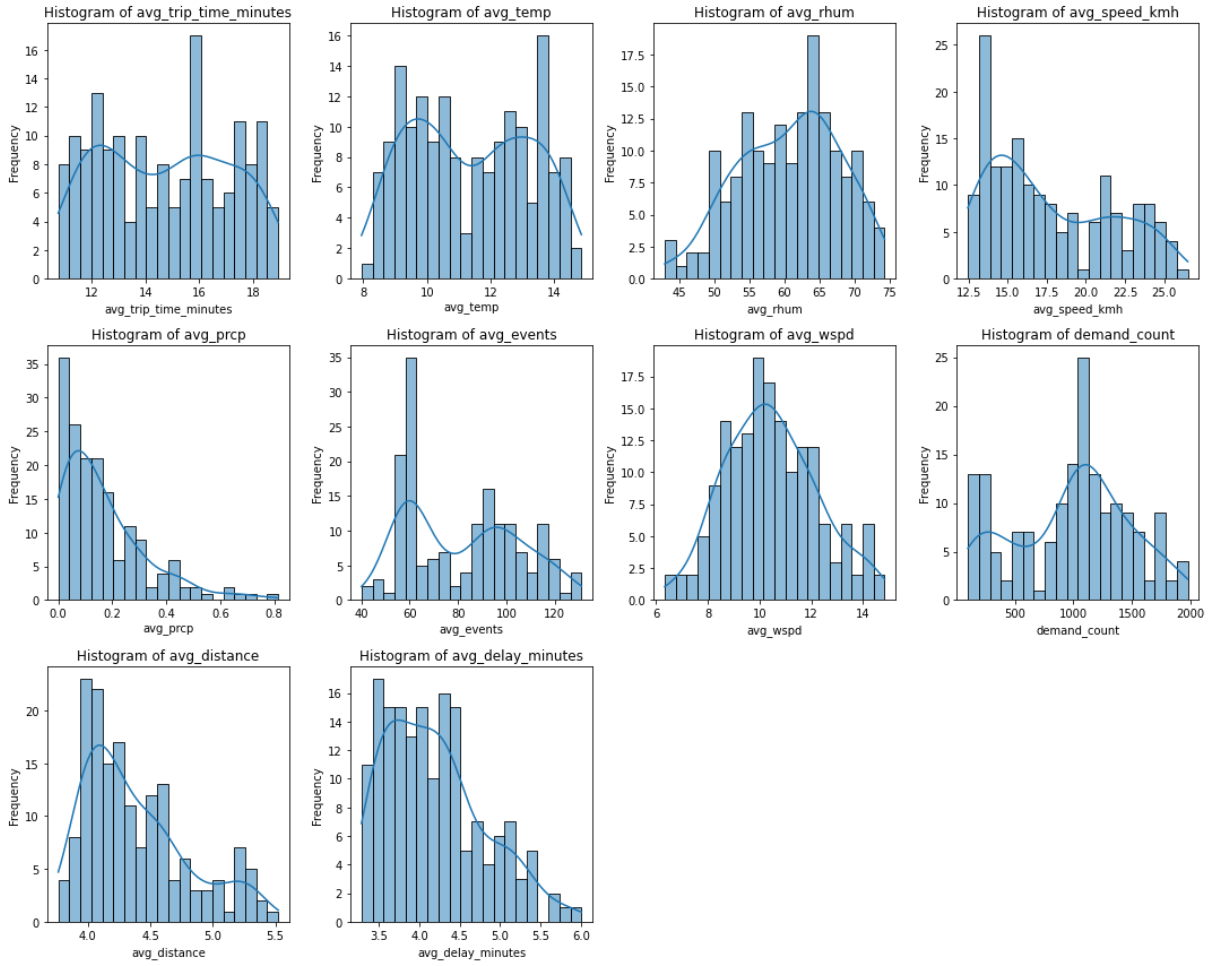
$$\mathrm{rs} = \frac{\Sigma((x_{ri} - (N/2 + 0.5))(y_{ri} - (N/2 + 0.5)))}{\sqrt{(\Sigma(x_{ri} - (N/2 + 0.5))^2 \Sigma(y_{ri} - (N/2 + 0.5))^2}}$$

Donde:

- d_i es la diferencia entre los rangos de las dos variables para el i-ésimo par de observaciones.
- n es el número total de observaciones.

Para poder determinar que método de cálculo del coeficiente de correlación elegir, es pertinente analizar la distribución de las variables que van a ser objeto de estudio (De Winter et al., 2016). Ver **Figura 13**.

Figura 13. *Histograma de cada una de las variables objeto de análisis.*



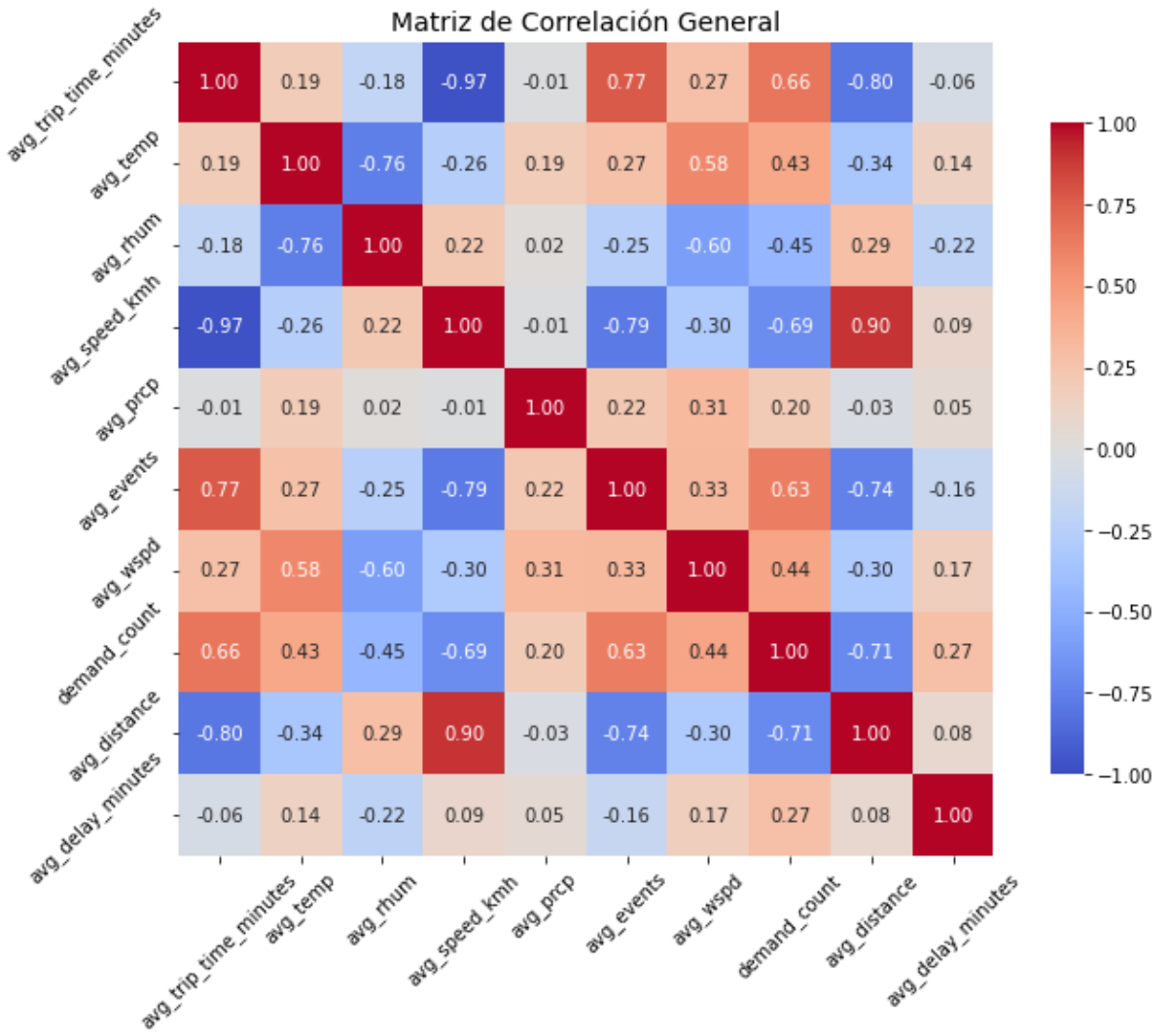
Fuente: Elaboración propia.

Como no todas las distribuciones de las variables se asemejan a una distribución normal, hay valores atípicos en la distribución y, además, se pretenden encontrar relaciones no lineales en los datos, se opta por un enfoque con el método *Spearman* (Reberik et al., 2015).

Durante el análisis de los datos de movilidad y su relación con variables contextuales, se observaron diferencias significativas en los valores de correlación al trabajar con diferentes niveles de agregación. Este problema surge al comparar dos enfoques: uno que agrupa los datos únicamente por variables temporales (hora y día de la semana) y otro que incluye además las coordenadas de inicio y fin de los trayectos (*pickup* y *dropoff*). Estas diferencias reflejan cómo la granularidad de los datos impacta en la variabilidad de las métricas y, por tanto, en las correlaciones calculadas.

En el primer enfoque, los datos se agrupan exclusivamente por hora y día de la semana, lo que elimina las variaciones geoespaciales específicas de cada trayecto. Este nivel de agregación reduce la dispersión de las variables analizadas, como la velocidad promedio (*stddev* = 3.93 km/h) y la precipitación promedio (*stddev* = 0.15 mm). La menor desviación estándar indica que los valores están más centralizados alrededor de la media, lo que facilita la identificación de patrones generales. Por ejemplo, la relación entre la velocidad y la precipitación o entre la demanda y el clima se ve más clara porque se diluyen las diferencias locales que podrían introducir ruido. En este caso, las correlaciones son más altas y consistentes, reflejando un patrón global. Ver **Figura 14.**

Figura 14. *Matriz de correlación con mayor nivel de agregación.*



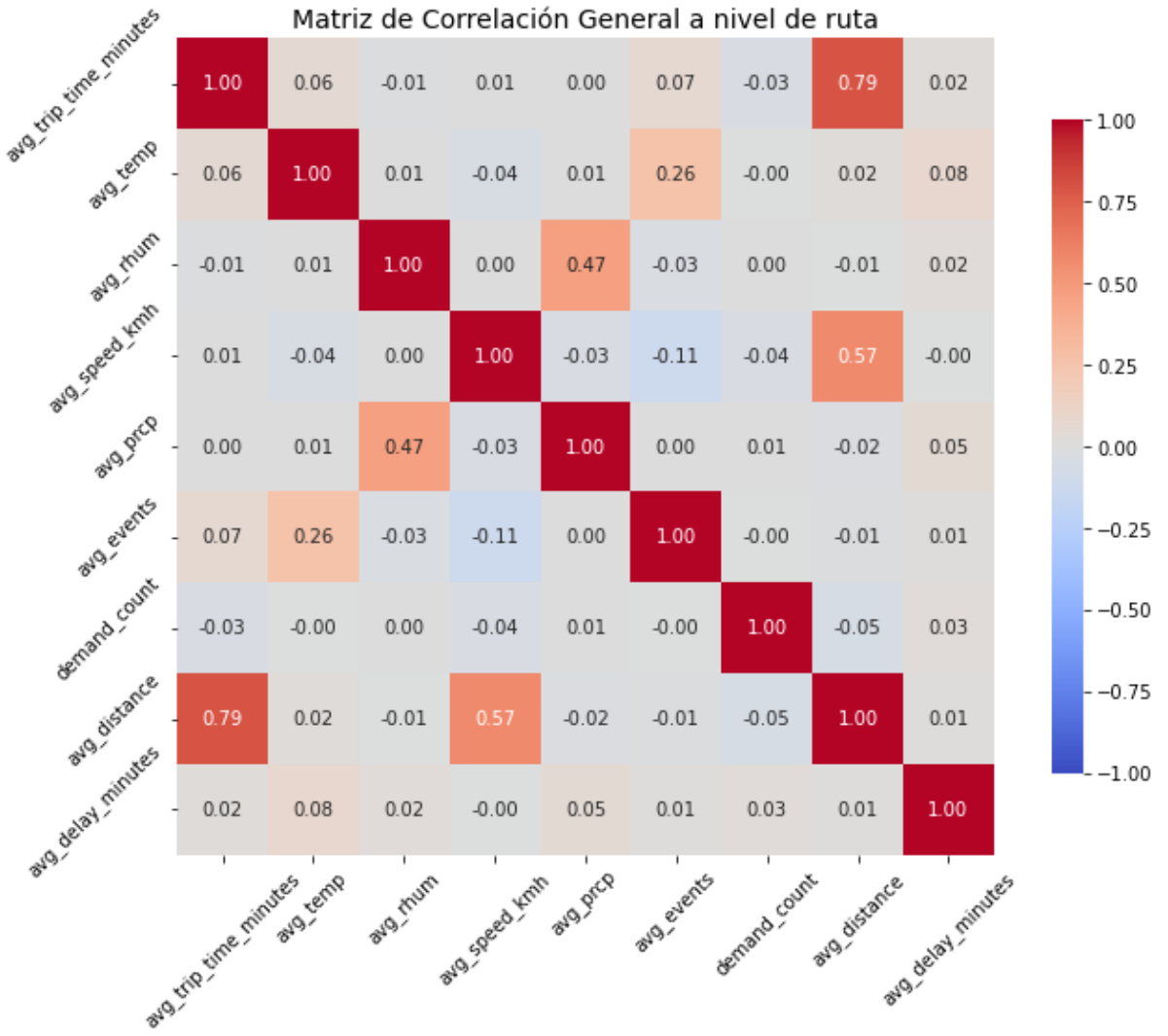
Fuente: Elaboración propia.

