Sistema basado en analítica de datos para apoyar la toma de decisiones en la planificación de la oferta del servicio de transporte público masivo en una pandemia

Universidad del Cauca

Trabajo Final de Grado
Modalidad: Trabajo de Investigación

Jonathan Valencia Bolaños

Director: PhD. José Luis Arciniegas Herrera
Universidad del Cauca
Co-Director: PhD. Hugo Alexer Parada Gelvez
Universidad Politécnica de Madrid, España

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Programa de Ingeniería Electrónica y Telecomunicaciones
Popayán – Cauca
2023

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todas las personas que han contribuido de manera significativa en la realización de este trabajo.

En primer lugar, agradezco a mi director de tesis, José Luis Arciniegas Herrera, y a mi codirector, Hugo Alexer Parada Gelvez, por su invaluable apoyo, dedicación y disposición a lo largo de este proceso. Sus valiosos comentarios y sugerencias han sido fundamentales para enriquecer y mejorar este trabajo.

También quiero expresar mi profunda gratitud a mi familia, quienes han sido el pilar fundamental en mi desarrollo personal y académico. En especial, mi hermana, quien ha sido un ejemplo a seguir, así como a Laura y a todas aquellas personas cercanas que han brindado su constante motivación en cada desafío que he afrontado.

Por último, no puedo dejar de agradecer a la Universidad del Cauca, en especial a todos los profesores y compañeros que han sido parte de mi formación profesional. Su enseñanza y conocimientos compartidos han sido invaluables para lograr alcanzar las metas que me he propuesto.

Índice

CAPÍTULO 1

Introducción

1.1. Planteamiento del problema

El sistema masivo de transporte es uno de los sectores más importantes en la infraestructura de las ciudades del mundo ya que miles de personas diariamente requieren de un medio para desplazarse de un punto a otro dentro de una ciudad, ya sea por motivos de educación, de trabajo o de ocio, el transporte público desempeña un rol muy importante en la sociedad urbana actual. Además es importante resaltar la importancia que tienen este tipo de medios de transporte en la movilidad sostenible ya que en términos de ocupación, consumo de energía, congestión y contaminación los medios masivos de transporte son mucho más eficientes que los carros particulares.

Sin embargo, la pandemia originada por el virus SARS-COV-2 produjo una afectación en distintos sectores económicos, entre ellos el sector de transporte [1]. Por tales circunstancias se plantean grandes retos que apuntan a una disminución en la demanda de pasajeros dentro de los sistemas de transporte público. Una de las principales razones de disminución de la demanda se debe a las regulaciones impuestas por los gobiernos para contener la propagación del virus [2]. Dichas regulaciones implican restricciones en la libre circulación de las personas, imposición de nuevas reglas de distanciamiento físico que afectan la capacidad de transporte ofrecida, e incluso en algunos casos se presentan cierres totales de los servicios de transporte.

Por consiguiente, un gran reto al cual se enfrentaron los proveedores de servicio de transporte público en la pandemia, fue el limitado conocimiento que se tenía para hacer frente a los cambios en los patrones de movilidad de las ciudades inducidos por la llegada del Covid-19. Esta carencia de información sólida, conllevó a que la toma de decisiones y la planificación se viera obligada a seguir procedimientos que en su momento no tenían evidencia de ser efectivos en la contención del virus [2]. Teniendo muchas veces que reducir la duración de sus operaciones diarias, permitir cierres temporales en algunas estaciones y limitar las capacidades de los vehículos para cumplir con las regulaciones de distanciamiento social. Estas medidas de prevención generaron también un impacto negativo sobre la calidad del servicio de transporte, al no satisfacer las necesidades y expectativas del usuario.

En este contexto, se plantea en esta investigación, examinar diversas estrategias basadas en predicción y análisis de datos que puedan contribuir a una mejora significativa en la eficiencia y la calidad del servicio de transporte público en una pandemia. Este enfoque se origina a partir de la comprensión de que la toma de decisiones basada en evidencia en la planificación del transporte público está estrechamente vinculada a la capacidad para anticipar y reaccionar ante las fluctuaciones en la demanda de pasajeros.

Con este propósito, diferentes países del mundo empiezan a recolectar y a compartir datos de movilidad durante la pandemia, para que a partir de estos datos recolectados surjan numerosas soluciones tecnológicas. En Colombia de igual manera existe una iniciativa desde el año 2011 que tiene como objetivo promover el uso y generación de datos

mediante el portal de datos abiertos del estado colombiano [3]. Sin embargo al realizar la búsqueda de datos relacionados al transporte masivo de las grandes ciudades como Bogotá, Medellín y Cali se encontraron datos globales que se limitan a proporcionar un total de pasajeros movilizados por día e ignorando datos relevantes para esta investigación como la movilidad entre zonas. De igual forma al ingresar al portal de datos abiertos de Transmilenio de la ciudad de Bogotá [4], o bien el portal del metro de Medellín [5] o el portal Metro Cali S.A [6], se hallan los mismos datos generales que indican tarifas de transporte, ubicación de las estaciones, recorridos y horarios de los distintos modos de transporte.

En razón de lo expuesto, se tiene que la implementación de este tipo de estudios representa un desafío al no encontrar información abierta fácilmente. En este sentido y pese a que los datos en el contexto Colombiano no permiten realizar un análisis a detalle de la movilidad, se consideró llevar el estudio de caso a un contexto internacional teniendo en cuenta que la propuesta que surja de este caso de estudio se podría replicar para otras ciudades siempre y cuando se tenga disponibilidad de los datos.

En el contexto internacional, particularmente en España, se ha visto como establecen un plan para responder al impacto del Covid-19, para ello el Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA) publica datos de movilidad generados a partir del posicionamiento de los teléfonos móviles. Esta base de datos abierta presenta los datos en un periodo de estudio que va desde el día 29 de febrero de 2020 al día 18 de marzo de 2021 y son ordenados por día, por hora y por tramos de distancias [7].

Dado que se dispone de datos para todas las ciudades de España, el enfoque principal de esta investigación se centrará exclusivamente en la ciudad de Madrid. Según los datos proporcionados por el portal de transparencia de Madrid [8], el transporte público de la ciudad presentó una disminución en la media anual de viajes cerca al 50 % en el año 2020. En cuanto al tráfico privado, se redujo en un 32 % en la red viaria metropolitana y en un 27 % en las vías urbanas. Por su parte, la movilidad peatonal se redujo en menor medida con una media de 25 % y la movilidad mediante bicicletas presentan un aumento superior al 20 %. De esta manera, en el año 2020, la demanda en los servicios de transporte público se redujo drásticamente a la mitad debido a las restricciones de movilidad, confinamientos y aumento de teletrabajo. En particular el metro de Madrid registró una disminución del 49 %, mientras que el servicio de Renfe cercanías experimentó una caída del 43 %.

Considerando el contexto de la movilidad en la ciudad de Madrid y aprovechando la gran cantidad de datos abiertos que el ministerio de transporte de España ha publicado, se plantea la posibilidad de desarrollar una herramienta de predicción que sirva de apoyo a los proveedores de servicios de transporte público durante una pandemia. La propuesta resultante de este estudio de caso busca anticipar el número de pasajeros que entran en una estación del sistema de transporte de metro y tren se cercanías de Madrid.

La predicción de la demanda de pasajeros en las estaciones de servicio es fundamental, ya que permite que el sistema de transporte público tome decisiones informadas para mejorar la planificación. Mediante la predicción de pasajeros en una estación, es posible asignar recursos de manera más eficaz, como la adición de trenes durante horas pico o la asignación de personal de seguridad en estaciones con alta afluencia. Además, el sistema de transporte podría ajustar la capacidad y frecuencia de los trenes para evitar aglomeraciones en momentos de alta demanda y minimizar el desperdicio de recursos en períodos de baja demanda.

En virtud de lo señalado, el presente trabajo de grado busca dar respuesta a la siguiente pregunta de investigación: ¿Cómo crear un sistema de predicción de afluencia en estaciones de metro y tren de cercanías que soporte la toma de decisiones operativas, como el reajuste de horarios y frecuencias, mediante el análisis de datos?

1.2. Objetivos

1.2.1. Objetivo General

Construir un sistema basado en analítica de datos para apoyar la toma de decisiones en la planificación de la oferta del servicio del transporte público masivo en la ciudad de Madrid (España).

1.2.2. Objetivos Específicos

1. Analizar los datos abiertos de movilidad y pandemia de Covid-19, estudio de caso en la ciudad de Madrid (España).

2. Desarrollar un modelo a través de técnicas de Machine Learning que ayude a la planificación del servicio de transporte público mediante el contraste de patrones de movilidad y tasas de incidencia de covid-19 publicadas por la autoridad sanitaria.

3. Evaluar el grado de precisión de la salida del modelo mediante herramientas estadísticas, estudio de caso transporte masivo de tren de cercanías y metro de Madrid (España).

1.3. Estado del arte

1.3.1. Movilidad en el marco de la pandemia por Covid-19

En esta sección se hace una revisión del impacto generado por la pandemia Covid-19 en temas de movilidad y transporte público.

"Public transport planning adaption under the COVID-19 pandemic crisis: literature review of research needs and directions"[2]

El presente artículo a partir de una revisión literaria busca identificar las necesidades que tiene el sector de transporte. Una de las necesidades que se destaca es la escasez de conocimiento que existe en este sector para contrarrestar los efectos de la pandemia, esto conlleva a que muchas de las decisiones tomadas por parte de los proveedores de servicio de transporte, sean decisiones que no tienen evidencia de ser efectivas. En esta revisión literaria las medidas impuestas por los gobiernos para enfrentar la propagación del virus representan una dificultad para el sector de transporte, ya que estas obligan a que se tenga que operar a una capacidad mínima de pasajeros. Por tal razón y con el fin de mejorar la gestión de multitudes dentro del sistema de transporte público, se plantea una propuesta que permita adaptar o ajustar la planificación del sector de transporte de acuerdo a las necesidades que aparezcan durante la pandemia. Algunas de las medidas tomadas en cuenta dentro de la planificación implican determinar una frecuencia óptima de servicio, cambios de horarios de los vehículos, reducción de las operaciones diarias, conseguir una carga uniforme dentro de cada vehículo y la suspensión temporal de algunas estaciones.

"Role of transport during outbreak of infectious diseases: Evidence from the past"[1]

Este artículo se encarga de revisar la literatura científica relacionada con el sector de transporte y las enfermedades infecciosas. Se encontró que el sector de transporte fue uno de los sectores más afectados en la pandemia por Covid-19, ya que al haber una reducción de los desplazamientos de las personas se tiene un efecto directo en las características del tráfico, como los patrones, los volúmenes, velocidad y nivel de servicio. En el transporte público la proporción de vehículos disminuyó y la proporción de vehículos particulares aumentó. También debido a la pandemia se modificaron los propósitos de viaje de las personas, sin embargo el trabajo se mantuvo como propósito más importante. Por último se menciona que las restricciones de viaje son efectivas si son impuestas en una etapa temprana del brote y sin ellas la propagación del Covid-19 podría verse acelerada.

"Effects of the Covid-19 lockdown on urban mobility: Empirical evidence from the city of Santander (Spain)"[9]

Este artículo analiza el impacto de las medidas de confinamiento impuestas en la ciudad de Santander, España. Se evidencia en este estudio que la caída en la movilidad fue mucho más crítica en el sistema de transporte público, por lo que se ha observado que hay una caída de un 85 % en demanda para las principales rutas de transporte público de la ciudad y la disminución de la demanda de pasajeros es atribuida en gran medida a la percepción de riesgo de contagio. En la última sección del artículo se destaca la dificultad que tiene el sistema de transporte en el ámbito económico, ya que los costos operacionales tienden a aumentar al mismo tiempo que se disminuye la ocupación en

cada vehículo por medidas de distanciamiento físico, es por ello que se sugiere el aumento de subsidios por parte de la administración pública y una adopción de medidas que garanticen la viabilidad del sistema de transporte público. Por último el autor sugiere un escalonamiento en tiempos de entrada a lugares de trabajo y establecimientos de educación para minimizar las horas pico dentro de las ciudades.

"COVID-19 and public transportation: Current assessment, prospects, and research needs"[10]

En este estudio se indica que el sistema de transporte público es un lugar de alto riesgo de contagio debido al espacio limitado y a las múltiples superficies de contacto que transmiten gérmenes, pero se encuentra evidencia de que el uso adecuado de cubre bocas junto con medidas de distanciamiento social menores a 2 metros son efectivas para evitar la propagación del virus. Además se consideran otros aspectos como la equidad social durante la crisis por Covid-19. Desde este punto de vista se observa que el trabajo en casa es más bien un privilegio para las personas que desempeñan trabajos con ingresos altos y aquellas personas que tienen ingresos más bajos requieren de un medio de transporte que los pueda llevar hasta sus trabajos. También es identificado en esta investigación que el cumplimiento del distanciamiento físico genera una demanda de pasajeros desatendidos por el sistema de transporte público, en el caso del metro de Washington se tiene una reducción de más del 80 % de pasajeros si se asume un distanciamiento de 1.5 m. Por último se plantean alternativas que ayuden a gestionar la capacidad limitada dentro de los vehículos de transporte público, para ello una opción es dirigir la demanda de pasajeros desatendida hacia los servicios de transporte a pedido (Taxis). De igual forma se tienen en cuenta alternativas donde se restringen las entradas mediante sistemas de reserva u otras alternativas que involucran el ofrecimiento de descuentos para estimular a los pasajeros a viajar en periodos de menor actividad.

"Tools for the monitoring, user characterization, and their applications to the public integrated transport system due to the COVID-19 disease effects: A case study in Bogotá TRANSMILENIO company"[11]

En el entorno Colombiano se encuentra un estudio donde se revisan las políticas implementadas por la compañía de transporte, TRANSMILENIO S.A. Estas políticas apuntan a cumplir con un límite de 35 % de capacidad, para lograrlo se propone organizar los horarios de los empleados públicos en diferentes turnos. Gráficamente el artículo muestra que si todos los empleados públicos que utilizan el servicio de BRT (Autobús de Tránsito Rápido) salen en la misma hora, se excede el límite de capacidad impuesto por el gobierno pero si los empleados salen de sus trabajos en 2 o 3 diferentes horarios este límite de 35 % no se supera, por lo que TRANSMILENIO S.A recomienda retornar las actividades presenciales a las instituciones públicas cercanas a la estación de donde se realizó el estudio, implementado 2 o 3 diferentes horarios y evitando las horas pico. Finalmente se hace una comparativa tomando en cuenta la variación en la demanda del sistema de transporte versus los casos de Covid-19 confirmados, de esto se concluye que no hay una relación directa entre la demanda y el contagio, siendo esto evidencia de que el uso del transporte público no aumenta los casos de Covid-19.

1.3.2. Soluciones tecnológicas

En esta segunda sección se hace una revisión literaria que posibilite la identificación de diferentes alternativas tecnológicas que den solución al problema abordado en este

trabajo.

“Survey of air exchange rates and evaluation of airborne infection risk of COVID-19 on commuter trains”[12]

Se muestra un estudio realizado en Japón donde se evalúa el riesgo de infección por Covid-19 en un tren suburbano. Para este estudio se tuvo en cuenta la tasa de intercambio de aire en un vagón del tren y la proporción de personas infectadas en la ciudad de Tokio. Además para este estudio de caso se asume que el número de pasajeros dentro del vagón varía entre 30-300 personas y la duración del viaje está en un intervalo de 7-60 minutos. De acuerdo a los datos anteriormente mencionados se propone un modelo que logre calcular el riesgo de infección de los pasajeros. Finalmente se muestran los resultados obtenidos teniendo en cuenta dos situaciones diferentes, para el caso en donde las personas infectadas permanecieron en silencio durante el viaje se observó que el riesgo de infección no era tan alto, caso contrario para el caso donde los pasajeros infectados continuamente estuvieron hablando y el riesgo de infección aumentó. Los resultados sugieren que medidas como abrir la ventana y hacer funcionar el aire acondicionado podría reducir el riesgo de contagio aéreo. Por último, al hacer una comprobación del modelo se logra estimar un factor de infección de 1.5x10-7, considerando un tren lleno con 150 pasajeros y una proporción de personas infectadas en Tokio equivalente al 0.3 %. En otras palabras y de acuerdo al modelo que se ha propuesto en este artículo, si se tienen 8.4 millones de viajeros diarios en Tokio, se tendrá que 1.3 de esos viajeros de tren se infectarán cada día.

“Sistema basado en analítica predictiva para ofrecer recomendación sobre uso de transporte público en Madrid durante la pandemia”[13]

El documento corresponde a un trabajo de grado de la Universidad Politécnica de Madrid y tiene como objetivo ofrecer al usuario información con antelación acerca de la ocupación dentro del vehículo de transporte público (Renfe y metro de Madrid). Para el desarrollo de la propuesta se implementa una combinación de árboles de decisión conocida como Random Forest, este algoritmo de Machine Learning fue seleccionado ya que demostró tener una alta efectividad para el tipo de datos que se están teniendo en cuenta en este trabajo. La principal fuente de datos que se tomó en cuenta para el desarrollo del sistema, corresponde a la base de datos de movilidad publicada por el Ministerio de Transporte. De igual forma se tiene en cuenta información meteorológica para obtener un mejor resultado en la predicción de la ocupación. Por último se tiene que el sistema fue desarrollado a partir de librerías de Python y permite observar información de ocupación entre 2 estaciones dependiendo de la ruta elegida, dado el caso de una ocupación elevada el sistema ofrece un servicio alternativo de bicicletas para poder desplazarse.

“Análisis y diseño de un sistema para apoyar un modelo de transporte público seguro basado en los datos de movilidad durante la pandemia en Madrid”[14]

El presente documento se encarga de diseñar una solución para las entidades de transporte público de Madrid (Renfe Cercanías, Metro, y la Empresa Municipal de Transportes de Madrid). En esta investigación se tiene como objetivo brindar información al sistema de transporte público para evitar superar los aforos permitidos dentro de los vehículos de transporte. Para lograr el objetivo planteado se utiliza el modelo predictivo de Machine Learning conocido como Prophet, este modelo fue desarrollado por Facebook y corresponde a un modelo de regresión no lineal que se enfoca en la realización de pronósticos mediante series temporales. El conjunto de datos utilizados para el desarrollo de la investigación principalmente fueron tomados de la base de datos abierta que suministró

el ministerio de transporte de España y datos relacionados a los servicios de transporte público que se encuentran disponibles en plataformas web. Finalmente a partir de los datos y las técnicas implementadas este trabajo construye un sistema que es capaz de brindar recomendaciones de frecuencias en los servicios de Renfe y Metro, realizar predicciones sobre desplazamientos con 30 días de antelación y mostrar al usuario en qué líneas hay problemas de aforo.

1.3.3. Recopilación de artículos de investigación

En la Tabla 1 se realiza una recopilación de los aportes y las brechas de cada uno de los artículos revisados. Este trabajo estará enfocado en dar solución a las brechas que han sido resaltadas en la tabla y aquellas brechas que no son resaltadas quedarán por fuera del alcance de este proyecto.

Título del artículo	Aportes	Brechas
"Public transport planning adaption under the COVID-19 pandemic crisis: literature review of research needs and directions" [2]	Propone una adaptación de las medidas de planificación del transporte público para hacer frente a los nuevos retos que ha dejado la pandemia Covid-19. Además se logra identificar algunas de las medidas que mejor se adaptan al considerar el distanciamiento social.	***No propone herramientas que sirvan de apoyo para tomar decisiones más acertadas en el sector de transporte público.*** No propone alternativas de movilidad para aquellas personas que serán desatendidas por el sistema de transporte público.
"Role of transport during outbreak of infectious diseases: Evidence from the past" [1]	Este artículo se centra en identificar los roles del transporte público durante una pandemia. Los resultados encontrados indican que el sector de transporte debe controlar la propagación de la infección y al mismo tiempo evaluar el impacto que genera la pandemia.	Hacen falta indicadores que sugieran la implementación de medidas dentro del sistema de transporte para cada etapa de la pandemia. No se presenta información para identificar las regulaciones que obtienen mejores resultados en la contención del virus.
"Effects of the Covid-19 lockdown on urban mobility: Empirical evidence from the city of Santander (Spain)"[9]	Se resalta de este artículo la gran cantidad de datos de diferentes fuentes que la ciudad de Santander recolecta, entre las técnicas que el autor utiliza para hacer el análisis de los datos se encuentra el procesamiento de imágenes y el análisis de datos del posicionamiento del transporte público. A partir del análisis de estos datos se muestra la evolución del transporte durante la cuarentena y los cambios de los patrones de movilidad.	***No proponen modelos predictivos que ayuden al proveedor de transporte en la gestión y organización del servicio.*** No se proponen medidas que logren minimizar las pérdidas económicas de los proveedores del servicio de transporte.

“COVID-19 and public transportation: Current assessment, prospects, and research needs”[10]	En este artículo se considera el rediseño del servicio de transporte para adaptarse a los nuevos patrones de movilidad. Además se evalúa la efectividad de las medidas de distanciamiento menores a 2 metros y el uso de cubre bocas dentro de vehículos de transporte público.	Se desconoce la probabilidad de infección dentro de un vehículo o una estación de transporte público. Hace falta priorizar el servicio de transporte para garantizar que todas las personas tengan un medio para desplazarse. No se proponen medidas para solucionar el desborde de la capacidad del servicio de transporte.
“Tools for the monitoring, user characterization, and their applications to the public integrated transport system due to the COVID-19 disease effects: A case study in Bogotá TRANSMILENIO company”[11]	En este estudio se pretende evitar que se supere el límite de ocupación dentro de los vehículos del sistema de transporte público de Bogotá. Se logra concluir que al organizar los horarios de empleados en 2 o 3 diferentes horarios no se supera el límite de 35 % de ocupación. Se encuentra además evidencia que sugiere que el uso de transporte público no aumenta los casos de Covid-19.	***Hace falta la implementación de más herramientas de visualización para dar seguimiento a los efectos del Covid-19 sobre el transporte público.*** ***No se proponen herramientas que faciliten la identificación de patrones de movilidad.***
“Survey of air exchange rates and evaluation of airborne infection risk of COVID-19 on commuter trains”[12]	Propone un modelo que permite calcular el riesgo de infección de los pasajeros dentro de un vagón de tren, para ello tiene en cuenta las tasas de intercambio de aire, el número de pasajeros dentro del vagón y la duración del viaje.	***De acuerdo a los resultados obtenidos, no se proponen herramientas de apoyo al sistema de transporte.*** Hace falta un estudio para estimar el riesgo de infección por micro partículas expulsadas al estornudar o toser.
“Sistema basado en analítica predictiva para ofrecer recomendación sobre uso de transporte público en Madrid durante la pandemia”[13]	Propone un sistema de predicción que informa al usuario la ocupación dentro del vehículo de transporte público, y en el caso de que la ocupación sea elevada ofrece un sistema de transporte alternativo.	***Hace falta proporcionar información útil a los proveedores del transporte masivo para la planificación de la oferta del servicio.***

“Análisis y diseño de un sistema para apoyar un modelo de transporte público seguro basado en los datos de movilidad durante la pandemia en Madrid”[14]	El sistema propuesto, utiliza modelos predictivos de Machine Learning para suministrar información al sistema de transporte público a través de una herramienta visual. Adicionalmente sugiere la frecuencia óptima del servicio para evitar aglomeraciones y mostrar las rutas que tienen problemas de aforo.	***Hace falta contrastar datos de movilidad con datos relacionados al Covid-19 para obtener información de valor que pueda servir en la toma de decisiones.***

Tabla 1: Cuadro comparativo de los diferentes artículos [Fuente propia]

Después de haber llevado a cabo la revisión documental del estado actual del conocimiento, se ha encontrado que muchos de los artículos se enfocan particularmente en identificar las nuevas necesidades que surgen en el sector de transporte público a partir de la aparición de la pandemia de Covid-19, pero son pocos los artículos y trabajos académicos que a partir de estas necesidades observadas construyen una solución direccionada a este sector. Se concluye además que son muy pocas las ciudades del mundo que a partir de técnicas en analítica de datos logran dar una herramienta de apoyo al sistema de transporte público, siendo esto una evidencia de la falta de datos abiertos para realizar este tipo de estudios. En razón de lo expuesto el presente trabajo de grado pretende dar un aporte al sistema de transporte público a partir de técnicas en analítica de datos,

1.4. Marco Teórico

1.4.1. Planificación del transporte público

El artículo titulado ”La planificación como herramienta en la movilidad del transporte Urbano“[15] subraya la importancia de la planificación como componente esencial de la administración que cumple diversos propósitos, un propósito fundamental consiste en coordinar esfuerzos y recursos dentro de las organizaciones para facilitar el logro de los objetivos empresariales.

Del mismo modo, en este artículo se resaltan las clasificaciones propuestas por Wilburg Jiménez Castro, quien ha identificado tres categorías esenciales de la planificación:

- Planificación Operativa: Se enfoca en garantizar que el transporte urbano se desarrolle de manera adecuada. Es decir, que sea de bienestar para el habitante, para la ciudad y para el medio ambiente.

- Planificación Económica y Social: Se realiza el inventario de recursos, necesidades y se determinan las metas. En el transporte urbano se cuenta con recursos financieros y talento humano capacitado. Además, se tiene identificado las necesidades de la población, de la ciudad y del medio ambiente.

- Planificación Física o Territorial: Se define como la adopción de programas y normas adecuadas, para el desarrollo de los recursos naturales y para el crecimiento de ciudades. Esta planificación es responsabilidad de las municipalidades, ya que de ellos dependen los planes de crecimiento ordenado para la movilidad responsable.

En línea con lo mencionado previamente, este trabajo de investigación se enfoca en brindar una herramienta de apoyo para la toma de decisiones en la planificación operativa. De acuerdo con la perspectiva de J.A.Quirós Alés (2015) [16], la planificación operativa se dedica a establecer los recursos necesarios para garantizar el funcionamiento eficiente del sistema ferroviario a corto plazo. Esta forma de planificación busca comprender el comportamiento de los viajeros para optimizar la oferta del transporte, anticipándose a posibles situaciones extraordinarias. Algunas de las actividades realizadas en la planificación operativa son mencionadas a continuación.

- Reajustar los horarios y frecuencia en caso de que se produzca alguna situación no esperada, de modo que el sistema pueda adaptarse a estos cambios sin alterar el funcionamiento.

- Reajustar la composición de los trenes, como son ensamblajes con otros trenes o vagones para dar servicio a una mayor o menor demanda.

- En situaciones extraordinarias como escenarios de climatología adversa se priorizan unos servicios sobre otros, decidiendo que trenes tienen preferencia en el paso por estaciones y vías, recolocando viajeros mediante transbordos, apartando unidades averiadas para su mantenimiento, etc.

- Asignación del personal a un puesto concreto y revisar que todas las tareas estén cubiertas con el personal necesario.

1.4.2. Sistema de apoyo a la toma de decisiones

De acuerdo con la perspectiva de Nick Tyler [17], la toma de decisiones está orientada a las acciones necesarias para limitar los factores negativos y ampliar los positivos. Por tal razón el sistema que se construye en esta investigación pretende servir de soporte para que los proveedores del servicio puedan llevar a cabo las acciones pertinentes al afrontar los retos operativos en la oferta del servicio de transporte en una pandemia.

La función principal del sistema de apoyo a la toma de decisiones es anticipar de forma horaria el número de pasajeros que ingresan a una estación de la red de transporte público, utilizando un modelo predictivo.

Este sistema se enfoca en proporcionar una herramienta para que los proveedores de servicio tomen sus propias decisiones basada en la evidencia que se suministra en este trabajo investigativo. En la sección previa 1.4.1 se describen las actividades o acciones que los proveedores de servicio podrían llevar a cabo de manera informada al contar con una herramienta de predicción.

Es importante subrayar que la herramienta propuesta en este estudio de caso pretende conseguir que la población mundial pueda estar mucho más preparada ante cualquier tipo de incidente que pueda venir en un futuro.

1.4.3. Tecnologías y herramientas software

1.4.3.1. Estándar GTFS

El estándar GTFS (General Transit Feed Specification) es un formato de datos comúnmente utilizado para describir información sobre servicios de transporte público, como rutas, paradas, horarios y tarifas. Fue desarrollado por Google en colaboración con agencias de transporte público de todo el mundo [18].

1.4.3.2. Framework Dash

Para el desarrollo de la aplicación web se utiliza el framework conocido como Dash, el cual se enfoca en la elaboración de aplicaciones web sobre analítica de datos. Es diseñado por la empresa Plotly y se emplea para desplegar prototipos o incluso llevar a producción aplicaciones de alto nivel en pocas línea de código. Dash permite que el programador se desentienda de la gestión de múltiples conexiones al servidor y de la elaboración de la interfaz Web a través de lenguajes como HTML, CSS y Javascript [19].

1.4.3.3. Computación paralela Ray

Ray es un framework de código abierto para escalar aplicaciones de inteligencia artificial y aplicaciones desarrolladas en Python. En ciencia de datos es una herramienta muy útil por lo que minimiza la complejidad y ofrece soluciones de procesamiento paralelo que permiten distribuir las cargas en varios nodos y GPUs [20].

1.4.4. Modelo de Predicción

1.4.4.1. Prophet

El modelo Prophet, desarrollado por Sean J. Taylor y Benjamin Letham en colaboración con Facebook [21], es una herramienta de análisis de series temporales que se enfoca en la predicción mediante una regresión no lineal modular. Este modelo combina diferentes componentes, los cuales pueden ser configurados de manera independiente. Cada componente busca modelar la tendencia, la estacionalidad (por ejemplo, patrones semanales, mensuales, etc.) y el impacto de días festivos.

Desde una perspectiva matemática, el modelo se puede resumir a partir de la ecuación de Prophet, la cual describe la forma en como se aborda la predicción de series temporales mediante la combinación de varios componentes.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

$g(t)$: Función que representa los cambios en la tendencia general de la serie temporal, modela variaciones no periódicas (e.g. tendencia creciente, decreciente o constante).
$s(t)$: Función que modela los cambios periódicos, es decir, estacionalidades diarias, semanales, mensuales y anuales (patrones que se repiten en los datos).
$h(t)$: Función que representa los regresores (variables adicionales) en prophet que influyen en la variable objetivo (e.g. días festivos, temperatura).
ϵ_t: Término de error dentro de función $y(t)$ que trata de ajustarse a los cambios que no han sido acomodados por el modelo.

1.4.5. Métricas de evaluación

1.4.5.1. Error Absoluto Medio

El error absoluto medio, conocido como MAE por sus siglas en inglés (Mean Absolute Error), es una métrica de evaluación que calcula la media de los errores absolutos [22].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

1.4.5.2. Raíz del Error Cuadrático Medio

La raíz del error cuadrático medio, conocido como RMSE por sus siglas en inglés (Root Mean Squared Error), es una métrica de evaluación que calcula la raíz cuadrada del error cuadrático promedio de las predicciones. Esta métrica se utiliza para medir la precisión de un modelo, donde un valor más bajo indica una mayor precisión del modelo para realizar predicciones [22].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

1.4.5.3. R cuadrado

También llamado coeficiente de determinación, muestra el rendimiento relativo del modelo, no importa si los valores de salida son muy grandes o muy pequeños, R^2 siempre estará entre $-\infty$ y 1 [22].

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)}$$

$$MSE(model) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$MSE(baseline) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$$

y_i: Valor real
$\hat{y}_i$: Predicción del modelo
$\bar{y}_i$: Valor constante que puede ser la media de los valores reales

1.4.5.4. Error absoluto medio porcentual

El error absoluto medio porcentual, conocido como MAPE por sus siglas en inglés (Mean Absolute Percentage Error), es una métrica de evaluación que calcula el promedio en porcentaje de los errores absolutos [23].

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

1.5. Metodología CRISP-DM

CRISP-DM es una metodología en el campo de la ciencia de datos y minería de datos que incluye descripciones de las fases normales de un proyecto de minería de datos, así como las tareas necesarias en cada una de las fases. Esta metodología se ha convertido en un marco de referencia ampliamente utilizado en proyectos de minería de datos y análisis predictivo.

Esta metodología proporciona un enfoque estructurado en seis fases que sintetiza el ciclo de vida de un proyecto. A continuación se describen las fases de la metodología CRISP-DM.

- **Fase 1 - Comprensión del Negocio:** Se enfoca en la comprensión de los objetivos del proyecto con el fin de obtener un plan preliminar que permita alcanzar dichos objetivos.
- **Fase 2 - Entendimiento de los datos:** Se centra en el entendimiento de los datos, realizando actividades que permitan familiarizarse con estos.
- **Fase 3 - Preparación de los datos:** Abarca todas las actividades necesarias para obtener el conjunto final de datos que se utilizarán en las herramientas de modelado.
- **Fase 4 - Modelado:** Se hace la selección y aplicación de las técnicas de modelado que permiten resolver el problema en cuestión.
- **Fase 5 - Evaluación:** Antes de proceder al despliegue final del modelo se evalúa el modelo obtenido comparándolo con los objetivos del proyecto.
- **Fase 6 - Despliegue:** Se organiza el conocimiento adquirido para presentar los resultados de la investigación.

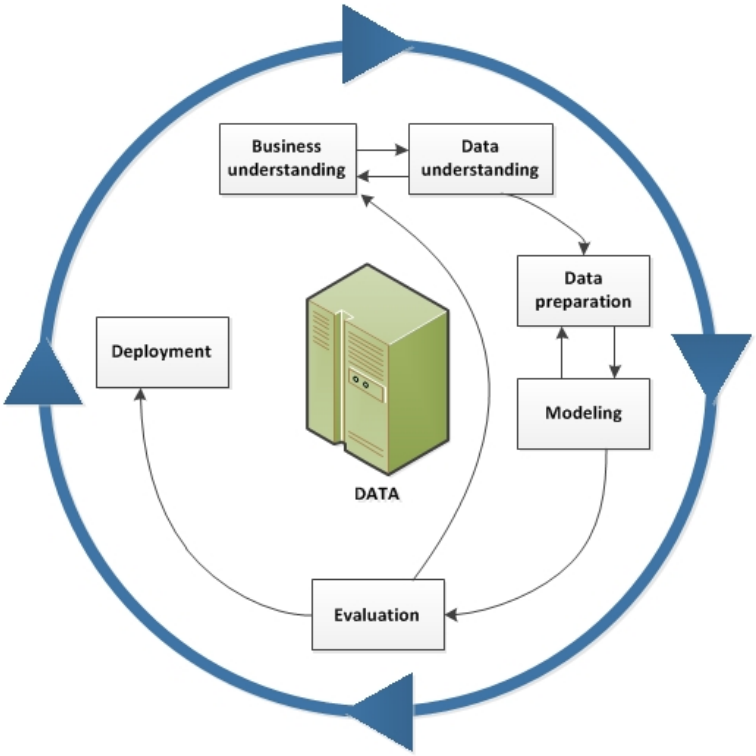

Figura 1: Ciclo de vida de minería de datos [24]

1.6. Diseño del sistema

Considerando que el objetivo del proyecto de investigación es apoyar la toma de decisiones en la planificación de la oferta de los servicios de cercanías y metro de Madrid, se plantea desarrollar un sistema capaz de anticipar el número de entradas de pasajeros en una de las estaciones de los servicios de transporte y en una hora determinada. Si el sistema posee esta capacidad de predicción es posible planificar de forma precisa los viajes realizados por parte del transporte público masivo en la ciudad de Madrid, ya que los proveedores de servicio tomarán sus propias decisiones de acuerdo a las necesidades que se identifiquen de forma anticipada en dicha estación. Aquellas decisiones incluyen el reajuste de horarios, frecuencias, composición de los trenes, asignación de personal y otras actividades, como se detalla en la sección 1.4.1.