La velocidad del tráfico presenta una alta correlación negativa con el número promedio de eventos en la ciudad, lo que sugiere que actividades como conciertos, eventos deportivos o manifestaciones generan un impacto significativo al incrementar la congestión en las vías cercanas. Asimismo, la demanda muestra una correlación negativa elevada, lo que podría sugerir que, o bien un mayor volumen de solicitudes disminuye la velocidad promedio debido a la saturación de las ruta, o el propio tráfico conduce a un incremento de la demanda o que ambas suceden al mismo tiempo. Además, la demanda también se ve influenciada fuertemente por el número de eventos de la ciudad, con una correlación positiva alta. Entre las variables meteorológicas, la velocidad del viento presenta una correlación negativa moderada, reflejando que condiciones ventosas también afectan la fluidez del tráfico, mientras que la temperatura tiene una relación similar, aunque de menor intensidad, lo que implica que su influencia, aunque existente, es menos significativa. Por otro lado, la precipitación no parece ser un factor determinante, ya que su correlación con la velocidad del tráfico es mínima. Estos hallazgos destacan cómo factores como los eventos y la demanda de *ride-hailing* tienen un impacto directo en la movilidad urbana, mientras que las condiciones meteorológicas juegan un papel menos significativo, pero no despreciable.

Es muy interesante el hecho de que la correlación entre la distancia y la velocidad de circulación sea negativa y prácticamente lineal, pues esto puede indicar que existen vías más rápidas dentro del área metropolitana que escapan a la congestión.

En el segundo enfoque, se añaden las coordenadas de inicio y fin, lo que fragmenta los datos en grupos más específicos. Esto aumenta considerablemente la desviación estándar de las variables analizadas (*stddev* = 7.42 km/h para velocidad y *stddev* = 0.82 mm para precipitación). La mayor dispersión refleja la influencia de factores locales, como características específicas de las rutas o condiciones ambientales, que no se capturan en el nivel agregado. Sin embargo, esta granularidad también reduce el tamaño efectivo de los grupos, lo que introduce mayor ruido y disminuye la robustez estadística de las correlaciones. Como resultado, las relaciones entre variables se vuelven más débiles y menos evidentes, ya que los patrones generales se diluyen por las variaciones locales. Ver **Figura 15.**

Figura 15. *Matriz de correlación general eliminando el nivel de agregación.*



Fuente: Elaboración propia.

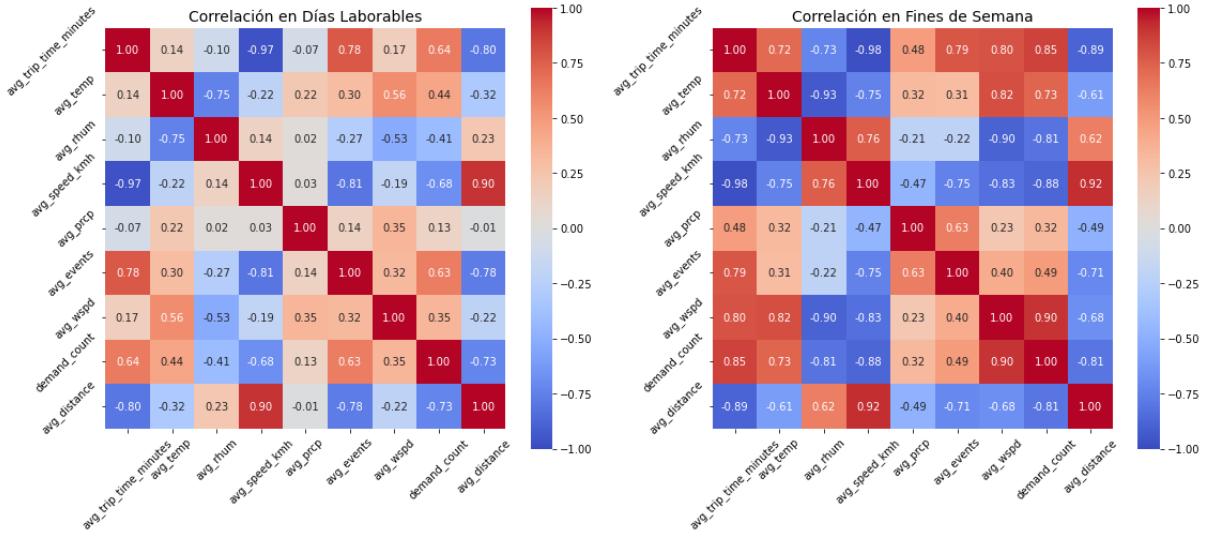
Esta diferencia pone de manifiesto un desafío clave en el análisis de datos de transporte: equilibrar la granularidad necesaria para captar variaciones locales con la simplicidad requerida para identificar patrones globales. Mientras que el nivel agregado es útil para estudios generales y planificación estratégica, el nivel desagregado es más adecuado para optimizar rutas o analizar zonas específicas. Por tanto, la elección del nivel de agregación debe depender del objetivo final del análisis.

La disminución en las correlaciones al desagregar los datos puede explicarse por dos factores principales. Primero, al incrementar el nivel de detalle, se captura una mayor variabilidad dentro de los subconjuntos, cada combinación de ubicación de inicio, fin, hora y día puede

tener comportamientos únicos, lo que refleja variaciones específicas, como rutas particulares o patrones de tráfico localizados, dispersando los valores de las variables y debilitando las correlaciones. Segundo, al desagregar los datos en grupos más pequeños, se reduce la cantidad de datos por grupo, lo que afecta la estabilidad estadística al calcular promedios o relaciones, ya que el tamaño de cada grupo disminuye, diluyendo las correlaciones. Un enfoque posible para mitigar este problema es añadir más precisión georreferenciada a los atributos de las rutas, como información más detallada sobre el tráfico local, las condiciones climáticas o la infraestructura vial, lo que ayudaría a mejorar la calidad y la robustez de las correlaciones a nivel de ruta.

Para profundizar en el análisis de los factores que influyen en la movilidad urbana, se ha generado una matriz de correlación adicional en la que se distingue entre días laborables y no laborables, siguiendo el primer enfoque de agregación. Ver **Figura 16**. Esta diferenciación es crucial, ya que se presupone que los patrones de movilidad varían significativamente entre ambos escenarios debido a las diferencias en actividades diarias, horarios laborales y eventos de ocio. Este enfoque permite identificar cómo las relaciones entre variables como el tiempo de viaje, la velocidad y la demanda se ven afectadas por el contexto temporal.

Figura 16. *Matriz de correlación diferenciando por tipo de día.*



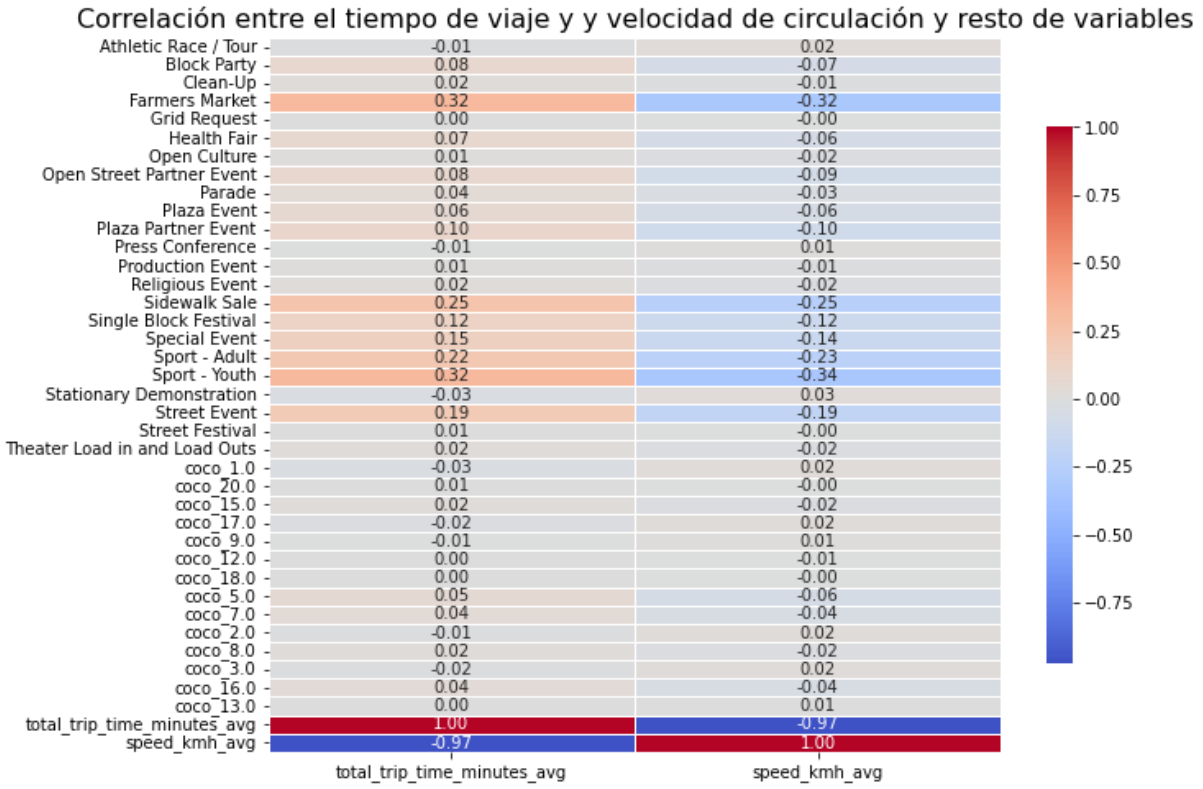
Fuente: Elaboración propia.

Los resultados obtenidos a partir de la matriz de correlación para días laborables y no laborables muestran que las diferencias en los patrones de movilidad, si bien son similares,

parecen ser significativas. La influencia del número de eventos es similar, pero las variables meteorológicas parecen afectar de manera distinta si es fin de semana o si es día laborable, donde las precipitaciones, la velocidad el viento y la temperatura influyen más fuertemente los fines de semana. Estos resultados podrían indicar que es relevante incluir el tipo de día (laborable, fin de semana) en modelos predictores.

Para profundizar en la influencia de eventos y condiciones meteorológicas en la movilidad, se ha creado una matriz de correlación avanzada que evalúa no solo el impacto del número promedio de eventos, sino también cómo cada tipo específico de evento (como conciertos, eventos deportivos o manifestaciones) y cada condición meteorológica particular (como lluvia, nieve o niebla) afectan las variables clave de movilidad. Ver **Figura 17**. Este análisis permite identificar patrones únicos en la interacción entre estos factores y el comportamiento de la movilidad, proporcionando una visión más granular y precisa de su influencia.