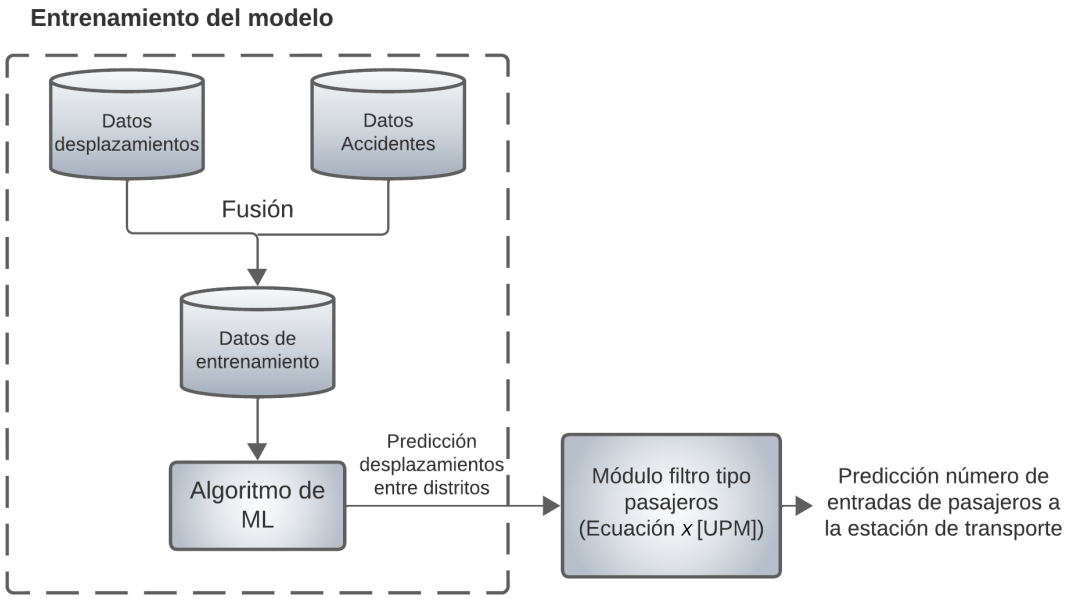

Figura 2: Entrenamiento del modelo [Fuente propia]

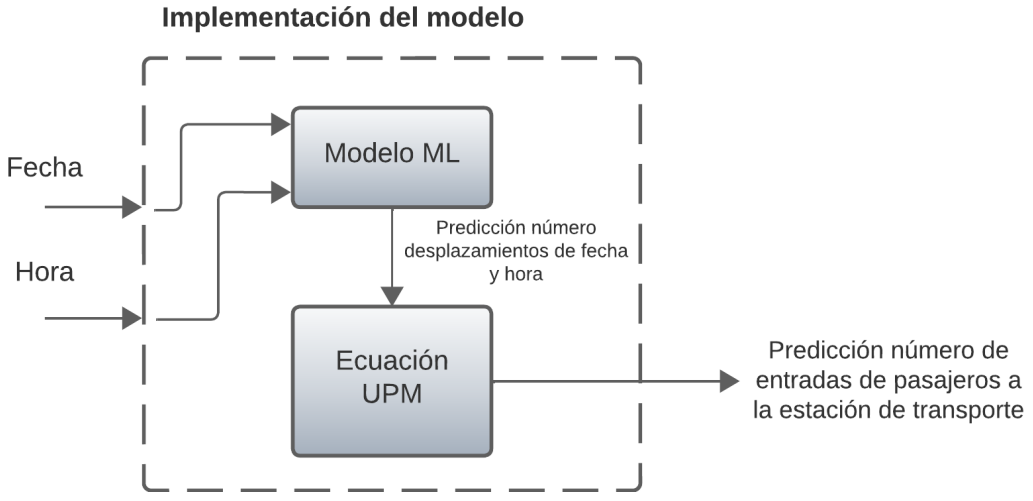

Figura 3: Implementación del modelo [Fuente propia]

La estructura del sistema se puede resumir en tres capas: datos, procesamiento y visualización. La arquitectura del sistema de la figura 4 muestra cada uno de los módulos, donde cada uno de ellos corresponde a un script de python.

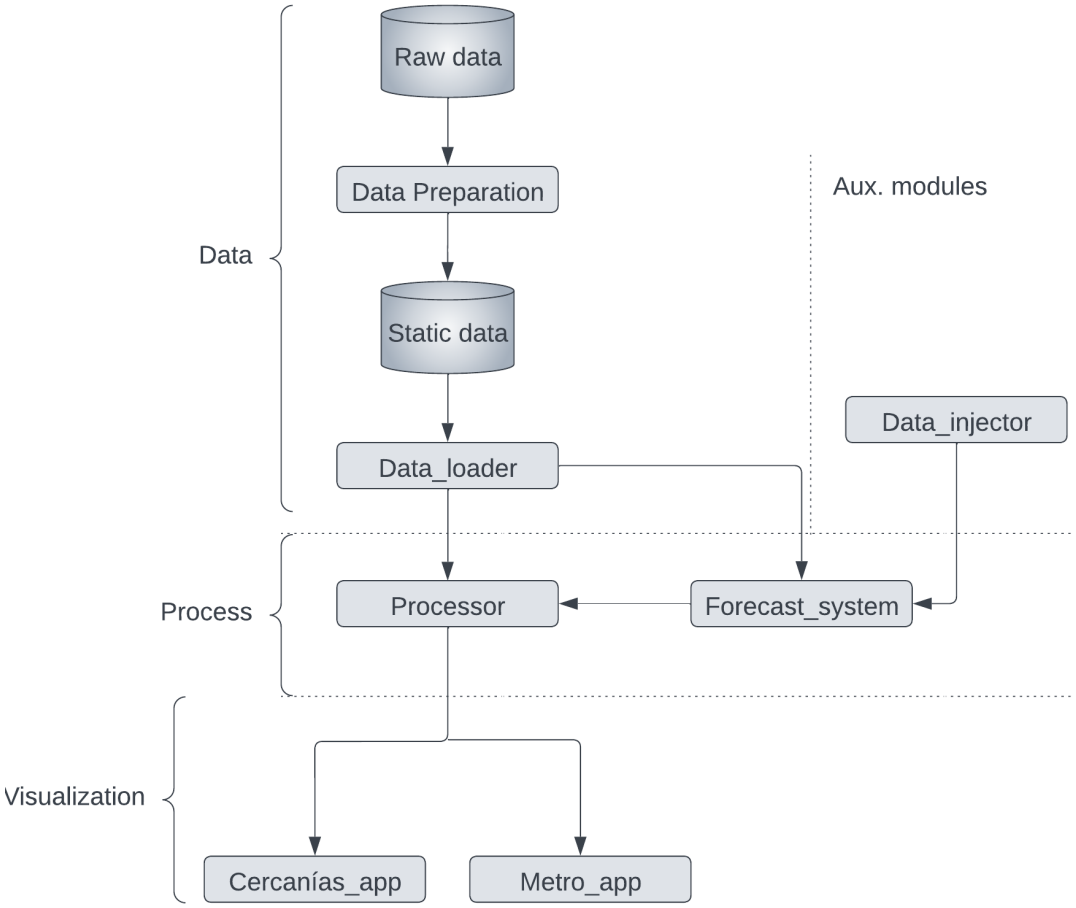

Figura 4: Arquitectura del sistema [Fuente propia]

- Raw data: Este módulo hace referencia a los datos que se utilizan para entrenar el modelo de predicción. Principalmente se emplean los datos de movilidad entre distritos. También y con el propósito de mejorar el modelo se emplean datos meteorológicos y datos de accidentes de tránsito.

- Data Preparation: Este módulo refleja el proceso de la Fase 3 de la metodología CRISP-DM, en el cual se llevan a cabo las actividades necesarias para adquirir el conjunto definitivo de datos que se utiliza para entrenar el modelo.

- Static data: Este módulo hace referencia al conjunto final de datos utilizado para entrenar los modelos.

- Data loader: Mediante este módulo se cargan los archivos de extensión “csv”, que corresponden a los datos necesarios para entrenar los modelos y los datos necesarios para realizar la estimación del número de entradas de pasajeros a una estación.

- Forecast System: En este módulo, se encuentran los modelos que han sido en-

trenados utilizando los datos de desplazamientos entre distritos y los datos de accidentalidad. Si el sistema se está ejecutando por primera vez, estos modelos se entrenan en un único ciclo de entrenamiento.

- Processor: Este módulo es muy importante en el sistema, ya que permite obtener la estimación del número de usuarios que entran a una estación a partir de las predicciones obtenidas por los modelos.
- Cercanías App: Este módulo corresponde al archivo python ejecutable que contiene la aplicación de predicción para el servicio de tren de Cercanías de Madrid.
- Metro App: Este módulo corresponde al archivo python ejecutable que contiene la aplicación de predicción para el servicio de Metro de Madrid.
- Data Injector: Este módulo auxiliar simula la inyección de datos en tiempo real, proporcionando información sobre movilidad y accidentalidad. Su propósito es sustituir la funcionalidad del sistema para inyectar datos en una futura pandemia.

1.7. Estructura del documento

Este documento ha sido organizado de acuerdo a la metodología CRISP-DM. Esta metodología, ampliamente utilizada en proyectos de análisis de datos, proporciona un enfoque estructurado por seis fases que permite descubrir conocimiento a partir de datos. A continuación se describe la estructura del documento.

- **Capítulo 1 - Introducción:** En este capítulo, se realiza el planteamiento del problema y se definen los objetivos del proyecto. Además, se presenta el estado actual del conocimiento a través de la sección de estado del arte. Asimismo, se incluye el marco teórico, que describe conceptos clave, tecnologías y herramientas utilizadas en el desarrollo del proyecto. Por último se presenta el diseño del sistema que muestra la arquitectura del sistema desarrollado.

- **Capítulo 2 - Fase 1: Comprensión de la información relevante para la planificación del servicio de transporte público en Madrid** En este capítulo, se desarrolla la fase 1 de la metodología, la cual tiene como finalidad la recopilación información para abordar el problema identificado. A través de esta información, se logra identificar los tipos de datos disponibles para el análisis y dar mayor claridad a los objetivos planteados.

- **Capítulo 3 - Fase 2: Entendimiento de los datos y su relación con la planificación del servicio de transporte público** En este capítulo, se recopilan y se estudian de cerca los datos disponibles para la minería de datos. En esta fase, se describen y se exploran los datos con la ayuda de tablas y de gráficos. Finalmente se realiza una verificación de la calidad de los datos, ya que la mayoría de los datos contienen errores de codificación.

- **Capítulo 4 - Fase 3: Preparación de los datos utilizados para la planificación del servicio de transporte público en Madrid** En este capítulo, se realiza la selección y limpieza de los datos relevantes. Además, se generan nuevos datos a partir de las características previamente seleccionadas. Finalmente se integran los diferentes conjuntos de datos en uno solo, obteniendo así el conjunto final de datos que se utilizan en la fase de modelado.

- **Capítulo 5 - Fase 4: Desarrollo y entrenamiento de un modelo de Machine Learning para la planificación del servicio de transporte público** En este capítulo, se exploran diversos algoritmos de regresión con el objetivo de abordar el problema en cuestión. Luego, se procede a entrenar los modelos y se lleva a cabo la validación de cada uno de estos modelos, con el propósito de seleccionar el modelo final que demuestre mejor rendimiento.

- **Capítulo 6 - Fase 5: Evaluación de la precisión del modelo en el estudio de caso de tren de cercanías y Metro de Madrid** En este capítulo, se evalúa si los resultados del modelo contribuyen a la toma de decisiones en la planificación de la oferta del servicio en el estudio de caso de tren de cercanías y Metro de Madrid. Debido a que el

- **Capítulo 7 - Fase 6: Visualización de los datos mediante una aplicación web** En este capítulo, se organiza el conocimiento adquirido para construir el sistema predictivo mediante una aplicación web.

- **Capítulo 8 - Conclusiones y líneas de trabajo futuras** Se presentan las conclusiones del trabajo de investigación y los trabajos futuros.

En la tabla 2 se presenta la relación entre las fases de la metodología y los objetivos del proyecto. Se logra apreciar en esta tabla que al implementar la metodología CRISP-DM se le da cumplimiento a todos los objetivos del proyecto.

Objetivos / Fases	OE1	OE2	OE3
Fase 1	X		
Fase 2	X	X	
Fase 3	X		
Fase 4		X	
Fase 5			X
Fase 6			X

Tabla 2: Relación de la metodología y el cumplimiento de los objetivos [Fuente propia]

CAPÍTULO 2

Fase 1: Comprensión de la información relevante para la planificación del servicio de transporte público en Madrid

En este capítulo, se desarrolla la fase 1 de la metodología CRISP-DM, la cual tiene como finalidad la recopilación información para abordar el problema identificado. A través de esta información, se logra identificar los tipos de datos disponibles para el análisis y dar mayor claridad a los objetivos planteados.

2.1. Organización territorial

España está compuesta por 17 comunidades autónomas, entre las cuales se encuentran Cataluña, Andalucía, Comunidad de Madrid, Comunidad Valenciana y Galicia. La Comunidad Autónoma de Madrid, situada en el centro del país, tiene a Madrid como capital, siendo esta también la capital de España. Además, en esta comunidad autónoma se ubican 179 municipios [25]. El objetivo de este trabajo de investigación es analizar en detalle el municipio de Madrid, una entidad administrativa que abarca la extensa ciudad de Madrid. Como punto de partida, es importante destacar que el municipio se divide en 21 distritos, que son unidades administrativas más pequeñas encargadas de la gestión de los servicios locales [26].

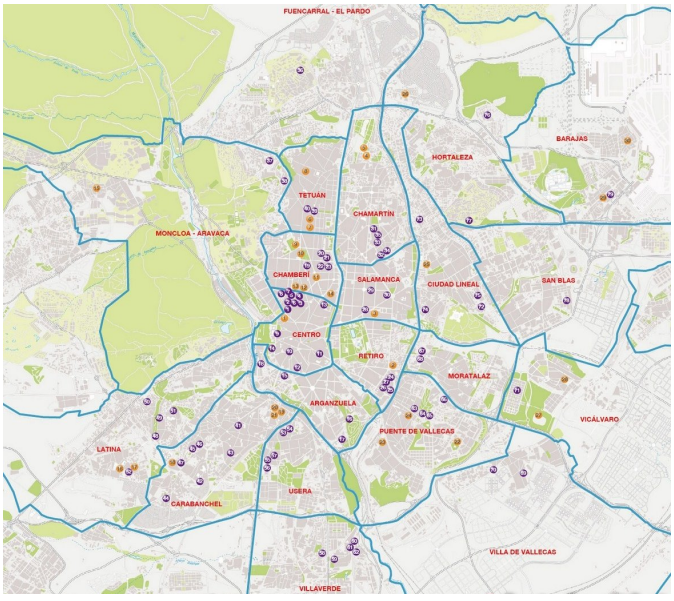

Figura 5: Mapa de distritos en la ciudad de Madrid [27]

2.2. Transporte Público en Madrid

El servicio de transporte público en Madrid es uno de los sistemas de transporte más extensos y desarrollados de España y Europa. La red de transporte público en Madrid está compuesta por una combinación de autobuses urbanos e interurbanos, metro, trenes de cercanías, metro ligero y taxis.

El Metro de Madrid es una parte fundamental del sistema de transporte público de la ciudad. Es uno de los metros más extensos del mundo, con una red de aproximadamente 300 kilómetros y 302 estaciones que conectan toda la ciudad y sus alrededores. El metro cuenta con 13 líneas que cubren áreas urbanas y suburbanas, lo que permite un transporte rápido y eficiente [28]. Además del metro, la ciudad cuenta con un sistema de trenes de cercanías, conocido como Cercanías, es otra opción de transporte público muy utilizada en Madrid. Conecta la ciudad con las áreas periféricas y los municipios cercanos mediante 10 líneas, facilitando los desplazamientos diarios de los residentes y trabajadores de la región [29].

Complementando estas opciones de transporte público, la ciudad también cuenta con una red de 35 kilómetros de metro ligero desde mediados de 2007, tras la desaparición de los tranvías de Madrid en 1972 [30] y con una amplia red de autobuses urbanos e interurbanos que cubren todas las áreas de Madrid y sus alrededores. Los autobuses urbanos son operados por la Empresa Municipal de Transportes de Madrid (EMT), mientras que los autobuses interurbanos conectan Madrid con otras ciudades y pueblos de la Comunidad de Madrid y provincias cercanas [31].

2.3. El panorama de la pandemia

La pandemia conllevo a un descenso de la movilidad en general pero el modo de transporte que más resultó afectado por la pandemia fue el transporte público con una disminución del 46 % en la media anual del número de viajeros, dicha disminución se debe a restricciones de movilidad, confinamientos y aumento de teletrabajo [8].

A continuación en la gráfica de la figura 6 se analiza el comportamiento de la movilidad, comparando el volumen de desplazamientos con el mes de febrero previo a la pandemia.

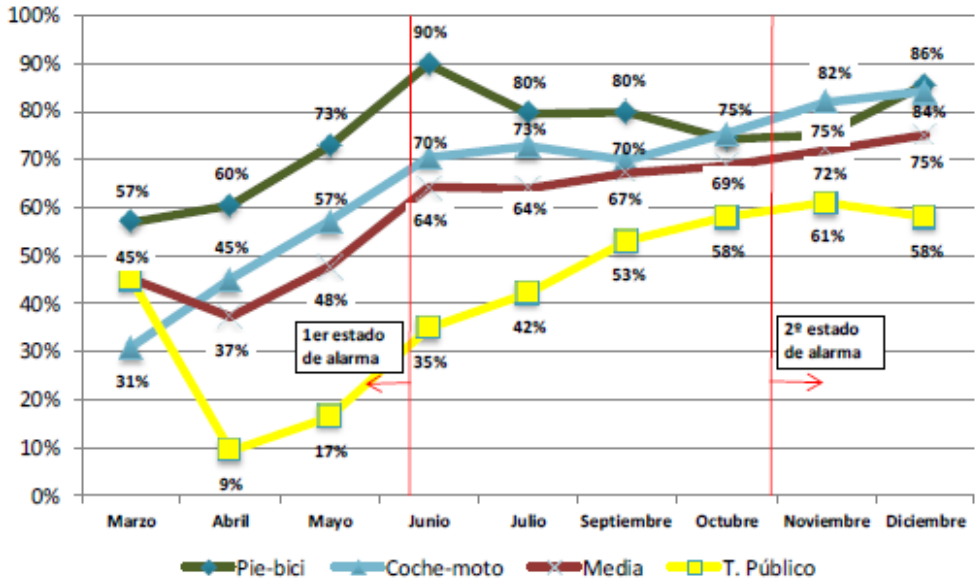

Figura 6: Porcentaje de desplazamientos respecto a los realizados antes de la pandemia (febrero 2020) [8]

A partir de la figura 6, se puede observar una drástica caída en los desplazamientos de los diferentes modos de transporte en marzo, alcanzando un valor del 45 % en el caso del transporte público. En abril, se registra la cifra más baja de desplazamientos en este modo de transporte, con una disminución del 9 % con respecto a febrero de 2020. Además, se nota que a lo largo de 2020, los desplazamientos no logran recuperar los niveles previos a la pandemia, especialmente en el caso del transporte público, que alcanza un máximo de desplazamientos del 61 % durante este año.

En el artículo ”Los datos de una pandemia en tres olas“ del portal web EDICIONES EL PAIS, se detallan las tres olas de la pandemia de COVID-19 en España [32].

La primera ola comenzó en marzo de 2020 y duró hasta junio del mismo año. Durante esta ola, España tuvo uno de los peores brotes de COVID-19 en Europa y el mundo. El número de casos y muertes aumentó rápidamente y el sistema sanitario se vio abrumado. El mayor pico de ingresos hospitalarios se presentó en abril.

La segunda ola comienza a finales de junio y dura hasta diciembre del mismo año. Durante esta ola, España implementó medidas más estrictas para controlar la propagación del virus. A pesar de esto, el número de casos y muertes aumentó nuevamente.

Durante la tercera ola de la pandemia en España, que comenzó en diciembre de 2020, se observó un aumento significativo en los desplazamientos antes de navidad. También se observa un aumento en los ingresos hospitalarios y un aumento de las personas que son ingresadas a la unidad de cuidados intensivos (UCI).

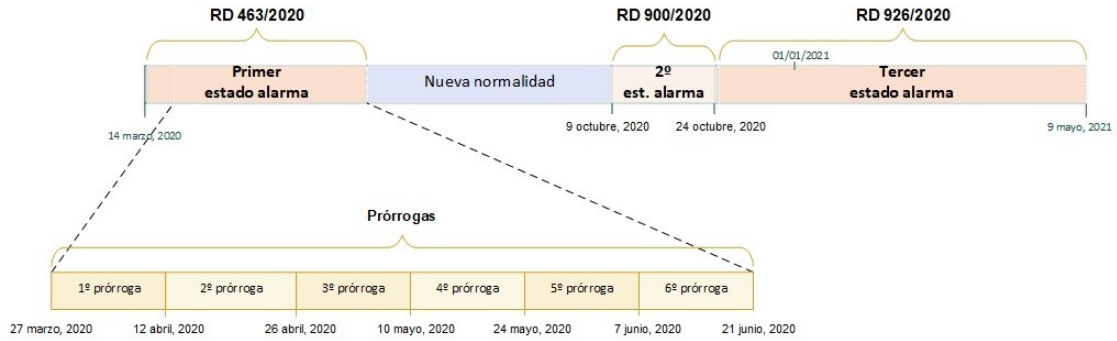

Figura 7: Diagrama cronológico de los estados de alarma durante la pandemia en Madrid [33]

El artículo “Pandemia de covid-19 en España ” detallan la evolución de la pandemia y las medidas tomadas por el gobierno español [34]. El 14 de marzo de 2020, el gobierno español declaró el estado de alarma en todo el territorio nacional y limitó la libre circulación de los ciudadanos a actos esenciales como la adquisición de alimentos y medicamentos o acudir a centros médicos o al lugar de trabajo, resultando en un confinamiento de la población en sus lugares de residencia.

Después de 42 días de confinamiento, es decir, el 26 de abril inicia la primera medida de alivio de la cuarentena y el 28 de abril se aprueba un plan de desescalada que se dividió en tres fases. Durante estas fases, se levantaron gradualmente las restricciones impuestas durante el confinamiento. Si bien el gobierno había estipulado que cada una de las fases debía durar como mínimo dos semanas, algunas comunidades autónomas deciden ir más rápido por lo que omiten la fase tres. No obstante, algunas zonas de España tienen que adoptar medidas de prevención rápidamente para evitar rebrotes.

El estado de alarma se prolongó hasta el 21 de junio de 2020. A partir del 21 de junio, España entró en una fase de nueva normalidad. En la comunidad de Madrid, el 9 de octubre de 2020 se decreta un estado de alarma en nueve municipios para hacer frente a una nueva escalada de casos por coronavirus. El estado de alarma en la comunidad de Madrid acaba el 24 de octubre.

El 25 de octubre de 2020, el gobierno español declaró un nuevo estado de alarma a nivel nacional debido al aumento del número de casos. Se impuso el toque de queda desde las 11 de la noche hasta las 6 de la mañana y se prohibieron las reuniones de más de seis personas. Adicionalmente, el estado de alarma se estableció como instrumento legal para permitir a las comunidades autónomas cerrar sus fronteras y realizar confinamientos parciales o totales. Este estado de alarma se prolongó hasta el 9 de mayo de 2021.

El 27 de diciembre de 2020 comienza la vacunación en España, con la vacuna de Pfizer y BioNTech. Posteriormente se sumaron la vacuna de Moderna en enero de 2021, la vacuna de AstraZeneca-Oxford en febrero y la vacuna de Jannsen en abril. El 9 de mayo de 2021 finaliza el estado de alarma en España, levantando todas las restricciones impuestas durante la pandemia y se permitió la libre circulación dentro del territorio nacional.

2.4. Información del sistema de transporte público de metro y cercanías en Madrid

Con el propósito de visualizar y analizar la estructura del sistema de transporte público en la ciudad de Madrid, se empleó el software QGIS para trazar las líneas del metro y tren de cercanías, permitiendo así identificar los distritos que atraviesan dichas líneas.

La información geográfica que permite visualizar la división territorial por distritos se encuentra contenida en el archivo shapefile del directorio “zonificación_distritos” presentado en el anexo A.2. Además, la información correspondiente al transporte público, se extrae de los datos GTFS mencionados en el anexo A.1. Estos datos permitirán visualizar la ubicación geográfica de las estaciones, así como también la geometría de los trayectos de cada una de las líneas.

2.4.1. Líneas y estaciones de la red de metro y cercanías de Madrid

Gracias al análisis de los datos GTFS, se han generado mapas detallados que muestran tanto la red del metro de Madrid como la red del tren de cercanías. En estos mapas también se representan los distritos de la ciudad de Madrid, así como algunos municipios cercanos dentro de la Comunidad de Madrid.

En el mapa de la red de metro representado en la figura 8, se puede apreciar claramente cómo el servicio de metro no se limita solo a la ciudad de Madrid, sino que se extiende hacia municipios cercanos. Esta extensión abarca localidades como Móstoles, ubicado al suroeste de la capital, Fuenlabrada y Getafe al sur, Arganda del Rey al sureste y San Sebastián de los Reyes al norte.

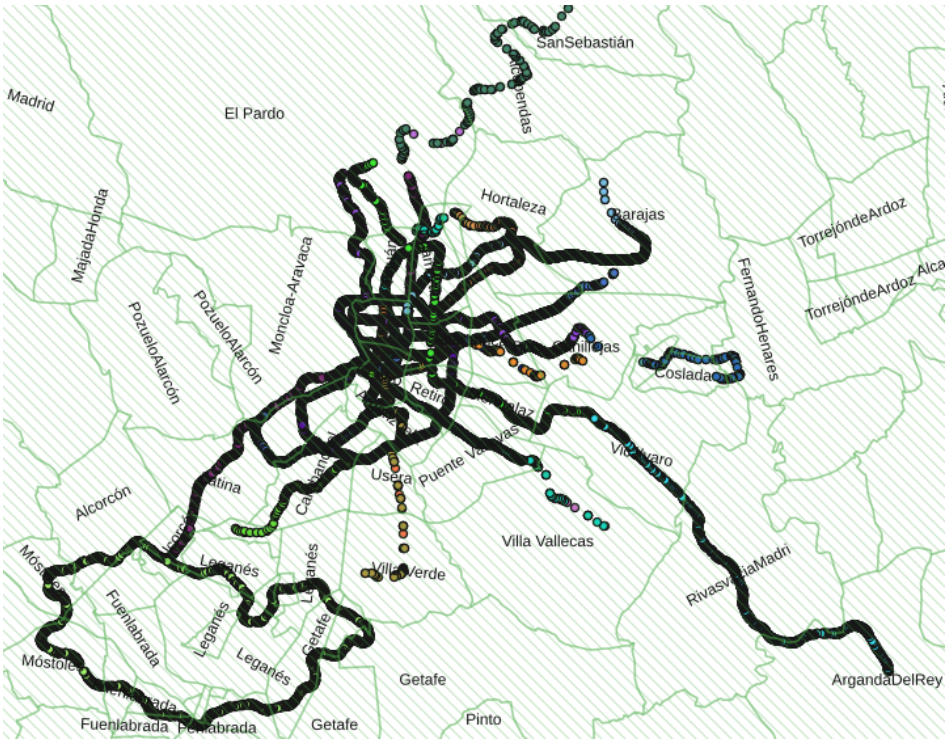

Figura 8: Mapa de la red de metro de Madrid y su contexto territorial [Fuente propia]

El mapa de la red de trenes de cercanías, presentado en la figura 9, muestra la conexión de diferentes municipios con la capital. El servicio de tren de cercanías se extiende hacia municipios más alejados de la capital, como lo son el caso de Cercedilla, Guadalajara y Aranjuez.

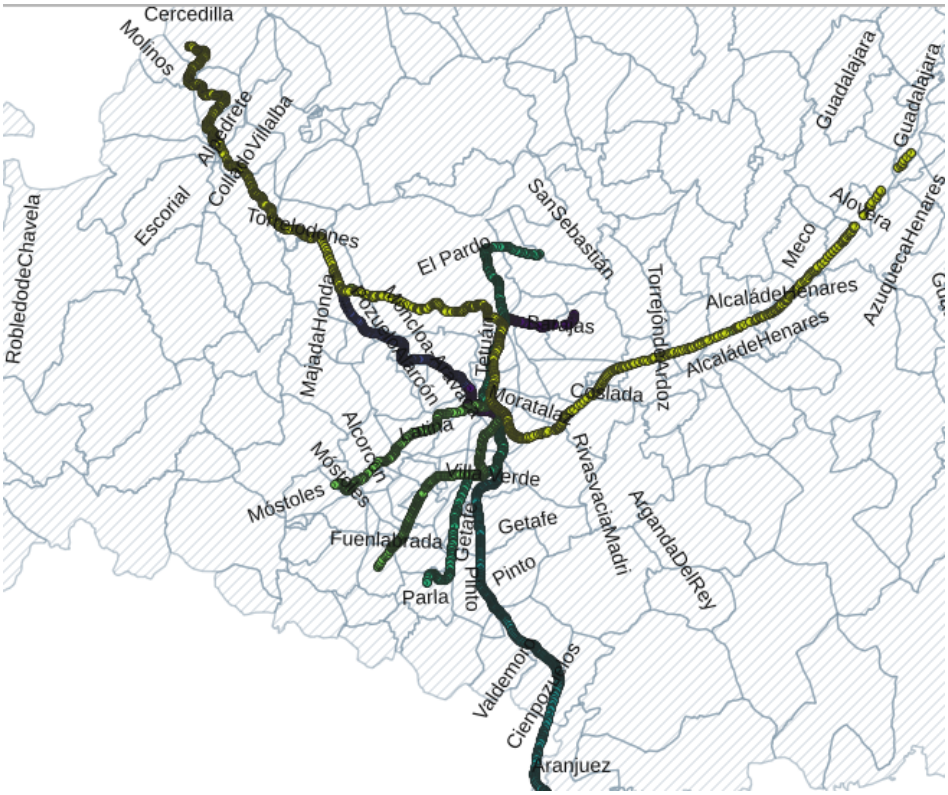

Figura 9: Mapa de la red de cercanías de Madrid y su contexto territorial [Fuente propia]

Elaboración propia: Los archivos que generan las figuras 8, 9, 10, 11 y 12, se encuentran disponibles en el repositorio de github [35] dentro del directorio GTFS – 2023

Con una visión completa de toda la red de metro y tren de cercanías de Madrid, se observa aquellas líneas más importantes para la movilidad de la ciudad, a partir de esta observación se decide enfocar el estudio en una de las estaciones de la línea 1 del metro y en una de las estaciones de la línea C4 del tren de cercanías de Madrid.

La línea 1 y sus estaciones son un segmento importante de la red metro, ya que en su recorrido atraviesa zonas residenciales, comerciales y turísticas, lo que la convierte en una opción ideal para el desplazamiento rápido por la ciudad. A través del mapa de la figura 10, se puede observar que la línea comienza en el norte de la ciudad en la estación Pinar de Chamartín y se extiende hacia el sureste, llegando hasta la estación de Valdecarros.

Figura 10: Mapa de la línea 1 del metro de Madrid y sus estaciones [Fuente propia]

La línea C4 tiene un recorrido que va desde la estación Alcobendas, ubicada al norte de Madrid y finaliza en la estación de Parla, al sur de la capital. En su trayecto atraviesa distintas estaciones importantes como Chamartín, Sol y Atocha, lo que la convierte en una línea relevante para aquellas personas que desean desplazarse desde las zonas periféricas hacia el centro de Madrid. Asimismo, ofrece una conveniente conexión para aquellos que residen en los municipios cercanos a la capital y desean desplazarse al área central de la ciudad.

En el mapa de la figura 11 se muestra la extensión de la línea C4 y todas sus estaciones. En su recorrido, esta línea atraviesa un total de 15 estaciones, algunas de las cuales ofrecen la posibilidad de hacer la conexión con el servicio de metro.

Elaboración propia: Los archivos que generan las figuras 8, 9, 10, 11 y 12, se encuentran disponibles en el repositorio de github [35] dentro del directorio GTFS – 2023

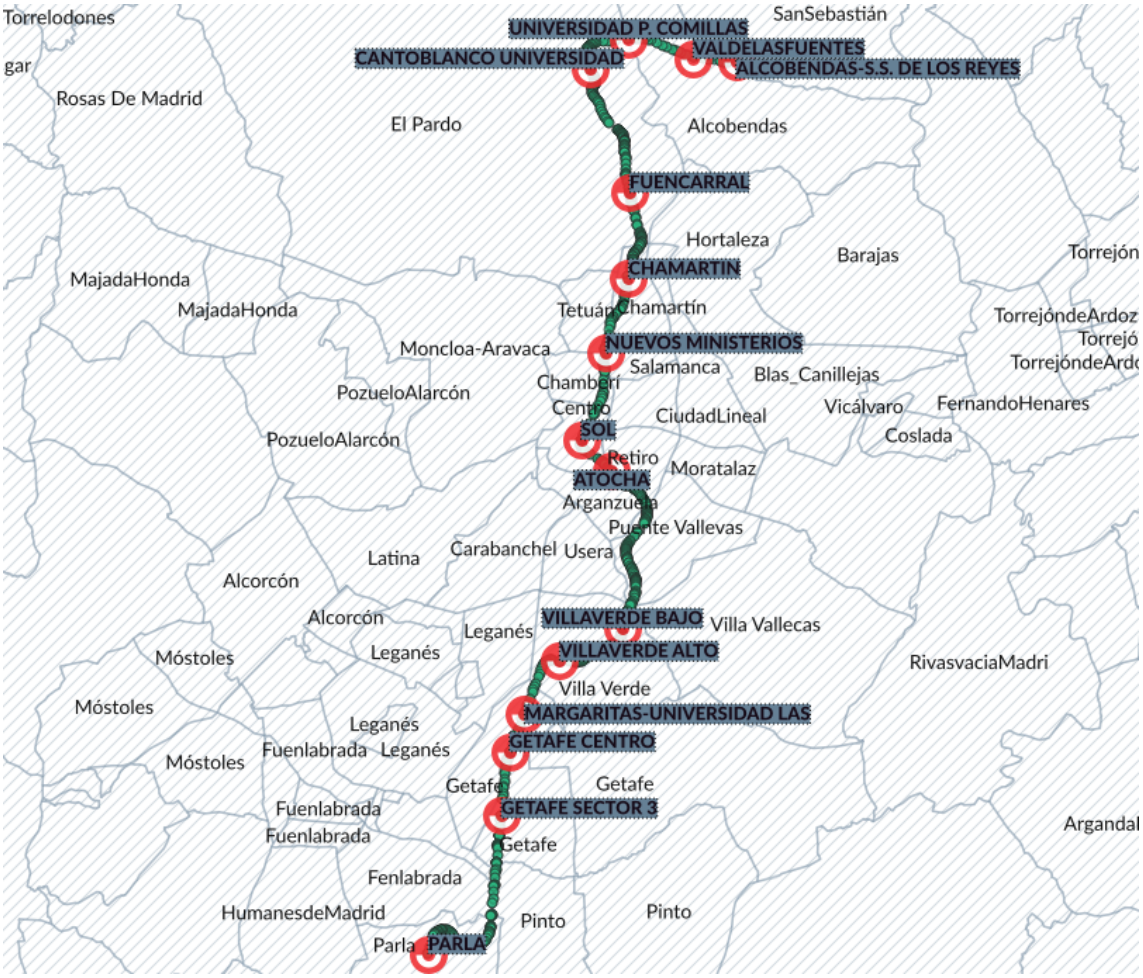

Figura 11: Mapa de la línea C4 del tren de cercanías de Madrid y sus estaciones [Fuente propia]

2.4.2. Selección de la estación para la predicción del número de entradas de pasajeros en los servicios de tren de cercanías y metro

En el mapa de la figura 12 se muestran las estaciones que tienen conexión entre los servicios de metro y tren de cercanías. La línea 1 de metro es representada por el color azul y la línea C4 de Cercanías se visualiza con un color rojo. Se puede observar de esta figura que las estaciones Chamartín, Puerta del Sol y Atocha tienen la misma ubicación geográfica para ambos servicios de transporte. Esto es clave en la red de transporte de la ciudad al permitir conexiones convenientes entre ambas líneas.