Figura 17. *Matriz de correlación para la evaluación de la influencia de cada tipo de evento y condición meteorológica.*



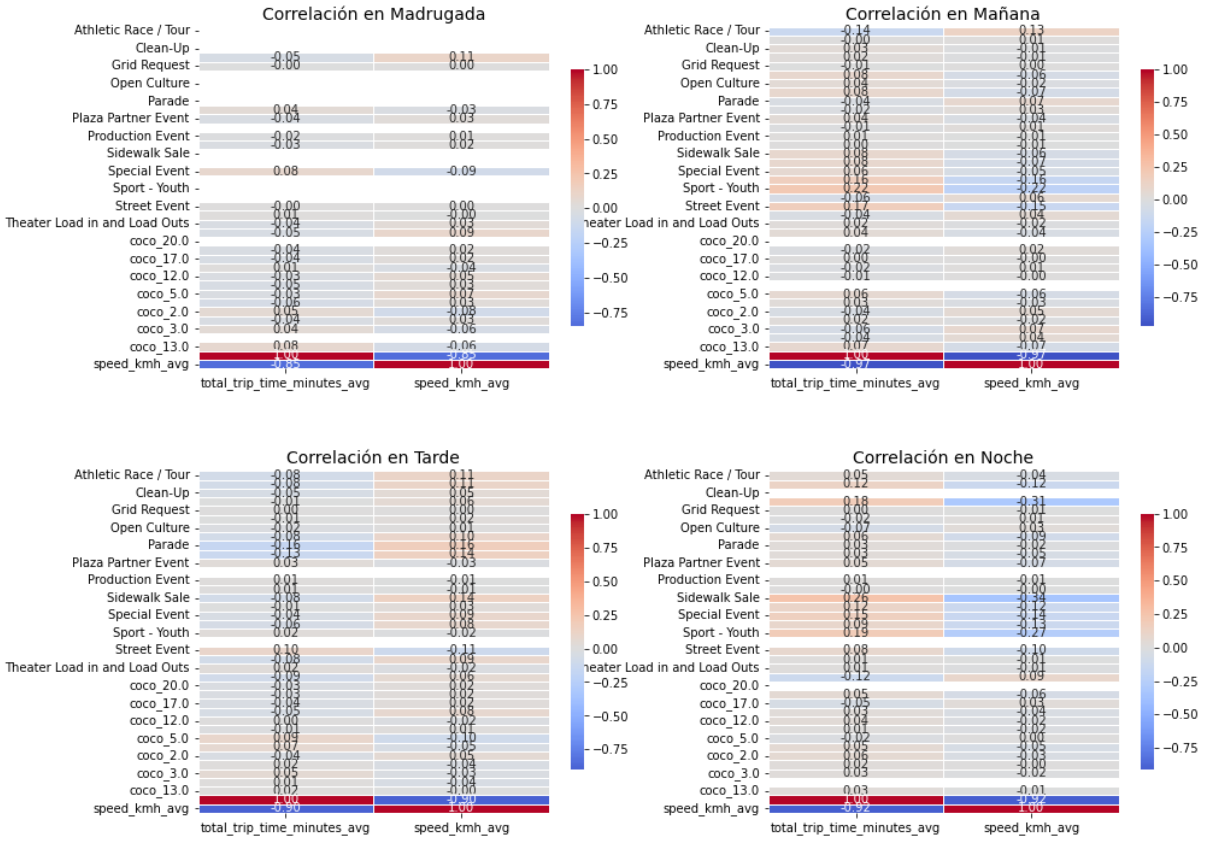
Fuente: Elaboración propia.

Los resultados muestran que, si bien las correlaciones son débiles en general, existen eventos que afectan en mayor medida al tráfico que otros, posiblemente debido a la afluencia de personas que acuden a estos eventos. Es remarcable que las correlaciones en su inmensa mayoría son en negativo frente a la velocidad de circulación, es decir, que la tendencia es que a mayor número de eventos mayor va a ser el tráfico. Del mismo modo, funciona a la inversa con el tiempo de viaje, pues un atributo es combinación lineal del otro. Los atributos que parecen mostrar una mayor influencia en el tráfico son:

- *Farmers Market* (-0.32)
- *Sidewalk Sale* (-0.25)
- *Sport – Youth* (-0.34)

Para ampliar el análisis y ofrecer una visión más detallada de la influencia de los trayectos y las diferentes franjas horarias en la movilidad, se ha creado una nueva matriz de correlación que las distintas partes del día: mañana (06H-11H), tarde (12H-17H), noche (18H-23H), y madrugada (00H-05H). Esta matriz permite evaluar cómo las variables de movilidad se ven afectadas de manera distinta según la duración del trayecto y el momento del día, proporcionando una comprensión más precisa de cómo estos factores interactúan y afectan los tiempos de viaje y la congestión del tráfico en distintos escenarios. Ver **Figura 18.**

Figura 18. *Matriz de correlación para la evaluación de la influencia de cada tipo de evento y condición meteorológica en distintos contextos.*



Fuente: Elaboración propia.

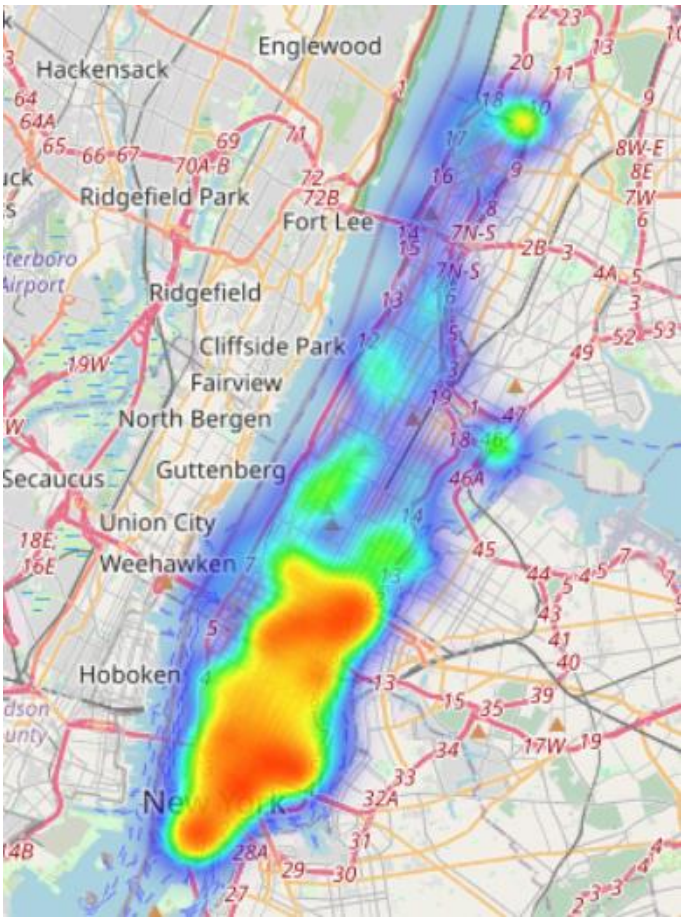
Las variables con mayor correlación con el tráfico, es decir, con la velocidad de circulación promedio (*speed_kmh*) son los eventos callejeros, los mercados y algunos eventos de producción. La figura confirma que la magnitud de estas correlaciones varía según el momento del día, siendo la mañana y la noche los períodos con mayor impacto. Es interesante el hecho de que en función del momento del día algunos eventos se correlacionen positivamente con el tráfico (por ejemplo, por la tarde). Asimismo, la distancia recorrida también muestra comportamientos diferentes en lo que a las correlaciones se refiere, siendo los trayectos cortos los más sensibles a estas variables. Estas correlaciones también indican que podría ser interesante la incorporación del momento del día (mañana, tarde, noche y madrugada) como atributo predictor en modelos de predicción de tráfico.

Con el objetivo de identificar patrones y problemáticas clave en la movilidad del área metropolitana de Manhattan, se han desarrollado mapas de calor que analizan dos aspectos

fundamentales: los retrasos en el transporte y la demanda de taxis. Estas visualizaciones permiten comprender mejor las dinámicas urbanas, facilitando la identificación de cuellos de botella en el tráfico y zonas de alta demanda, información crucial para diseñar estrategias que optimicen el transporte en la ciudad.

El análisis de los retrasos en el transporte muestra que las áreas con mayor congestión se concentran en el centro y sur de Manhattan, especialmente en torno a los distritos financieros y comerciales, específicamente, en áreas cercanas al *Central Business District* y las zonas alrededor de *Battery Park*. Otra región con retrasos significativos es la zona central, específicamente *Midtown Manhattan*, que incluye *Times Square*, el área del *Empire State Building* y alrededores. En cuanto al norte se refiere, si bien se presentan retrasos, se aprecia que el volumen de cogestión es mucho menor. Ver **Figura 19.**

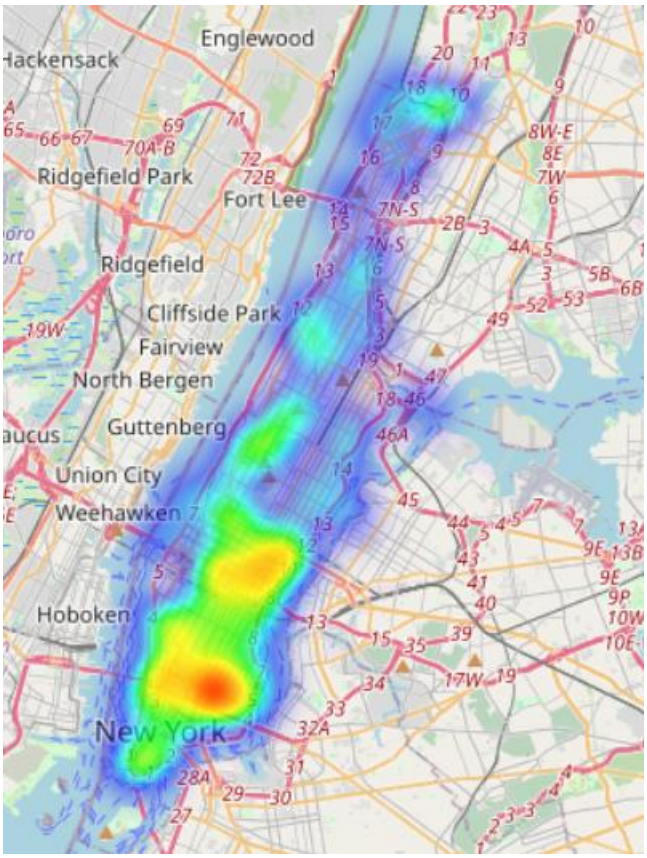
Figura 19. *Mapa de calor con la distribución geográfica de los retrasos en la recogida de ride-hailing.*



Fuente: Elaboración propia.

El análisis de la demanda sugiere correlación geográfica entre ambos factores, pues es la *Central Business District* la que concentra el mayor volumen de la demanda y sucede algo similar con *Midtown Manhattan.* Ver **Figura 20.**

Figura 20. *Mapa de calor con la distribución geográfica de la demanda en la recogida de ride-hailing.*

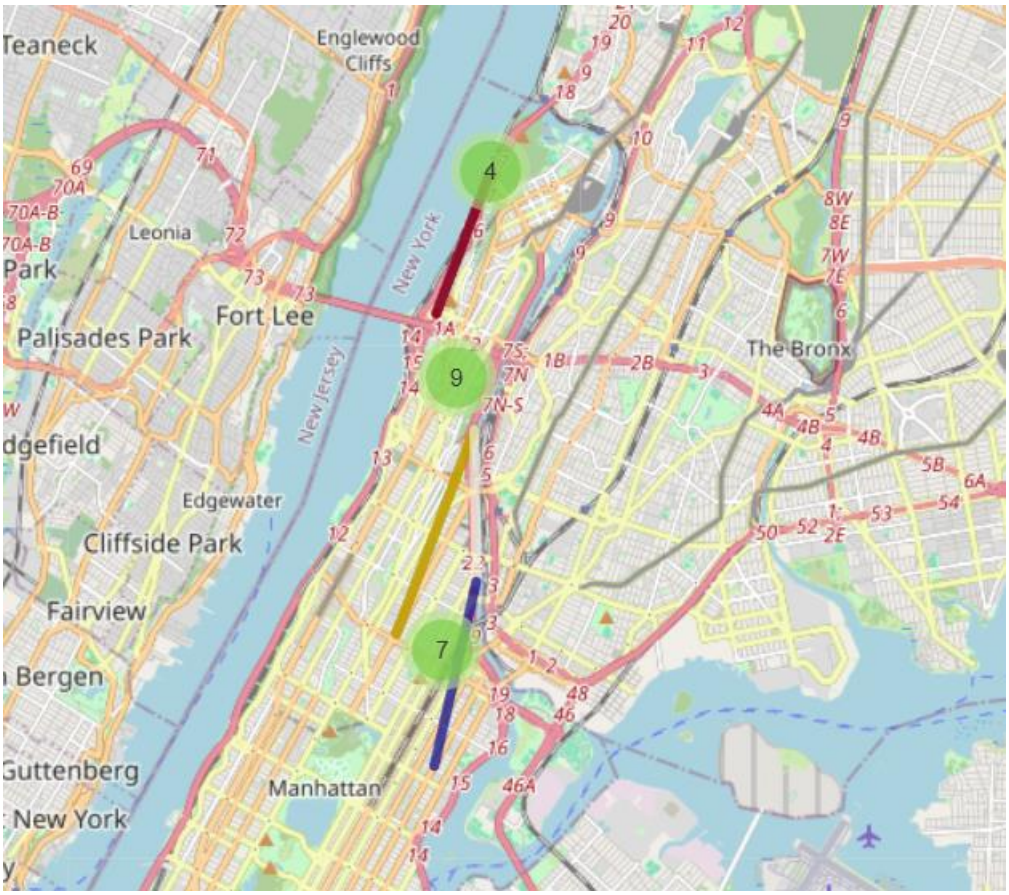


Fuente: Elaboración propia.

La superposición de los mapas de retrasos y demanda permite identificar una correlación clara entre ambos factores: las áreas con mayor demanda de taxis tienden a coincidir con los puntos críticos de congestión.

En la siguiente sección se incluye un mapa que representa las rutas con mayor demanda en el área de estudio, resaltando las conexiones más utilizadas y su ubicación geográfica. Este análisis gráfico permite identificar los patrones de uso y las zonas donde se concentra un flujo elevado de pasajeros, ofreciendo una visualización clara de las dinámicas de movilidad en la región. Ver **Figura 21.**

Figura 21. *Rutas de ride-hailing más demandadas.*



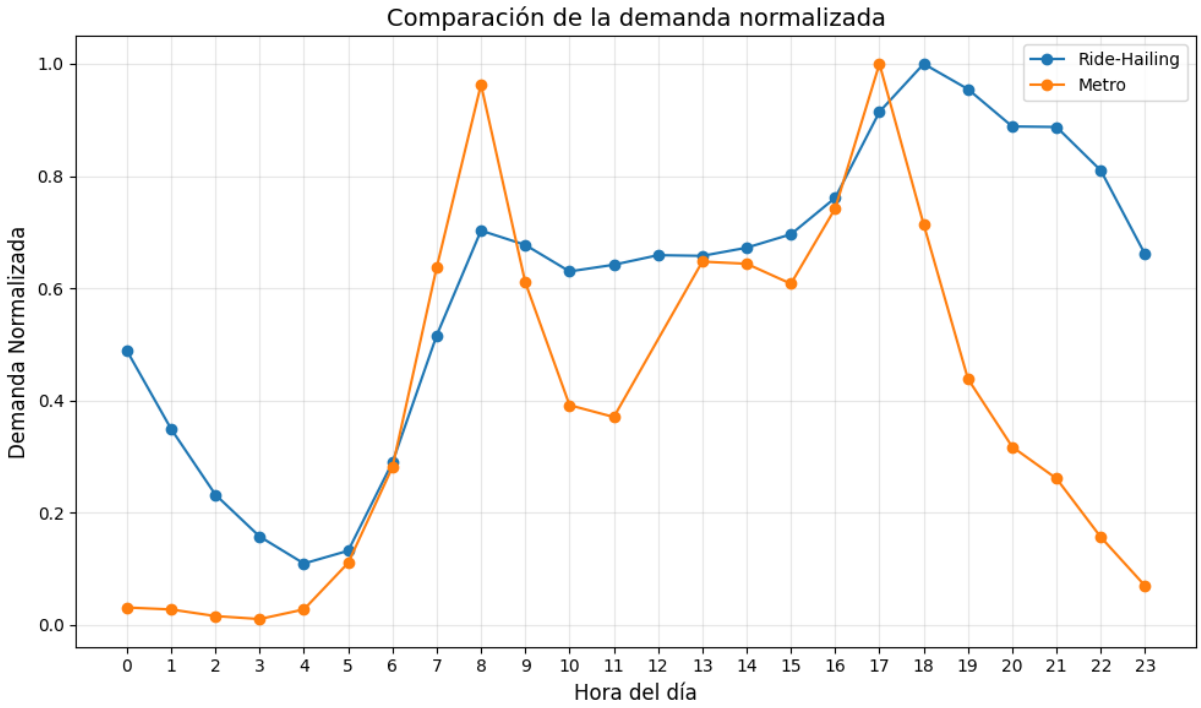
Fuente: Elaboración propia.

El análisis de las rutas más demandadas constata que estas se encuentran principalmente en el norte de Manhattan, lo que sugiere un flujo constante de usuarios. Es interesante este hecho, puesto que esto podría sugerir que hay determinados grupos que podrían utilizar estas rutas de manera constante por falta de cobertura de otros medios de transporte público.

5.2.3.2. Metro de Manhattan

En primer lugar, se ha llevado a cabo un análisis de la influencia de los atributos temporales en la demanda, identificando las horas punta. Los resultados demuestran que el comportamiento de la demanda de metro y *ride-hailing* no funcionan de la misma manera, ya que las horas punta no coinciden. Ver **Figura 22**.

Figura 22. *Comparativa de la evolución de la demanda por hora del día.*



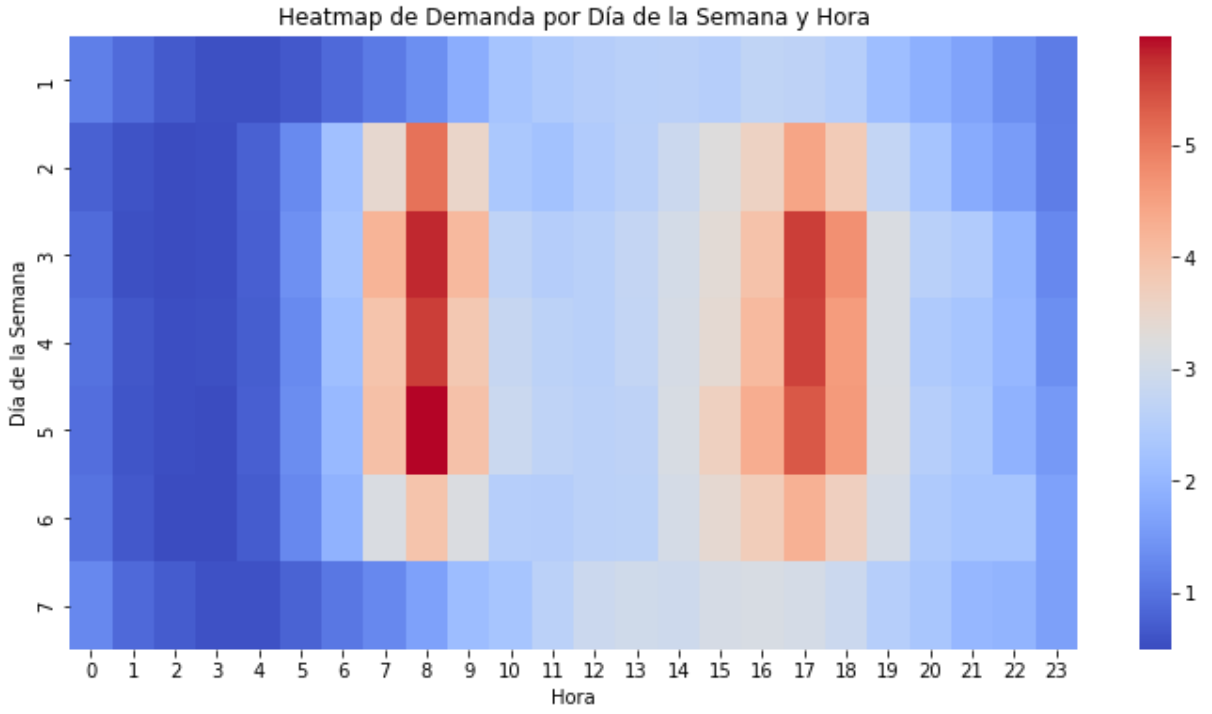
Fuente: Elaboración propia.

La gráfica ilustra que, si bien las horas pico la demanda de movilidad de ambos servicios se comporta de manera similar, los rangos horarios de alta demanda no se comportan de la misma manera, lo que sugiere que el metro no es tan eficiente por la noche como el *ride-hailing.*

Conviene remarcar que, tal y como están configuradas las solicitudes a la *api*, no siempre se disponen de todos los datos para todas las horas en el caso del metro. El problema radica en que, al obtener los datos de la *api* del metro, cada ejecución devuelve un conjunto diferente de registros, lo que genera una variabilidad en los datos disponibles en función de la hora del día. Debido a que la capacidad computacional limita la cantidad de datos que se pueden descargar por ejecución (máximo 3 millones de registros), no siempre se obtiene información completa de todas las horas del día. Además, al estar representada la demanda de todas las rutas posibles dentro de Nueva York, el número de registro por cada hora y día es elevadísimo, reduciendo el número de días y horas disponibles. Esto provoca que ciertos intervalos, no se encuentren siempre disponibles en los conjuntos de datos descargados, resultando en la falta de datos en esos períodos en las visualizaciones y análisis.

Para hacer frente a esta casuística, se trabajó con una muestra de un conjunto de datos previamente descargados de la propia *api*, de forma que había muchos días distintos representados, desde marzo hasta junio. De esta manera, se logró generar un mapa de calor de la demanda con todas las horas representadas y todos los días. Ver **Figura 23**.

Figura 23. *Mapa de calor de la demanda de metro en Manhattan por hora y día de la semana.*



Fuente: Elaboración propia.

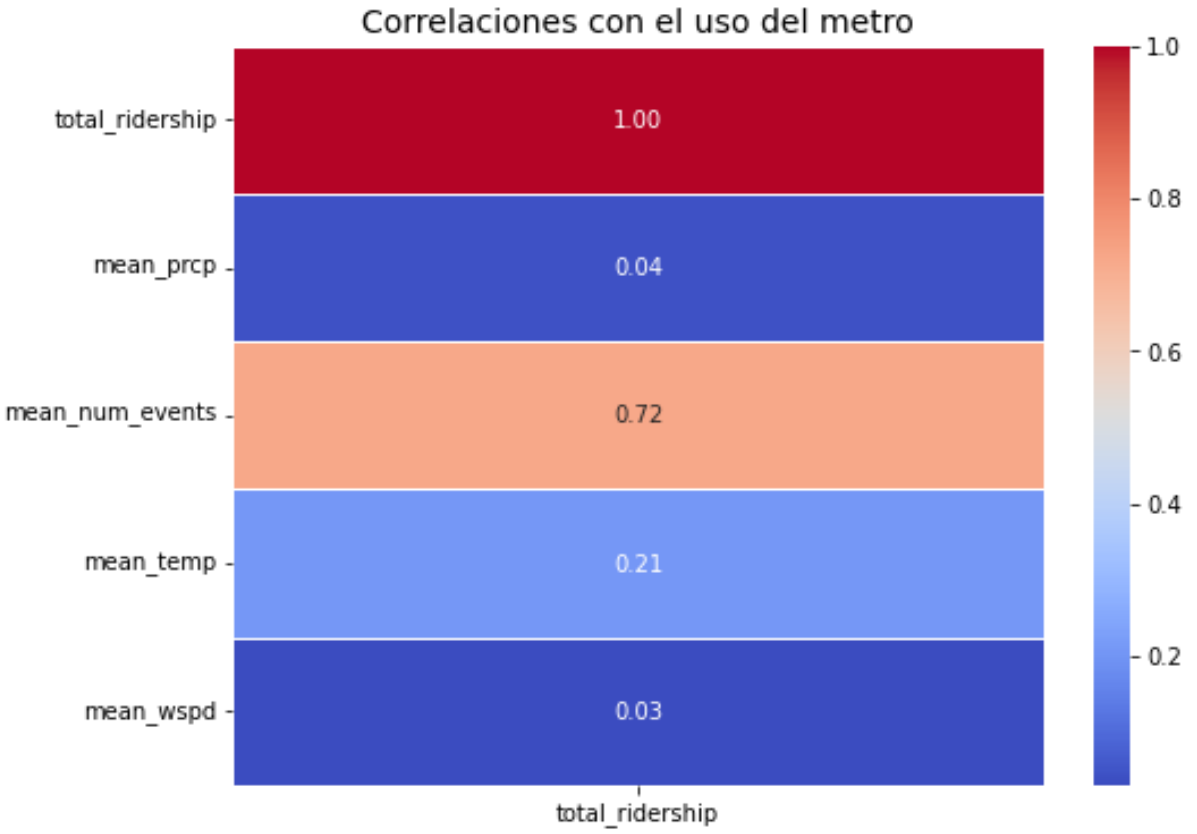
A pesar de la limitación computacional, se observan claramente las horas pico y se identifican los días laborables como los periodos de mayor uso del transporte público.

Al igual que para el caso de *ride-hailing*, se ha llevado a cabo un análisis de la influencia de variables meteorológicas y de eventos en la demanda de metro. Los resultados evidencian que existe correlación entre los eventos y la demanda de metro. Con respecto a las condiciones meteorológicas, su influencia parece ser menor. Ver **Figura 24**. Si bien las tendencias son leves, se evidencia que:

- El número de eventos se correlaciona positivamente (0.72) con el uso del metro.
- Las precipitaciones y la velocidad del viento prácticamente no muestran correlación. (<0.1)

- La temperatura se correlaciona positivamente (0.21) con la demanda.