Los desplazamientos entre el distrito Chamartín y el distrito Centro son de gran relevancia, por un lado, en el distrito Chamartín se encuentra una de las principales estaciones de transporte de la ciudad, la cual recibe el mismo nombre del distrito. Esta estación se destaca por su carácter intermodal, ya que integra diversos modos de transporte en un mismo lugar. Por otro lado, en el distrito centro se encuentra la estación “Puerta del Sol”, la cual recibe gran afluencia de personas debido a la gran actividad comercial, cultural y turística de esta localidad.

Elaboración propia: Los archivos que generan las figuras 8, 9, 10, 11 y 12, se encuentran disponibles en el repositorio de github [35] dentro del directorio GTFS – 2023

Figura 12: Estaciones de intercambio entre la línea 1 del metro y la línea C4 de Cercanías [Fuente propia]

De acuerdo al análisis de la información geográfica del sistema de transporte de Madrid, se ha definido el alcance del trabajo de investigación de la siguiente manera: Predicción horaria del número de entradas de pasajeros en la estación Chamartín, para lo cual se llevará a cabo el entrenamiento de un modelo predictivo para el trayecto comprendido entre el distrito Chamartín y el distrito Centro.

Es importante destacar que cada trayecto (e.g. Chamartín-Centro, Chamartín-Chamberí, Centro-Retiro, etc.), sigue un patrón de desplazamiento específico, influenciado por la naturaleza de cada uno de los distritos, ya sea residencial, turística o comercial.

En consecuencia, para llevar a cabo una predicción horaria precisa en la estación Chamartín a partir de los datos de desplazamientos entre distritos, fue esencial llevar a cabo un análisis específico por trayecto, en este caso se analiza el trayecto Chamartín-Centro. Por lo que para extender el análisis a otra estación de servicio, resulta necesario entrenar otros modelos para los trayectos que involucren a la estación de estudio.

Elaboración propia: Los archivos que generan las figuras 8, 9, 10, 11 y 12, se encuentran disponibles en el repositorio de github [35] dentro del directorio GTFS – 2023

CAPÍTULO 3

Fase 2: Entendimiento de los datos y su relación con la planificación del servicio de transporte público

En este capítulo, se recopilan y se estudian de cerca los datos disponibles para la minería de datos. En esta fase, se describen y se exploran los datos con la ayuda de tablas y de gráficos. Finalmente se realiza una verificación de la calidad de los datos, ya que la mayoría de los datos contienen errores de codificación.

3.1. Datos reales de desplazamiento entre distritos

La recopilación de los datos de desplazamiento entre distritos se hizo a partir de los datos que proporciona el gobierno de España (ver Anexo A.2). Estos datos de desplazamiento corresponden a datos diarios de movilidad en las ciudades y municipios de todo el país, para la recolección de estos datos se utiliza una metodología relativamente novedosa que utiliza el posicionamiento de los teléfonos móviles, por lo cual los datos de desplazamientos representados en este dataset corresponden a desplazamientos realizados en vehículos particulares, bicicletas, taxis, transporte urbano, etc.

Los datos de desplazamientos están disponibles para descargar de acuerdo al mes que se vaya a estudiar. En esta fase se describe la estructura de los datos de un día del mes de febrero del año 2020, considerando que todos los archivos tienen la misma estructura.

Los archivos descargados tienen la estructura de nombre “AAAAMMDD_maestra_1_mitma _distrito.txt”, donde los números al inicio del nombre indican el año, mes y día. Cada uno de los archivos tiene un peso aproximado de 400 MB y contiene los desplazamientos entre los diferentes distritos y municipios de todo el país. Para simplificar el manejo de los datos, se lleva a cabo la conversión de estos datos en archivos CSV.

Con el objetivo de proporcionar una comprensión clara de los datos de estudio, a continuación, se presenta la estructura del archivo “20200214_maestra_1_mitma _distrito.csv” a través de tablas y gráficos. En la figura 13 se observan todas las columnas y cinco registros mediante el método head() de la librería pandas.

	fecha	origen	destino	actividad_origen	actividad_destino	residencia	edad	periodo	distancia	viajes	viajes_km
0	20200214	01001_AM	01001_AM	casa	otros	1	NaN	0	002-005	9.775	35.601
1	20200214	01001_AM	01001_AM	casa	otros	1	NaN	0	005-010	8.754	72.488
2	20200214	01001_AM	01001_AM	casa	otros	1	NaN	1	002-005	8.431	30.339
3	20200214	01001_AM	01001_AM	casa	otros	1	NaN	1	010-050	9.775	100.837
4	20200214	01001_AM	01001_AM	casa	otros	1	NaN	2	002-005	6.007	23.952

Figura 13: Estructura de los datos de desplazamiento entre distritos [Fuente propia]

En la figura 14 se muestra la dimensión del archivo que corresponde al 14 de febrero del

Elaboración propia: El archivo que genera la figura 49 se encuentra disponible en el repositorio de github [35] con el nombre de Tablas_Covid_19_vs_Desplazamientos_1.ipynb – 2023

año 2020. Además se puede observar el tipo de dato de cada una de las columnas del archivo.

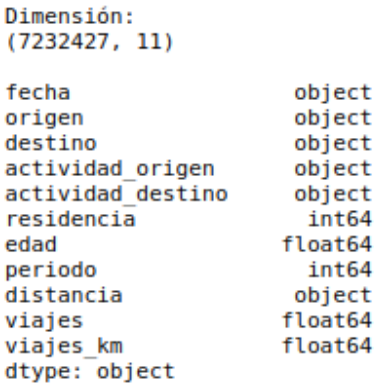

Figura 14: Características del archivo CSV [Fuente propia]

Teniendo en cuenta la información presentada en las figuras 13 y 14, se realiza a continuación una descripción detallada de las columnas presentes en el archivo CSV.

- **fecha:** En esta columna se registran el año, mes y día en que se produce el desplazamiento entre distritos.
- **origen:** En esta columna se registra el código del distrito de origen del desplazamiento. Se identifican 2839 códigos únicos en este campo.
- **destino:** En esta columna se registra el código del distrito de destino del desplazamiento. Se identifican 2839 códigos únicos en este campo.
- **actividad_origen:** En esta columna se describe la actividad asociada al desplazamiento de origen, la cual puede ser “casa”, “trabajo” u “otra”.
- **actividad_destino:** En esta columna se describe la actividad asociada al desplazamiento de destino, la cual puede ser “casa”, “trabajo” u “otra”.
- **residencia:** No se especifica en el formato de archivo.
- **edad:** Edad de la persona asociada al desplazamiento. Todos los registros de esta columna tienen el valor NaN.
- **periodo:** Esta columna corresponde a intervalos de una hora e indica la hora en que se efectúa el desplazamiento, con valores que van desde 0 hasta 23. Siendo 3 el periodo que corresponde con el intervalo entre 03:00 y las 04:00.
- **distancia:** Esta columna hace referencia a la distancia recorrida y se clasifica en seis intervalos: 500m – 2km, 2km-5km, 5km-10km, 10km-50km, 50km-100km y +100km.
- **viajes:** Esta columna corresponde al número de viajes realizados.
- **viajes_km:**Esta columna corresponde al número de viajes por kilómetro.

Es importante resaltar que los campos están separados por ’|’ (barra vertical) y los valores numéricos tienen ’.’ (punto) como separador de decimales.

Al realizar una primera exploración de los datos, se logra identificar que los campos más relevantes para esta investigación son: “fecha”, “origen”, “destino”, “periodo”, “viajes” y “distancia”. Estas variables adquieren relevancia en este estudio, ya que permiten construir una serie temporal que proporciona información detallada sobre los patrones presentes en los datos.

Sin embargo en la figura 15 se lleva a cabo la medición de la correlación entre las diferentes variables numéricas del conjunto de datos. Esto se realiza con el propósito de determinar si alguna de las variables que no se están considerando tienen una relación importante con la variable “viajes”, que es la de mayor relevancia.

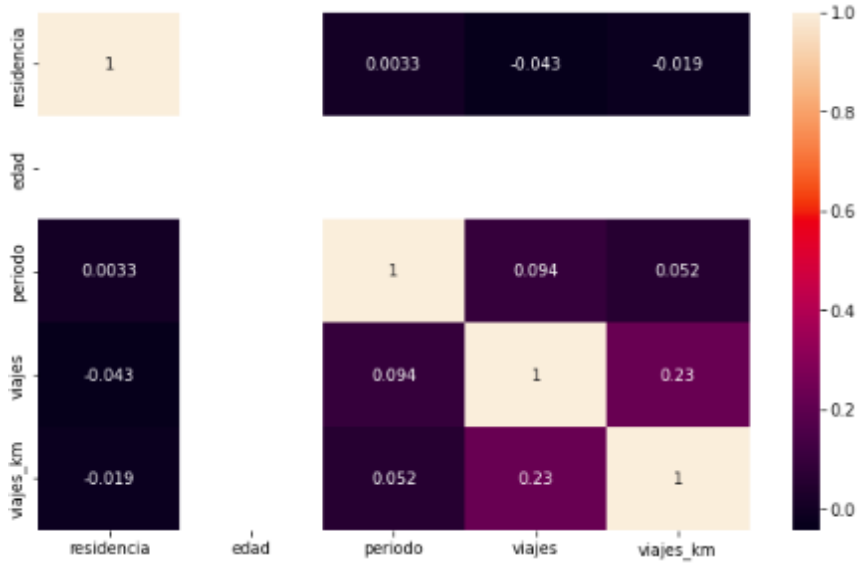

Figura 15: Correlación de las variables numéricas del conjunto de datos [Fuente propia]

Al observar la figura 15, se logra determinar que no existe correlación entre las variables numéricas del conjunto de datos, por lo que se decide descartar las variables “viajes_km”, “edad” y “residencia”, las cuales no son útiles para construir la serie temporal y no guardan una relación importante con los desplazamientos entre distritos.

Teniendo en cuenta lo anterior, en las figuras 16 y 17 se elaboran diagramas de caja o “BoxPlot” para los siete millones de registros que contiene el archivo CSV. A través de este diagrama, se busca analizar los valores numéricos del campo “’viajes’ y su relación con los campos “periodo” y “distancia”.

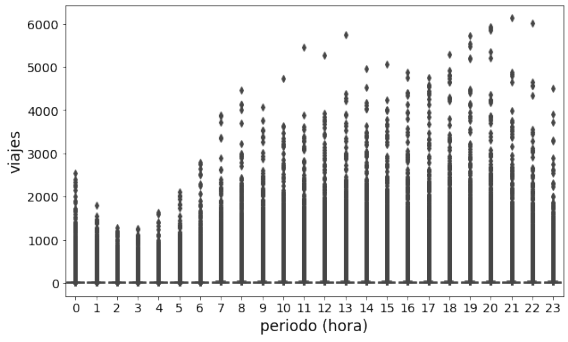

Figura 16: Diagrama de caja: Número de viajes por hora - todos los trayectos [Fuente propia]

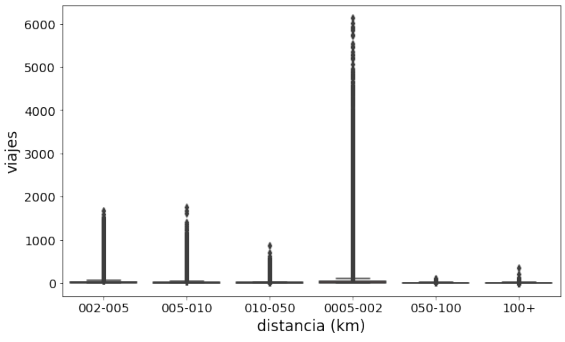

Figura 17: Diagrama de caja: Número de viajes por distancia - todos los trayectos [Fuente propia]

El diagrama de caja proporciona medidas estadísticas como la mediana y permite comprender la dispersión de los datos de manera visual. Sin embargo, en las figuras 16 y 17, la representación gráfica de este diagrama no es óptima, por lo que se observan solamente los valores que el diagrama identifica como atípicos. Este comportamiento se debe a que se están incluyendo datos de desplazamientos de todos los trayectos en España, lo que resulta en una mayor variabilidad de los datos y dificulta su apreciación.

En razón de lo expuesto anteriormente se identifica que los datos deben ser analizados por trayectos para lograr identificar información clave dentro de la gran cantidad de datos que se tiene. En las figuras 18 y 19 se puede apreciar el diagrama de caja para un trayecto origen-destino específico. Sin embargo, al analizar estas dos gráficas, se puede apreciar que detecta como valores atípicos una gran cantidad de datos, observando que la mediana de los datos no supera los 50 viajes. La gran cantidad de valores detectados como atípicos se debe a que se está explorando los datos de un único día (14 febrero 2020) en un trayecto específico, lo que conlleva a una distorsión de los resultado obtenidos.

Por lo descrito anteriormente, es necesario unir todos los archivos de los diferentes días para realizar la detección de valores atípicos. Esta detección se realiza en la sección 4.1 de la fase de preparación de datos.

En la figura 18, se representa los datos horarios del 14 de febrero en una trayecto específico. Por otro parte, en la figura 19, se representa el número de viajes en relación con la distancia recorrida. Este último gráfico muestra únicamente tres intervalos de distancia, ya que es un trayecto entre dos distritos cercanos en la ciudad de Madrid. Se observa que los trayectos dentro del intervalo de 500 metros a 2 kilómetros muestran un menor número de viajes en comparación con los otros dos intervalos.

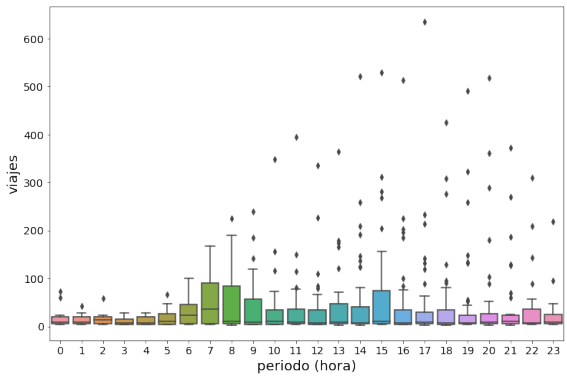

Figura 18: Diagrama de caja: Número de viajes por hora - trayecto específico [Fuente propia]

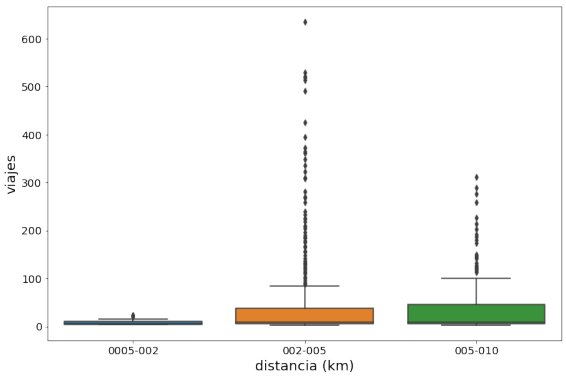

Figura 19: Diagrama de caja: Número de viajes por distancia - trayecto específico [Fuente propia]

En esta fase 2 de entendimiento de los datos, se identificaron campos de mayor relevancia y se concluyó que para descubrir los patrones de movilidad en los datos es fundamental realizar un análisis por trayectos. Esto implica la transformación de los datos en la fase 3 de la metodología. Por esta razón, en la tabla de la figura 20, se exploran los datos de un trayecto y una hora específica con el objetivo de identificar las modificaciones necesarias que serán aplicadas en una fase posterior del análisis.

	fecha	origen	destino	actividad_origen	actividad_destino	residencia	edad	periodo	distancia	viajes	viajes_km
2828962	20200601	2807905	2807901	casa	otros	28	NaN	1	002-005	11.008	41.319
2828963	20200601	2807905	2807901	casa	otros	28	NaN	1	005-010	5.504	29.072
2829012	20200601	2807905	2807901	casa	trabajo	28	NaN	1	002-005	5.504	19.440
2829083	20200601	2807905	2807901	otros	otros	28	NaN	1	002-005	5.504	17.783
2829160	20200601	2807905	2807901	trabajo	casa	28	NaN	1	005-010	4.895	28.698

Figura 20: Tabla para un trayecto y hora específica [Fuente propia]

A partir de la figura 20 se identifican algunas tareas que se deben de realizar en la fase 3 de preparación de los datos:

- Unir en un solo archivo todos los archivos diarios.
- Seleccionar las columnas relevantes para la investigación: “fecha”, “origen”, “destino”, “periodo”, “distancia” y “viajes”.
- Transformar la estructura de los datos de tal forma que los datos de las columnas origen y destino conformen un solo campo que represente el trayecto.
- Transformar la estructura de los datos para conformar un dataset que represente una serie temporal.
- Calcular la suma de los valores del campo "viajes"para que la serie temporal refleje el número de viajes o desplazamientos en una fecha y hora específica para cada trayecto.

Es importante en el análisis de datos verificar la calidad de los datos, con este propósito en se realiza un proceso de revisión de valores vacíos para los 169 archivos que serán objeto de estudio en este trabajo. Cada uno de estos archivos corresponde a un día dentro del periodo de análisis, comprendido desde el 14 de febrero hasta el 31 de julio del 2020. En la tabla 3 se ve reflejado con una x las columnas en donde se encontraron valores vacíos.

Columna / Mes	fecha	origen	destino	actividad origen	actividad destino	residencia	edad	periodo	distancia	viajes	viajes km
Febrero							x				
Marzo							x				
Abril							x				
Mayo							x				
Junio							x				
Julio							x				

Tabla 3: Verificación de valores vacíos en los 169 archivos CSV [Fuente propia]

3.2. Datos reales accidentalidad

Los datos reales reportados de accidentalidad son tomados en cuenta en este trabajo de investigación debido a que en análisis durante la fase de modelado, se ha identificado diferentes fluctuaciones a lo largo del día que no hacen parte de un patrón estacional. Estas observaciones indican una alta probabilidad de que los datos de accidentalidad sean un factor importante para explicar los cambios repentinos en el patrón de movilidad. En el anexo A.4 se brinda más información acerca del archivo 2020_Accidentalidad.csv que se explora en esta sección.

Con el objetivo de proporcionar una comprensión clara de los datos de accidentes, a continuación, en las figura 21 y 22, se presenta la estructura original del archivo "2020_Accidentalidad.csv". En la figura 21 se observan las primeras 10 columnas y en la figura 22 se muestran las 9 columnas restantes.

	num_expediente	fecha	hora	localizacion	numero	cod_distrito	distrito	tipo_accidente	estado_meteorológico	tipo_vehiculo
0	2019S040008	07/09/2020	23:00:00	CALL. SAN MAXIMILIANO, 38	38	15.0	CIUDAD LINEAL	Choque contra obstáculo fijo	Despejado	Turismo
1	2019S040008	07/09/2020	23:00:00	CALL. SAN MAXIMILIANO, 38	38	15.0	CIUDAD LINEAL	Choque contra obstáculo fijo	Despejado	VMU eléctrico
2	2020S000001	01/01/2020	1:15:00	AVDA. CANILLEJAS A VICALVARO / CALL. SILFIDE	1	20.0	SAN BLAS-CANILLEJAS	Colisión fronto-lateral	NaN	Turismo

Figura 21: Estructura de los datos de accidentes de tránsito 1 [Fuente propia]

tipo_persona	rango_edad	sexo	cod_lesividad	lesividad	coordenada_x_utm	coordenada_y_utm	positiva_alcohol	positiva_droga
Conductor	De 21 a 24 años	Hombre	NaN	NaN	444578,153	4475148,102	N	NaN
Conductor	De 25 a 29 años	Mujer	NaN	NaN	444578,153	4475148,102	N	NaN
Conductor	De 18 a 20 años	Hombre	NaN	NaN	447894,521	4476691,236	N	NaN

Figura 22: Estructura de los datos de accidentes de tránsito 2 [Fuente propia]

En la figura 23 se muestra la dimensión del archivo y el tipo de dato que contiene cada columna.

```
Dimensión:
(32433, 19)

num_expediente          object
fecha                   object
hora                    object
localizacion            object
numero                  object
cod_distrito           float64
distrito                object
tipo_accidente          object
estado_meteorológico    object
tipo_vehiculo           object
tipo_persona            object
rango_edad              object
sexo                    object
cod_lesividad          float64
lesividad               object
coordenada_x_utm        object
coordenada_y_utm        object
positiva_alcohol        object
positiva_droga         float64
dtype: object
```

Figura 23: Características del archivo CSV [Fuente propia]

En las figuras 24 y 25, se puede identificar cada una de las categorías que contiene las columnas “distrito” y “tipo_accidente”. En la columna “distrito” se identifican los 21 distritos de la ciudad de Madrid, mientras que en la columna “tipo_accidente” se identifican 13 tipos de accidentes.

```
PUENTE DE VALLECAS      2635
SALAMANCA               2486
CIUDAD LINEAL           2127
CHAMARTÍN               2084
CARABANCHEL             2076
FUENCARRAL-EL PARDO     1786
SAN BLAS-CANILLEJAS     1778
MONCLOA-ARAVACA         1756
TETUÁN                  1598
RETIRO                  1530
CENTRO                  1518
HORTALEZA               1508
LATINA                  1505
CHAMBERÍ                1491
ARGANZUELA              1351
USERA                   1217
VILLA DE VALLECAS        976
VILLAVERDE               971
MORATALAZ                961
BARAJAS                  545
VICÁLVARO                532
Name: distrito, dtype: int64
```

Figura 24: Cantidad de datos por cada categoría en la columna “distrito” [Fuente propia]

```
Colisión fronto-lateral         8085
Alcance                         7294
Choque contra obstáculo fijo    4667
Colisión lateral                4386
Colisión múltiple               2231
Atropello a persona             2129
Caída                           2118
Colisión frontal                 899
Otro                             251
Solo salida de la vía            151
Vuelco                           145
Atropello a animal                75
Despeñamiento                      2
Name: tipo_accidente, dtype: int64
```

Figura 25: Cantidad de datos por cada categoría en la columna “tipo_accidente” [Fuente propia]

En la figura 26 se realiza un proceso de revisión de valores vacíos para todas las columnas del archivo de accidentes. En esta figura se pueden apreciar valores faltantes en las columnas “estado_meteorológico”, “tipo_vehículo”, “cod_lesividad”, “lesividad” y “positiva_droga”. Adicionalmente se realizó la búsqueda de caracteres especiales como “?”, “*”, espacios en blanco y palabras especiales como NaN, null. Encontrando 2 valores faltantes el 23 de noviembre de 2020 en las columnas “cod_distrito” y “distrito”.

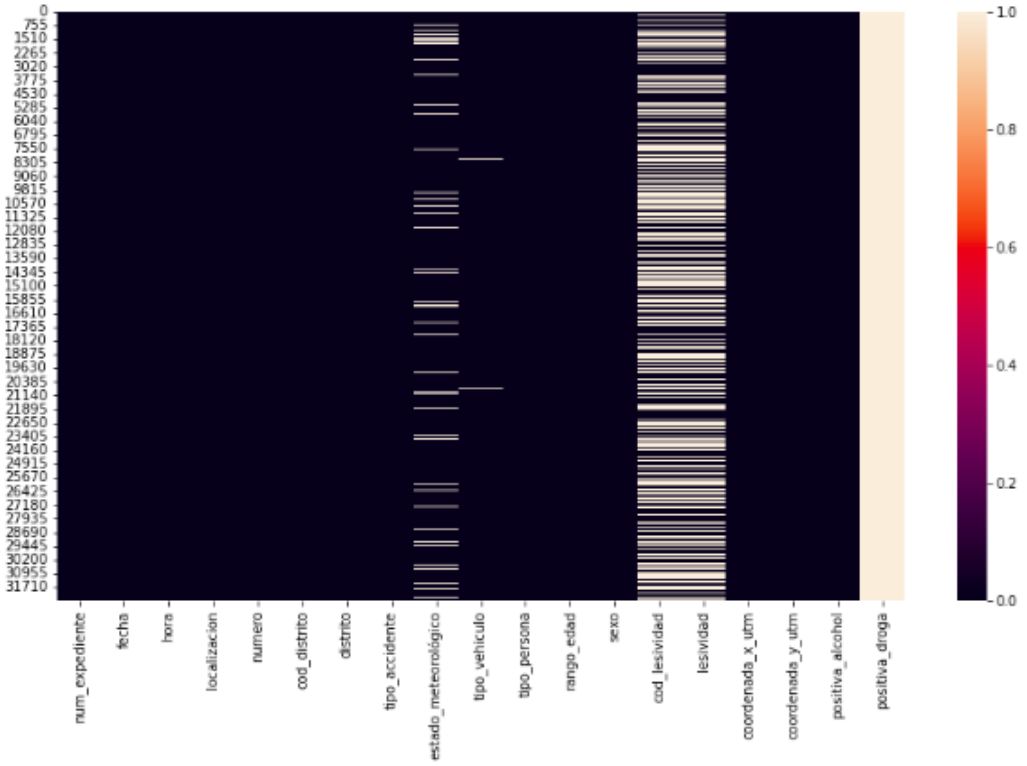

Figura 26: Mapa de calor para detectar valores faltantes [Fuente propia]

Después de realizar el análisis exploratorio de los datos de accidentes, se logra identificar que los campos más relevantes para esta investigación son: “num_expediente”, “fecha”, “hora”, “cod_distrito”, “distrito” y “tipo_accidente”. Estas variables adquieren relevancia en este estudio, ya que permiten identificar el momento y el lugar en el que ocurren los accidentes de tránsito.

A continuación se identifican algunas tareas que se deben de realizar en la fase 3 de preparación de los datos:

- Seleccionar las columnas relevantes para la investigación.
- Combinar la columna fecha con la columna hora para conformar un dataset que represente una serie temporal.
- Seleccionar solo aquellos distritos que pueden tener relación con el trayecto de estudio “Chamartín-Centro”.
- Seleccionar solo los tipos de accidentes que pueden generar una afectación en el tráfico, como por ejemplo “colisión múltiple”.
- Realizar imputación de datos faltantes si es necesario.

CAPÍTULO 4

Fase 3: Preparación de los datos utilizados para la planificación del servicio de transporte público en Madrid

La fase de preparación de los datos de la metodología CRISP-DM es parte esencial en este trabajo de investigación. Para llevar a cabo esta fase, se realizan actividades de limpieza, selección de características y transformación de los datos, que permitirán obtener el conjunto final de datos reales que se utilizarán en las herramientas de modelado.

4.1. Preparación de datos para el modelado

En esta sección se obtienen los dataset necesarios para entrenar el modelo de machine learning. Los datos que se emplean son datos reales de desplazamiento y datos de accidentalidad, los cuales fueron investigados durante la segunda fase de entendimiento de los datos. En esta sección, se llevaron a cabo actividades de reestructuración y limpieza de datos. Para llevar a cabo estas tareas, los autores en [36] proponen un proceso específico para la limpieza de datos en modelos de regresión. Este proceso incluye las siguientes etapas de limpieza: verificación de valores faltantes, imputación, detección de valores atípicos (outliers), eliminación de instancias duplicadas y reducción de dimensionalidad. El código fuente desarrollado para las actividades en esta fase se encuentra disponibles en el repositorio de GitHub [35].

4.1.1. Datos reales de desplazamiento entre distritos

Los datos de desplazamiento entre distritos son fundamentales para el desarrollo de los modelos de predicción que se entrenan en la siguiente fase de modelado. Por tal motivo, en esta sección se realizan las actividades de limpieza de datos.

En la sección 3.1 se realizó la verificación de valores vacíos o faltantes, mediante la búsqueda de caracteres especiales como “?”, “*”, espacios en blanco y palabras especiales como NaN, null. De esta forma, se descubrió que la columna “edad” contiene numerosos valores vacíos; sin embargo, las demás columnas no presentan ningún valor nulo. Dado que la columna “edad” no es relevante para este estudio, no es necesario realizar imputación de valores.

Debido a que se cuenta con una gran cantidad de archivos diarios, la siguiente tarea consiste en reducir la dimensionalidad basados en los coeficientes de correlación de la figura 15 y fusionar todos los archivos de desplazamientos desde el 14 de febrero hasta el 31 de julio del 2020 en un único archivo. Para llevar a cabo este proceso, se emplea el cuaderno de Jupyter llamado “Reduc_Dimensionalidad_Despl.ipynb”, el cual da lugar a la creación del archivo “trayectoFeb_Jul.csv”. En este nuevo archivo se tienen únicamente los registros que tienen como origen el código de distrito 2807905 (Chamartín) y como

destino el código de distrito 2807901 (Centro). El peso de este archivo se reduce a 2.2 MB con una dimensión de 43261 filas y 6 columnas. En la figura 27 se puede observar la estructura de este nuevo archivo.

fecha	origen	destino	periodo	distancia	viajes
20200621	2807905	2807901	0	002-005	16.335
20200621	2807905	2807901	1	002-005	16.335
20200621	2807905	2807901	1	005-010	5.445
20200621	2807905	2807901	2	002-005	16.335
20200621	2807905	2807901	3	002-005	16.335

Figura 27: Estructura del archivo “trayectoFeb_Jul.csv” [Fuente propia]

La siguiente actividad de limpieza que se realiza corresponde a la identificación y eliminación de valores atípicos para la columna viajes. Esta actividad permite encontrar comportamientos anormales en los registros. Para ello, primero se realiza una verificación mediante diagramas de caja, las figuras 28, 29, 30 y 31 muestran los diagramas para cada una de las etapas de la pandemia desde el 14 de febrero hasta el 31 de julio del 2020. Estas gráficas ilustran la cantidad de viajes realizados en función del periodo del día (hora) en que ocurrieron. En el eje vertical se representa el número de viajes, mientras que en el eje horizontal se encuentra la columna periodo.

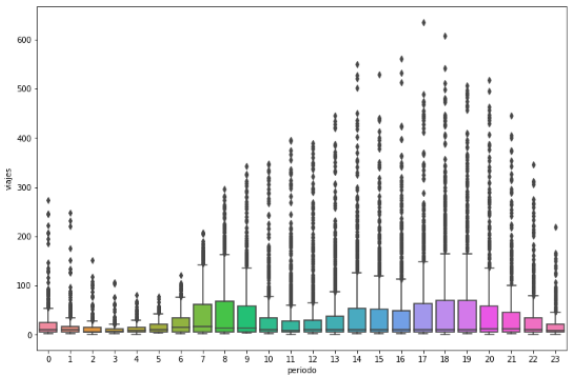

Figura 28: Diagrama de caja - Periodo referencia [Fuente propia]

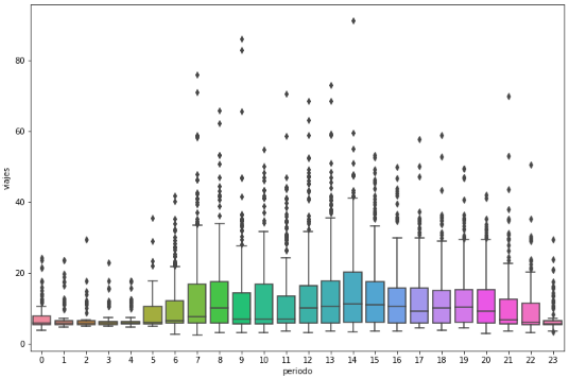

Figura 29: Diagrama de caja - Periodo Confinamiento [Fuente propia]

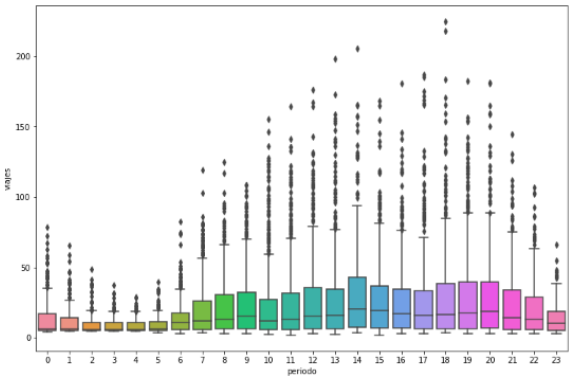

Figura 30: Diagrama de caja - Periodo Desescalada [Fuente propia]

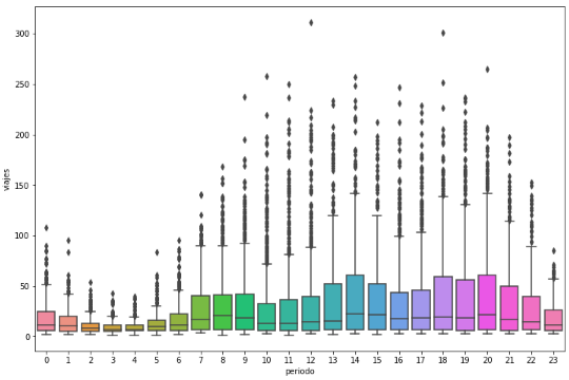

Figura 31: Diagrama de caja - Periodo Nueva normalidad [Fuente propia]

Al analizar cada etapa de la pandemia, se obtiene una representación más precisa de los datos, ya que el número de viajes o desplazamientos varía a lo largo de todo el período pandémico. Sin embargo, para eliminar los valores atípicos, se utiliza el algoritmo “Local Outlier Factor” (LOF), que se basa en la distancia para detectar valores atípicos. Según los autores en [36], este algoritmo demuestra un buen rendimiento en la identificación de valores atípicos positivos, con un mínimo de falsos positivos.

En las figuras 32 a 39, se presenta la detección de valores atípicos. Para este análisis, además de segmentar los datos según la etapa de la pandemia, se considera la segmentación basada en la distancia recorrida durante el trayecto específico de Chamartín a Centro, que abarca distancias de 2 a 5 km y de 5 a 10 km. Es importante señalar que los valores que se encuentran en una región menos densa que sus vecinos se representan en las gráficas mediante una circunferencia de mayor diámetro, indicando la presencia de valores atípicos.

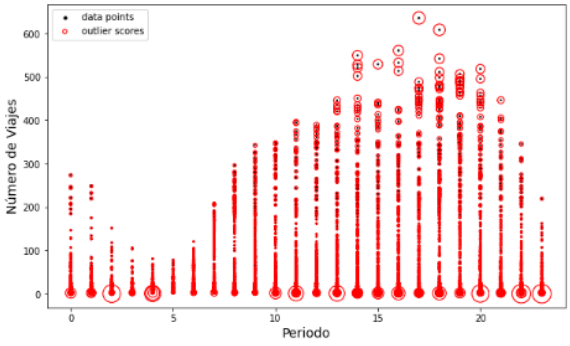

Figura 32: Valores atípicos detectados por LOF - Referencia (002-005) [Fuente propia]

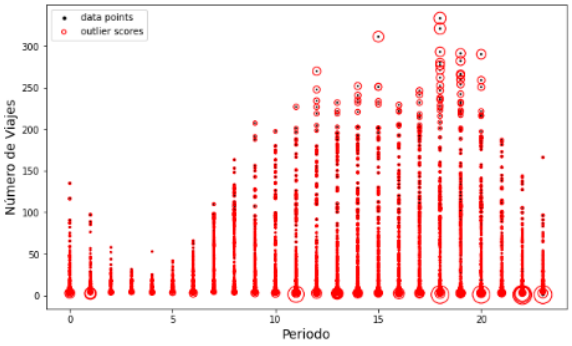

Figura 33: Valores atípicos detectados por LOF - Referencia (005-010) [Fuente propia]

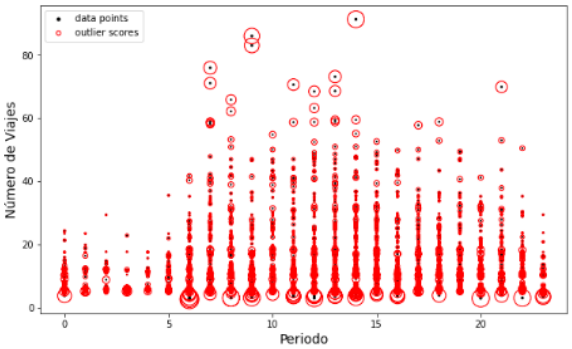

Figura 34: Valores atípicos detectados por LOF - Confinamiento (002-005) [Fuente propia]

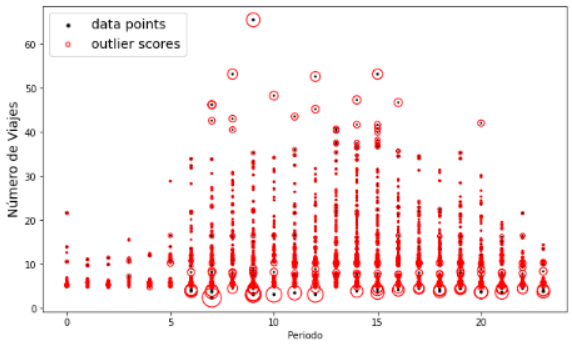

Figura 35: Valores atípicos detectados por LOF - Confinamiento (005-010) [Fuente propia]

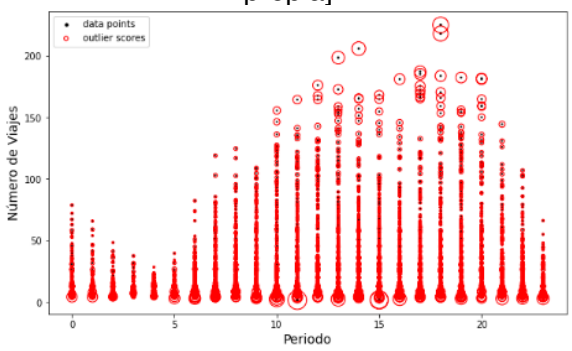

Figura 36: Valores atípicos detectados por LOF - Desescalada (002-005) [Fuente propia]

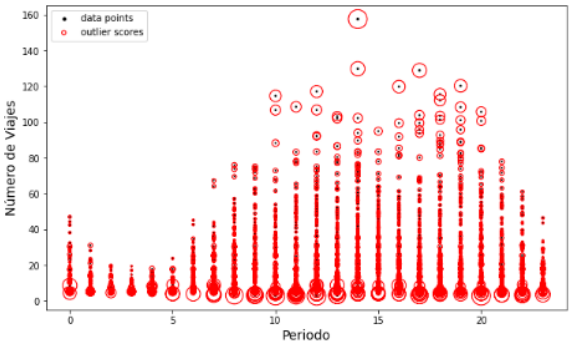

Figura 37: Valores atípicos detectados por LOF - Desescalada (005-010) [Fuente propia]

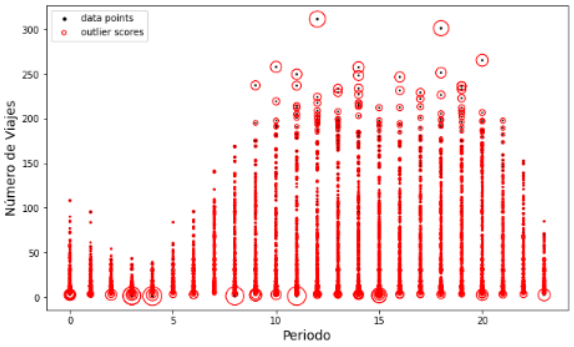

Figura 38: Valores atípicos detectados por LOF - Nnormalidad (002-005) [Fuente propia]

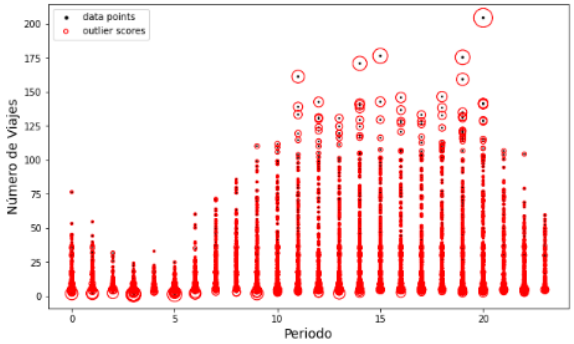

Figura 39: Valores atípicos detectados por LOF - Nnormalidad (005-010) [Fuente propia]

De acuerdo a las gráficas presentadas previamente, se comprende que los valores atípicos detectados son observaciones reales pero no representan los datos típicos de los días habituales. Por esta razón, se eliminan para evitar que estos datos introduzcan interferencias al entrenar el modelo. De esta manera el número de registros eliminados se presenta a continuación.

- **Referencia (002-005):** 76 registros eliminados
- **Referencia (005-010):** 65 registros eliminados
- **Confinamiento (002-005):** 98 registros eliminados
- **Confinamiento (005-010):** 116 registros eliminados
- **Desescalada (002-005):** 130 registros eliminados
- **Desescalada (005-010):** 139 registros eliminados
- **Nueva normalidad (002-005):** 50 registros eliminados
- **Nueva normalidad (005-010):** 86 registros eliminados

A partir de la limpieza de datos atípicos se genera un nuevo archivo csv de nombre “trayectoFeb_Jul_NOoutliers” con una dimensión de 42501 filas y 6 columnas.

La próxima tarea de limpieza implica eliminar duplicados. Durante esta revisión, se identificaron 2247 duplicados. Para determinar estos duplicados, se evaluaron las seis columnas de interés. Si un registro era idéntico a otro en las seis columnas, se clasificaba como duplicado. Después de realizar esta limpieza se crea un nuevo archivo de nombre “trayectoFeb_Jul_limpio” con unas dimensiones de 41361 filas y 6 columnas. A partir de este archivo con los datos limpios se genera el archivo “SerieTempFeb_Jul” que representa una serie temporal.

Para obtener el archivo limpio de serie temporal se combinan las columnas fecha y hora para formar una sola columna en formato “Año/Mes/Día Hora”. Utilizando el método pivot_table() de la biblioteca Pandas, se reconfigura la estructura de los datos para que el trayecto Chamartín-Centro se convierta en un nuevo atributo del conjunto de datos. Los valores de este atributo representan la suma de los viajes para cada fecha y hora específicas. En la figura 40 se representa el proceso de suma de los valores de viajes.

fecha	periodo	origen	destino	distancia	viajes
2020-02-14	0	2807905	2807901	002-005	23.751
2020-02-14	0	2807905	2807901	002-005	9.500
2020-02-14	0	2807905	2807901	002-005	60.195
2020-02-14	0	2807905	2807901	002-005	4.718
2020-02-14	0	2807905	2807901	002-005	9.610
...	...	...	...	...	...
2020-07-31	0	2807905	2807901	005-010	17.470
2020-07-31	0	2807905	2807901	005-010	13.193
2020-07-31	0	2807905	2807901	005-010	5.808
2020-07-31	0	2807905	2807901	005-010	25.685
2020-07-31	0	2807905	2807901	005-010	2.947

Figura 40: Suma de la columna viajes para dos fechas diferentes en la hora 0 [Fuente propia]

De esta forma se obtiene el dataset limpio para los desplazamientos entre el distrito chamartín y el distrito centro. Este archivo recibe un nombre de “SerieTempFeb_Jul”, el cual tiene un peso de 111.6 kB y una dimensión de (4056, 2). En la figura 41 se observan los primeros registros de este conjunto de datos.

ds	2807905-2807901
2020-02-14 00:00:00	308.104
2020-02-14 01:00:00	165.496
2020-02-14 02:00:00	155.760
2020-02-14 03:00:00	126.662
2020-02-14 04:00:00	123.672
2020-02-14 05:00:00	230.861
2020-02-14 06:00:00	393.165
2020-02-14 07:00:00	955.119
2020-02-14 08:00:00	1226.425
2020-02-14 09:00:00	1058.345
2020-02-14 10:00:00	1006.363
2020-02-14 11:00:00	1129.488
2020-02-14 12:00:00	1165.153

Figura 41: Dataset limpio de los desplazamientos del trayecto Chamartín-Centro [Fuente propia]

4.1.2. Datos reales de accidentes de tránsito

Es crucial llevar a cabo un proceso de limpieza de datos para obtener el conjunto final de datos reales sobre accidentalidad utilizado en la fase de modelado. En la sección 3.2, se llevó a cabo la primera actividad de limpieza, donde se identificaron varios valores vacíos en el conjunto de datos. Sin embargo, las columnas que contienen los valores

vacíos no son relevantes para este estudio. Además, los valores vacíos encontrados en las columnas “distrito” y “cod_distrito” no son relevantes debido a que corresponden al mes de noviembre, que está fuera del periodo de análisis. Por lo anterior no es necesario realizar un proceso de imputación de datos.

La segunda actividad de limpieza corresponde a la detección de valores atípicos. Sin embargo, las columnas num_expediente, fecha, hora, cod_distrito, distrito y tipo_accidente son identificadores categóricos únicos y no son representados como valores continuos. Debido a lo anterior la presencia de valores atípicos en estas columnas no son un problema relevante.

La tercera actividad de limpieza corresponde a la identificación y eliminación de registros duplicados. En este dataset, se encontraron 2568 registros duplicados, los cuales fueron eliminados para asegurar la integridad y precisión de los datos.

La cuarta actividad de limpieza corresponde al proceso de reducción de dimensionalidad, en donde se reduce el número de atributos, seleccionando las columnas más relevantes. Las columnas fecha y hora se combinan para formar una sola columna en formato “Año/Mes/Día Hora”. A continuación en la figura 42 se muestra la nueva estructura de los datos. La dimensión de este archivo es (30947, 4).

ds	num_expediente	cod_distrito	distrito	tipo_accidente
2020-01-01 01:00:00	2020S000014	3.0	RETIRO	Colisión fronto-lateral
2020-01-01 01:00:00	2020S000014	3.0	RETIRO	Colisión fronto-lateral
2020-01-01 01:00:00	2020S000004	15.0	CIUDAD LINEAL	Choque contra obstáculo fijo
2020-01-01 01:00:00	2020S000004	15.0	CIUDAD LINEAL	Choque contra obstáculo fijo
2020-01-01 01:00:00	2020S000004	15.0	CIUDAD LINEAL	Choque contra obstáculo fijo

Figura 42: Reducción de dimensionalidad de los datos de accidentes [Fuente propia]

El siguiente procedimiento se basó en la selección de los registros en la columna “cod_distrito”. Se seleccionaron los registros que incluyeran el valor “Chamartín” y los distritos adyacentes a él. Por lo tanto, se seleccionaron todos aquellos registros que incluyeran los distritos de “Chamartín”, “Chamberí”, “Hortaleza” y “El Pardo”. Posteriormente se realiza una nueva selección para los registros de la columna “tipo_accidente” todos aquellos accidentes que podrían afectar la circulación normal del tráfico. Por lo que se incluyen los registros de accidentes por atropello a persona, choque contra obstáculo fijo, colisión múltiple, colisión fronto-lateral, colisión lateral, alcance y vuelco.

La siguiente tarea se enfocó en redondear la hora al valor más cercano en punto, ya que los datos de los desplazamientos están registrados en horas exactas. Tras el procedimiento anterior, se genera una nueva columna llamada .Accidente"de tipo booleano, donde se establecen todos sus valores como "True". De esta forma se obtienen los siguientes archivos CSV.

- **Acc_ds1.csv**: Serie temporal con todos los datos reales de accidentes del año 2020. En este archivo, la columna “Accidente” indica la hora aproximada en la que ocurre un accidente. Este archivo explica la disminución repentina de los desplazamientos. Dimensión de 2009 filas y 2 columnas.

- **Acc_ds2.csv**: Serie temporal con todos los datos reales de accidentes del año 2020. En este archivo, la columna “Accidente” indica la hora siguiente a la ocurren-

cia de un accidente. Este archivo explica el aumento repentino de los desplazamientos (Se observa un aumento en los desplazamientos después de la ocurrencia de accidente). Dimensión de 2009 filas y 2 columnas.

	Accidente
ds1	
2020-01-01 02:00:00	True
2020-01-01 10:00:00	True
2020-01-01 13:00:00	True
2020-01-01 15:00:00	True
2020-01-01 17:00:00	True
...	...

Figura 43: Estructura del archivo Acc_ds1.csv [Fuente propia]

	Accidente
ds2	
2020-01-01 03:00:00	True
2020-01-01 11:00:00	True
2020-01-01 14:00:00	True
2020-01-01 16:00:00	True
2020-01-01 18:00:00	True
...	...

Figura 44: Estructura del archivo Acc_ds2.csv [Fuente propia]

4.1.3. Conjunto final de datos para el modelado

Para obtener los datos finales utilizados en el entrenamiento de los modelos, se emplean los dos archivos resultantes de la preparación de datos detallada en las secciones 4.1.1 y 4.1.2. Este procedimiento se encuentra en el cuaderno de jupyter con el nombre de “Dataset_final_despl” [35]. A continuación las figuras 45, 46, 47 y 48 muestran la estructura de los datos de entrenamiento de los modelos de franjas horarias.

	ds	y	Lunes0_5	t-168Mod
0	2020-02-14 00:00:00	308.104	False	0.0000
1	2020-02-14 01:00:00	165.496	False	0.0000
2	2020-02-14 02:00:00	155.760	False	0.0000
3	2020-02-14 03:00:00	126.662	False	0.0000
4	2020-02-14 04:00:00	123.672	False	0.0000
...	...	...	...	...
4051	2020-07-31 19:00:00	505.915	False	483.7714
4052	2020-07-31 20:00:00	549.821	False	774.2273
4053	2020-07-31 21:00:00	455.421	False	352.1119
4054	2020-07-31 22:00:00	294.034	False	523.9080
4055	2020-07-31 23:00:00	240.226	False	235.8420

Figura 45: Dataset final para modelo 0-5 [Fuente propia]

	ds	y	Lunes6_11	t-168Mod
0	2020-02-14 00:00:00	308.104	False	0.0000
1	2020-02-14 01:00:00	165.496	False	0.0000
2	2020-02-14 02:00:00	155.760	False	0.0000
3	2020-02-14 03:00:00	126.662	False	0.0000
4	2020-02-14 04:00:00	123.672	False	0.0000
...	...	...	...	...
4051	2020-07-31 19:00:00	505.915	False	483.7714
4052	2020-07-31 20:00:00	549.821	False	774.2273
4053	2020-07-31 21:00:00	455.421	False	352.1119
4054	2020-07-31 22:00:00	294.034	False	523.9080
4055	2020-07-31 23:00:00	240.226	False	235.8420

Figura 46: Dataset final para modelo 6-11 [Fuente propia]

	ds	y	Lunes12_17	t-168Mod
0	2020-02-14 00:00:00	308.104	False	0.0000
1	2020-02-14 01:00:00	165.496	False	0.0000
2	2020-02-14 02:00:00	155.760	False	0.0000
3	2020-02-14 03:00:00	126.662	False	0.0000
4	2020-02-14 04:00:00	123.672	False	0.0000
...	...	...	...	...
4051	2020-07-31 19:00:00	505.915	False	483.7714
4052	2020-07-31 20:00:00	549.821	False	774.2273
4053	2020-07-31 21:00:00	455.421	False	352.1119
4054	2020-07-31 22:00:00	294.034	False	523.9080
4055	2020-07-31 23:00:00	240.226	False	235.8420

Figura 47: Dataset final para modelo 12-17 [Fuente propia]

	ds	y	t-168Mod
0	2020-02-14 00:00:00	308.104	0.0000
1	2020-02-14 01:00:00	165.496	0.0000
2	2020-02-14 02:00:00	155.760	0.0000
3	2020-02-14 03:00:00	126.662	0.0000
4	2020-02-14 04:00:00	123.672	0.0000
...	...	...	...
4051	2020-07-31 19:00:00	505.915	483.7714
4052	2020-07-31 20:00:00	549.821	774.2273
4053	2020-07-31 21:00:00	455.421	352.1119
4054	2020-07-31 22:00:00	294.034	523.9080
4055	2020-07-31 23:00:00	240.226	235.8420

Figura 48: Dataset final para modelo 18-23 [Fuente propia]

Para obtener los cuatro conjuntos finales de datos de las figuras 45 a 48, lo primero que se realiza es renombrar la columna 2807905-2807901 del dataset limpio de desplazamientos de la sección 4.1.1. Es necesario cambiar el nombre por “y” ya que el modelo prophet para realizar las predicciones requiere de dos atributos, “ds” y “y”.

Las columnas “Lunes0_5”, “Lunes6_11” y “Lunes12_17” están compuestas de valores booleanos y son necesarios para ajustar una estacionalidad personalizada por hora en los modelos. Cuando su valor está en True indica que el registro es un día lunes dentro del intervalo horario específico.

La columna “t-168Mod” representa el valor de la semana pasada corregido, se hace de esta forma para conservar el patrón de desplazamiento que es alterado por un accidente. Para proporcionar una comprensión más clara del regresor "t-168Mod", se ha creado la tabla 5 que explica la construcción de la columna "t-168Mod".

Ds	t-168Mod
2020-07-27 07:00:00	404.82
2020-07-27 08:00:00	381.84
2020-07-27 15:00:00	588.04
2020-07-27 17:00:00	461.62

Tabla 4: Columna t-168Mod con desplazamientos corregidos de la semana pasada [Fuente propia]

Ds	y	Accidente1	Accidente2	Factor	y_corregido
2020-07-20 07:00:00	368.02	True	False	1.1	404.82
2020-07-20 08:00:00	545.49	False	True	0.7	381.84
2020-07-20 15:00:00	588.04	True	True	1	588.04
2020-07-20 17:00:00	461.62	False	False	1	461.62

Tabla 5: Corrección de los desplazamientos [Fuente propia]

De acuerdo a lo observado en la tabla 5, los factores de aumento o reducción de la columna ’factor’ se determinaron a partir de las observaciones realizadas en los datos y se seleccionaron los factores que producían los mejores resultados, además se considera un factor de uno para los casos en que no hay incidencias de tránsito o cuando hay tanto un aumento como una disminución en la misma hora.

Al seguir el procedimiento anterior, se logra mantener el patrón de movilidad para los

días lunes en el periodo de nueva normalidad. Si ocurre un accidente, el número de desplazamientos se ve afectado y altera el patrón de movilidad. En ese caso, el factor interviene para modificar el valor afectado, rectificando las caídas o subidas provocadas por los accidentes. De esta manera, se obtienen valores ideales que no se ven afectados por incidencias de tráfico. Considerando lo mencionado anteriormente y al incorporar la nueva columna "t-168Mod", el modelo mejora su capacidad predictiva. Las evidencias de esta afirmación se muestran en la sección 5.3.

4.2. Preparación datos reales de tasa de incidencia

Según el libro "Indicadores de salud. Aspectos conceptuales y operativos" de la Organización Panamericana de la Salud, se define la tasa de incidencia como "el número de casos nuevos de una enfermedad u otra condición de salud dividido por la población en riesgo de la enfermedad (población expuesta) en un lugar específico y durante un período especifico" [37]. En el contexto de la investigación, estos datos pueden ser relevantes para comprender cómo la incidencia de enfermedades, como el COVID-19, puede afectar los patrones de desplazamiento de las personas.

En el portal web de datos abiertos de la Comunidad de Madrid, es posible acceder a la información epidemiológica del Covid-19 detallada por municipios y distritos de Madrid [38]. Esta información está disponible en dos archivos CSV que contienen datos reales reportados sobre los casos confirmados y las tasas de incidencia acumulada.

- Archivo covid19FEB_JUL.csv:

  Este archivo CSV abarca el período desde el 26 de febrero de 2020 hasta el 1 de julio de 2020 contiene datos diarios. La estructura de este archivo CSV se presenta a continuación.

  municipio_distrito | fecha_informe | casos_confirmados_ultimos_14dias | tasa_incidencia_acumulada_ultimos_14dias | casos_confirmados_totales | tasa_incidencia_acumulada_total | codigo_geometria

- Archivo covid19JUL_2022.csv:

  El archivo CSV abarca el período desde 2 julio de 2020 hasta el 29 de marzo de 2022 contiene datos semanales. La estructura de este archivo CSV se presenta a continuación.

  municipio_distrito | fecha_informe | casos_confirmados_activos_ultimos_14dias | tasa_incidencia_acumulada_activos_ultimos_14dias | casos_confirmados_ultimos_14dias | tasa_incidencia_acumulada_ultimos_14dias | casos_confirmados_totales | tasa_incidencia_acumulada_total | codigo_geometria

Con el fin de obtener el conjunto de datos de incidencia que se analizarán en este trabajo de investigación, se lleva a cabo un proceso de selección de las columnas relevantes. En este proceso se seleccionan las siguientes columnas: "fecha_informe", "municipio_distrito" y "tasa_incidencia_acumulada_ultimos_14dias".

En el siguiente paso, se lleva a cabo la transformación de la estructura de los datos. Con

el fin de lograr esta tarea, se procede a reorganizar los registros de la columna "municipio_distrito", de manera que cada uno de ellos se convierta en una columna individual. Este proceso resulta en una nueva estructura para los archivos CSV generados, la nueva estructura de datos se muestra a continuación.

|fecha_informe|Alcobendas|Chamartín|Centro| ... |distritos_municipios restantes|

Dado que los datos en los dos archivos se presentan de manera diferente, uno en formato diario y otro en formato semanal, se generan dos archivos CSV distintos:

- **TasaIncidenciaD.csv**: Serie temporal que contiene los datos reales de incidencia desde el 26 de febrero de 2020 hasta el 30 de junio de 2020.
- **TasaIncidenciaS.csv**: Serie temporal que contiene los datos reales de incidencia desde el 01 de Julio de 2020 hasta el 09 de mayo de 2021.

Los procedimientos previamente descritos son realizados en el notebook “Tablas _Covid _19.ipynb”.

4.3. Búsqueda de patrones

Como conclusión de la fase de preparación de datos se crean las gráficas de las secciones 4.3.1 y 4.3.2, utilizando los datos que han sido preparados en esta fase. Estas gráficas son fundamentales en el trabajo de investigación ya que con ellas se toman decisiones de modelado que se verán reflejadas en la sección 5.3.

Por lo anterior, en esta sección de búsqueda de patrones se describen los patrones de movilidad que se han logrado encontrar en este trabajo de investigación. En la sección 4.3.1 se analiza la gráfica de contraste entre el patrón de movilidad de la pandemia y las tasas de incidencia reportadas durante este estado de excepción. Por otra parte en la sección 4.3.2 se crean diferentes gráficas que permitan observar el comportamiento de movilidad de las personas durante la pandemia. Para realizar el análisis de patrones, se utilizan los dataset resultantes de la sección 4.1.

Cabe resaltar que en este trabajo de investigación el análisis de patrones es de gran relevancia, por tal motivo se han explorado diferentes herramientas que faciliten esta identificación. En el Anexo I se detalla una de estas herramientas, que, aunque no se utiliza directamente en este trabajo, se considera valiosa por su potencial en la identificación de patrones.

4.3.1. Contraste patrones de movilidad y tasas de incidencia de Covid-19

El objetivo en esta sección es contrastar los datos reales reportados de movilidad urbana y la incidencia de COVID-19 en la ciudad de Madrid para obtener una comprensión más profunda de la relación entre datos. Este análisis permitirá extraer conclusiones importantes que servirán como base para la planificación del servicio de transporte público en Madrid. En esta sección se incluye la gráfica de la figura 49, la cual muestra los datos reales de desplazamientos diarios para el trayecto “Chamartín-Centro” y los datos reales correspondientes a la tasa de incidencia de los dos distritos involucrados en dicho trayecto. En la gráfica, al representar los datos de movilidad por día, es posible observar el patrón semanal que fue identificado en gráficas anteriores. En este patrón, los desplazamientos de los días laborales se mantienen en el mismo rango, mientras que los fines

de semana se aprecia una disminución.

La figura 49 muestra de manera gráfica el patrón de movilidad de la pandemia. Este patrón está determinado por las diferentes etapas establecidas por las autoridades estatales. A continuación se describen algunas características de este patrón al ser contrastado con los datos reales reportados de tasa de incidencia.

Como punto inicial, se observa una relación inversa entre las tasas de incidencia y los desplazamientos durante el periodo de confinamiento. Durante este periodo, se observa la aparición de la primera ola de infectados, la cual presenta la menor cantidad de casos confirmados en contraste con las otras dos olas representadas en la figura 49. Se puede observar que a medida que las restricciones aumentan y se limitan los desplazamientos, se produce una disminución en la tasa de incidencia de la enfermedad. Esto sugiere que las medidas de confinamiento contribuyeron a reducir la propagación del virus al restringir la movilidad de las personas.

Como segunda observación, se puede apreciar que a medida que las restricciones se alivian y la actividad social y económica se reanuda, los desplazamientos comienzan a aumentar, lo que provoca un aumento en la interacción entre las personas y, por ende, a un aumento en la propagación del virus. Esto se ve reflejado en la segunda ola de contagio que se produce después de el periodo de desescalada y durante la nueva normalidad.

También se puede apreciar del gráfico de la figura 49 una disminución en los desplazamientos cuando la tasa de incidencia comienza a aumentar. Esta apreciación es evidente durante la etapa de nueva normalidad y el periodo conocido como estado de alarma. Esto puede deberse a diferentes factores, como cambios en las medidas de control o cambios en el comportamiento individual. Esta disminución en los desplazamientos afectan la propagación del virus, provocando que las curvas de tasas de incidencia comiencen a disminuir.

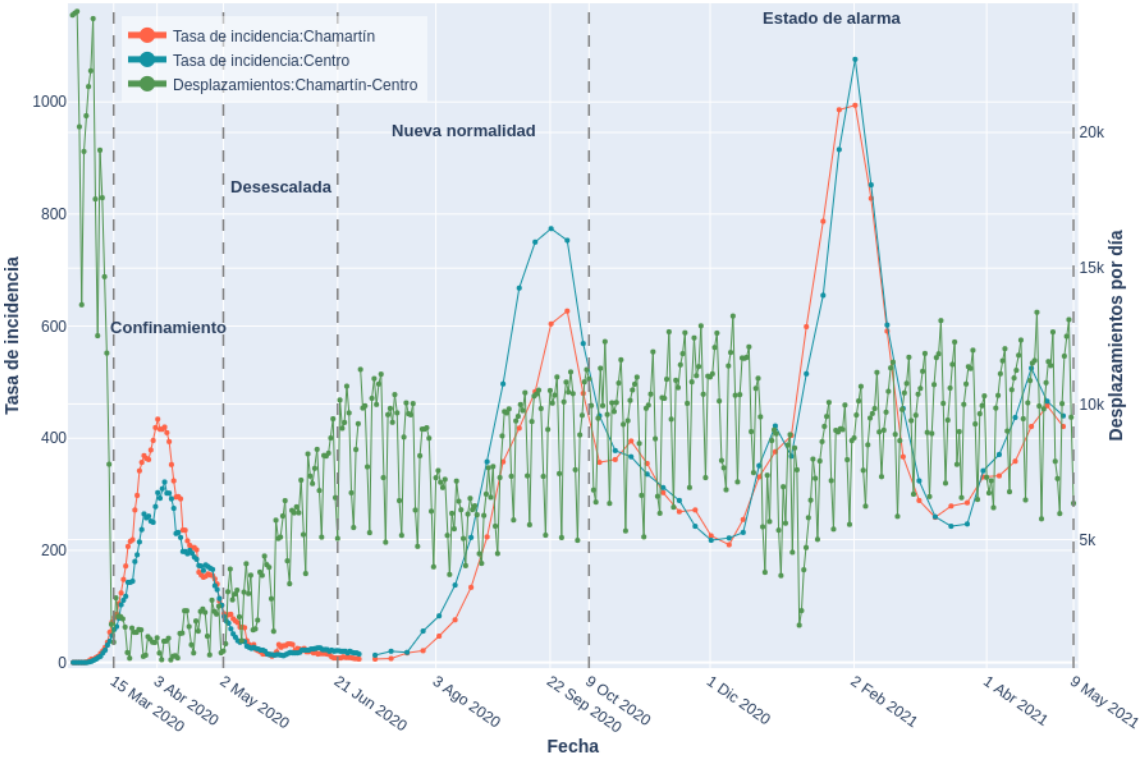

Figura 49: Contraste de patrones de movilidad y tasas de incidencia [Fuente propia]

Elaboración propia: El archivo que genera la figura 49 se encuentra disponible en el repositorio de github [35] con el nombre de Tablas_Covid_19_vs_Desplazamientos_1.ipynb – 2023

Por último, durante el estado de alarma se aprecia una relativa estabilidad en los desplazamientos, aunque a un nivel más elevado en comparación con otras etapas. Esto podría atribuirse a una flexibilización mayor de las restricciones, lo que permite una mayor movilidad de las personas dentro de los límites establecidos. Como resultado de este incremento en los desplazamientos durante dicho periodo, se produce la tercera ola, caracterizada por un aumento significativo en el número de casos confirmados.

Estas observaciones resaltan la relación entre los desplazamiento y las tasas de incidencia durante la pandemia, demostrando que la movilidad de las personas puede influir en la propagación del virus. Esto sugiere, que al examinar y comprender los patrones de movilidad durante una pandemia, es posible tomar decisiones informadas para adaptar y optimizar el transporte público en función de las necesidades cambiantes de los usuarios. Esto puede implicar ajustes en las rutas, horarios, frecuencias y capacidad de los medios de transporte, así como la implementación de medidas de seguridad y distanciamiento social.

Tomando en consideración todo lo expuesto con anterioridad , en este trabajo de investigación se buscará construir un modelo predictivo para el periodo de nueva normalidad. Se elige este periodo específico debido a que, junto con el periodo de desescalada, es responsable del incremento en las tasas de incidencia durante la segunda ola de infecciones. Si se implementan ajustes en el sistema de transporte durante este periodo, es posible lograr una rápida disminución en el pico máximo de la segunda ola.

Sin embargo, la figura 49 revela de manera notoria que durante este periodo de nueva normalidad, los datos reales reportados de desplazamiento continúan presentando variaciones significativas, debido a la relajación o intensificación de las restricciones. Estas variaciones en los patrones de desplazamiento representan un desafío para lograr predicciones precisas, es por ello que se determina un periodo de estudio más reducido dentro de la nueva normalidad. Este periodo estará comprendido entre el día 21 de junio de 2020 hasta el 31 de julio de 2020.

4.3.2. Análisis de datos de movilidad para la búsqueda de patrones de comportamiento

En una primera aproximación para descubrir los patrones de comportamiento de la movilidad en el trayecto ”Chamartín-Centro“, se lleva a cabo la identificación de los principales distritos de destino que parten del distrito de Chamartín. Para realizar este análisis se emplea el dataset obtenido en el anexo A.5 y se tienen en cuenta tres periodos específicos comprendidos entre el 14 de febrero y el 31 de julio del año 2020. Para la visualización de las curvas en cada uno de los tres periodos se realiza una sumatoria de las 24 horas, para de esta manera observar los datos reales de desplazamientos por día. Esta forma de representar los datos permite observar un estacionalidad semanal en donde los días sábados y domingos los desplazamientos disminuyen considerablemente.

En la figura 50 se representa gráficamente el periodo de referencia comprendido entre el 14 de febrero y el 15 de Marzo del año 2020, este periodo refleja el comportamiento de los desplazamientos antes de la aparición de la pandemia por Covid-19. En esta figura se presentan gráficamente cinco trayectos que tienen como origen al distrito Chamartín, estos cinco trayectos son el resultado de un proceso de filtrado en donde se descartan los trayectos que no tengan una media de desplazamientos superior a 15 mil, por lo que

Elaboración propia: Los archivos que generan las figuras 50, 51, 52, 53, 54 y 55 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov1_Referencia.ipynb y PatronesMov1 _Pandemia.ipynb – 2023

se puede decir que la gráfica de las figura 50 representa los principales destinos que tienen como origen el distrito chamartín.

Al realizar el análisis de la figura 50 se puede apreciar que el trayecto ”Chamartín-El Pardo“ presenta la mayor afluencia de personas, seguida por ”Chamartín-Hortaleza“. Además, los destinos ”Chamberí“, ”Ciudad Lineal“ y ”Centro“ se encuentran entre los cinco destinos más frecuentados desde el distrito de ”Chamartín“.

Figura 50: Trayectos del distrito Chamartín con mayor flujo de personas durante el periodo de referencia [Fuente propia]

El segundo periodo de análisis está comprendido entre el 16 de marzo y el 31 de mayo del año 2020. Los principales trayectos para este periodo se pueden observar en la gráfica de la figura 51. En esta gráfica se logra identificar que el trayecto “Chamartín-Centro” es uno de los más afectados por la aparición de la enfermedad. La curva de color azul que representa este trayecto se desplaza hasta el extremo final de la gráfica, lo que indica una marcada disminución en los desplazamientos diarios para este destino. Antes de la pandemia, el destino ”Centro“ se encontraba entre los cinco principales trayectos, pero durante este periodo específico, dejó de estar en esa categoría. También se puede observar en la figura 51 que al inicio de la pandemia los desplazamientos diarios para el trayecto ”Chamartín-Centro“ eran inferiores a 2 mil, mostrando el impacto significativo que tuvo la enfermedad en la movilidad de este trayecto.

De igual manera en la figura 51 se logra apreciar la aparición de otros distritos con un mayor número de desplazamientos que el destino centro, tales como el destino de Tetuán, Salamanca y Puente de Vallecas que no aparecían en la gráfica del periodo de referencia.

Elaboración propia: Los archivos que generan las figuras 50, 51, 52, 53, 54 y 55 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov1_Referencia.ipynb y PatronesMov1 _Pandemia.ipynb – 2023

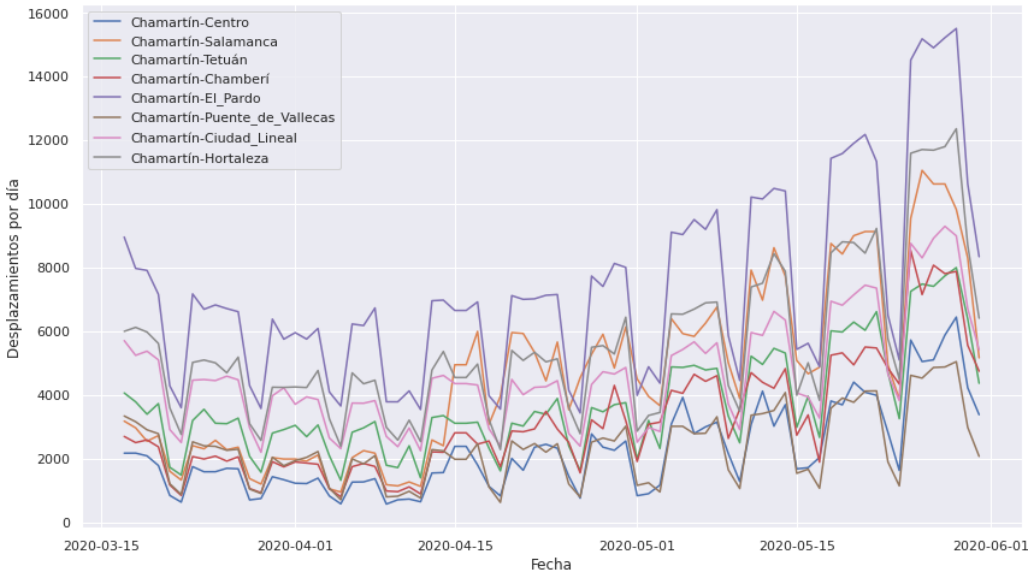

Figura 51: Trayectos del distrito Chamartín con mayor flujo de personas durante el periodo del 16 de marzo al 31 de mayo del año 2020 [Fuente propia]

En la gráfica de la figura 52 se observa el tercer período de análisis, comprendido desde el 1 de junio hasta el 31 de julio del año 2020. En este periodo el principal trayecto sigue siendo “Chamartín-El Pardo” y aunque en el periodo de referencia el trayecto “Chamartín-Centro” se mantuviera al nivel de los trayectos “Chamartín-Chamberí” y “Chamartín-Ciudad_Lineal”, se puede observar que en este periodo el destino Centro continua apareciendo con un menor número de desplazamientos.