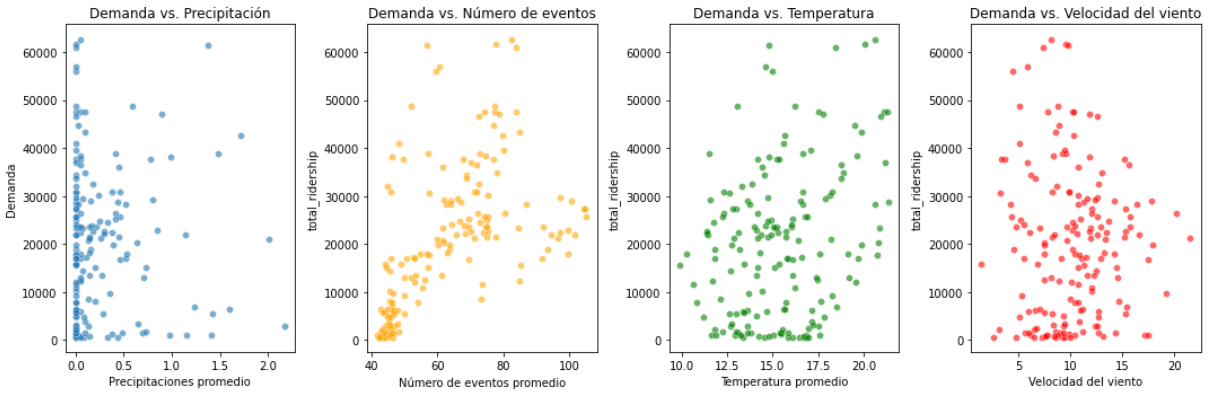
Figura 24. *Mapa de correlación de los atributos meteorológicos y contextuales.*



Fuente: Elaboración propia.

Estas correlaciones se pueden visualizar más claramente en un gráfico de dispersión. Ver **Figura 25**.

Figura 25. *Gráficos de dispersión de los atributos meteorológicos y contextuales.*



Fuente: Elaboración propia.

Los gráficos de dispersión muestran que existe una correlación positiva evidente con número de eventos, mientras que las condiciones climatológicas no parecen mostrar un patrón claro de correlación.

5.3 Modelado de los datos

Para abordar el problema de predicción de categorías de tráfico (fluido o denso), se llevaron a cabo experimentos utilizando tanto redes neuronales de grafos como el modelo *Random Forest*, variando los atributos de entrada para evaluar su impacto en la precisión del modelo.

Se optó por una GNN debido a su capacidad para modelar relaciones complejas y dinámicas entre nodos (puntos de recogida y destino) y para integrar información contextual de las aristas (como hora del día, eventos y condiciones meteorológicas). A diferencia de los modelos tradicionales, las GNNs son particularmente efectivas para capturar patrones espaciales y temporales en estructuras de grafo, donde las conexiones entre los nodos son esenciales para comprender el comportamiento del sistema. La arquitectura de la red desarrollada en este trabajo permite la propagación de información a través del grafo, aprovechando tanto los atributos locales como las dependencias globales, lo que resulta en predicciones más precisas y una mayor capacidad para abordar la complejidad inherente a los sistemas urbanos de transporte.

En primer lugar, se calculó una categoría de tráfico binaria basada en si la velocidad promedio de un trayecto estaba por encima o por debajo de la media histórica para ese trayecto específico. Este atributo es el que finalmente será predicho por el modelo.

De forma general, para construir el grafo se parte de un *dataframe* con información de rutas, eventos y características de tráfico. Primero, se identifican todas las ubicaciones únicas (nodos) y se asigna un índice único a cada una. Luego, se generan las aristas (*edges*) a partir de las relaciones entre las ubicaciones de origen y destino, formando pares de índices que se convierten en un tensor con *PyTorch Geometric*. Las características de los nodos, como hora, día, eventos y demanda, se escalan utilizando *StandardScaler*. Finalmente, se combinan los nodos, aristas y etiquetas de tráfico en el objeto *Data*, que representa el grafo y se divide en conjuntos de entrenamiento y prueba.

En una primera configuración experimental, los datos del grafo se procesaron para obtener un conjunto inicial de características, incluyendo atributos temporales como la hora del día y el día de la semana, atributos espaciales (ID's de recogida y de destino) y número de eventos por cada tipo específico. Se utilizaron los datos de cada evento específico y cada condición meteorológica específica (*coco_*), al entender que eso arrojaría mayor precisión y una información más detallada al modelo. Para las GNN, las ubicaciones de origen y destino se mapearon como nodos, y las aristas representaron los trayectos entre ubicaciones, mientras que las características del nodo incluyeron el resto de los atributos disponibles. El modelo GNN, implementado con *PyTorch Geometric*, se entrenó para clasificar el tráfico con base en estas características, utilizando una arquitectura de dos capas de convolución gráfica y funciones de activación *ReLU*. Por último, conviene indicar que se dividió el conjunto de datos en dos distintos: entrenamiento (80% de registros sobre el total) y prueba (20% de datos restantes).

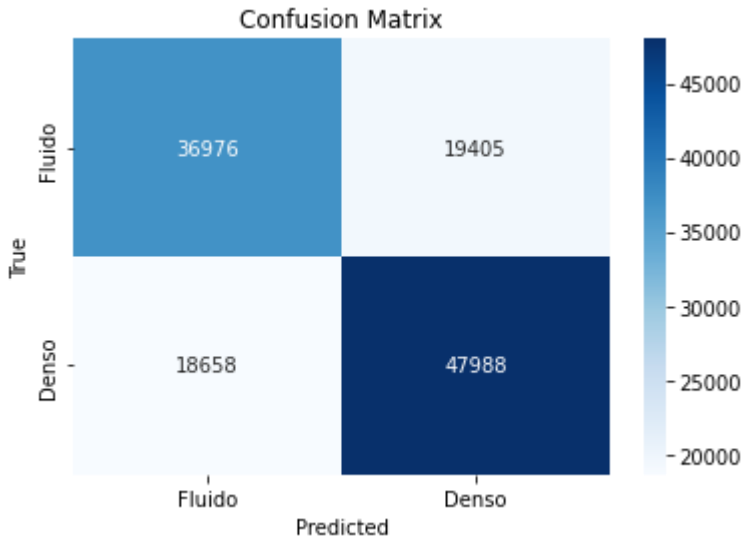
Simultáneamente, se entrenó un modelo *Random Forest* utilizando el mismo conjunto de características. Este modelo no considera explícitamente las relaciones topológicas del grafo, pero permite analizar la importancia de las variables en la predicción. Ambas configuraciones mostraron buenos resultados en términos de precisión y capacidad para diferenciar entre tráfico fluido y denso, destacando la importancia de los atributos relacionados con eventos en el rendimiento del modelo.

En un segundo enfoque, tanto para la red neuronal como para el *Random Forest*, se redujo el conjunto de atributos, promediando los valores de eventos totales en lugar de usar eventos individuales y tomando solo una serie de atributos meteorológicos como la temperatura y las precipitaciones. Este conjunto simplificado permitió evaluar si era posible obtener resultados comparables con un modelo más simple y con menor carga computacional.

5.4 Evaluación de los modelos

A continuación, se muestra la matriz de confusión obtenida tras las predicciones realizadas por la red neuronal con el primer enfoque. Ver **Figura 26**.

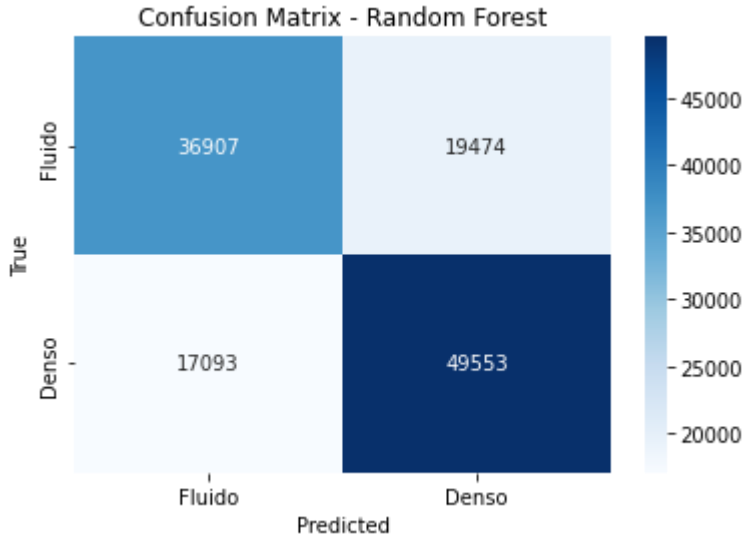
Figura 26. *Matriz de confusión GNN, primer enfoque.*



Fuente: Elaboración propia.

A continuación, se muestra la matriz de confusión obtenida tras las predicciones realizadas por el *Random Forest* para el primer enfoque. Ver **Figura 27**.

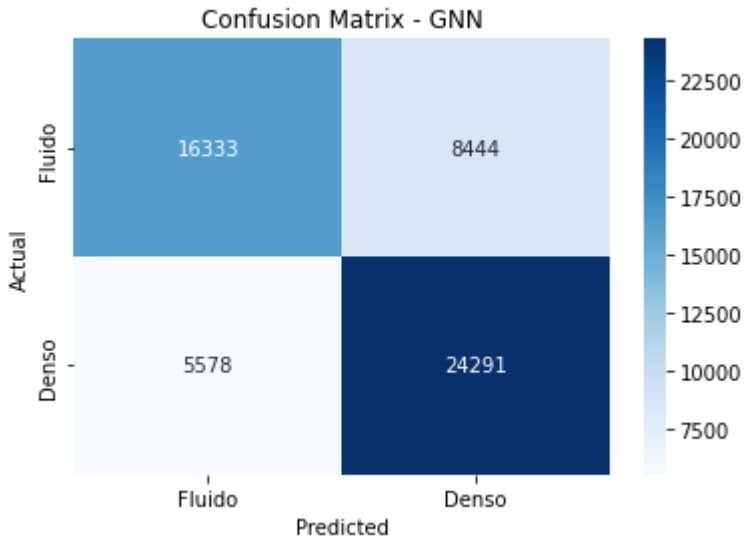
Figura 27. *Matriz de confusión Random Forest, primer enfoque.*



Fuente: Elaboración propia.

A continuación, se muestra la matriz de confusión obtenida tras las predicciones realizadas por la red neuronal para el segundo enfoque. Ver **Figura 28**.

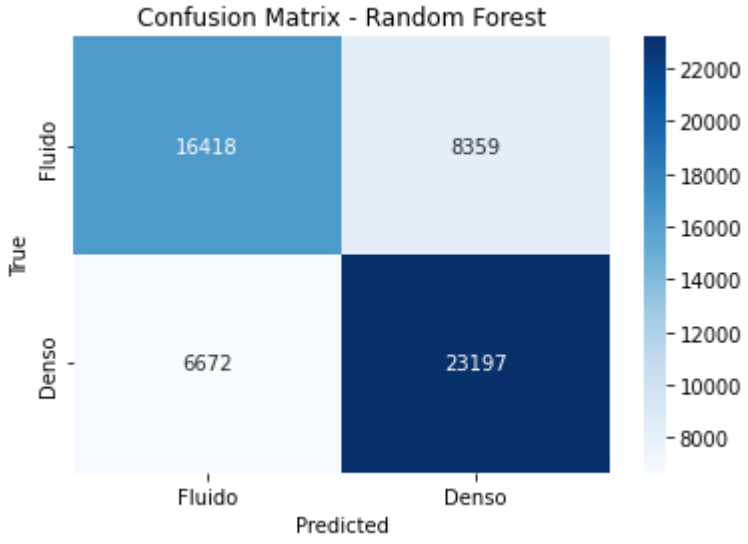
Figura 28. *Matriz de confusión GNN, segundo enfoque.*



Fuente: Elaboración propia.

A continuación, se muestra la matriz de confusión obtenida tras las predicciones realizadas por el *Random Forest* para el segundo enfoque. Ver **Figura 29**.

Figura 29. *Matriz de confusión Random Forest, segundo enfoque.*

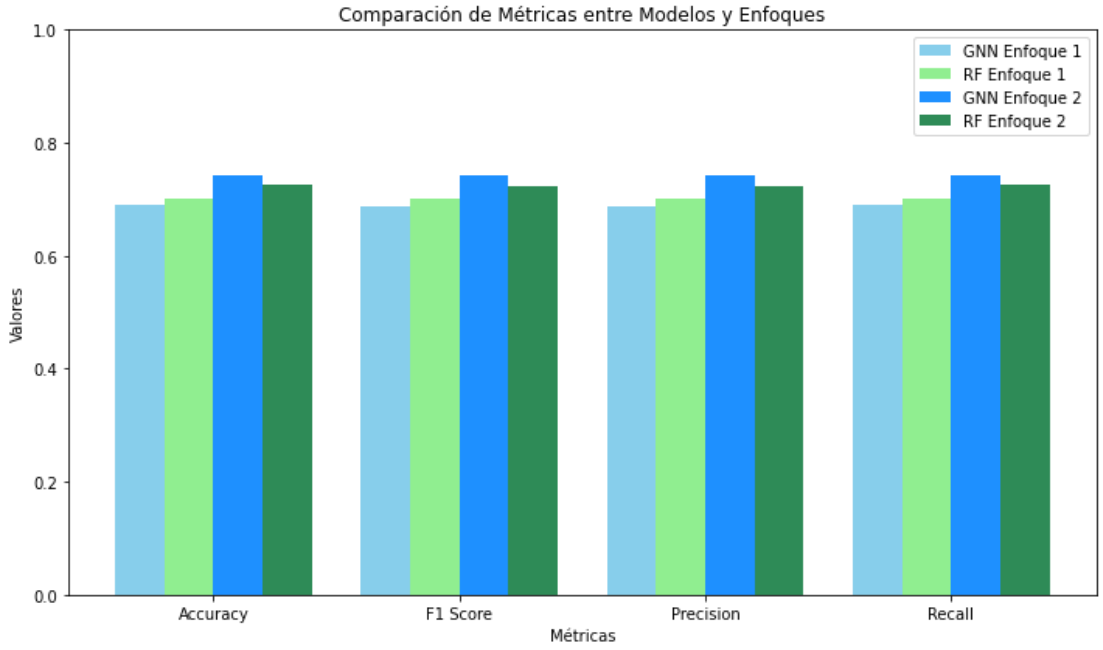


Fuente: Elaboración propia.

Los resultados muestran que, si bien los modelos se comportan de manera similar para los distintos modelos y enfoques, la simplificación del modelo para ambos modelos resulta positivo, es decir, la reducción de atributos. Esto puede indicar que la introducción de muchos

atributos podría estar introduciendo ruido en los modelos, reduciendo su capacidad. Asimismo, la red neuronal basada en grafos en el segundo enfoque (menos atributos) presenta el mejor comportamiento para clasificar ambas instancias. Ver **Figura 30**.

Figura 30. *Comparativa de métricas entre modelos y enfoques.*



Fuente: Elaboración propia.

5.2.4. Herramientas utilizadas

En el desarrollo de este proyecto se ha empleado un conjunto diverso y avanzado de herramientas y librerías que han sido esenciales para abordar todas las fases del proyecto, desde la ingestión y preprocesamiento de datos hasta la modelización y visualización de resultados.

Para la manipulación y el procesamiento de datos, se utilizó *PySpark*, que resultó clave para trabajar con grandes volúmenes de información de manera eficiente gracias a su capacidad de procesamiento distribuido. Módulos como *SparkSession*, *functions* y *types* permitieron la creación de esquemas, transformaciones de datos, operaciones SQL y la optimización de cálculos en entornos distribuidos. Además, se complementó con *Pandas* y *Numpy* para realizar análisis rápidos en *datasets* de tamaño moderado y manejar estructuras de datos tabulares y matriciales. Asimismo, se empleó la librería *Meteostat* para obtener datos meteorológicos históricos, que fueron integrados como variables relevantes en el análisis.

En cuanto a la visualización de datos, las librerías *Matplotlib* y *Seaborn* desempeñaron un papel fundamental en la creación de gráficos descriptivos y visualizaciones avanzadas, ayudando a identificar patrones y relaciones en los datos, así como a comunicar los resultados obtenidos. Por otro lado, *Folium*, junto con sus complementos, se utilizó para la representación geoespacial, permitiendo crear mapas interactivos, *clusters* de marcadores y mapas de calor que facilitaron la exploración visual de patrones de movilidad urbana.

Para las tareas de *machine learning* y *deep learning*, se recurrió a *Scikit-learn*, que proporcionó herramientas para el preprocesamiento de datos, como la estandarización mediante *StandardScaler*, y para la implementación de algoritmos supervisados, como el *Random Forest*. También se utilizó en la evaluación de los modelos mediante métricas como la precisión, la matriz de confusión y los puntajes F1. En el ámbito del aprendizaje profundo, *PyTorch* y su extensión *PyTorch Geometric* resultaron fundamentales para la creación y entrenamiento de modelos basados en redes neuronales. En particular, estas herramientas facilitaron el trabajo con datos estructurados en forma de grafos, optimizando los modelos de aprendizaje profundo para abordar problemas específicos del proyecto.

Además de estas librerías, se emplearon herramientas auxiliares como *Requests*, para la obtención de datos a través de peticiones a *apis*, y *Glob*, que facilitó la localización y gestión de múltiples archivos en el sistema de ficheros. También se utilizaron utilidades nativas de *Python*, como *functools* y *collections*, para crear funciones personalizadas y manejar estructuras de datos de manera eficiente.

Finalmente, en el ámbito de la infraestructura y almacenamiento de datos, se aprovechó *PySpark StorageLevel* para optimizar la persistencia y el uso eficiente de recursos durante el procesamiento distribuido, así como *Torch Geometric Data*, que permitió gestionar datos en formato de grafos, aspecto esencial para los modelos de optimización de redes de transporte.

En conjunto, estas herramientas y librerías han permitido abordar un análisis robusto y escalable, integrando técnicas de Big Data, aprendizaje automático y visualización en un flujo de trabajo cohesivo y eficaz.

6. Código fuente y datos analizados

6.1. Código fuente

El código fuente realizado en este trabajo está disponible en: puertanaliza/TFM_Op_NYC

6.2. Datos Analizados

Los datos analizados están disponibles en: Optimización de redes de transporte urbano NYC

Asimismo, en estos enlaces se puede obtener el URL API para los conjuntos de datos de eventos y demanda de metro:

MTA Subway Origin-Destination Ridership Estimate: 2024 | State of New York

NYC Permitted Event Information - Historical | NYC Open Data

7. Conclusiones

En el presente trabajo se han analizado diversos factores que afectan al tráfico y la movilidad en Manhattan, proporcionando una visión integral sobre la influencia de las variables temporales, espaciales y contextuales en la congestión y la demanda de transporte. Los resultados obtenidos permiten identificar patrones relevantes y abrir líneas futuras para el análisis y la optimización de la movilidad urbana.

El análisis confirma que el tráfico y la congestión dependen tanto de variables temporales, como el día de la semana y la hora, como de variables espaciales relacionadas con la georreferenciación de las ubicaciones. Estas variables determinan los niveles de congestión, que se concentran de forma más intensa en el sur y centro de Manhattan, siendo el sur la zona más afectada. Además, se ha observado cómo el comportamiento del tráfico y las variables que lo afectan, varían significativamente según el momento del día (mañana, tarde, noche o madrugada) y según la distancia de los trayectos, mostrando dinámicas diferenciadas entre desplazamientos largos y cortos.

Respecto a las condiciones meteorológicas, su influencia sobre la movilidad no parece ser tan relevante, ya que muestran una correlación más débil con las velocidades de viaje. Sin embargo, el estudio evidencia que las condiciones positivas, como la temperatura, tienden a aumentar ligeramente la velocidad de circulación, mientras que las condiciones adversas tienden a disminuirla (velocidad del viento, precipitaciones). Además, la influencia de estas variables es mayor durante los fines de semana que durante los días laborables. Por otro lado, el número y tipo de eventos en la ciudad presentan una mayor influencia sobre el tráfico en comparación con las condiciones meteorológicas. El análisis demuestra que, aunque no todos los eventos afectan de la misma manera (posiblemente debido a la capacidad de congregación de público de cada tipo de evento), a mayor número de eventos, menor es la velocidad de circulación. Finalmente, se ha observado que la demanda de *ride-hailing* esta correlacionada negativamente con el tráfico, es decir, la tendencia es que, a mayor demanda de taxis, menor es la velocidad de circulación.

Otro hallazgo significativo es que las rutas más demandadas no se concentran necesariamente en las zonas con mayor congestión. Esto sugiere que, en dichas zonas, el transporte público

podría no ser tan eficiente o que existen otras variables no consideradas, como la percepción de seguridad, que influyen en esta demanda.

Por último, se destaca que los momentos de mayor demanda de servicios de *ride-hailing* sucede durante las noches de los fines de semana (entre las 19:00 horas y las 23:00 horas). Es remarcable que este hecho no coincide con los momentos de mayor congestión y las horas punta. Este fenómeno podría estar relacionado con factores como la comodidad o la seguridad percibida. Además, esta hipótesis coincide con los resultados de la demanda de metro obtenidos, donde se refleja que la demanda de metro disminuye significativamente los fines de semana por la noche, por lo que se evidencia que la eficiencia del transporte público en esa franja horaria disminuye, hecho que inclina a los usuarios a demandar en mayor medida los servicios de *ride-hailing.*

De igual manera, los resultados del análisis de uso de metro que se ha realizado apuntan a conclusiones similares. Se evidencia que la demanda de metro tiene un comportamiento influenciado por las variables temporales, siendo los días laborables al inicio y final de la jornada de trabajo las horas pico de demanda. Asimismo, se ha demostrado que el número de eventos promedio tiene una alta correlación con el uso del metro, de lo que se deduce que esta variable ha de ser considerada en la planificación del transporte. Por otro lado, las precipitaciones y la velocidad del viento no parecen tener la misma influencia en el uso del metro que en el transporte privado, siendo sus correlaciones muy bajas. En cualquier caso, para obtener unos resultados más representativos en cuanto a la influencia de estas variables, se debe aumentar la capacidad computacional, de forma que se recojan un número elevado de fenómenos extremos que pudieran influir en la movilidad.

Se ha diseñado una red neuronal basada en grafos capaz de predecir el estado del tráfico, clasificándolo como denso o fluido a partir de variables temporales, espaciales y contextuales (la influencia de condiciones meteorológicas y eventos locales), con una efectividad del 75% en métricas como *accuracy*, *recall*, *precisión* y *F1-score*. Esta herramienta presenta un gran potencial para optimizar las redes de transporte, ya que permite prever patrones de congestión y accionar posibles palancas de cambio para evitarlo, contribuyendo así con el medio ambiente y el ahorro de tiempo de los ciudadanos de Nueva York.

Los resultados de los modelos sugieren que, en el contexto del transporte de *ride-hailing*, las GNN superan a los *Random Forest* en la tarea de clasificar el tráfico como denso o fluido,

porque son capaces de modelar mejor las relaciones complejas entre puntos de recogida y destino dentro de una red. Estas conexiones pueden incluir patrones de tráfico, interacciones entre rutas y la influencia de un punto sobre otro. Los *Random Forest*, en cambio, analizan los datos de forma aislada, sin considerar estas dependencias. Por ello, las GNN son más precisas al capturar la dinámica del tráfico y ofrecer una clasificación más precisa y equilibrada en escenarios complejos.

Estos experimentos ilustran cómo la selección de atributos y el modelo utilizado impactan directamente en el rendimiento de las predicciones. La capacidad de evaluar diferentes configuraciones permite optimizar el equilibrio entre precisión y eficiencia computacional en contextos prácticos.

Estos modelos pueden resultar de utilidad a la hora de predecir la movilidad de la ciudad de Nueva York, ayudando a las administraciones públicas a planificar el transporte urbano en función del tráfico y la demanda.

8. Limitaciones y prospectiva

8.1. Limitaciones

En el desarrollo de este trabajo se han identificado varias limitaciones que afectan a la precisión y el alcance de los resultados obtenidos. Estas limitaciones están relacionadas con la naturaleza de los datos utilizados, así como con restricciones técnicas y de capacidad computacional que influyen en la profundidad del análisis y las comparativas realizadas.

En este trabajo se ha empleado información basada en ubicaciones zonales aproximadas, lo que implica que las áreas geográficas utilizadas no necesariamente representan las coordenadas exactas de origen y destino de los viajes. Esta aproximación, aunque útil para un análisis general, reduce la precisión de los resultados, ya que no se consideran las variaciones específicas dentro de cada zona, lo que podría influir en la caracterización real de la demanda y el comportamiento del transporte. Las condiciones meteorológicas y los eventos especiales no han sido georreferenciados con precisión para cada ubicación dentro de Manhattan. En su lugar, se ha tomado un único punto de referencia para Manhattan debido a la falta de atributos específicos en el conjunto de datos de eventos. Esta limitación disminuye la precisión de los resultados y la significancia del análisis estadístico a nivel de ruta, ya que no se consideran las posibles variaciones locales en las condiciones climáticas o la distribución geográfica de los eventos, las cuales podrían tener un impacto en la movilidad y la demanda.