Posterior al análisis de los tres periodos se concluye que el destino centro es uno de los destinos más afectados por la pandemia. Esto representa una razón adicional para enfocar la investigación en el trayecto “Chamartín-Centro”. De esta manera, un modelo predictivo específico para esta ruta clave de la ciudad permitiría que el sistema de transporte tome medidas concretas para mejorar la seguridad y confianza de las personas al viajar hacia el destino ”Centro“ durante una pandemia.

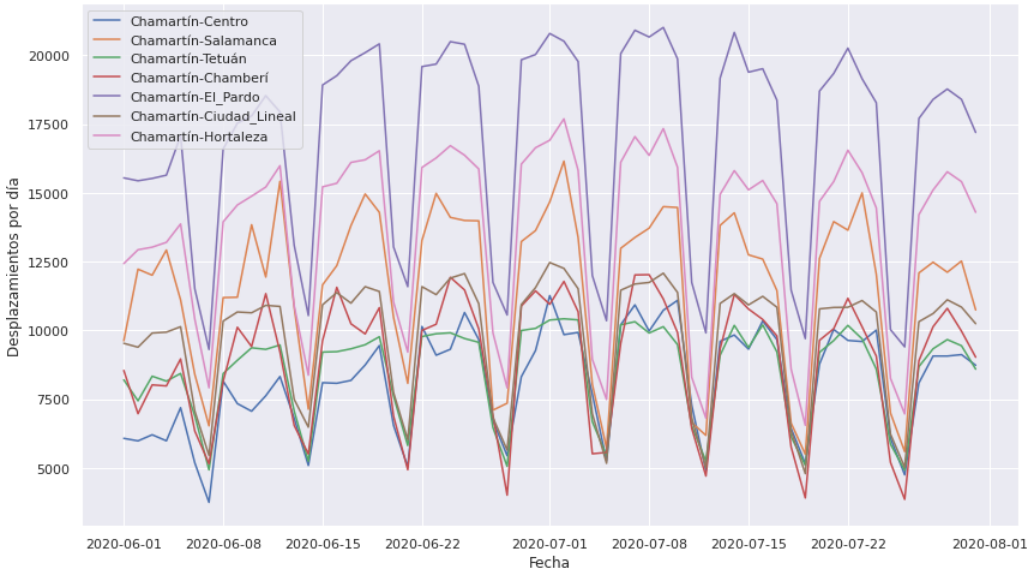

Figura 52: Trayectos del distrito Chamartín con mayor flujo de personas durante el periodo del 1 de junio al 31 de julio del año 2020 [Fuente propia]

Elaboración propia: Los archivos que generan las figuras 50, 51, 52, 53, 54 y 55 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov1_Referencia.ipynb y PatronesMov1 _Pandemia.ipynb – 2023

La exploración previa desempeña un papel fundamental en la investigación, ya que permite comprender los flujos de movilidad entre los distritos durante una pandemia. Esta exploración facilita la identificación de los trayectos más concurridos y aquellos que experimentan una disminución en su afluencia. Estos hallazgos son de gran importancia para la planificación del sistema de transporte público, ya que ayudan a mejorar la eficiencia del servicio.

Como segunda parte de la exploración de los patrones de movilidad, se llevó a acabo un análisis del artículo sobre ciencia de datos titulado “Stop aggregating away the signal in your data”[39]. En este artículo, el autor resalta la importancia de mantener el nivel de detalle inicial de los datos y evitar la reducción de su tamaño. Es común que al procesar grandes volúmenes de datos que se encuentran representados por horas, se tienda a resumir la información de forma diaria, semanal o mensual. Esta forma de procesar los datos genera la pérdida del contexto y puede llevar a la pérdida de información valiosa.

A partir de lo mencionado previamente y con el propósito de identificar información clave dentro de los datos reales reportados de movilidad, se ha generado la gráfica de la figura 53. En esta representación, se visualizan los datos horarios de todos los días durante las cinco semanas del periodo de referencia (una línea para cada semana del periodo). Esta forma alternativa de mostrar los datos de movilidad, permite una observación más detallada de los datos, revelando así patrones recurrentes que corresponden a una estacionalidad diaria.

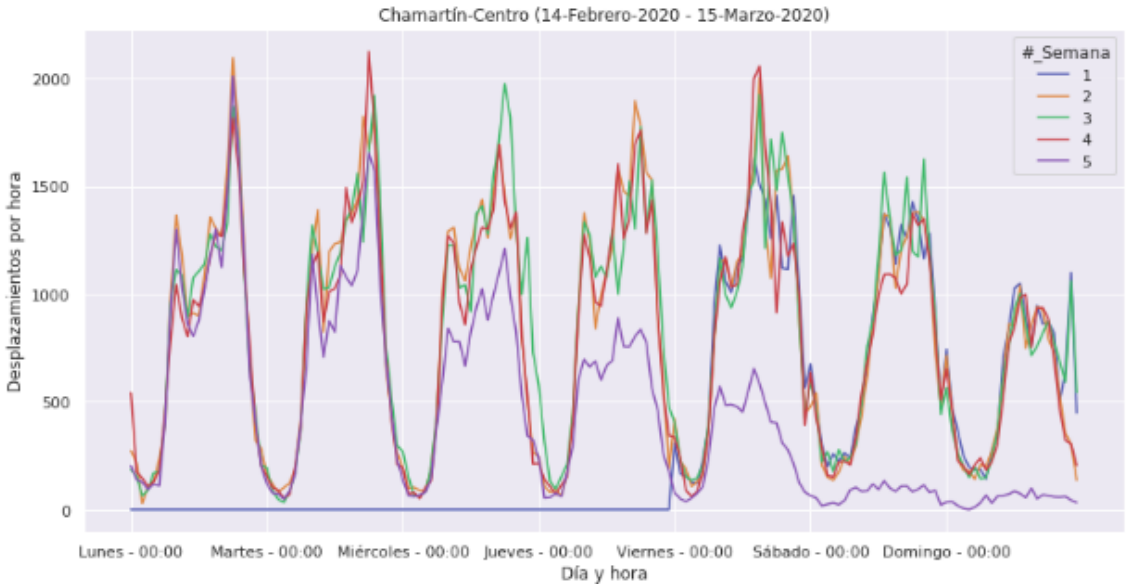

Figura 53: Representación horaria de los desplazamientos a lo largo de las cinco semanas del periodo de referencia [Fuente propia]

En la gráfica de la figura 54, se realiza nuevamente la representación de las cinco semanas pero esta vez considerando el periodo del mes de abril. La primera observación que se extrae de esta gráfica es la notable disminución de los desplazamientos, lo cual es consecuencia de las medidas implementadas para frenar la propagación de la enfermedad de Covid-19. Como segunda observación, se identifica una variación significativa de los datos en diferentes horarios del día y a lo largo de las cinco semanas, provocando que la estacionalidad diaria que se había identificado en el periodo de referencia ya no sea tan notable.

Elaboración propia: Los archivos que generan las figuras 50, 51, 52, 53, 54 y 55 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov1_Referencia.ipynb y PatronesMov1 _Pandemia.ipynb – 2023

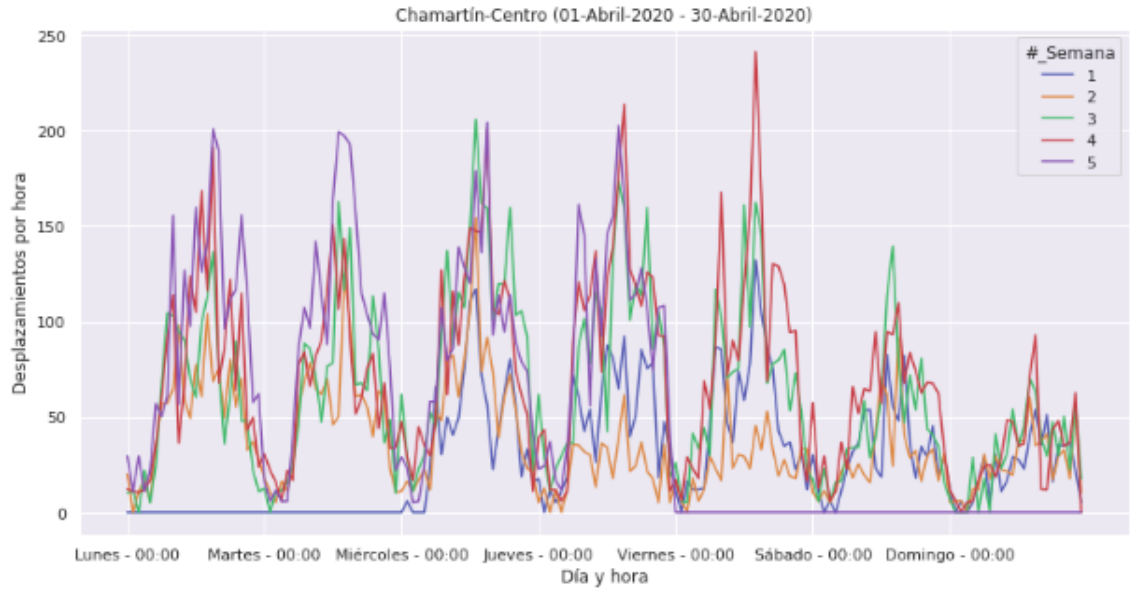

Figura 54: Representación horaria de los desplazamientos a lo largo de las cinco semanas del periodo del mes de abril [Fuente propia]

Para el periodo del mes de julio, representado en la gráfica de la figura 55, se observa una mayor regularidad en el número de desplazamientos a lo largo de las cinco semanas. En contraste con el mes de abril, los datos muestran una menor dispersión, lo que indica una mayor estabilidad en los patrones de movilidad. Sin embargo, el patrón diario que se había identificado en el periodo de referencia sigue siendo desdibujado por las variaciones en diferentes horas del día.

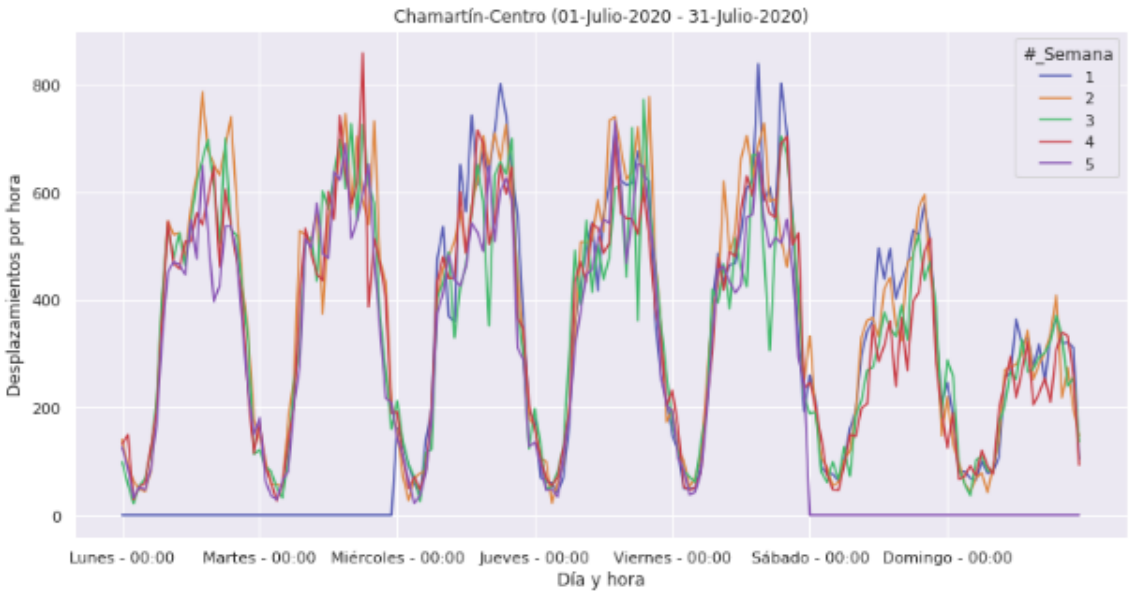

Figura 55: Representación horaria de los desplazamientos a lo largo de las cinco semanas del periodo del mes de julio [Fuente propia]

Con la intención de identificar los patrones por día de la semana y siguiendo las directrices del artículo sobre ciencia de datos, se elabora una nueva gráfica de ticks (figuras 56 a 62). Los ticks son representados en la gráfica mediante barras verticales y permiten una mejor visualización de los desplazamientos por hora. Además, al utilizar una paleta de colores, se logra la identificación de cada fecha específica en los días de la semana.

Elaboración propia: Los archivos que generan las figuras 50, 51, 52, 53, 54 y 55 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov1_Referencia.ipynb y PatronesMov1 _Pandemia.ipynb – 2023

En la figura 56 se visualiza la gráfica de ticks para cada día de la semana en el periodo de referencia. Esta representación visual nos permite observar un patrón característico de movilidad durante las horas de la mañana, cuando las personas comienzan a salir de sus hogares para dar inicio a sus actividades diarias. Además, en los días laborales se pueden identificar algunos picos que aparecen de forma recurrente en las horas 08:00, 14:00 y 18:00. Sin embargo, es interesante destacar que existe una variación en el pico de los días viernes, donde el pico de las 18:00 se desplaza hacia la hora 19:00.

En las gráficas presentadas en la figura 56, también se puede observar una notable disminución en los desplazamientos durante los últimos días del periodo de referencia. Esta reducción en la movilidad es atribuible a la aparición del COVID-19 y a las medidas de contención implementadas para mitigar su propagación.

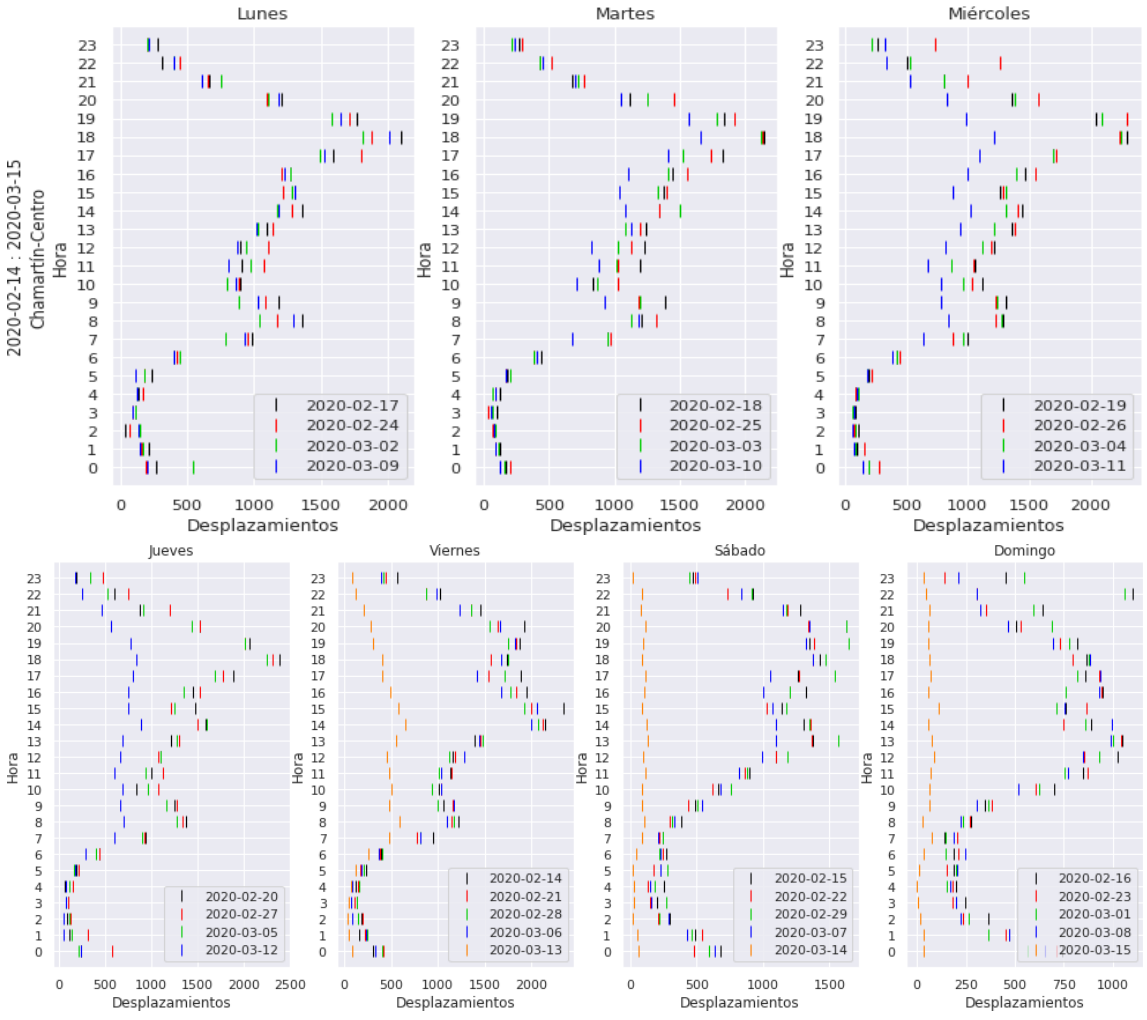

Figura 56: Gráfica de ticks para cada día de la semana en el periodo de referencia [Fuente propia]

Si bien es cierto que los desplazamientos entre Chamartín y Centro presentan patrones similares en los días laborales de la semana, es importante destacar que cada día tiene sus propias características distintivas. Por lo expuesto anteriormente y con el objetivo de desarrollar un modelo que capture de manera precisa los patrones de movilidad, se determina que es necesario contar con un modelo individual para cada día de la semana.

Con el fin de delimitar el alcance de la investigación, se ha designado al día lunes como

Elaboración propia: Los archivos que generan las figuras 56 a 62 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov2_FebMay.ipynb y PatronesMov2_JunJul.ipynb – 2023

el principal enfoque de la investigación. Esta selección se debe a que el día lunes representa el inicio de la semana laboral y muchas personas retoman sus actividades después del fin de semana, lo que genera patrones de movilidad específicos. En las siguientes figuras se analiza en profundidad las características de este día.

En la figura 57 se presentan los días lunes del periodo de referencia y en la figura 58 se visualizan los dos últimos lunes del mes marzo. Al realizar una comparación de los dos periodos representados en las gráficas 57 y 58, se puede observar claramente la caída del número de desplazamientos de un periodo a otro para el trayecto "Chamartín-Centro". También se puede apreciar que en el periodo de referencia las horas 17:00, 18:00 y 19:00 registran el mayor número de desplazamientos. Sin embargo, a partir del 15 de marzo, cuando se implementó el primer estado de alarma debido a la enfermedad, se observa que el pico de desplazamientos en estas horas desaparece.

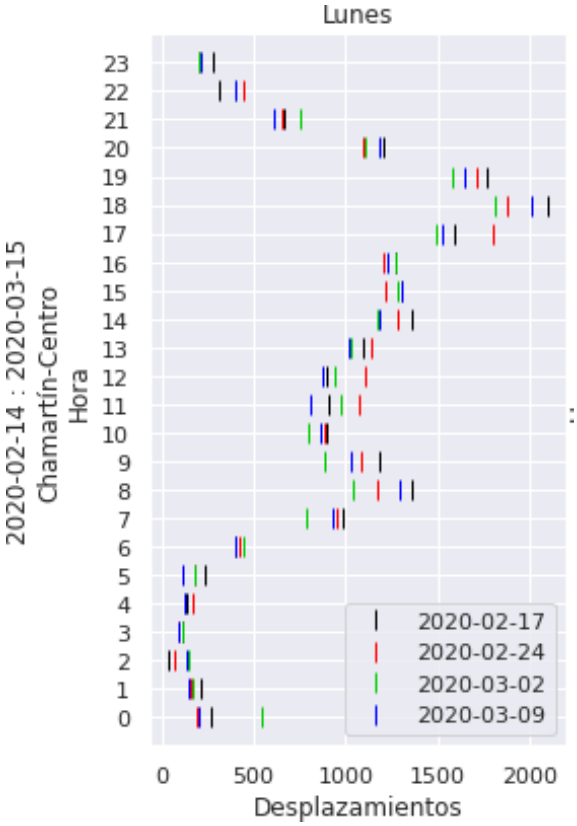

Figura 57: Gráfica de ticks de los días lunes en el periodo de referencia [Fuente propia]

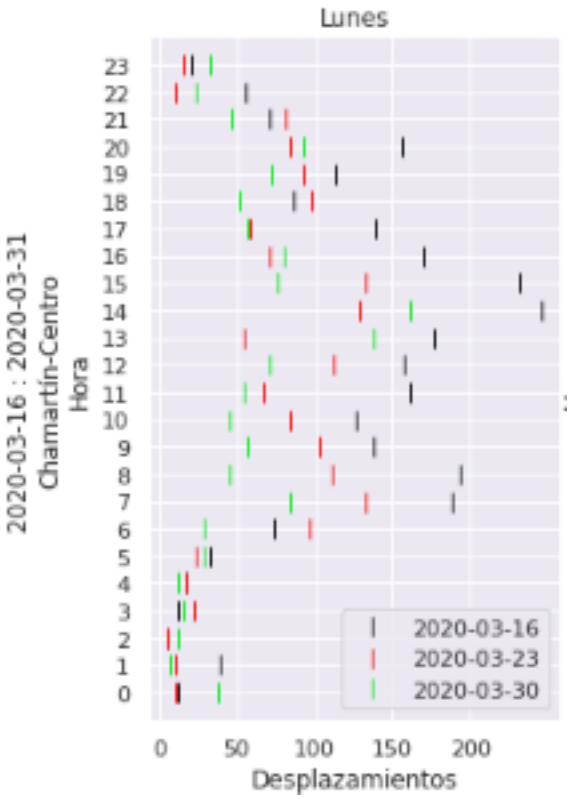

Figura 58: Gráfica de ticks de los días lunes en el periodo de marzo [Fuente propia]

En la figura 59 se presenta la gráfica del periodo del mes de abril y en la figura 60 el periodo del mes de mayo, en estas gráficas se puede apreciar un aumento progresivo de los desplazamientos a medida que avanzan las semanas. Es importante destacar que, si nos enfocamos en el intervalo horario de 07:00 a 20:00, se puede apreciar una mayor dispersión de los datos reales reportados.

Al analizar este intervalo horario de 07:00 a 20:00 durante el mes de abril, los valores de desplazamiento se encuentran dispersos en un rango aproximado de 50 a 200 desplazamientos. Por otro lado, en el mes de mayo, se observa una ampliación en el rango de desplazamientos, que se sitúa entre 150 y 400 desplazamientos. El aumento del rango de desplazamientos en el mes de mayo podría explicarse debido al inicio de la etapa de desescalada que comienza el 2 de mayo. A partir de esta fecha los desplazamientos

Elaboración propia: Los archivos que generan las figuras 56 a 62 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov2_FebMay.ipynb y PatronesMov2_JunJul.ipynb – 2023

empiezan a aumentar progresivamente en la ciudad, lo que conlleva a un aumento en la dispersión de los datos reales. El comportamiento de los desplazamientos en la etapa de desescalada se logró observar en la figura 49 de la sección previa.

También se puede observar de las figuras 59 y 60 que el intervalo horario con menor dispersión de los datos reales durante la pandemia corresponde al intervalo horario de 0 a 6, intervalo donde se presenta el menor número de desplazamientos del día.

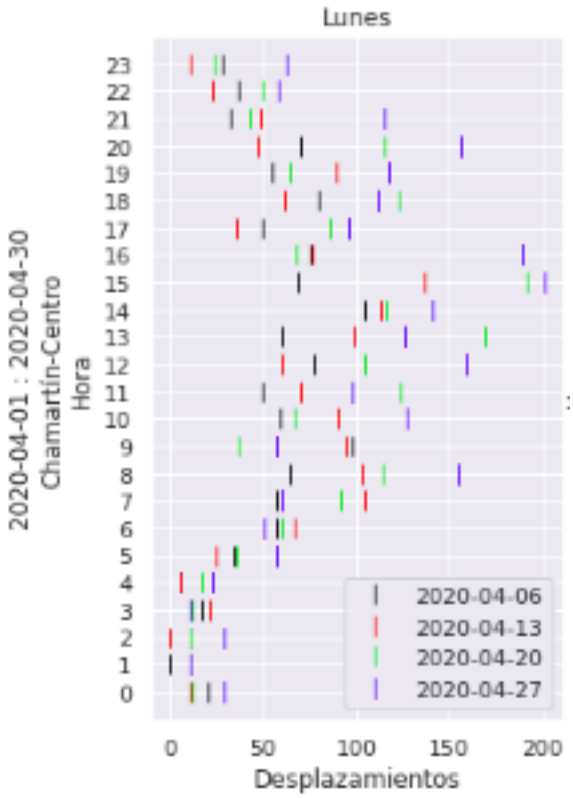

Figura 59: Gráfica de ticks de los días lunes en el periodo de abril [Fuente propia]

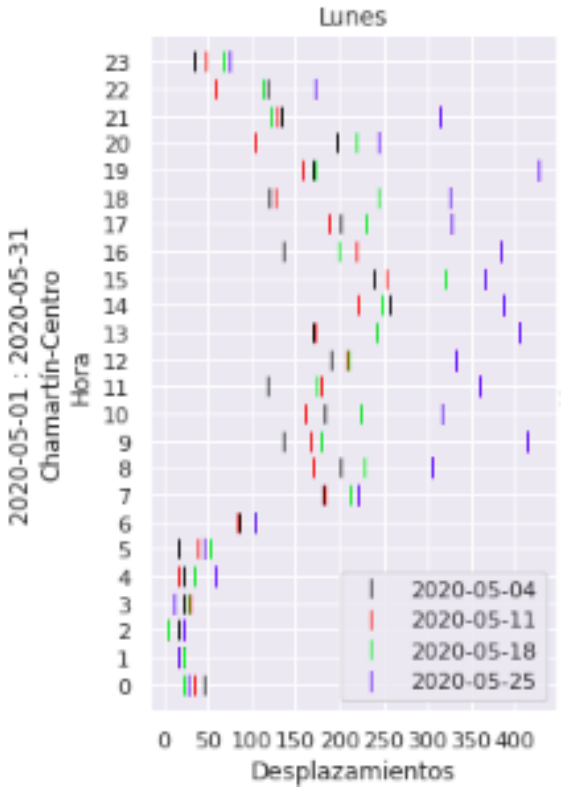

Figura 60: Gráfica de ticks de los días lunes en el periodo de mayo [Fuente propia]

Continuando con el análisis del intervalo horario de 07:00 a 20:00, en las gráficas de la figuras 61 y 62 se visualizan los periodos correspondientes al mes de junio y al mes de julio respectivamente. En el mes de junio, se puede observar un nuevo aumento en la dispersión de los datos reales con respecto al mes de mayo, con un rango de desplazamientos que oscila entre 300 y 650. Sin embargo, en el mes de julio los datos parecen estar más próximos en comparación con el mes de junio, ya que los datos se sitúan en su mayoría entre 400 y 700, lo cual refleja una menor diferencia entre los valores máximos y mínimos de los desplazamientos.

La disminución en el rango de desplazamientos del intervalo horario de 07:00 a 20:00 durante el mes de julio refleja una menor dispersión de los datos reales en comparación con el mes de junio. Para explicar esta reducción en la dispersión de los datos entre ambos meses, es necesario examinar detenidamente la figura 49 de la sección previa. En dicha figura, se puede observar una clara disminución en los desplazamientos que ocurren al inicio de julio entre el distrito Chamartín y el distrito Centro. Esta reducción en el número de desplazamientos en este trayecto específico parece estar generando un efecto directo en la disminución de la dispersión de los datos en el intervalo horario analizado.

Elaboración propia: Los archivos que generan las figuras 56 a 62 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov2_FebMay.ipynb y PatronesMov2_JunJul.ipynb – 2023

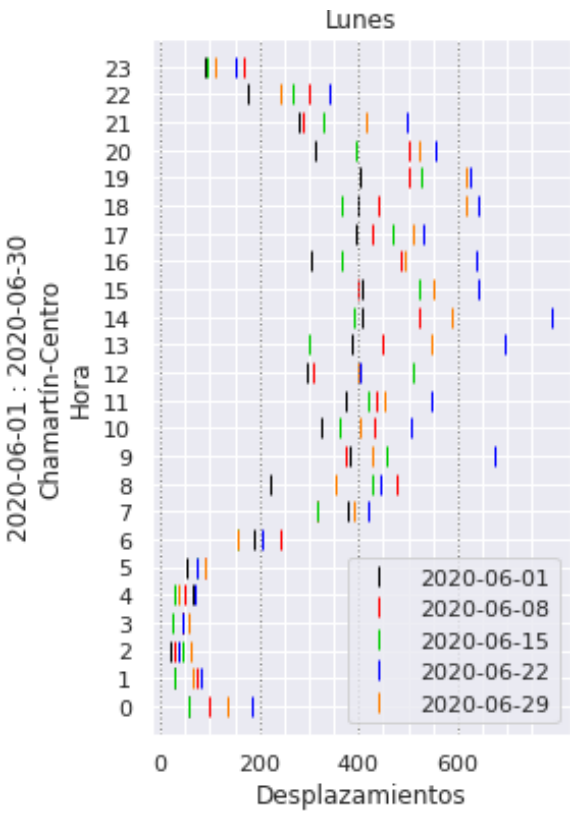

Figura 61: Gráfica de ticks de los días lunes en el periodo de junio [Fuente propia]

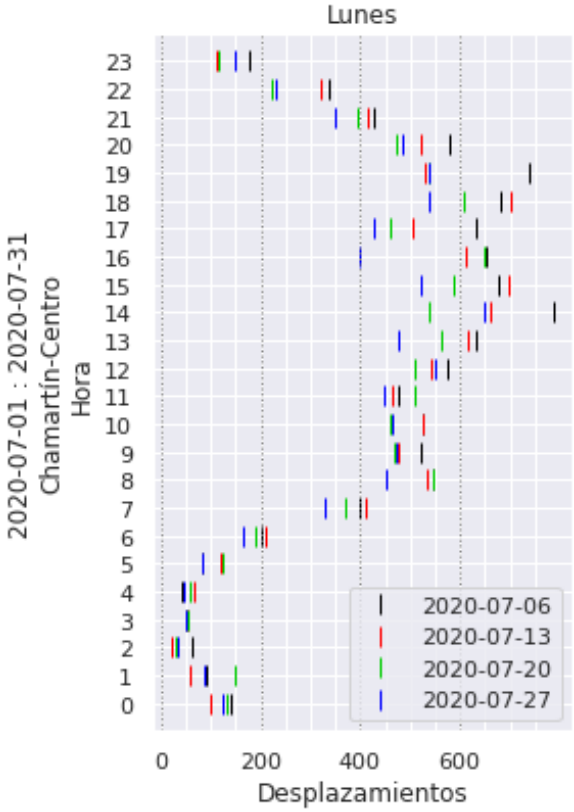

Figura 62: Gráfica de ticks de los días lunes en el periodo de julio [Fuente propia]

A partir del análisis gráfico realizado previamente se puede afirmar que los datos reales reportados se encuentran más dispersos en el intervalo horario de 07:00 a 20:00. Es en este intervalo donde puede ser más complejo realizar pronósticos para un modelo predictivo, por lo que las predicciones podrían tener un margen de error más amplio y ser menos precisas dentro de este intervalo. De igual manera se logra identificar que la diferencia de los rangos de desplazamientos en el intervalo horario de estudio, comienza a tener un aumento desde marzo hasta junio. Sin embargo, en julio los datos parecen estar más cercanos entre sí, lo que sugiere una variación en los patrones de movilidad de un mes a otro. Esto podría atribuirse a factores como modificaciones en las medidas de control de la pandemia o cambios en el comportamiento de las personas.

De acuerdo con lo mencionado anteriormente y a la evidencia proporcionada por las gráficas de esta sección, se concluye que el modelo de predicción debe ser segmentado según las horas del día. Esto se debe a que la estacionalidad diaria durante el período de estado de alarma no sigue un patrón claramente definido, como se hizo evidente en la figura 54 de esta sección. De igual manera, este análisis sugiere la necesidad de desarrollar un modelo específico para cada etapa de la pandemia, ya que los patrones de movilidad y comportamiento de las personas cambian a medida que avanza el contexto epidemiológico.

Elaboración propia: Los archivos que generan las figuras 56 a 62 se encuentran disponibles en el repositorio de github [35] con el nombre de PatronesMov2_FebMay.ipynb y PatronesMov2 _JunJul.ipynb – 2023

CAPÍTULO 5

Fase 4: Desarrollo y entrenamiento de un modelo de Machine Learning para la planificación del servicio de transporte público

En este capítulo, se exploran diversos algoritmos de regresión con el objetivo de construir el modelo de predicción para el trayecto “Chamartín-Centro” en el periodo de nueva normalidad. Los algoritmos considerados en este trabajo de investigación son lo modelos ARIMA, Random Forest y el modelo Prophet. Posterior a la selección del algoritmo, se procede a entrenar los modelos y se lleva a cabo la validación de cada uno de estos modelos, con el propósito de seleccionar el modelo final que demuestre mejor rendimiento.

5.1. Selección del algoritmo de predicción para el modelado

En el marco de la investigación realizada, se llevó a cabo una revisión para seleccionar el modelo adecuado para el análisis de datos. Este proceso de revisión es detallado en la tabla 6.

Revisión del modelo	Referencia	Hallazgos
ARIMA/Prophet	“Time Series Forecasting Model for Supermarket Sales using FB-Prophet” [40]	En este estudio de predicción de ventas de supermercados, se evaluaron los modelos ARIMA y Prophet. Mientras ARIMA necesita parámetros específicos y tiene limitaciones en el manejo de datos faltantes y valores atípicos, Prophet se destacó por su flexibilidad, capacidad para manejar datos incompletos y su detección automática de tendencias. Los resultados indicaron que Prophet ofreció predicciones cercanas a la realidad, mostrando un buen rendimiento y alta precisión en comparación con ARIMA.

Elaboración propia: Los archivos de las figuras 63 a 66 y las tablas 7 a 10 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23), y producen los mismos resultados con el modelo ajustado por m=Prophet() – 2023

ARIMA/Prophet	"Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET" [41]	En este estudio sobre la enfermedad COVID-19 en Indonesia, se compararon los modelos ARIMA y Prophet de Facebook para analizar y prever series temporales. Mientras que ARIMA requería transformación de datos y resultaba complicado de usar, Prophet se destacó por su naturaleza automatizada, facilidad de uso y resistencia a valores atípicos y cambios de tendencia. Además, Prophet mostró ser más preciso, especialmente en situaciones con fuertes efectos estacionales, lo que lo convirtió en la elección preferida para este análisis.
Random Forest	Anexo H	Se evaluó el desempeño del algoritmo de aprendizaje automático Random Forest para el análisis de series temporales, se ajustaron 1000 árboles para construir el modelo. Sin embargo, a pesar de los ajustes, las predicciones de Random Forest no mostraron mejoras significativas en comparación con los modelos ajustados con Prophet. Por lo que Prophet demuestra ser más adecuado y versátil permitiendo ajustar varios parámetros como tendencia y estacionalidad, así como incorporar variables complementarias.

Tabla 6: Revisión modelos de predicción [Fuente propia]

De acuerdo a la revisión de los modelos de predicción de la tabla 6, se selecciona el modelo Prophet para el entrenamiento del modelo predictivo por su capacidad para manejar datos incompletos y su detección automática de tendencias en series temporales. Su flexibilidad, combinada con su precisión superior en comparación con otros modelos, como ARIMA y Random Forest, lo convierte en una elección ideal para el análisis y pronóstico de datos temporales.

5.2. Modelos Prophet de referencia ajustados con parámetros por defecto

En esta sección, se presentan gráficas que muestran la comparación entre los datos reales reportados y las predicciones generadas por un modelo ajustado con valores por defecto. Asimismo, se incluyen tablas con algunas métricas de evaluación para proporcionar un punto de referencia y permitir comprender el desempeño de los modelos que serán presentados más adelante.

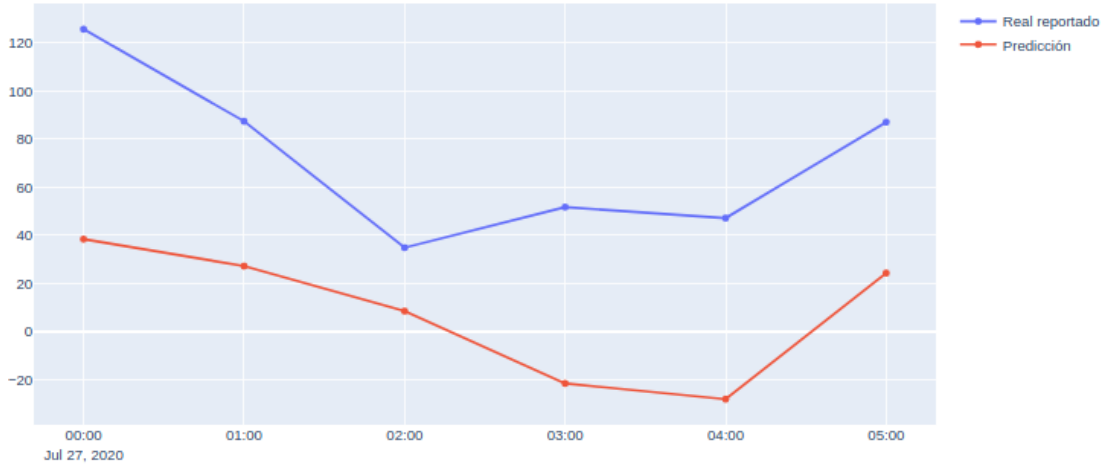

Figura 63: Comparación entre datos reales reportados y predicciones del modelo por defecto - Bloque horario 0-5 [Fuente propia]

Bloque Horario	MAE	RMSE	R2	MAPE
0-5	64.1	66.8	-3.6	97.6

Tabla 7: Cálculo de métricas - Bloque horario 0-5 [Fuente propia]

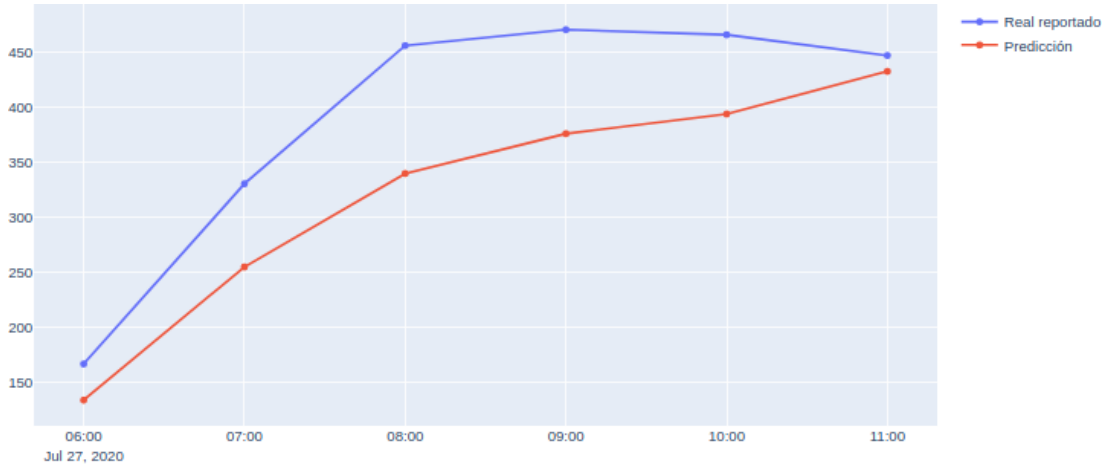

Figura 64: Comparación entre datos reales reportados y predicciones del modelo por defecto - Bloque horario 6-11 [Fuente propia]

Elaboración propia: Los archivos de las figuras 63 a 66 y las tablas 7 a 10 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23), y producen los mismos resultados con el modelo ajustado por m=Prophet() – 2023

Bloque Horario	MAE	RMSE	R2	MAPE
6-11	67.6	75.9	0.5	17.8

Tabla 8: Cálculo de métricas - Bloque horario 6-11 [Fuente propia]

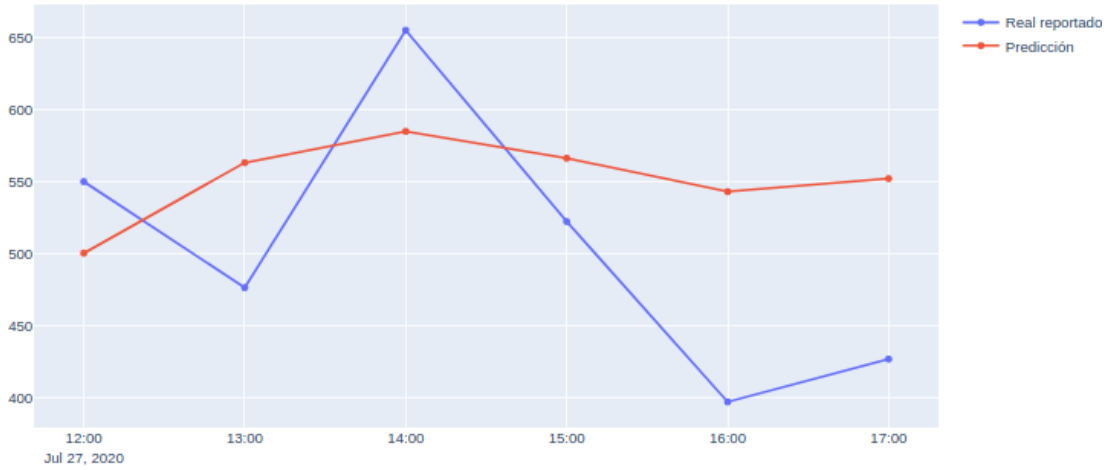

Figura 65: Comparación entre datos reales reportados y predicciones del modelo por defecto - Bloque horario 12-17 [Fuente propia]

Bloque Horario	MAE	RMSE	R2	MAPE
12-17	86.9	94.7	-0.2	18.7

Tabla 9: Cálculo de métricas - Bloque horario 12-17 [Fuente propia]

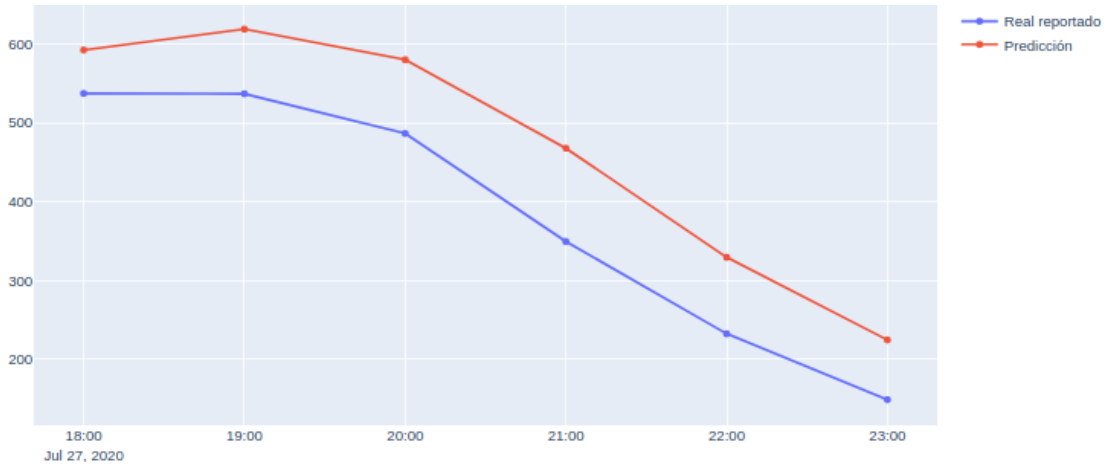

Figura 66: Comparación entre datos reales reportados y predicciones del modelo por defecto - Bloque horario 18-23 [Fuente propia]

Bloque Horario	MAE	RMSE	R2	MAPE
18-23	86.9	89.0	0.6	28.5

Tabla 10: Cálculo de métricas - Bloque horario 18-23 [Fuente propia]

Elaboración propia: Los archivos de las figuras 63 a 66 y las tablas 7 a 10 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23), y producen los mismos resultados con el modelo ajustado por m=Prophet() – 2023

5.3. Predicción de los desplazamientos en base a los datos reales reportados de accidentes

En esta sección se presenta el desarrollo del modelo de predicción final, teniendo en cuenta los datos reales de accidentes de la ciudad de Madrid. El modelo aquí presentado implementa la técnica de pronóstico “Rolling forecasting”, la cual implica actualizar constantemente un modelo de pronóstico a medida que se obtienen nuevos datos. Este modelo ajusta una ventana de predicción horaria y se entrena para una etapa específica de la pandemia. Para el desarrollo de este modelo se incluyen los regresores “t-168Mod’.

5.3.1. Modelado

En esta sección se describe el diseño de los modelos predictivos de movilidad de la ciudad de Madrid, España. En estos modelos, se han incorporado variables complementarias como los accidentes de tránsito o los días previos para comprender el comportamiento de la movilidad en trayectos y días específicos. También, se plantea entrenar múltiples modelos de acuerdo a a la segmentación del día por bloques horarios. La necesidad de ajustar los parámetros del modelo en función del bloque horario en cuestión se debe a que los datos pueden presentar diferentes patrones y tendencias en diferentes momentos del día.

En la figura 67 se presenta el diseño del modelo que pretende pronosticar la franja horaria 0 a 5. Este modelo ajusta una estacionalidad diaria y una estacionalidad personalizada para los días lunes dentro del periodo de nueva normalidad. Para la estacionalidad diaria se observa una mejor representación si los parámetros son configurados manualmente mediante la función de Prophet “add_seasonality”, por lo que se ajusta la estacionalidad con un periodo igual a 1 día y un orden de fourier igual a 4. Con respecto a la estacionalidad personalizada que tiene como nombre “lunes0_5”, se establece un orden de fourier igual a 6 y un periodo equivalente a 1/4 de día que se utiliza para representar las 6 horas presentes en la franja horaria 0 a 5. Por último se incorporan los regresores “t-168Mod” que ayudarán a representar las variaciones horarias de movilidad.

Modelo Final Prophet #: Regresor de la variable accidente para la predicción de la franja horaria 0-5 de los días lunes - Nueva normalidad

```
m=Prophet(changepoint_range=0.8,seasonality_prior_scale=0.3,weekly_seasonality=False,
          daily_seasonality=False, seasonality_mode='additive',changepoint_prior_scale=0.1)
m.add_seasonality(name='Daily', period=1, fourier_order=4,prior_scale=0.007)
m.add_seasonality(name='Lunes0_5', period=1/4, fourier_order=6, condition_name='Lunes0_5',prior_scale=0.01)
m.add_regressor('t-168Mod',mode='additive',prior_scale=0.5) .
```

Figura 67: Configuración de parámetros del Modelo A [Fuente propia]

La gráfica de componentes presentada en la figura 68 ayuda a entender cómo se comportan los datos en función del tiempo y las diferentes fuentes de variación que influyen en ellos, en ella se evidencia la tendencia general, las estacionalidades y el efecto del regresor implementado.

Elaboración propia: Los archivos que generan las figuras 67 a 79 y los resultados de las tablas 11 a 14 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23) – 2023

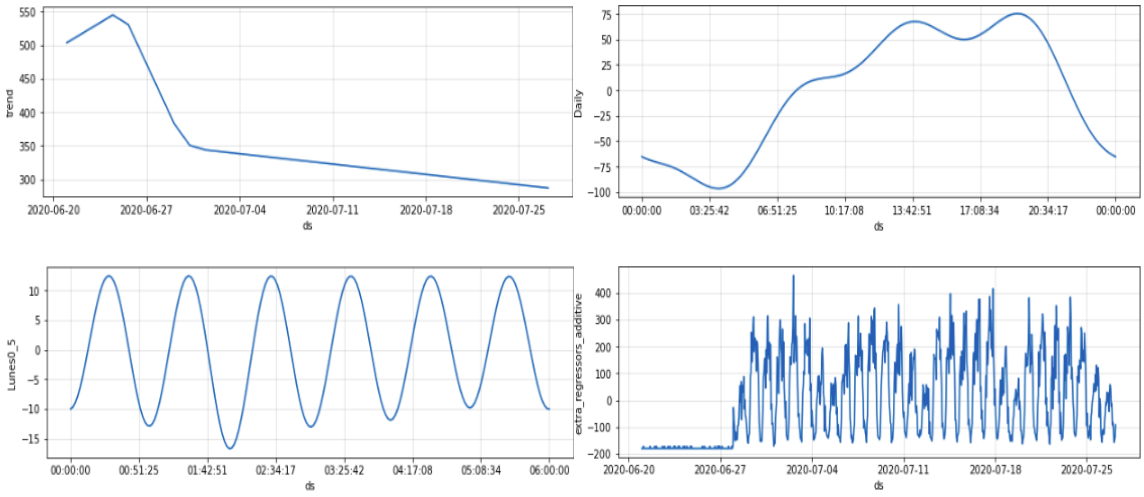

Figura 68: Gráfica de componentes del Modelo A [Fuente propia]

En la figura 69 se muestra la configuración del modelo que realiza la predicción del bloque horario 6 a 11 de la mañana. A diferencia del modelo anterior se ajusta una estacionalidad semanal y se ajusta el parámetro “changepoint_prior_scale” con un valor de uno. Este parámetro determina cuánto cambia la tendencia en los puntos de cambio, por lo que se establece un valor alto para que el modelo sea más sensible a los cambios en los datos y logre capturar variaciones diarias más específicas.

Modelo Final Prophet #: Regresor de la variable accidente para la predicción de la franja horaria 6-11 de los días lunes - Nueva normalidad

```
m=Prophet(changepoint_range=0.8,seasonality_prior_scale=0.3,weekly_seasonality=False,
          daily_seasonality=False, seasonality_mode='additive',changepoint_prior_scale=1)
m.add_seasonality(name='Lunes6_11', period=1/4, fourier_order=6, condition_name='Lunes6_11',prior_scale=0.05)
m.add_seasonality(name='Weekly', period=7, fourier_order=2,prior_scale=0.1)
m.add_seasonality(name='Daily', period=1, fourier_order=4,prior_scale=0.1)
m.add_regressor('t-168Mod',mode='additive',prior_scale=1) .
```

Figura 69: Configuración de parámetros del Modelo B [Fuente propia]

A continuación en la figura 70 se muestra la gráfica de componentes del modelo B, en ella se puede observar la contribución de la tendencia, las estacionalidades y los regresores ajustados.

Elaboración propia: Los archivos que generan las figuras 67 a 79 y los resultados de las tablas 11 a 14 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23) – 2023

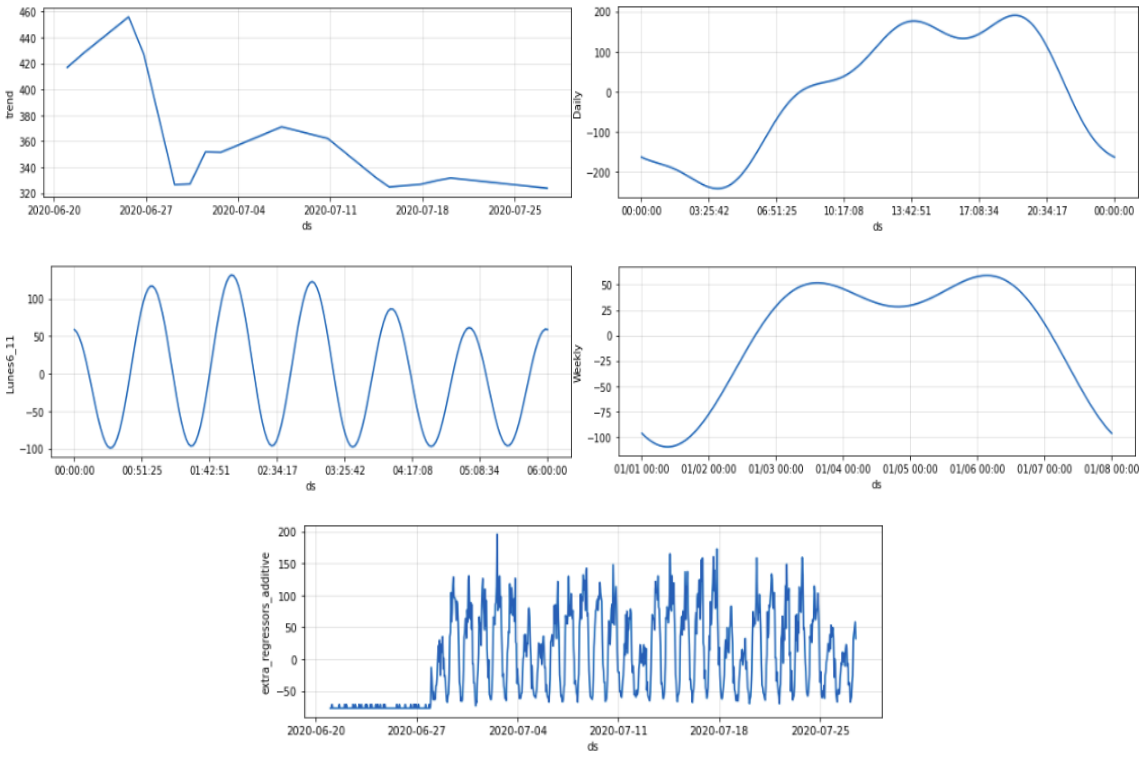

Figura 70: Gráfica de componentes del Modelo B [Fuente propia]

En la figura 71 se muestra la configuración del modelo de predicción de la franja horaria 12 a 17, este modelo logra representar los patrones de movilidad mediante una tendencia, una estacionalidad y el regresor implementado.

Modelo Final Prophet #: Regresor de la variable accidente para la predicción de la franja horaria 12-17 de los días lunes - Nueva normalidad

```
m= Prophet(changepoint_range=0.8,seasonality_prior_scale=0.3,weekly_seasonality=False,
           daily_seasonality=False, seasonality_mode='additive',changepoint_prior_scale=0.1)
m.add_seasonality(name='Lunes12_17', period=1/4, fourier_order=6, condition_name='Lunes12_17',prior_scale=0.008)
m.add_regressor('t-168Mod',mode='additive',prior_scale=0.5) .
```

Figura 71: Configuración de parámetros del Modelo C [Fuente propia]

La gráfica de componentes de la figura 72, además de presentar la tendencia general y la contribución de los regresores, también detalla la estacionalidad personalizada para el día lunes, representada mediante una serie de Fourier que describe minuciosamente las variaciones en la movilidad durante esa franja horaria. Es importante destacar que en este bloque horario, se presentan variaciones en los datos que no pueden ser representadas con precisión mediante las estacionalidades diaria y semanal. Por ende, utilizar estas estacionalidades resulta inadecuado para capturar las fluctuaciones horarias en este bloque específico.

Elaboración propia: Los archivos que generan las figuras 67 a 79 y los resultados de las tablas 11 a 14 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23) – 2023

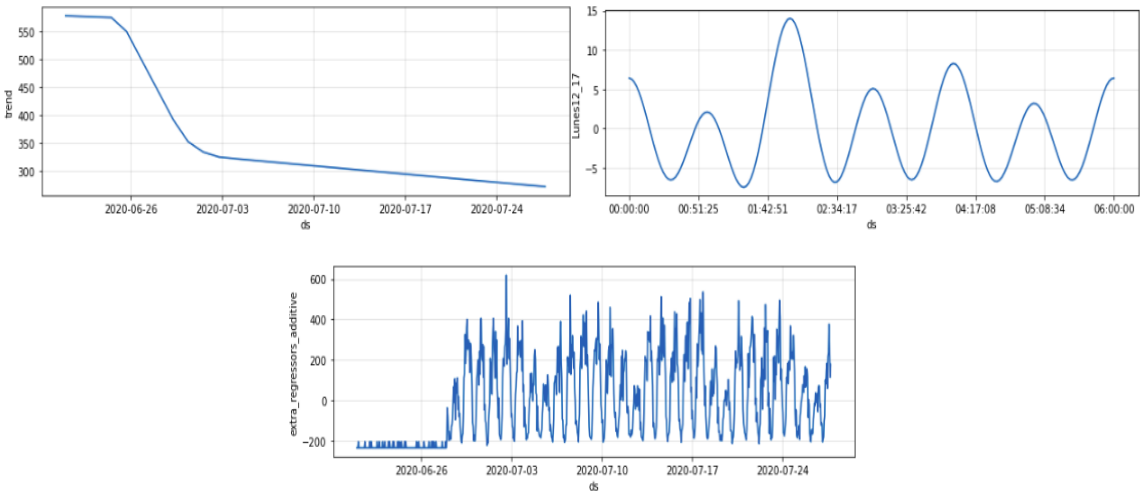

Figura 72: Gráfica de componentes del Modelo C [Fuente propia]

En la figura 73 se muestra la configuración del modelo de predicción para la franja horaria de 18 a 23. En este modelo, se ajustan la estacionalidad semanal y la diaria, pero se excluye la estacionalidad personalizada para el día lunes. Además, se ajusta el regresor que permiten una mejor representación de las variaciones horarias en los datos.

Modelo Final Prophet #: Regresor de la variable accidente para la predicción de la franja horaria 18-23 de los días lunes - Nueva normalidad

```
m=Prophet(changepoint_range=0.8,seasonality_prior_scale=0.3,weekly_seasonality=False,
daily_seasonality=False, seasonality_mode='additive',changepoint_prior_scale=0.1)
m.add_seasonality(name='Weekly', period=7, fourier_order=2,prior_scale=0.005)
m.add_seasonality(name='Daily', period=1, fourier_order=4,prior_scale=0.005)
m.add_regressor('t-168Mod',mode='additive',prior_scale=0.5).
```

Figura 73: Configuración de parámetros del Modelo D [Fuente propia]

De la gráfica de componentes de la figura 74, se obtiene una buena representación de los patrones mediante las estacionalidades semanal y diaria. Esto se debe a que en este bloque no se presentan muchas variaciones y el patrón de movilidad es más estable.

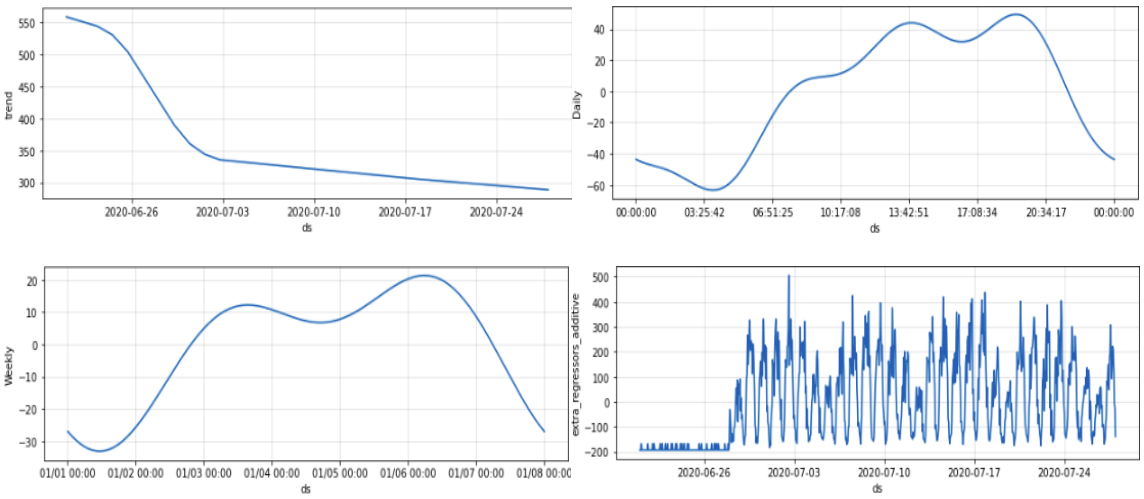

Figura 74: Gráfica de componentes del Modelo D [Fuente propia]

Elaboración propia: Los archivos que generan las figuras 67 a 79 y los resultados de las tablas 11 a 14 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23) – 2023

5.3.2. Validación del modelo

Con la validación se puede verificar que los modelos construidos son adecuados para la predicción de los datos y para los objetivos del proyecto. Por tal razón en esta sección se realiza la validación de los modelos previamente presentados, esto implica dividir los datos en un conjunto de entrenamiento y en un conjunto de validación. El conjunto de validación utilizado para obtener las predicciones se compone del bloque horario correspondiente al último lunes del periodo de nueva normalidad, específicamente el 27 de julio del 2020. Por lo tanto, el conjunto de entrenamiento está compuesto por los datos previos a esta fecha. En la gráfica de la figura 75 se muestra el proceso de validación cruzada. En esta se representa los puntos de corte (cutoffs), el horizonte de predicción de una hora y los datos de entrenamiento para el modelo de bloque horario 0 a 5.

En la gráfica de la figura 75, se aprecia la predicción para el día 27 de julio del año 2020. Las predicciones son representadas mediante los puntos que se encuentran dentro de la franja de color azul, los "cutoffs" representados mediante líneas de color rojo, el horizonte de predicción representado después del último "cutoff" mediante una línea de color gris y los datos de entrenamiento que son representados por medio de los puntos anteriores al primer "cutoff".

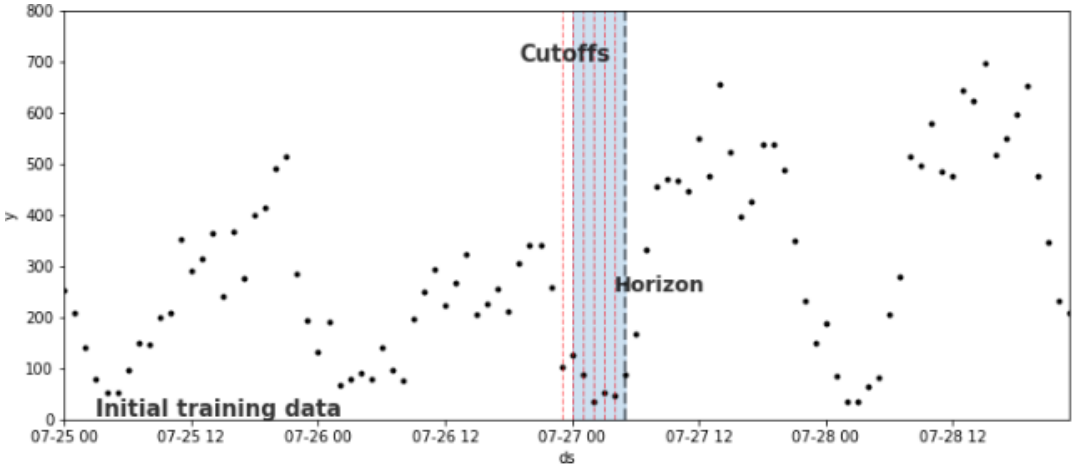

Figura 75: Validación cruzada de la franja horaria 0 a 5 - 27 de julio 2020 [Fuente propia]

El proceso de validación cruzada descrito anteriormente es similar para los otros modelos que realizan predicciones por bloques horarios. En esencia, se utiliza la misma metodología de validación cruzada para evaluar la eficacia predictiva de los modelos, lo que permite seleccionar el más adecuado para la predicción de la movilidad.

En la figura 76 se presenta la comparación entre datos reales y predicciones del bloque horario 0 a 5. El modelo para este intervalo horario ha demostrado ser efectivo, obteniendo buenos resultados en la representación de los patrones de movilidad

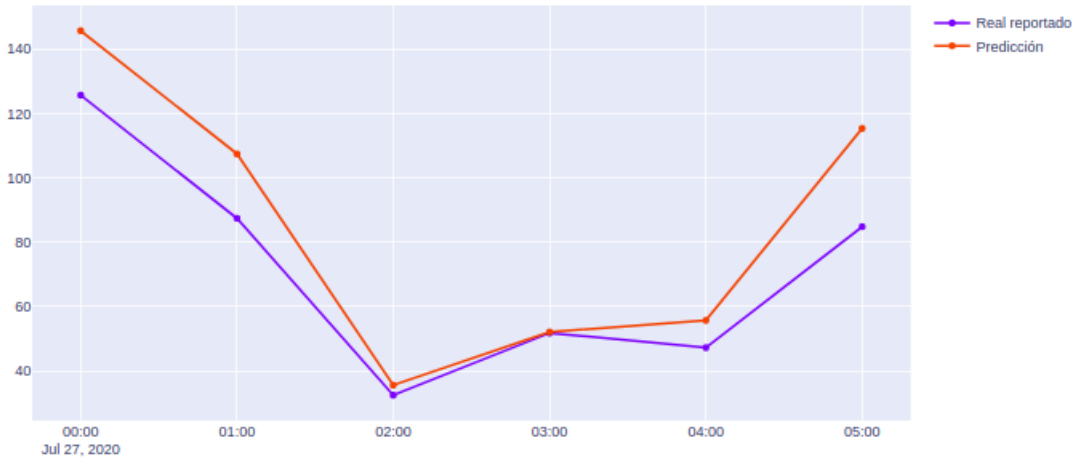

Figura 76: Comparación entre datos reales reportados y predicciones del Modelo A [Fuente propia]

En esta sección, se calculan métricas de desempeño con el objetivo de obtener un modelo que pueda realizar predicciones precisas. En la tabla 11 se presentan los datos reales de movilidad, las predicciones, el error absoluto de la predicción y las métricas de desempeño MAE, RMSE y R2.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-27 00:00:00	125.6	145.6	20.0	13.73	17.37	0.69	17.14 %
2020-07-27 01:00:00	87.4	107.4	20.0				
2020-07-27 02:00:00	32.4	35.5	3.1				
2020-07-27 03:00:00	51.7	52.1	0.31				
2020-07-27 04:00:00	47.2	55.6	8.45				
2020-07-27 05:00:00	84.8	115.3	30.5				

Tabla 11: Validación del modelo A [Fuente propia]

El conjunto de validación para el modelo B está comprendido entre las horas 06:00 y 11:00, en la figura 77 se puede observar una gran similitud entre las dos curvas.

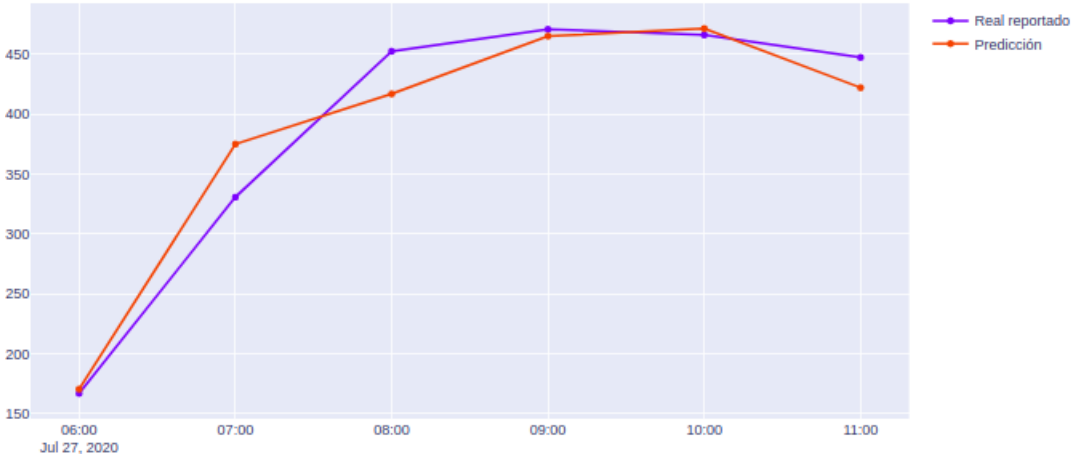

Figura 77: Comparación entre datos reales reportados y predicciones del Modelo B [Fuente propia]

En la tabla 12, la métrica R2 destaca al mostrar un valor cercano a 1.0, lo que indica que

Elaboración propia: Los archivos que generan las figuras 67 a 79 y los resultados de las tablas 11 a 14 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23) – 2023

el modelo tiene una capacidad de predicción excelente.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-27 06:00:00	166.8	170.4	3.52	19.92	25.61	0.94	5.22 %
2020-07-27 07:00:00	330.6	374.8	44.2				
2020-07-27 08:00:00	452.5	416.8	35.6				
2020-07-27 09:00:00	470.7	465.1	5.62				
2020-07-27 10:00:00	466.1	471.3	5.29				
2020-07-27 11:00:00	447.2	422.05	25.24				

Tabla 12: Validación del modelo B [Fuente propia]

El bloque horario comprendido entre las horas 12:00 a 17:00, es el bloque que más variaciones presenta durante el día, a pesar de esto, los resultados presentados en la figura 78 demuestran una gran similitud entre las dos curvas. De esta manera se valida el modelo y se evidencia su capacidad para comprender las variaciones de los datos.

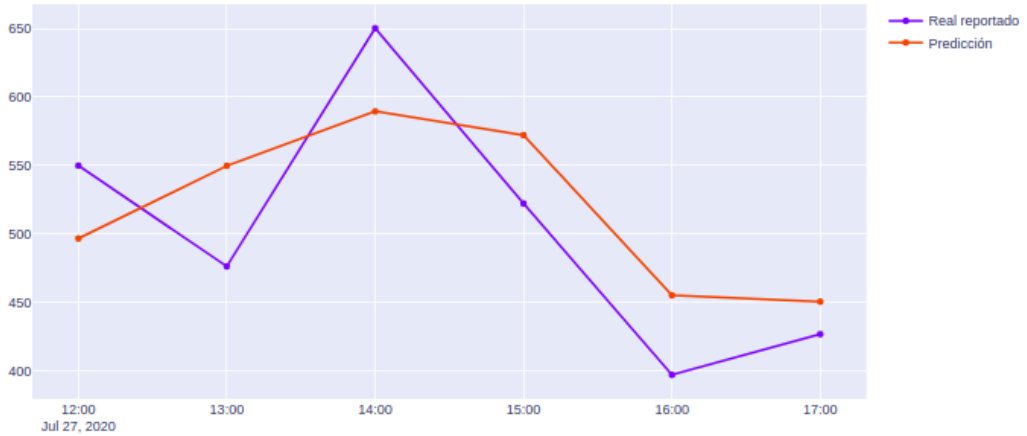

Figura 78: Comparación entre datos reales reportados y predicciones del Modelo C [Fuente propia]

Las métricas de desempeño en la tabla 13 corroboran los resultados que se observaron previamente en la curva de predicción del modelo C.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-27 12:00:00	549.8	496.5	53.3	53.28	55.38	0.56	10.71 %
2020-07-27 13:00:00	476.1	549.7	73.53				
2020-07-27 14:00:00	650.3	589.4	60.82				
2020-07-27 15:00:00	522.0	572.2	50.11				
2020-07-27 16:00:00	396.8	455.0	58.16				
2020-07-27 17:00:00	426.5	450.00	23.75				

Tabla 13: Validación del modelo C [Fuente propia]

Finalmente en la figura 79 se presenta la gráfica comparativa del bloque horario 18:00 a 23:00. En este bloque horario, al no presentar grandes variaciones, se logra representar adecuadamente el patrón de movilidad mediante las estacionalidades diaria y semanal. Asimismo, se destaca la importancia del regresor “t-168Mod” para capturar las características del lunes previo a la predicción.