Debido a restricciones de capacidad computacional, se ha trabajado con volúmenes de datos más pequeños comparado con el total disponible, pues se ha utilizado *Azure Databricks Community Edition*, una herramienta con limitaciones importantes en comparación con versiones de mayor capacidad. Además, a consecuencia de esta falta de capacidad, se debe tener en cuenta que siempre se ha trabajado con registros de viajes desde y hacia Manhattan, es decir, no se han contemplado los viajes con origen o destino fuera de los límites del área metropolitana de Manhattan, tanto para los datos de *ride-hailing* como para los de metro. Estas restricciones han impedido el análisis de conjuntos de datos más grandes, lo que limita el alcance del estudio y el potencial de los resultados obtenidos, particularmente en términos de robustez y representatividad.

Aunque los dos conjuntos de datos principales utilizados en este estudio son similares, existen diferencias notables en la cantidad y tipo de atributos que contienen. En particular, el conjunto de datos de metro tiene una menor dimensión de atributos en comparación con el *dataframe* de la TLC, lo que restringe la posibilidad de análisis en aspectos como los tiempos de viaje. Esto resulta en una comparativa más limitada, centrada únicamente en la demanda y sin considerar otros factores clave como los tiempos asociados a los trayectos.

8.2. Trabajo futuro

De cara a futuros trabajos, se proponen diversas líneas de investigación y desarrollo que abordan las limitaciones identificadas en este estudio. Estas líneas están orientadas a mejorar la calidad de los datos, fortalecer las capacidades analíticas y fomentar una integración más robusta entre los distintos modos de transporte urbano para ampliar el alcance y precisión de los análisis.

Sería valioso integrar información meteorológica y eventos georreferenciados en todo Manhattan para considerar las variaciones locales que impactan en la movilidad. Bases de datos como *NOAA National Weather Service* (NOAA, 2024), para condiciones climáticas, y acceder mediante *API* a plataformas de gestión de eventos, como *Eventbrite* (Eventbrite, 2024), e incorporar información del calendario de la ciudad (festividades) pueden proporcionar esta información con mayor detalle espacial y temporal (Belhadi et al., 2020). Estas incorporaciones mejorarían la capacidad de modelar cómo los factores externos, como el clima o los eventos, afectan la congestión y la demanda.

Desde el punto de vista técnico, una mejora crucial sería superar las limitaciones computacionales mediante el uso de plataformas con mayor capacidad, como versiones avanzadas de *Azure Databricks, Google BigQuery*, o servicios en la nube de *Amazon Web Services* (AWS) con infraestructura escalable. Estas herramientas permitirían manejar volúmenes de datos significativamente mayores, aumentando la representatividad de los resultados y permitiendo realizar simulaciones más precisas y complejas.

En cuanto a los modelos de inteligencia artificial, los futuros pasos deberían ir en la línea del incremento de los datos de entrenamiento, aumentando como bien se ha comentado previamente la capacidad computacional. Asimismo, se debería profundizar en distintas arquitecturas de la red neuronal, así como la optimización de los parámetros del modelo de

Random Forest. Además, sería recomendable probar distintos modelos de *machine* y *deep learning* a los propuestos en este trabajo. Cabe destacar que, para que el modelo desarrollado sea útil, las administraciones neoyorkinas deberían adaptar los conjuntos de datos de las predicciones meteorológicas a la estructura de *meteostat*, para poder construir los *dataframes* objeto de predicción de forma correcta, es decir, se deben disponer de las predicciones meteorológicas y la expectativa de eventos de antemano, construyendo *dataframes* que constituyan los *inputs* de entrada del modelo. Por último, como se ha comentado previamente, la incorporación de datos más precisos podría incrementar el rendimiento de los modelos.

La importancia de los hallazgos del proyecto radica en su capacidad para mejorar los modelos predictivos de inteligencia artificial aplicados a la optimización del transporte urbano. La identificación de patrones temporales, espaciales y contextuales podría permitir el desarrollo de modelos más precisos y adaptativos, capaces de anticipar dinámicas de tráfico y demanda de transporte en diferentes escenarios (Cheong-Chan et al., 2025; Licheng et al., 2019). Estos modelos pueden ser de gran utilidad para las administraciones públicas a la hora de organizar la movilidad y el transporte urbano, siendo capaces de predecir el tráfico y la demanda largo plazo, ofreciendo soluciones personalizadas para mejorar la eficiencia del transporte público y privado, reducir los tiempos de viaje y promover una movilidad urbana más sostenible.

El modelo propuesto está diseñado para ser una herramienta en la toma de decisiones de las administraciones públicas, con el objetivo de optimizar el transporte urbano y la asignación de recursos. A medida que el modelo se refine y logre predicciones más precisas, las administraciones podrán gestionar de manera más eficiente la distribución de recursos, la programación de frecuencias y la planificación de infraestructuras, maximizando el uso de recursos y reduciendo costes. Sin embargo, un margen de error en el modelo puede tener consecuencias, como asignaciones ineficientes de recursos que agraven la congestión en áreas críticas o sean infrautilizados. Esto subraya la importancia de continuar mejorando la precisión del modelo e incorporar mecanismos de retroalimentación para corregir decisiones en tiempo real: los eventos inesperados, como manifestaciones, accidentes o condiciones climáticas anormalmente atípicas, representan un desafío crítico para los modelos de predicción en redes de transporte urbano. Una estrategia clave para mitigar sus limitaciones es la combinación de datos históricos con datos en tiempo real. Los datos históricos permiten

identificar patrones y comportamientos habituales en la movilidad urbana, mientras que los datos en tiempo real, obtenidos a través de sensores *IoT*, aplicaciones móviles y redes sociales, proporcionan información inmediata sobre el estado actual del sistema. Este enfoque híbrido, respaldado por investigaciones como la de Liu et al. (2018), considero que mejoraría la precisión. Además, la implementación de sistemas de aprendizaje automático adaptativo puede procesar estas fuentes de datos de manera simultánea, permitiendo predicciones más robustas incluso en situaciones de alta incertidumbre.

En cuanto la tarificación dinámica de los peajes en Manhattan, una posible mejora en sería la incorporación de variables climáticas y eventos urbanos en el cálculo de los precios de peaje. Actualmente, el mecanismo de peaje considera factores como la hora del día, el tipo de vehículo, el método de pago y el cruce de créditos, pero no integra información sobre condiciones meteorológicas o eventos en la ciudad (MTA, 2025). Los resultados mostrados en el presente trabajo muestran que se han identificado correlaciones claras entre la velocidad de circulación y estos factores, lo que sugiere que su inclusión en el modelo de tarificación podría mejorar la eficiencia del sistema. Al incorporar estas variables, se lograría una tarificación más precisa y adaptativa, permitiendo ajustar los precios en función de condiciones que impactan directamente en la congestión y el flujo vehicular.

Una propuesta estratégica clave para el análisis transversal (transporte público y privado) sería el desarrollo de una política de datos conjunta entre los diferentes medios de transporte de Manhattan, con el fin de facilitar el análisis intermodal de los tiempos de viaje. Un ejemplo relevante es el sistema de datos de *SmartCards* utilizado en Pekín, que permite promediar tiempos de viaje en transporte público desde un punto de origen hasta un destino final dentro de una misma franja horaria, utilizando grafos para modelar el transporte (Ahmad et al., 2020). En este contexto, se podría explorar la creación de una base de datos intermodal que integre información sobre el transporte público, como metro y autobuses, con modos de transporte privado, como taxis o servicios de *ride-hailing*. Este enfoque requeriría revisar cuidadosamente las implicaciones de privacidad y garantizar que los datos sean anonimizados para cumplir con las normativas vigentes, como el Reglamento General de Protección de Datos.

Para abordar las diferencias en la calidad de los conjuntos de datos utilizados, se sugiere enriquecer las bases existentes con datos adicionales que incluyan variables

sociodemográficas, incluir aforos en los datos de eventos y registros relacionados con la seguridad en la ciudad. Incorporar información sociodemográfica, como niveles de ingresos, densidad de población, patrones de uso del suelo, políticas de transporte (NYC OpenData, 2024a) o distribución de actividades económicas, disponibles en plataformas como el *Census Bureau* o *NYC OpenData*, permitiría analizar cómo las características de las diferentes comunidades afectan la demanda de transporte y la elección modal. Asimismo, integrar datos sobre la seguridad en la ciudad, como tasas de criminalidad por zonas o incidentes reportados, podría ofrecer una perspectiva clave para comprender las decisiones de movilidad, especialmente durante horarios nocturnos o en zonas específicas donde la percepción de seguridad puede influir significativamente en la elección del modo de transporte. Finalmente, integrar datos sobre las medidas y políticas de transporte podría contribuir a la evaluación de las políticas públicas y su desempeño. Estos datos de seguridad y políticas de transporte están disponibles de forma pública en la página *NYC OpenData.* Estas variables no solo enriquecerían el análisis, sino que también ayudarían a diseñar estrategias *data-driven* más inclusivas y seguras para la optimización de las redes de transporte urbano.

En resumen, implementar estas mejoras permitirá, no solo aumentar la precisión y representatividad del análisis, sino también avanzar hacia una integración más efectiva entre los distintos modos de transporte. Esto sentará las bases para desarrollar modelos más completos que optimicen la movilidad urbana, al tiempo que se respetan los derechos de privacidad de los usuarios y se aprovechan al máximo los recursos tecnológicos y de datos disponibles.

Finalmente, cabe mencionar que, de cara a futuros desarrollos en la optimización de las redes de trasporte urbano, sería interesante fomentar la creación de políticas de extracción y diccionarios de datos conjuntos entre países. Esto facilitaría una mejor comparación y análisis transversal entre diferentes sistemas de transporte, permitiendo identificar patrones y mejores prácticas a nivel global. En lugar de tratar cada ciudad o país de manera aislada, la integración de datos estandarizados entre distintas regiones permitiría la creación de modelos más robustos y precisos, con un enfoque más globalizado. La estandarización de los datos de movilidad y transporte permitiría, por ejemplo, un análisis más coherente de la eficiencia de las redes de transporte intermodales en diversas ciudades, ayudando a identificar las mejores políticas públicas y estrategias para reducir la congestión, mejorar la sostenibilidad y optimizar

la demanda de transporte en el futuro. Implementar estas políticas también contribuiría a la creación de soluciones tecnológicas interoperables, que faciliten la integración de datos en tiempo real y mejoren la capacidad de predicción y análisis en redes de transporte urbanas de todo el mundo.

Este enfoque permitiría establecer sinergias entre distintas ciudades, garantizando que los sistemas de transporte compartan un marco de referencia común para la toma de decisiones, lo que podría tener un impacto positivo en la eficiencia del transporte público y privado. Además, al contar con un diccionario de datos estandarizado, se mejoraría la interoperabilidad entre plataformas de transporte, lo que facilitaría la colaboración entre diferentes proveedores de servicios de movilidad y las autoridades encargadas de la gestión urbana.