Elaboración propia: Los archivos que generan las figuras 67 a 79 y los resultados de las tablas 11 a 14 están en el repositorio de GitHub [35] como 1-0-ModeloLunes0-5.ipynb (6-11, 12-17 y 18-23) – 2023

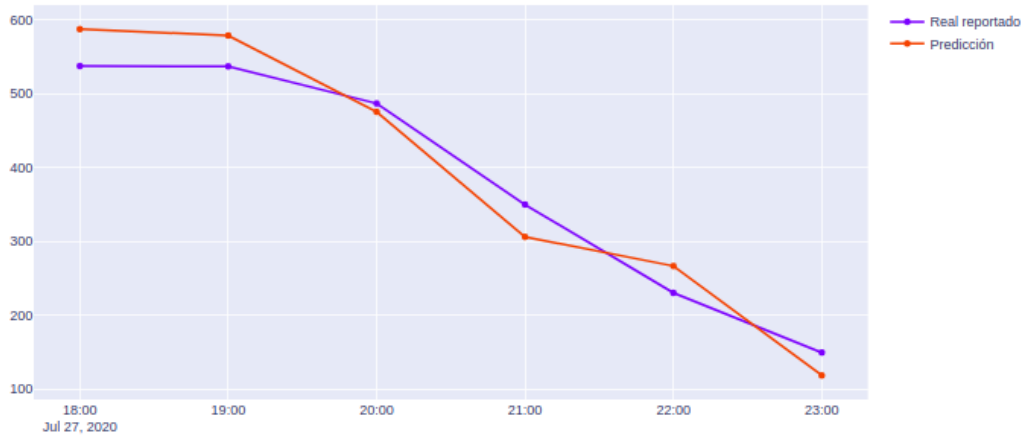

Figura 79: Comparación entre datos reales reportados y predicciones del Modelo D [Fuente propia]

Las métricas de desempeño para este modelo confirman una vez más una buena representación de los patrones de movilidad, como se observó en los modelos anteriores. La métrica R2 destaca por un valor cercano a 1.0, lo cual indica una alta precisión del modelo.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-27 18:00:00	537.2	587.2	49.98	35.73	37.82	0.937	11.43 %
2020-07-27 19:00:00	536.8	578.7	41.90				
2020-07-27 20:00:00	486.5	475.2	11.37				
2020-07-27 21:00:00	349.5	305.8	43.69				
2020-07-27 22:00:00	229.9	266.3	36.44				
2020-07-27 23:00:00	149.1	118.1	31.01				

Tabla 14: Validación del modelo D [Fuente propia]

5.4. Variantes de modelos Prophet entrenados en esta fase

Es importante destacar que los resultados presentados en esta sección son el producto de un proceso de exploración y evaluación de diversos modelos, y es a través de este proceso que se obtiene el modelo final que se presenta en la sección 5.3 de este capítulo. Todos los modelos que se presentan a continuación se encuentran en la sección de Anexos de este documento.

Título	Referencia	Hallazgos
Predicción de los desplazamientos en base a factores meteorológicos	Anexo B	Se incorporan dos variables complementarias relacionadas con las condiciones meteorológicas: temperatura y precipitación. Sin embargo, los resultados de la validación del modelo indican que las variables meteorológicas no son relevantes para considerarlas como variables predictoras del modelo final.
Análisis del conjunto de datos de entrenamiento para la predicción de bloques horarios	Anexo C	Se evalúan tres modelos con diferentes conjuntos de datos para encontrar el mejor conjunto de datos de entrenamiento. Los resultados sugieren que una ventana de datos más amplia y representativa es crucial para capturar los patrones de manera más efectiva
Variables complementarias para mejorar predicción	Anexo D	Se ajustan dos variables complementarias que pretenden explicar los cambios repentinos en el patrón de movilidad. No se consideran en el modelo final para no sobreajustar el modelo.
Modelo con estacionalidad personalizada por día de la semana	Anexo E	Se entrena un modelo que utiliza una estacionalidad personalizada para cada día de la semana, permitiendo una predicción precisa de Lunes a Domingos y en cualquier hora del día. Estos resultados fundamentan la decisión de entrenar el modelo final con estacionalidades personalizadas para cada día de la semana e intervalo horario.

Modelos ajustados para la etapa de desescalada y Nueva normalidad	Anexo F	Se ajustan los modelos para una ventana de predicción de un solo día y se entrenan modelos para dos etapas diferentes de la pandemia (Desescalada y Nueva normalidad). Estos resultados son la base para que el modelo final implemente “Rolling forecasting” para una ventana de predicción de una hora. De igual manera justifica la decisión de entrenar un modelo diferente para cada etapa de la pandemia.
Ajuste de regresores adicionales para los modelos de desescalada y Nueva normalidad	Anexo G	Se realiza una primera exploración del impacto en el entrenamiento del modelo al adicionar variables complementarias. Los resultados encontrados son la base para considerar adicionar variables meteorológicas y de tipo accidente.

Tabla 15: Revisión modelos de predicción [Fuente propia]

CAPÍTULO 6

Fase 5: Evaluación de la precisión del modelo en el estudio de caso de tren de cercanías y Metro de Madrid

En este capítulo, se abordan la fase 5 de la metodología, la cual comprende las actividades relacionadas con la evaluación del modelo en el estudio de caso del tren de cercanías y Metro de Madrid.

Este trabajo de investigación se enfoca en analizar una parte específica de la extensa red de los servicios de transporte público, debido a que el comportamiento puede variar considerablemente en diferentes áreas de la red de transporte. Por lo tanto, para evaluar el modelo en el estudio de caso del tren de cercanías y el Metro de Madrid, es necesario seleccionar una estación dentro de dicha red para llevar a cabo la estimación del número de entradas de pasajeros. En este sentido, la estación Chamartín ha sido elegida como el punto de enfoque principal, considerando su relevancia y flujo constante de pasajeros. En la siguiente sección, se detallan los métodos implementados para llevar a cabo la estimación del número de entradas de pasajeros de dicha estación.

6.1. Estimación del número de usuarios que ingresan a la estación Chamartín en Cercanías y Metro de Madrid

Para realizar la estimación del número de usuarios que ingresan a la estación Chamartín, son necesarios los archivos que se describen en el anexo A.3 y se aplica el método 1, descrito en el trabajo de investigación titulado “Análisis y Diseño de un Sistema para Apoyar un Modelo de Transporte Público Seguro Basado en los Datos de Movilidad Durante la Pandemia en Madrid” [14]. En dicho estudio, se lleva a cabo la estimación del número de usuarios que entran y salen de las estaciones a partir de los datos reales reportados de desplazamientos entre distritos. El actual trabajo investigativo se enfoca en la estimación del número de usuarios que ingresan, por lo que a continuación se describe el método propuesto para llevar a cabo la estimación de interés.

- Método 1: Como punto de partida se calcula el factor Pe_i, el cual relaciona el número de entradas de pasajeros de una determinada estación con el número de desplazamientos realizados en el distrito al que pertenece dicha estación. De tal manera que, si en un distrito se ha realizado una cantidad n de desplazamientos hacia otros distritos, se pueda averiguar que fracción de esos desplazamientos corresponde a viajes que se han realizado a través de la estaciones que se encuentren en el interior de ese distrito.

  Expresado matemáticamente se tiene:

  $$Pe_i = \frac{e_{\mathrm{ref}_i}}{do_{\mathrm{ref}_i}}$$

  Pe_i: Factor para las entradas de pasajeros en cada estación i.
  e_{ref_i}: Número de entradas de pasajeros en cada estación i durante el periodo de

referencia.
do_{ref_i}: Número de desplazamientos de un distrito hacia el resto para cada estación i en el periodo de referencia.

Teniendo en cuenta el factor Pe_i anterior, el cálculo del número de entradas de pasajeros se expresaría de la siguiente manera:

$$e_i = do_i Pe_i$$

e_i: Número de entradas de pasajeros en cada estación i.
do_i: Desplazamientos de un distrito hacia el resto para cada estación i.

Finalmente, para que el número de usuarios que entran a las estaciones coincida con el número de usuarios reales que han utilizado el transporte público durante un periodo de tiempo (e.g. 1 mes), se va a aumentar o disminuir homogéneamente la cantidad de usuarios que entran en las estaciones por medio de un factor P'.

$$e_T = \sum_{i=1}^{n} e_i$$

$$\frac{e_T}{P'} = T$$

e_T: Total de entradas de pasajeros después de realizar la sumatoria de las n estaciones.
P': Factor para aumentar o disminuir la cantidad de usuarios que entran en las estaciones.
T: Número de usuarios reales en un periodo de tiempo concreto.

6.2. Evaluación del modelo predictivo

Para realizar la evaluación del modelo predictivo en el estudio de caso, es necesario estimar el número de entradas de pasajeros utilizando las predicciones de desplazamientos entre el distrito “Chamartín-Centro” y los valores reales de las entradas de pasajeros por hora en la estación Chamartín. Sin embargo, debido a la falta de datos por hora de las entradas de pasajeros a las estaciones de los servicios de tren de cercanías y metro de Madrid durante la pandemia, se procede a evaluar el modelo utilizando los valores estimados de las entradas de pasajeros en la estación Chamartín en lugar de los datos reales reportados. A continuación, se presenta un diagrama que explica el procedimiento llevado a cabo en esta sección.

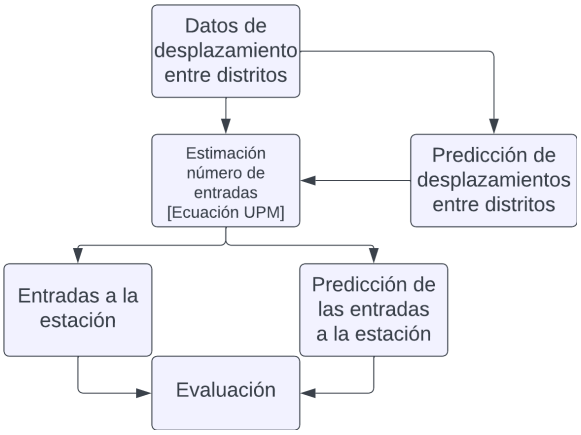

Figura 80: Procedimiento para la evaluación del modelo [Fuente propia]

En esta sección se procede a evaluar el modelo para el servicio de Cercanías y para Metro en dos fechas distintas: Fecha 1 (20 de julio de 2020) y Fecha 2 (27 de julio de 2020). Al igual que en el capítulo anterior, se utilizan las mismas métricas de evaluación que se emplearon en la validación del modelo.

6.2.1. Estudio de caso: Tren de Cercanías

En la tabla 16 se presentan los resultados de la evaluación, en donde la columna “y” representa el número estimado de entradas de pasajeros a la estación Chamartín y la columna “yhat” representa la predicción de las entradas de pasajeros a la estación. En cuanto a la columna “Error Absoluto”, se puede observar que hay un error más significativo en las horas 14:00 y 16:00. Estas horas corresponden a momentos en los que el modelo tuvo dificultades para representar el patrón observado el 20 de julio.

Por otro lado, las métricas de evaluación evidencian una alta precisión del modelo en la predicción del número de entradas en la estación Chamartín. En particular, el coeficiente R2 muestra un valor cercano a 1, lo que indica un ajuste muy bueno del modelo a los datos observados. Además, el valor de la métrica MAPE es inferior al 1 %, lo que señala un bajo porcentaje de error medio absoluto. Es importante destacar que estas métricas se calculan considerando las 24 horas de la Fecha 1.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-20 00:00:00	222.0	219.7	2.3				
2020-07-20 01:00:00	139.1	135.3	3.8				
2020-07-20 02:00:00	96.8	98.5	1.7				
2020-07-20 03:00:00	92.1	93.3	1.2				
2020-07-20 04:00:00	109.6	110.3	0.7				
2020-07-20 05:00:00	198.4	198.2	0.2				
2020-07-20 06:00:00	423.7	423.3	0.4				
2020-07-20 07:00:00	666.0	668.2	2.2				
2020-07-20 08:00:00	726.3	721.7	4.6				
2020-07-20 09:00:00	672.1	674.8	2.7				
2020-07-20 10:00:00	624.8	627.5	2.7				
2020-07-20 11:00:00	704.6	702.9	1.7	3.66	6.01	0.99	0.69 %
2020-07-20 12:00:00	877.6	876.8	0.8				
2020-07-20 13:00:00	1164.1	1162.8	1.3				
2020-07-20 14:00:00	1424.0	1445.0	21.0				
2020-07-20 15:00:00	1218.3	1221.0	2.7				
2020-07-20 16:00:00	950.7	934.0	16.7				
2020-07-20 17:00:00	983.8	986.8	3.0				
2020-07-20 18:00:00	1030.7	1026.9	3.8				
2020-07-20 19:00:00	981.6	985.8	4.2				
2020-07-20 20:00:00	892.8	895.9	3.1				
2020-07-20 21:00:00	674.5	675.4	0.9				
2020-07-20 22:00:00	473.2	477.0	3.8				
2020-07-20 23:00:00	238.6	241.2	2.6				

Tabla 16: Evaluación de la predicción para el servicio de tren de cercanías - Fecha 1 [Fuente propia]

Elaboración propia: Resultados de la tabla 16 disponibles en el repositorio de github [35] con el nombre de Evaluacion_Entradas_Est.ipynb – 2023

La tabla 17 presenta la evaluación del servicio de tren de cercanías para la Fecha 2. En esta tabla, se puede observar que los valores pronosticados son muy cercanos a los valores estimados, de igual manera las métricas de evaluación arrojan mejores resultados si se comparan con los resultados de la Fecha 1. Lo anterior indica un alto nivel de ajuste de los modelos entrenados para los diferentes segmentos del día.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-27 00:00:00	220.0	219.8	0.2	1.33	1.68	0.99	0.31 %
2020-07-27 01:00:00	145.0	145.9	0.9				
2020-07-27 02:00:00	87.3	87.1	0.2				
2020-07-27 03:00:00	90.5	89.2	1.3				
2020-07-27 04:00:00	106.4	106.0	0.4				
2020-07-27 05:00:00	197.5	199.5	2.0				
2020-07-27 06:00:00	396.5	396.4	0.1				
2020-07-27 07:00:00	642.4	645.8	3.4				
2020-07-27 08:00:00	675.6	675.2	0.4				
2020-07-27 09:00:00	651.9	652.5	0.6				
2020-07-27 10:00:00	649.2	649.3	0.1				
2020-07-27 11:00:00	687.5	688.0	0.5				
2020-07-27 12:00:00	827.7	824.2	3.5				
2020-07-27 13:00:00	1124.9	1128.0	3.1				
2020-07-27 14:00:00	1339.3	1341.9	2.6				
2020-07-27 15:00:00	1124.0	1125.6	1.6				
2020-07-27 16:00:00	865.0	866.1	1.1				
2020-07-27 17:00:00	870.9	873.3	2.4				
2020-07-27 18:00:00	997.2	998.5	1.3				
2020-07-27 19:00:00	929.2	930.2	1.0				
2020-07-27 20:00:00	851.5	849.5	2.0				
2020-07-27 21:00:00	652.9	653.4	0.5				
2020-07-27 22:00:00	467.5	468.9	1.4				
2020-07-27 23:00:00	257.9	256.3	1.6				

Tabla 17: Evaluación de la predicción para el servicio de tren de cercanías - Fecha 2 [Fuente propia]

6.2.2. Estudio de caso: Metro

Para llevar a cabo la evaluación de los modelos de predicción en el servicio de metro, se procede de la misma manera que en el estudio de caso anterior. En primer lugar, se unen todas las predicciones correspondientes a la Fecha 1 en una tabla, y de igual manera se agrupan las predicciones de la Fecha 2 en otra tabla. Este enfoque permite calcular las métricas de evaluación pertinentes para cada una de las tablas generadas, tomando en consideración las 24 horas del día específico que se está evaluando en cada caso. Al dividir las predicciones en tablas separadas, se facilita el análisis comparativo y la obtención de métricas más precisas para cada fecha en particular.

Elaboración propia: Resultados de la tabla 17 disponibles en el repositorio de github [35] con el nombre de Evaluacion_Entradas_Est.ipynb – 2023

En la tabla 18 se presentan los resultados que se obtuvieron para la evaluación de la fecha 1 en el servicio de metro. Los resultados de la columna error absoluto nuevamente indican un error más significativo en las horas 14:00 y 16:00 ya que como se ha mencionado previamente, el modelo para este segmento del día no logra representar adecuadamente el patrón de comportamiento para estas horas. Por otro lado las métricas de evaluación siguen indicando un buen ajuste general del modelo para la fecha 1.

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-20 00:00:00	117.3	116.0	1.3				
2020-07-20 01:00:00	73.5	71.5	2.0				
2020-07-20 02:00:00	51.2	52.0	0.8				
2020-07-20 03:00:00	48.7	49.3	0.6				
2020-07-20 04:00:00	57.9	58.2	0.3				
2020-07-20 05:00:00	104.9	104.7	0.2				
2020-07-20 06:00:00	223.9	223.6	0.3				
2020-07-20 07:00:00	352.0	353.0	1.0				
2020-07-20 08:00:00	383.9	381.2	2.7				
2020-07-20 09:00:00	355.2	356.5	1.3				
2020-07-20 10:00:00	330.2	331.5	1.3				
2020-07-20 11:00:00	372.4	371.3	1.1	1.90	3.13	0.99	0.68 %
2020-07-20 12:00:00	463.9	463.2	0.7				
2020-07-20 13:00:00	615.3	614.3	1.0				
2020-07-20 14:00:00	752.7	763.3	10.6				
2020-07-20 15:00:00	644.0	645.0	1.0				
2020-07-20 16:00:00	502.5	493.4	9.1				
2020-07-20 17:00:00	520.0	521.3	1.3				
2020-07-20 18:00:00	544.8	542.5	2.3				
2020-07-20 19:00:00	518.8	520.7	1.9				
2020-07-20 20:00:00	471.9	473.3	1.4				
2020-07-20 21:00:00	356.5	356.8	0.3				
2020-07-20 22:00:00	250.1	252.0	1.9				
2020-07-20 23:00:00	126.1	127.4	1.3				

Tabla 18: Evaluación de la predicción para el servicio de metro - Fecha 1 [Fuente propia]

En la tabla 19 se evidencian los resultados encontrados para la fecha 2. Al igual que se pudo evidenciar en el servicio de cercanías, en esta fecha en particular, se destaca la precisión de las predicciones y evidencia un error absoluto mayoritariamente por debajo de 2. En términos de precisión en las predicciones, las métricas de esta tabla reflejan los mejores resultados obtenidos.

Elaboración propia: Resultados de la tabla 18 disponibles en el repositorio de github [35] con el nombre de Evaluacion_Entradas_Est.ipynb – 2023

Ds	y	y_hat	Error Absoluto	MAE	RMSE	R2	MAPE
2020-07-27 00:00:00	115.0	114.9	0.1				
2020-07-20 01:00:00	75.8	76.3	0.5				
2020-07-27 02:00:00	45.6	45.5	0.1				
2020-07-27 03:00:00	47.3	46.6	0.7				
2020-07-27 04:00:00	55.6	55.4	0.2				
2020-07-27 05:00:00	103.3	104.3	1.0				
2020-07-27 06:00:00	207.3	207.2	0.1				
2020-07-27 07:00:00	336.0	337.6	1.6				
2020-07-27 08:00:00	353.3	353.0	0.3				
2020-07-27 09:00:00	341.0	341.1	0.1				
2020-07-27 10:00:00	339.5	339.4	0.1				
2020-07-27 11:00:00	359.6	359.7	0.1	0.62	0.82	0.99	0.30 %
2020-07-27 12:00:00	432.9	430.9	2.0				
2020-07-27 13:00:00	588.4	589.7	1.3				
2020-07-27 14:00:00	700.5	701.6	1.1				
2020-07-27 15:00:00	587.9	588.4	0.5				
2020-07-27 16:00:00	452.4	452.8	0.4				
2020-07-27 17:00:00	455.5	456.6	1.1				
2020-07-27 18:00:00	521.5	522.0	0.5				
2020-07-27 19:00:00	486.0	486.3	0.3				
2020-07-27 20:00:00	445.3	444.1	1.2				
2020-07-27 21:00:00	341.5	341.6	0.1				
2020-07-27 22:00:00	244.5	245.1	0.6				
2020-07-27 23:00:00	134.9	134.0	0.9				

Tabla 19: Evaluación de la predicción para el servicio de metro - Fecha 2 [Fuente propia]

De los resultados presentados en esta sección, es importante destacar que las predicciones se han realizado a nivel horario para el trayecto “Chamartín-Centro”. Sin embargo, para estimar el número de entradas a la estación Chamartín, se ha empleado el archivo timeseries_o.csv, el cual contiene la sumatoria de todos los viajes con un mismo distrito origen en común. Esto implica que para obtener un número de entradas de pasajeros a la estación que sea completamente el resultado de un pronóstico, es necesario el entrenamiento de un modelo para cada uno de los trayectos involucrados, por ejemplo un modelo para Chamartín-Retiro otro modelo para Chamartín-ElPardo otro modelo para Chamartín-Salamanca, y así sucesivamente para todos los trayectos con origen en Chamartín.

Elaboración propia: Resultados de la tabla 19 disponibles en el repositorio de github [35] con el nombre de Evaluacion_Entradas_Est.ipynb – 2023

CAPÍTULO 7

Fase 6: Visualización de los datos mediante una aplicación web

Como parte de la fase 6 de la metodología, se organiza el conocimiento adquirido para presentar los resultados de la investigación mediante una aplicación web.

Para la visualización de los datos, se han desarrollado dos aplicaciones web, una aplicación para el servicio de transporte de tren de cercanías y otra aplicación para el servicio de metro. Las aplicaciones constan de un panel de control situado en el lado izquierdo, el cual le permite al usuario manipular la información que se presenta en la gráfica del lado derecho.

Los datos de las predicciones que se podrán visualizar en las aplicaciones son resultados generados a partir de la ejecución de la aplicación en un ambiente controlado, ya que los datos por hora que se le están inyectando a la aplicación corresponden a datos de prueba dentro del periodo de nueva normalidad que se ha estudiado. Se hace de esta manera debido a que no se tienen datos reales en una nueva pandemia. Esto quiere decir que el prototipo de la aplicación que se presenta en este trabajo está pensada para predecir el número de entradas de pasajeros en la estación Chamartín en una futura pandemia, por lo que en un sistema de producción se le deben de inyectar datos reales y en tiempo real para que el sistema realice las predicciones por hora.

Al iniciar la aplicación por primera vez, la gráfica se actualiza con los datos reales históricos del último lunes en el periodo de nueva normalidad y para realizar una predicción en la aplicación web, el usuario debe de seleccionar la fecha y la hora en el panel de control.

En la figura 81 se puede apreciar la interfaz de usuario de la aplicación después de haber realizado la estimación de las entradas de pasajeros en la estación chamartín para el lunes 13 de julio de 2020, esta fecha corresponde al último lunes con el que el modelo fue entrenado y las fechas posteriores corresponden a una predicción. También se puede observar en esta figura como se selecciona la primera fecha y hora que el sistema va a predecir (20 de julio de 2020 00:00:00). Como el sistema de predicción se esta ejecutando dentro de un ambiente controlado, se inician las predicciones a partir del lunes siguiente al 13 de julio de 2020 y en la hora 0.

Elaboración propia: La aplicación web de las figuras 81 y 82 están disponibles en el repositorio de github [35] dentro del directorio que tiene como nombre ”App“ – 2023

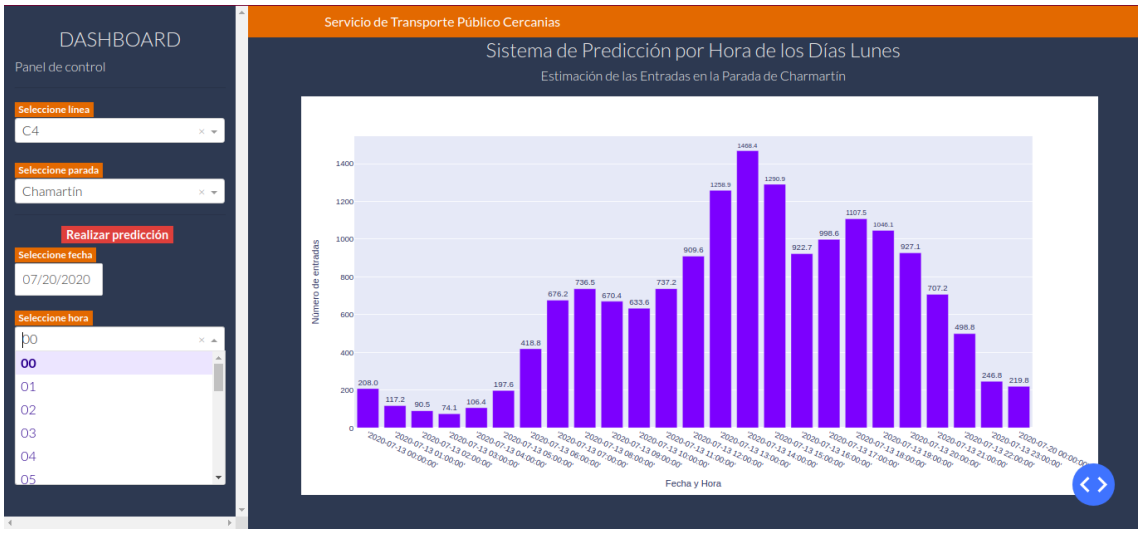

Figura 81: Interfaz de la aplicación web para la predicción de entradas de pasajeros en el servicio de tren de cercanías [Fuente propia]

En la figura 82 se puede apreciar la interfaz de usuario para el servicio de metro. De igual manera en la gráfica se logra apreciar las predicciones para las horas 0 y 1 del 20 de julio de 2020, se van seleccionando las horas una hora en adelante para simular la inyección de los datos en cada hora. Para realizar una predicción, el modelo requiere una actualización en tiempo real de los datos reales de accidentes ocurridos una hora atrás y los datos reales de desplazamiento del lunes previo, por lo que el modelo se estaría entrenando cada hora con la inyección de los nuevos datos.

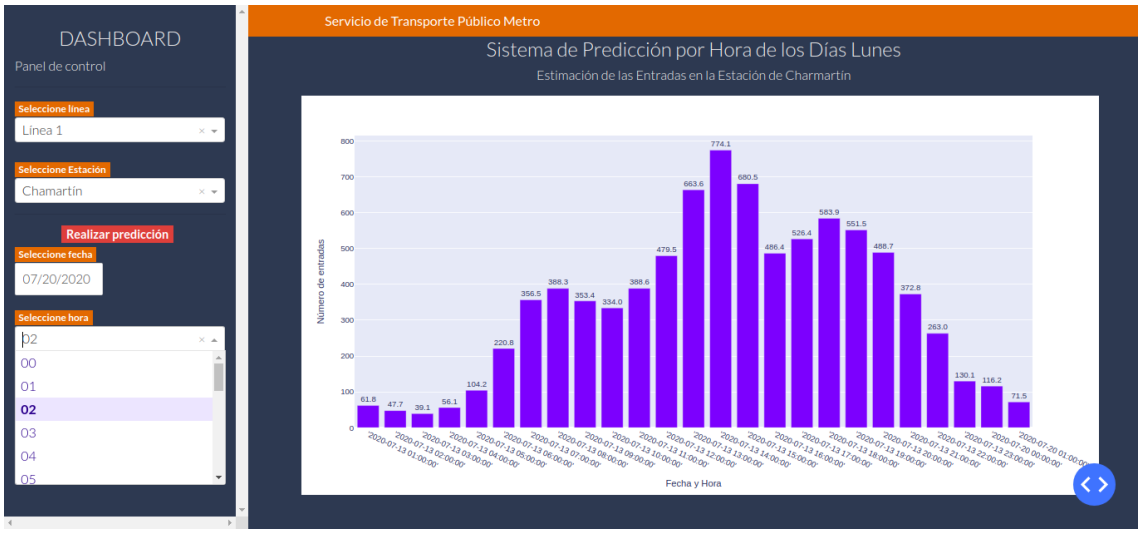

Figura 82: Interfaz de la aplicación web para la predicción de entradas de pasajeros en el servicio de metro [Fuente propia]

Las dos aplicaciones diseñadas han experimentado una optimización en sus tiempos de ejecución al emplear el framework de computación paralela denominado Ray. Los tiempos de ejecución son calculados en un dispositivo de cómputo con un procesador Intel Core i5 de 4 núcleos a 2,2 GHz y 12 GB de memoria RAM. En el caso de la aplicación para el servicio de cercanías, se estima un tiempo de ejecución de aproximadamente 2 minutos y 30 segundos la primera vez que se ejecuta. Una vez que los datos han sido cargados, los tiempos de ejecución mejoran considerablemente, reduciéndose a aproxi-

Elaboración propia: La aplicación web de las figuras 81 y 82 están disponibles en el repositorio de github [35] dentro del directorio que tiene como nombre ”App“ – 2023

madamente 35 segundos para realizar las predicciones. Por otro lado, la aplicación para el servicio de metro presenta tiempos de ejecución más prolongados. La primera vez que se ejecuta, se estima un tiempo de aproximadamente 5 minutos, y posteriormente oscila entre 3 y 4 minutos. Durante la ejecución de la predicción, la ventana del navegador web muestra el mensaje “Updating”, indicando que se está llevando a cabo la predicción.

CAPÍTULO 8

Conclusiones y líneas de trabajo futuras

8.1. Conclusiones

- Tras analizar los cambios en los patrones de movilidad a lo largo del avance de la pandemia, se logró identificar que para obtener mejores resultados en las predicciones horarias es mejor entrenar un modelo para cada etapa de la pandemia, ya que cada etapa presentaba patrones propios. Esto resalta la relevancia de identificar y seleccionar cuidadosamente el periodo de estudio en el desarrollo de modelos predictivos en una pandemia, garantizando así una mayor precisión y fiabilidad en las predicciones obtenidas.

- Para identificar información clave en grandes volúmenes de datos, es esencial conservar el nivel de detalle para no perder datos valiosos. En este estudio de búsqueda de patrones, se mantuvo la granularidad horaria de los datos, lo que resultó en decisiones cruciales durante el entrenamiento de los modelos. Por ejemplo, se decidió segmentar el modelo según los días de la semana y un intervalo horario específico.

- Durante el proceso de modelado se evidenció que para mejorar la capacidad de un modelo de predicción de desplazamientos es necesario incluir variables complementarias que permitan explicar las fluctuaciones en los datos. En este trabajo se exploraron las condiciones meteorológicas como la temperatura y la precipitación, así como también, las incidencias de tráfico en la ciudad de Madrid. Demostrando que variables como temperatura y precipitación solo tendrían un impacto significativo en la ciudad de Madrid en caso de condiciones extremas. Por el contrario las variables de accidentes de tráfico evidenciaron mejorar las predicciones de los modelos.

- La segmentación por bloques horarios utilizada para entrenar cuatro modelos predictivos para los días lunes demostró ser la metodología más efectiva en el contexto de una pandemia, ya que se logró capturar patrones de movimiento que varían durante diferentes momentos del día, como las horas pico de tráfico, los períodos de menor actividad y las tendencias nocturnas.

- Al analizar la eficacia de los modelos de predicción, se determinó que incluir los registros horarios de desplazamientos de un lunes de la semana anterior como variable complementaria resulta altamente eficiente para realizar predicciones en horas específicas. Esta observación llevó a la adaptación de la variable complementaria "t-168Mod", que combina la información sobre accidentes con los datos de desplazamientos del lunes inmediatamente anterior.

- Durante la evaluación de los modelos entrenados, se pudo notar que cada métrica de evaluación tiene una sensibilidad específica frente a los errores de predicción. En particular, las métricas RMSE y R2 demostraron ser especialmente sensibles a estos errores, mientras que la métrica MAPE mostró una variación mínima en los resultados cuando se presentaron errores. En este trabajo se emplean varias métri-

cas para asegurar una comprensión más completa del rendimiento de los modelos entrenados.

- Se identificó una mayor complejidad para representar los patrones de movilidad en el bloque horario 12 a 17 debido a que este intervalo horario presenta una mayor variabilidad de los datos. Esta variabilidad también se reflejó en las gráficas de la sección de análisis de patrones y podría estar relacionada con los accidentes de tráfico que provocan aumentos o disminuciones de los desplazamientos en ciertas horas específicas, tal como se justifica en este trabajo.

8.2. Líneas de trabajo futuras

- En un trabajo futuro se podría explorar más las características de la herramienta “Matrixprofile” para la búsqueda de patrones en el contexto de movilidad. En el anexo G del presente trabajo se realiza una exploración inicial.
- Para trabajos futuros en esta línea de investigación se podría continuar con el entrenamiento de modelos para otros trayectos que tengan como origen el distrito Chamartín. De esta forma, se obtendría un sistema de predicción robusto para pronosticar las entradas de pasajeros a la estación Chamartín por horas. Esta ampliación proporcionaría una visión más precisa de los patrones de desplazamiento en esa área específica.
- Se logró identificar que algunos parques importantes de la ciudad son cerrados debido a condiciones meteorológicas adversas. Por lo anterior, en un trabajo futuro, se podría profundizar en la posibilidad de que el cierre de estos parques tenga un impacto significativo en la movilidad de la ciudad.
- En un trabajo futuro se podrían explorar variables complementarias que mejoren la capacidad de los modelos para comprender las variaciones en los datos horarios. Asimismo, se podría mejorar el modelo presentado en este estudio para el bloque horario de 12 a 17, optimizando el comportamiento de los regresores “Accup” y “Accdown” para que reflejen adecuadamente las variaciones horarias.

A. Anexo - Recopilación de datos

A.1. Datos GTFS

Este formato de datos es fundamental en esta investigación para identificar las líneas de metro y tren de cercanías en la zona metropolitana de Madrid. En el portal de datos abiertos del consorcio regional de transportes de Madrid, se obtienen el conjunto de datos de la red de metro y la red de cercanías en formato GTFS [42].