Referencias bibliográficas

Abadi, M., Barhan, P. Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Chemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. USENIX Association, 265–283. DOI: https://doi.org/10.48550/arXiv.1605.08695

Adams, E. (2023). Connecting to the Core. Available at: Connecting to the Core | Safer, Greener, and More Convenient Access to the Manhattan Central Business District (nyc.gov)

AEDP. Agencia Española de Protección de Datos. (2022). Guía básica de anonimización. Available at: https://www.aepd.es/documento/guia-basica-anonimizacion.pdf

Alanazi, F. (2023). A systematic literature review of autonomous and connected vehicles in traffic management. Applied Sciences, 13(3), 1789. https://doi.org/10.3390/app13031789

Belhadi, A., Djenouri, Y., Djenouri, D. y Chun-Wei Lin, J. (2020). A recurrent neural network for urban long-term traffic flow forecasting. Applied Intellingence. 50: 3252–3265. DOI: https://doi.org/10.1007/s10489-020-01716-1

Borgi, T., Zoghlami, N. y Abed, M. (2017). Big data for transport and logistics: A review. IC_ASET, 44-49. DOI: 10.1109/ASET.2017.7983742

Chen, C., Ma, J., Susilo, Y., Liu, Y. y Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg, 68:285-299. DOI: https://doi.org/10.1016/j.trc.2016.04.005

Cheong-Chan, R. K., Mun-Yee Lim, J. y Parthiban, R. (2025). Long-term traffic speed prediction utilizing data augmentation via segmented time frame clustering. Knowledge-Based Systems. 308:112785. DOI: https://doi.org/10.1016/j.knosys.2024.112785

Cho, K., Van Merrienboear, B., Gulcechre, C., Bahdanau, D., Bougares, F., Schwenk, H. y Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv.* 1406-1078. DOI: https://doi.org/10.48550/arXiv.1406.1078

Chung, Y. G., Jeon, Y., Yoo, S., Kim, H., y Hwang, H. (2022). Big data analysis and artificial intelligence in epilepsy – Common data model analysis and machine learning-based seizure detection and forecasting. Clinical and experimental pediatrics, 65(6), 272–282. DOI: https://doi.org/10.3345/cep.2021.00766

Colak, S., Alexander, L. P., Alvim, B. y González, M. C. (2015). Analyzing cell phone location data for urban travel: current methods, challenges, and opportunities. Transp. Res. Part C Emerg., 58:161-181. DOI: https://doi.org/10.3141/2526-14

De Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, *21*(3), 273. DOI: https://doi.org/10.1037/met0000079

Department of Transportation. (2024). System Optimization Bureau. Available at: Intelligent Transportation Systems (ITS) (ny.gov)

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1(1), 269-271. DOI: https://doi.org/10.1007/BF01386390

Graph Everywhere. (2024). ¿Qué son los grafos?. Available at: https://www.grapheverywhere.com/que-son-los-grafos/

Eventbrite. (2024). Intro to APIs. Available at: https://www.eventbrite.com/platform/docs/introduction

Fang, S., Pan, X., Xiang, S. y Pan, C. (2021). Meta-MSNet: Meta-Learning Based Multi-Source Data Fusion for Traffic Flow Prediction. IEEE Signal Processing Letters, 28: 6-10. DOI: https://doi.org/10.1109/lsp.2020.3037527

Fitzpatrick, M. (2023). City Privacy Protection Policies and Protocols. Office and Technology and Innovation. Available at: https://www.nyc.gov/content/oti/pages/information-privacy

González, M. C., Hidalgo, C. A. y Barabási, A. L. (2008). Understanding individual human mobility patterns. Nature, 453(7196):779-782. DOI: https://doi.org/10.1038/nature06958

Hochreiter, S., y Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

IBM. (2024). ¿Qué son las redes neuronales? Available at: https://www.ibm.com/es-es/topics/neural-networks

ITF (2019), *What is the Value of Saving Travel Time?: Summary and Conclusions*, ITF Roundtable Reports, No. 176, OECD Publishing, Paris, https://doi.org/10.1787/eeb102ea-en.

Ke, J., Zheng, H., Yang, H. y Chen, X. (2018). Short-term forecasting of passenger demand under on-demand ride services using deep learning approaches. Transp. Res. Part C Emerg., 85:591-608. DOI: https://doi.org/10.1016/j.trc.2017.10.016

Kim, J., Zheng, K., Corcoran, J., Ahn, S., & Papamanolis, M. (2022). Trajectory Flow Map: Graph-based Approach to Analysing Temporal Evolution of Aggregated Traffic Flows in Large-scale Urban Networks. *arXiv,* 2212.02927. DOI: https://doi.org/10.48550/arXiv.2212.02927

Kingsford, C., y Salzberg, S. L. (2008). What are decision trees? Nature Biotechnology, 26(9), 1011–1013. DOI: https://doi.org/10.1038/nbt0908-1011

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. DOI: https://doi.org/10.1038/nature14539

Li, Y., Chai, S., Ma, Z. y Wang, G. (2021). A Hybrid Deep Learning Framework for Long-Term Traffic Flow Prediction. IEEE Access, 9: 11264-11271. DOI: https://doi.org/10.1109/ACCESS.2021.3050836

Licheng, Q., Wei, L., Wenjing, L., Dongfang, M. y Yinhai, W. (2019). Daily long-term traffic flow forecasting based on a deep neural network. Expert Systems with Applications. 121: 304-312. DOI: https://doi.org/10.1016/j.eswa.2018.12.031

Liu, Q., Zheng, X., Stanley, E., Xiao, F. y Liu, W. (2021). A Spatiotemporal Co-Clustering Framework for Discovering Mobility Patterns: A Study of Manhattan Taxi Data. IEEE Access, 9:34338-34351. DOI: 10.1109/ACCESS.2021.3052795

Liu, Z., Li, Z., Wu, K., & Li, M. (2018). Urban traffic prediction from mobility data using deep learning. IEEE Network, 32(4), 40–46. DOI: https://doi.org/10.1109/MNET.2018.1700411

Lv, Y., Duan, Y., Kang, W., Li, Z. y Wang, F. Y. (2015). Traffic flow prediction with big data: A deep learning approach. T-IT, 16(2):865-873. DOI: 10.1109/TITS.2014.2345663

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataram, S., Liu, D., Freeman, J., Tsai, DB., Made, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M. y Talwalkar, A. (2016). MLLib: Machine Learning in Apache Spark. Journal of Machine Learning Research, 17,1; 1235-1241. DOI: https://dl.acm.org/doi/10.5555/2946645.2946679

Meteostat. (2024). Meteostat Developers. Available at: https://dev.meteostat.net/

MTA. Metropolitan Transportation Authority. (2025). *Fewer cars, faster trips: Congestion Relief Zone tolling data is in.* Available at: Fewer cars, faster trips: Congestion Relief Zone

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Nair, V., y Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML), 807-814. DOI: https://dl.acm.org/doi/10.5555/3104322.3104425

New York City Department of Transportation. (2023). Notice of Adoption: Open Streets Program. Available at: NYC DOT Notice of Adoption - Open Streets Program

Newman, P. y Kenworthy, J. (2015). The end of Automobile Dependence. Island Press. DOI: https://doi.org/10.5822/978-1-61091-613-4

NOAA. National Oceanic and Atmospheric Administration. (2024). Climate Data Online Data Tools. Available at: https://www.ncei.noaa.gov/cdo-web/datatools

NYC.gov. (2024a). NYC.gov Privacy Policy. Collection of Information by NYC.gov. Available at: https://www.nyc.gov/home/privacy-policy.page

NYC.gov. (2024b). NYC Taxi & Limousine Comission. TLC Trip Record Data. Available at: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

NYC Environment and Health Data Portal. (2021). The Public Health Impacts of PM2.5 from Traffic Air Pollution. Available at: Traffic and PM2.5 air pollution (nyc.gov)

NYC OpenData. (2024a). NYC Rezoning Tracker. Available at: https://data.cityofnewyork.us/City-Government/NYC-Rezoning-Tracker/fd95-5ihz/about_data

NYC OpenData. (2024b). Open Data for All New Yorkers. Available at: https://opendata.cityofnewyork.us/

NYC OpenData. (2025). NYC Permitted Event Information – Historical. Available at: https://data.cityofnewyork.us/City-Government/NYC-Permitted-Event-Information-Historical/bkfu-528j/about_data

OMS. Organización Mundial de la Salud. (2024). Compendium of WHO and other UN guidance in health and environment, 2024 update. [Internet] Geneva. WHO. Available at: Compendium of WHO and other UN guidance in health and environment, 2024 update

Pelgrims, I., Devleesschauwer, B., Guyot M., Keune, H., Nawrot, T., Remmen, R., Saenen, D. N., Trabelsi, S., Thomas, I., Aerts, R. y De Clercq, M. E. (2021). Association between urban environment and mental health in Brussels, Belgium. BMC Public Health, 21(1):635. DOI: 10.1186/s12889-021-10557-7

Pishue, B. (2024). 2023 INRIX Global Traffic Scorecard with Q1 2024 update. Available at: Global Traffic Scorecard | INRIX Global Traffic Rankings

Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J. y Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. Int J Syst Assur Eng Manag., 13(Suppl.1):S16-S27. DOI: https://doi.org/10.1007/s13198-021-01130-x

Ranjan, N., Bhandari, S., Zhao, H. P., Kim, H. y Khan, P. (2020). City-Wide Traffic Congestion Prediction Based on CNN, LSTM and Transpose CNN. IEEE Access, 8: 81606-81620. DOI: 10.1109/ACCESS.2020.2991462

Rebekić, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or Spearman's correlation coefficient-which one to use? *Poljoprivreda*, *21*(2), 47-54. DOI: https://doi.org/10.18047/poljo.21.2.8

Rincón Pinzón, M. A., Mejía Rodríguez, C. A., Ramírez Camargo, E. A. y Arévalo Vergel, L. M. (2024). Análisis e implementación de clustering en casos de dengue mediante algoritmo de aprendizaje no supervisado. Revista colombiana de tecnologías de avanzada, 2(44), 104–111 DOI: https://doi.org/10.24054/rcta.v2i44.3021

Roberton, J., Schmidt, S. y Stiles, R. (2020). Emissions from the Taxi and For-Hire Vehicle Transportation Sector in New York City. Authorea. DOI: 10.31124/advance.12152844.v1

Rumelhart, D. E., Hinton, G. E., y Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536. DOI: https://doi.org/10.1038/323533a0

Russell, S., y Norvig, P. (2020). Artificial Intelligence: A Modern Approach. Pearson.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. y Monfardini, G. (2009). The Graph Neural Network Model. IEEE Transactions on Neural Networks, 20(1), 61–80. DOI: 10.1109/TNN.2008.2005605

Shahriari, S., Ghasri, M., Sisson, S. A., y Rashidi, T. (2020). Ensemble of ARIMA: combining parametric and bootstrapping technique for traffic flow prediction. Transportmetrica A: Transport Science, 16(3): 1552–1573. DOI: https://doi.org/10.1080/23249935.2020.1764662

Shi, X., Qi, H., Shen, Y., Wu, G. y Yin, B. (2021). A Spatial–Temporal Attention Approach for Traffic Prediction. IEEE Transactions on Intelligent Transportation Systems, 22(8):4909-4918. DOI: https://doi.org/10.1109/TITS.2020.2983651

Steel Guard Safety Products. (2024). Noisiest Cities in US 2024. Available at: Noisiest Cities in America (steelguardsafety.com)

The Official Website of New York State. (2024a). Freedom of Information Law Text. Available at: https://opengovernment.ny.gov/freedom-information-law

The Official Website of New York State. (2024b). MTA Subway Origin-Destination Ridership Estimate: 2024. Available at: https://data.ny.gov/Transportation/MTA-Subway-Origin-Destination-Ridership-Estimate-2/jsu2-fbtj/about_data

U.S. Department of Justice. (2022). Office of Privacy and Civil Liberties. Privacy Act of 1974. Available at: https://www.justice.gov/opcl/privacy-act-1974

Wang, Z., Su, X. y Ding, Z. (2021). Long-term traffic prediction based on LSTM encoder-decoder architecture. IEEE Transactions on Intelligent Transportation Systems, 22(10): 6561-6571. DOI: https://doi.org/10.1109/TITS.2020.2995546

World Population Review. (2024). New York City, New York Population, 2024. Available at: New York City, New York Population 2024 (worldpopulationreview.com)

Yao, S., Zhang, H., Wang, C., Zeng, D. y Ye, M. (2023). GSTGAT: Gated spatiotemporal graph attention network for traffic demand forecasting. IET Intell., 18:258-268. DOI: https://doi.org/10.1049/itr2.12449

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S. y Stoica, I. (2010). Spark: Cluster Computing with Working Sets. HotCloud, 10(10-10), 95. Available at: https://dl.acm.org/doi/10.5555/1863103.1863113

Zhang, J., Zheng, Y. y Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. AAAI, 1655-1661. DOI: https://doi.org/10.1609/aaai.v31i1.10735

Zhao, F., Zeng G. Q. y Lu, K. D. (2020). EnLSTM-WPEO: Short-Term Traffic Flow Prediction by Ensemble LSTM, NNCT Weight Integration, and Population Extremal Optimization. IEEE Transactions on Vehicular Technology, 69(1):101-113. DOI: https://doi.org/10.1109/TVT.2019.2952605

Zheng, Y., Capra, L., Wolfson, O. y Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. T-ITS, 5(3):38. DOI: http://dx.doi.org/10.1145/2629592

Zhu, L., Yu, F. R., Wang, Y., Ning, B. y Tang, T. (2019). Big data analytics in intelligent transportation systems: A survey. *T-ITS,* 20(1):383-398. DOI: 10.1109/TITS.2018.2815678

Zhu, W., Sun, Y., Yi, X., & Wang, Y. (2022). Correlation Information-based Spatiotemporal Network for Traffic Flow Forecasting. *ArXiv,* 2205.10365. DOI: https://doi.org/10.48550/arXiv.2205.10365