El estándar GTFS define una serie de archivos de texto que contienen información clave sobre el sistema de transporte público. Los archivos incluidos para la red de metro y la red de cercanías se muestran a continuación:

- **"agency.txt":** Información sobre la agencia responsable del servicio.
- **"calendar.txt":** Información sobre los días en los que los servicios están disponibles.
- **"calendar_dates.txt":** Información sobre fechas específicas en las que hay cambios en la operación del servicio de transporte.
- **"fare_attributes.txt"** Información sobre tarifas.
- **"fare_rules.txt":** Información sobre reglas tarifarias.
- **"feed_info.txt":** Información general sobre el conjunto de datos GTFS.
- **"frequencies.txt":** Información sobre la frecuencia con la que se ofrecen los servicios en una determinada ruta y horario.
- **"routes.txt":** Información sobre las rutas de transporte disponibles.
- **"shapes.txt":** Geometría de cada trayecto que dibuja el camino que ha de seguir el vehículo.
- **"stops.txt":** Información sobre las paradas y su ubicación geográfica.
- **"stop_times.txt":** Horarios de llegada y salida en cada parada para los viajes.
- **"trips.txt":** Información sobre los viajes o servicios específicos dentro de las rutas.

A.2. Datos reales de desplazamientos entre distritos

Este trabajo de investigación se ha llevado a cabo gracias a los datos abiertos proporcionados por el gobierno de España, a través del Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA). El MITMA ha publicado una gran cantidad de información sobre la movilidad en España durante el periodo de pandemia por el COVID-19 a nivel nacional, la información ha sido recopilada a través de los registros de ubicación de los teléfonos móviles. Estos datos brindan una perspectiva valiosa sobre los patrones de desplazamiento y comportamiento de la población.

Los datos han sido tomados de la página oficial del Ministerio de Transportes, Movilidad y Agenda Urbana [7]. En la página web se encuentra la información estructurada en los siguientes directorios:

- **Maestra1-mitma-distritos:**

  Dentro de este directorio se encuentran datos reales reportados sobre los desplazamientos entre distritos, organizados por días y por meses completos. Dentro de este directorio se encuentran archivos de texto comprimido con campos separados por "|" (barra vertical). A continuación se indica la estructura de cada uno de los archivos:

  fecha | origen | destino | actividad_origen | actividad_destino | residencia | edad | periodo | distancia | viajes | viajes_km

- **Maestra2-mitma-distritos:**

  Este directorio contiene los archivos con las matrices de viajes por persona. La estructura de cada uno de los archivos se muestra a continuación:

  fecha | distrito | numero_viajes | personas

  Estos archivos de texto comprimido contienen el número de personas que han realizado viajes en cada fecha y son separados de igual manera por "|" (barra vertical) y los valores numéricos tienen "." (punto) como separador de decimales.

- **Maestra1-mitma-municipios** y **Maestra2-mitma-municipios:** En estos directorios se ofrece la misma información con una división del territorio a nivel de municipios o agrupaciones de estos en el caso de zonas con poca población.

A su vez, en el directorio principal aparecen los siguientes archivos:

- **relaciones_distrito_mitma.csv.** Contiene la relación entre los distritos y la zonificación utilizada.
- **zonificación_distritos.zip.** Contiene un shapefile con la zonificación utilizada correspondiente a distritos o a una agrupación de estos.

En este trabajo de investigación se hará uso únicamente de los archivos que se encuentran dentro del directorio "Maestra1-mitma-distritos" y del archivo "zonificación_distritos.zip".

Los otros directorios previamente descritos, se mencionan en este apartado para proporcionar una visión completa de todos los datos disponibles.

A.3. Datos reales de servicios de transporte público

Los servicios de transporte de Cercanías y Metro Madrid, comparten información acerca de su infraestructura y su funcionamiento. Los siguientes archivos se obtienen de las páginas web de cada servicio de transporte y son necesarios dentro del proyecto de investigación para realizar la estimación del número de usuarios en una de las estaciones de los sistemas de transporte público.

- **07102000.XLS** [43]: Archivo que contiene un resumen mensual de los desplazamientos efectuados desde 2013 hasta marzo de 2021 en los servicios de Cercanías de Madrid y Barcelona. La estructura de sus datos se presenta a continuación.

  | Periodo | Total | Madrid | Barcelona | Urbanas e Interurbanas | Ancho métrico |

- **Ref._PA049_Demanda_Diaria_2023_04.xlsx** [44]: El archivo contiene el recuento diario de usuarios de todo el servicio de Metro desde 2020 hasta 2023. El archivo contiene varias hojas de cálculo, una para cada año. La estructura de sus datos se presenta a continuación.

  | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |

Asimismo, para lograr realizar la estimación del número de usuarios en una de las estaciones de los sistemas de transporte público, se utilizarán algunos archivos en formato CSV obtenidos del trabajo de investigación de la Universidad Politécnica de Madrid en el año 2021 [14]. A continuación se detallan los archivos relevantes para este proyecto de investigación.

1. **routes.csv**: Archivo de rutas de los servicios de transporte de Cercanías y Metro de Madrid generados a partir de archivos GTFS. A continuación se describen las columnas del archivo csv.
   - direction: Puede ser 0 (ida) o 1 (vuelta).
   - stop_district: Contiene el id del distrito asociado a la parada en la que se encuentra.
   - stop_id: Contiene la id de la parada (se mantienen las mismas ids que se suministraban en los GTFS)
   - stop_lat: Indica la latitud de la ubicación de la parada.
   - stop_lon: Indica la longitud de la ubicación de la parada.
   - stop_name: Contiene el nombre de la parada (Ej. Chamartín).
   - line: Indica el id de la línea (Ej. C3).

- service: Puede ser cualquiera de los tres servicios de transporte público: “cercanías” o “metro”.

2. **up_down_bystop.csv y metro_up_down_bystop.csv**: Archivos que contienen los datos reales sobre el número de usuarios por estación (Metro y Cercanías). La estructura de los datos es el siguiente.

   - stop_name: Nombre de la parada.
   - stop_id: Identificador de la parada.
   - up: Número de entradas de pasajeros para el periodo de tiempo establecido.
   - down: Número de salidas de pasajeros para el periodo de tiempo establecido.

3. **timeseries_o.csv**: Este archivo contiene todos aquellos viajes que tuvieron el mismo distrito origen en común. La primera columna “ds” representa la fecha y hora en la que se produjeron los viajes y el resto de columnas representan a todos los distritos de Madrid y Guadalajara. A continuación se describe la estructura del archivo.

   | ds | 2807905 | 2807901 | 2807907 | ... |Sumatoria viajes con origen común|

A.4. Datos reales de accidentalidad

En el portal de datos abiertos del ayuntamiento de Madrid se encuentran los datos reales de accidentes de tráfico de la ciudad de Madrid [45]. Los datos publicados se encuentran disponibles para descargar en formatos “.csv” y “.xlsx”. A continuación se muestra la estructura del archivo 2020_Accidentalidad.csv.

| num_expediente | fecha | hora | localizacion | numero | cod_distrito | distrito | tipo_accidente | estado_meteorológico | tipo_vehiculo | tipo_persona | rango_edad | sexo | cod_lesividad | lesividad | coordenada_x_utm | coordenada_y_utm | positiva_alcohol | positiva_droga |

Los datos relevantes para el actual trabajo de investigación corresponden a los datos presentes en las columnas:

| num_expediente | fecha | hora | cod_distrito | distrito | tipo_accidente

num_expediente: Los registros en esta columna están definidos de la siguiente forma: AAAASNNNNNN, donde AAAA es el año del accidente, S cuando se trata de un expediente con accidente y NNNNNN es un número correlativo por año.
fecha: En formato dd/mm/aaaa
hora: La hora se establece en rangos horarios de 1 hora
cod_distrito: Código de distrito
distrito: Nombre del distrito
tipo_accidente: Puede ser colisión doble, colisión múltiple, alcance, choque contra obstáculo o elemento de la vía, atropello a persona, Vuelco, caída y otras causas.

A.5. Dataset para gráfica de trayectos

En este anexo se detalla el proceso de obtención del conjunto de datos que posibilita la generación de las gráficas presentadas en las figuras 50, 51 y 52 de la sección 4.3.2. Los archivos que se obtienen constituyen la base de datos reales de desplazamientos para diferentes trayectos. Los archivos originales se encuentran en formato ".txt", por lo que es conveniente en este trabajo realizar una conversión de los datos reales reportados a un formato ".csv".

Una vez se obtienen todos los datos en formato ".csv", se reduce el tamaño de cada archivo al realizar un filtrado de los trayectos que no corresponden a combinaciones origen/destino realizadas en Madrid o Guadalajara (Municipio donde opera el servicio de cercanías). Este proceso de filtrado se divide en dos etapas. En la primera etapa, se realiza el filtrado de los datos según el distrito de origen, y en la segunda etapa se filtran los datos según el distrito de destino.

En el primer proceso de filtrado, se utiliza el script "Data_Filtrado_Madrid.py" (disponible en repositorio GitHub). Este script de python, además de seleccionar los registros que tienen un origen en Madrid o en el municipio de Guadalajara, también se encarga de concatenar los archivos diarios en un solo archivo por mes. Es decir, que mediante este proceso se obtienen nuevos archivos filtrados por origen con el formato de nombre **"DataFiltrado(Mes)(Año)_1.csv"** y **"DataFiltrado(Mes)(Año)_2.csv"**. Los nombres de los archivos resultantes sugieren que por cada mes se van a obtener 2 archivos CSV, a continuación se describe el tamaño en bytes y la dimensión de cada archivo resultante.

Archivo	Tamaño (MB)	Dimensiones
DataFiltradoFebrero2020.csv	775.8	(12000168, 11)
DataFiltradoMarzo2020_1.csv	662.1	(10245914, 11)
DataFiltradoMarzo2020_2.csv	325.9	(5066294, 11)
DataFiltradoAbril2020_1.csv	307.7	(4787883, 11)
DataFiltradoAbril2020_2.csv	316.2	(4915502, 11)
DataFiltradoMayo2020_1.csv	372.4	(5784187, 11)
DataFiltradoMayo2020_2.csv	458.2	(7106322, 11)
DataFiltradoJunio2020_1.csv	587.3	(9098793, 11)
DataFiltradoJunio2020_2.csv	574.6	(8903412, 11)
DataFiltradoJulio2020_1.csv	655.8	(10160252, 11)
DataFiltradoJulio2020_2.csv	616.1	(9555619, 11)

Tabla 20: Especificaciones de archivos mensuales [Fuente propia]

Para la segunda etapa de filtrado se utilizan los siguientes notebooks:

- "Trayectos_Periodo_Referencia.ipynb"
- "Trayectos_Periodo_Confinamiento1.ipynb"
- "Trayectos _Periodo_Confinamiento2.ipynb"

Los notebooks mencionados anteriormente no solo realizan el filtrado de los datos según el destino, sino que también se encargan de transformar la estructura original de los datos para generar nuevos archivos que representan series temporales. Estos nuevos archivos contienen información de los desplazamientos por trayectos origen-destino. A continuación se describe la estructura de los nuevos archivos csv.

ds	19024-19024	19024-19046_AM	19024-19053_AM	19024-19058_AM	19024-19071_AM	19024-19086_AM	19024-1913001	19024-1913002	...	28023-28060_AM	28053-2801302	2805809-28087	28059-2801301	28087-28141	28089-19326_AM
2020-02-14 00:00:00	0.000	114.811	0.000	91.364	0.000	0.0	10.236	10.619	...	0.0	0.0	0.0	0.0	0.0	0.0
2020-02-14 01:00:00	15.354	140.702	0.000	71.535	5.778	0.0	5.118	15.354	...	0.0	0.0	0.0	0.0	0.0	0.0
2020-02-14 02:00:00	10.236	96.146	0.000	82.323	0.000	0.0	15.701	10.619	...	0.0	0.0	0.0	0.0	0.0	0.0
2020-02-14 03:00:00	20.608	69.881	10.943	68.443	0.000	0.0	10.236	0.000	...	0.0	0.0	0.0	0.0	0.0	0.0
2020-02-14 04:00:00	20.472	107.192	0.000	64.517	0.000	0.0	15.354	15.737	...	0.0	0.0	0.0	0.0	0.0	0.0

Figura 83: Serie temporal de desplazamientos origen-destino en Madrid [Fuente propia]

Según se muestra en la tabla de la figura 83, la columna “ds” indica la fecha y hora en la que se registran los desplazamientos, mientras que las demás columnas representan los diversos trayectos de origen a destino realizados en la ciudad de Madrid. Es importante notar que los valores en 0 indican la ausencia de desplazamientos durante una hora específica. A continuación se presenta la estructura de la nuevos datos:

| ds | 2807905-2807901 | 2807901-2807905 | 2807907-2807901 | ... | origen-destino |

Como resultado de los procedimientos previamente descritos, se obtienen los siguientes archivos CSV:

- **“Trayectos_Periodo_Referencia.csv”**: Serie temporal que abarca el periodo desde el 26 de febrero de 2020 hasta el 16 de marzo de 2020. Tamaño de 122.1 MB y una dimensión de: (768, 33759).
- **“Trayectos_Periodo_Confinamiento1.csv”**: Serie temporal que abarca el periodo desde el 17 de marzo de 2020 hasta el 31 de mayo de 2020. Tamaño de 266.9 MB y una dimensión de: (1824, 32972).
- **“Trayectos_Periodo_Confinamiento2.csv”**: Serie temporal que abarca el periodo desde el 01 de junio de 2020 hasta el 31 de julio de 2020. Tamaño de 234.1 MB y una dimensión de: (1464, 34353).

B. Anexo: Predicción de los desplazamientos en base a factores meteorológicos

En esta sección se incorporan dos nuevas variables complementarias relacionadas con las condiciones meteorológicas. Estas variables pueden permitir al modelo capturar mejor las posibles relaciones entre las condiciones climáticas y los desplazamientos de las personas, lo cual puede mejorar la precisión y la capacidad predictiva del modelo. Los datos reales meteorológicos que se tendrán en cuenta en este apartado corresponden a los datos de temperatura y precipitación de la ciudad de Madrid.

Según los hallazgos presentados en la sección 4.3.1, donde se comparan los patrones de movilidad con las tasas de incidencia, se ha definido el periodo de estudio para un intervalo de la nueva normalidad. Sin embargo, en el actual análisis de variables meteorológicas, se incluirá el periodo de desescalada para obtener una comprensión más completa del comportamiento de las variables de temperatura y precipitación.

Modelado

En la figura 84 se presenta la configuración del modelo que se evaluará en esta sección. Sin embargo, para realizar el proceso de validación en el siguiente apartado, este modelo también se entrenará utilizando los regresores de forma individual. Es importante destacar que el modelo E se ha entrenado con datos reales de toda la semana (Lunes a domingo) y las 24 horas del día.

Modelo Prophet #: Regresor de variables metereológicas - Desescalada y Nueva normalidad

```
m = Prophet(weekly_seasonality=True,daily_seasonality=True)
m.add_regressor('Temp',mode='additive')
m.add_regressor('Prec',mode='additive')
```

Figura 84: Configuración de parámetros del Modelo E - Desescalada y Nueva normalidad [Fuente propia]

Validación del modelo

En la tabla 21 se presenta la validación del modelo, para lo cual se ajustan por separado los regresores de temperatura y precipitación.

Regresor	Etapa 1	RMSE	Etapa 2	RMSE
Ninguno	Desescalada	93.18	Nueva normalidad	79.08
Temp	Desescalada	93.21	Nueva normalidad	74.16
Prec	Desescalada	93.15	Nueva normalidad	78.82
Temp/Prec	Desescalada	93.18	Nueva normalidad	73.66

Tabla 21: Validación de las variables temperatura y precipitación en el modelo E [Fuente propia]

Elaboración propia: Los archivos que generan las figuras 84 a 90 y los resultados de la tabla 21 están en el repositorio de GitHub [35] como Metereologicos_Lunes09.ipynb – 2023

De acuerdo a los resultados de la validación del modelo, se puede observar que las variables meteorológicas no influyen significativamente en la predicción del periodo de desescalada, ya que al comparar las métricas de evaluación con los resultados obtenidos en la tabla 22, se observa una similitud en los valores, con un RMSE de 93. Al realizar el análisis para el periodo de nueva normalidad, se puede notar una mejora significativa al hacer uso principalmente de los datos reales de temperatura.

Con el objetivo de corroborar estos resultados, se lleva a cabo la generación de las gráficas de contribución para cada regresor, representadas en las figuras 85 y 86, como una manera adicional de validar los hallazgos obtenidos. En estas gráficas se presenta el impacto de la variables predictoras temperatura y precipitación para el periodo de desescalada.

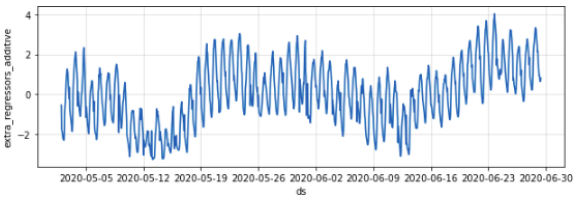

Figura 85: Contribución del regresor de temperatura al modelo predictivo - Desescalada [Fuente propia]

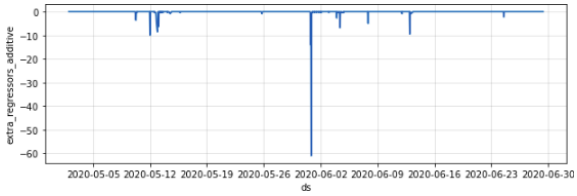

Figura 86: Contribución del regresor de precipitación al modelo predictivo - Desescalada [Fuente propia]

En las figuras 87 y 88 se presenta el impacto de las variables predictoras temperatura y precipitación para el periodo de nueva normalidad.

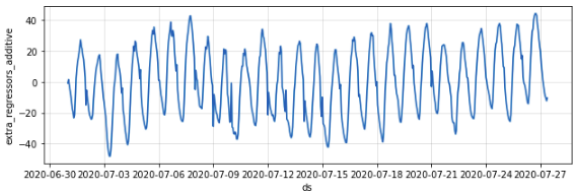

Figura 87: Contribución del regresor de temperatura al modelo predictivo - Nueva normalidad [Fuente propia]

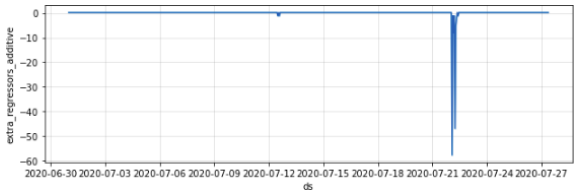

Figura 88: Contribución del regresor de precipitación al modelo predictivo - Nueva normalidad [Fuente propia]

De acuerdo con las gráficas de contribución de las figuras 87 y 88, se puede inferir que la temperatura tiene un mayor impacto en comparación con la precipitación en el modelo predictivo. Esto se debe a que los datos reales de temperatura tienen un efecto constante en la predicción, mientras que los datos reales de precipitación influyen en el modelo únicamente cuando se registra lluvia.

Por otra parte, se puede observar que la temperatura tiene un mayor impacto durante el periodo de nueva normalidad, evidenciado por un efecto de +/- 40 en la gráfica de la figura 87. En cambio, durante el periodo de desescalada, en la figura 85, se presenta un efecto moderado de aproximadamente +/- 4. Esto sugiere que la temperatura tiene una influencia más significativa en el modelo durante la nueva normalidad en comparación con la desescalada. Una posible explicación para este hallazgo podría estar relacionada con la forma en que las personas interactúan con su entorno en la nueva normalidad. Después del confinamiento, es probable que las personas prefieran actividades al aire libre, lo cual podría influir en su percepción y respuesta ante la temperatura, y a su vez, tener un impacto en su movilidad.

Elaboración propia: Los archivos que generan las figuras 84 a 90 y los resultados de la tabla 21 están en el repositorio de GitHub [35] como Metereologicos_Lunes09.ipynb – 2023

Sin embargo, al observar las gráficas de predicción de las figuras 89 y 90, se puede percibir que las variables predictoras meteorológicas no explican las fluctuaciones en los datos. Esto puede deberse a que las estaciones del tiempo en la ciudad de Madrid están bien definidas y las personas ya están preparadas para sobrellevar cada estación. Por lo tanto, para que las variables meteorológicas tengan un impacto significativo en la movilidad, sería necesario que se presente una temperatura o precipitación extrema que esté fuera de lo previsto.

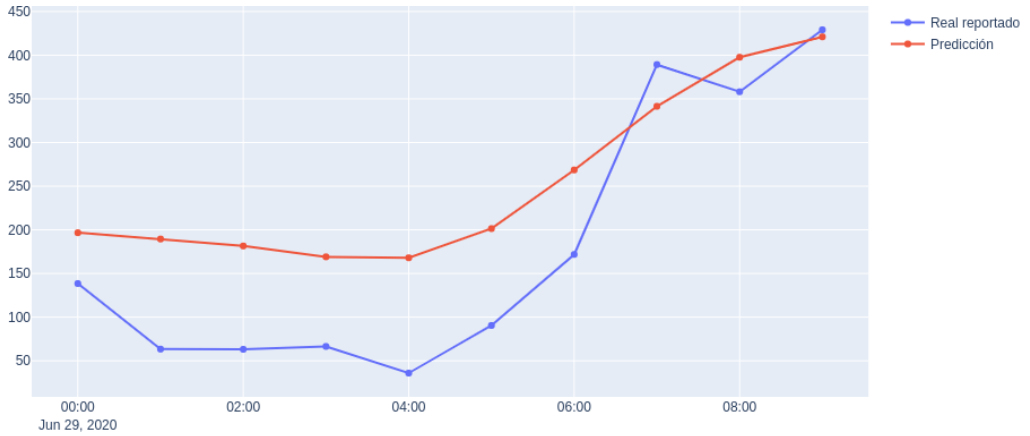

Figura 89: Comparación entre datos reales reportados y predicciones del Modelo E - Desescalada [Fuente propia]

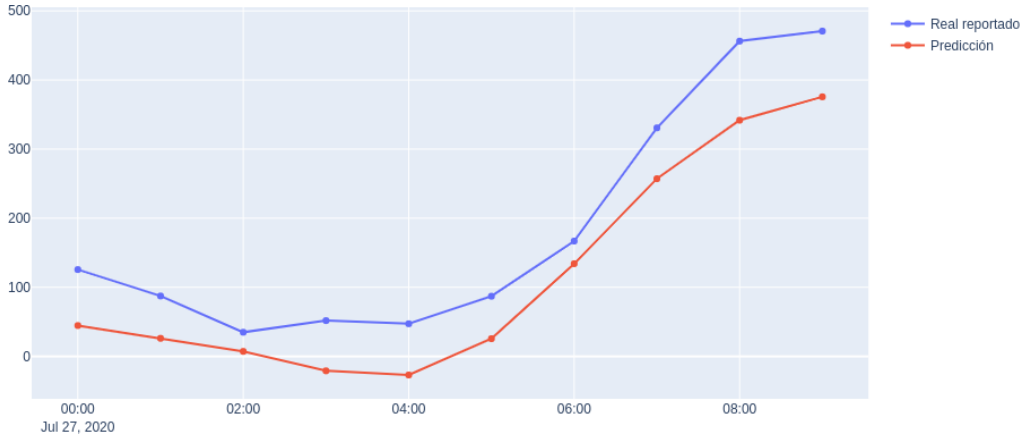

Figura 90: Comparación entre datos reales reportados y predicciones del Modelo E - Nueva normalidad [Fuente propia]

A modo de síntesis, las variables meteorológicas en el periodo de desescalada no evidencian un aumento o disminución significativo en la predicción de los modelos. Para el caso de el periodo de nueva normalidad se puede observar una leve mejora en la métrica RMSE, tal como se observa en la tabla 21.

En el modelo final de la sección 5.3 se ajustan las variables meteorológicas pero no se logra apreciar una mejora en las predicciones, por lo que se determina que estas variables no tienen ningún impacto en los modelos que fueron entrenados para el periodo de nueva normalidad. Por lo anterior se concluye que no es beneficioso para el modelo final utilizar estas variables.

El notebook que generó los resultados de la tabla 21 se encuentran disponibles en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `Metereologicos_Lunes09.ipynb`

Elaboración propia: Los archivos que generan las figuras 84 a 90 y los resultados de la tabla 21 están en el repositorio de GitHub [35] como Metereologicos_Lunes09.ipynb – 2023

C. Anexo: Análisis del conjunto de datos de entrenamiento para la predicción de bloques horarios

Se propone para esta nueva sección una estrategia que busca mejorar la precisión de las predicciones de patrones de movilidad mediante la reducción de la ventana de predicción del modelo. Además de esto, se considera la importancia de seleccionar cuidadosamente el conjunto de datos de entrenamiento, con el fin de mejorar la confiabilidad de las predicciones obtenidas.

La selección del conjunto de datos de entrenamiento para un modelo predictivo es un factor crítico que puede influir en la precisión de las predicciones. Por lo tanto, surge la duda de si es mejor utilizar datos de entrenamiento similares a la situación que se desea predecir, para que el modelo aprenda patrones específicos de comportamiento, o incluir un conjunto de datos más amplio para capturar características más complejas de estos patrones.

Modelado

En este apartado se examinan varios modelos y se aplican diferentes ajustes en el conjunto de datos de entrenamiento. El objetivo final es identificar el conjunto de datos que brinde mayor rendimiento al momento de realizar predicciones.

El modelo que se muestra en la figura 91 es entrenado con datos de toda la semana (Lunes a domingo) y las 24 horas del día. Los parámetros de configuración de este modelo corresponden al modelo por defecto de Prophet y se mantiene la misma configuración para ambos períodos de estudio (Desescalada y Nueva normalidad).

Modelo Prophet #1: Entrenado con todos los datos - Desescalada y Nueva normalidad

```
m = Prophet(weekly_seasonality=True,daily_seasonality=True)
```

Figura 91: Configuración de parámetros del Modelo E1 - Desescalada y Nueva normalidad

El modelo presentado en la figura 92 es entrenado con datos recopilados de lunes a domingo, pero restringiendo el rango horario únicamente de 0 a 9 am. Debido a la restricción horaria de los datos de entrenamiento, este modelo en particular permite únicamente ajustar una estacionalidad semanal y un patrón estacional que se repite cada 10 horas.

Modelo Prophet #2: Entrenado con datos diarios y bloque horario 00:00-09:00 - Desescalada y Nueva normalidad

```
m = Prophet(weekly_seasonality=True,daily_seasonality=False)
m.add_seasonality(name='Lunes0_9', period=1/2.4, fourier_order=6,
condition_name='Lunes0_9',prior_scale=0.5)
```

Figura 92: Configuración de parámetros del Modelo E2 - Desescalada y Nueva normalidad

El tercer modelo presentado en la figura 93 se entrena únicamente con los datos del día lunes y en el rango horario de 0 a 9 am. Al tener datos con días y horas restringidas no es posible ajustar una estacionalidad semanal o diaria, ya que no hay suficiente información para identificar las variaciones estacionales. Por consiguiente se opta por ajustar una estacionalidad que represente las 10 horas que se desean predecir.

Modelo Prophet #3: Entrenado con datos del día lunes y bloque horario 00:00-09:00 - Desescalada y Nueva normalidad

```
m = Prophet(weekly_seasonality=False,daily_seasonality=False)
m.add_seasonality(name='Lunes0_9', period=1/2.4, fourier_order=6,
condition_name='Lunes0_9',prior_scale=0.5)
```

Figura 93: Configuración de parámetros del Modelo E3 - Desescalada y Nueva normalidad

Validación del modelo

A continuación, se presentan las predicciones obtenidas al entrenar el modelo de desescalada con tres conjuntos de datos diferentes.

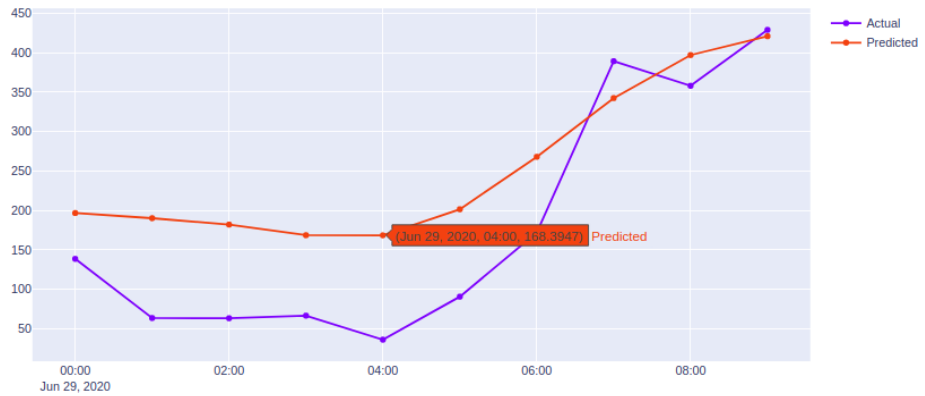

Figura 94: Predicción del modelo E1 entrenado con todos los datos - Desescalada

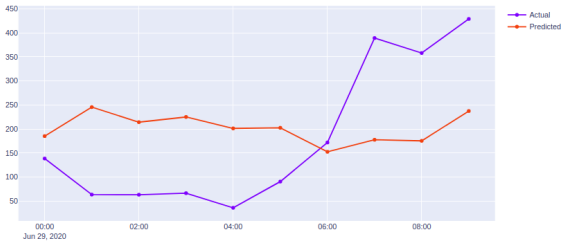

Figura 95: Predicción del modelo E2 entrenado con datos diarios (00:00 - 09:00) - Desescalada

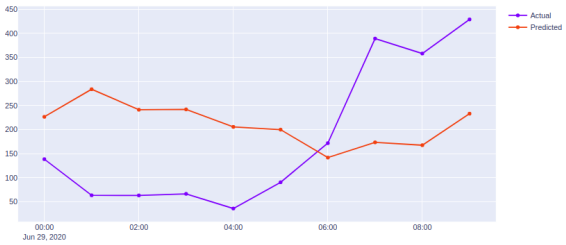

Figura 96: Predicción del modelo E3 entrenado con datos del día lunes (00:00 - 09:00) - Desescalada

Para el modelo de nueva normalidad se presentan las gráficas de predicción de las figuras 97, 98 y 99, obtenidas para los tres conjuntos de datos de entrenamiento.

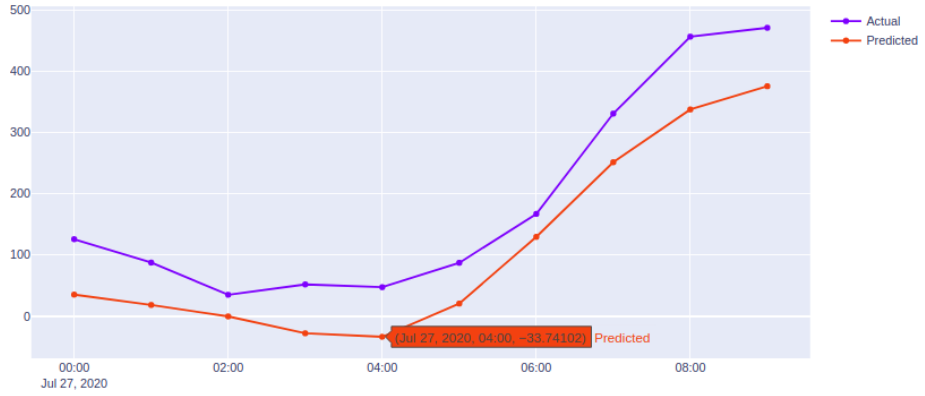

Figura 97: Predicción del modelo E1 entrenado con todos los datos - Nueva normalidad

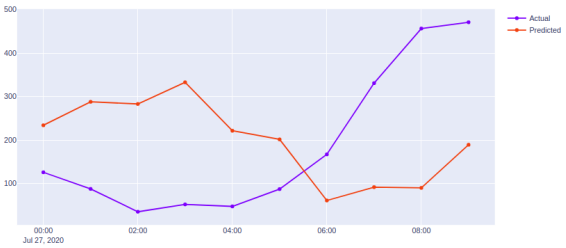

Figura 98: Predicción del modelo E2 entrenado con datos diarios (00:00 - 09:00) - Nueva normalidad

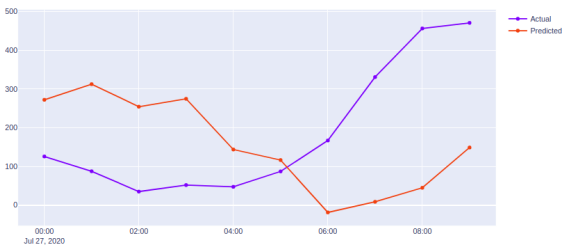

Figura 99: Predicción del modelo E3 entrenado con datos del día lunes (00:00 - 09:00) - Nueva normalidad

De acuerdo a las gráficas de predicción de las figuras 97, 98 y 99, se procede a evaluar los resultados de los diferentes conjuntos de entrenamiento del modelo mediante la métrica RMSE. Además, es importante destacar que las predicciones se realizan para el bloque horario de 0 a 9 am del último día lunes de los periodos de estudio (desescalada y nueva normalidad).

Modelo	Datos de entrenamiento	Etapa 1	RMSE	Etapa 2	RMSE
E1	Todos	Desescalada	93.2	Nueva normalidad	79.1
E2	Diarios 00:00 - 09:00	Desescalada	154.5	Nueva normalidad	227.5
E3	Lunes 00:00 - 09:00	Desescalada	167.8	Nueva normalidad	243.1

Tabla 22: Análisis de desempeño del modelo k con diferentes conjuntos de datos de entrenamiento

A partir de los resultados presentados en la tabla 22, se pueden obtener algunas conclusiones de acuerdo al conjunto de datos con el que se ha entrenado el modelo:

- Modelo E1: En este escenario el modelo se ha entrenado con todo el conjunto de datos. Se observa que este modelo es el que mejor resultados obtiene con respecto a la evaluación de la métrica RMSE. Estos resultados estarían indicando que el modelo encuentra una representación más acertada de los patrones estacionales al entrenar el modelo con todos los datos.

- Modelo E2: En este escenario, se observa que las predicciones son menos precisas según los valores obtenidos de la métrica RMSE. Esto podría deberse a que al entrenar el modelo únicamente con las horas que se desean predecir, se están omitiendo datos relevantes que podrían ayudar al modelo a representar adecuadamente los patrones de movilidad

- Modelo E3: El modelo es entrenado de acuerdo a las condiciones específicas que se desean pronosticar para evitar que el modelo se ajuste a un rango demasiado amplio de situaciones. Sin embargo, los resultados obtenidos hasta el momento indican que el enfoque de entrenar el modelo de acuerdo a las condiciones específicas que se desean pronosticar podría no ser válido para el conjunto de datos que se está analizando.

Es importante tener en cuenta que la estacionalidad se refiere a patrones recurrentes que ocurren en un conjunto de datos en momentos específicos del tiempo, como patrones semanales, mensuales o diarios. Por lo tanto, para capturar adecuadamente estos patrones estacionales en un modelo predictivo, es necesario contar con suficientes datos.

Si un modelo se entrena únicamente con datos del día lunes, no será capaz de capturar una estacionalidad semanal completa, ya que solo tiene información de un día de la semana. Del mismo modo, si se entrena con datos limitados, como el bloque horario de 0 a 9, no será capaz de ajustar una estacionalidad diaria completa.

Los notebooks que generaron los resultados de la tabla 22 se encuentran disponibles en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con los nombres Segmentacion_TrainALL, Segmentacion_TrainALL09 y Segmentacion_TrainLunes09.

D. Anexo: Variables complementarias para mejorar predicción

Funcionamiento del regresor “Accup”

El regresor conocido como “Accup” fue diseñado para ajustar las predicciones en función de los accidentes que ocurrieron en el mismo día de la predicción. Este regresor contrasta con el regresor “t-168Mod” que ajusta la predicción de acuerdo a los accidentes ocurridos una semana antes de la predicción. A continuación se evidencia en una tabla de verdad el funcionamiento del regresor de acuerdo a las incidencias de tráfico reportadas en el dataset de accidentes.

Ds	Accidente1	Accidente2	Accup
2020-07-27 11:00:00	True	True	False
2020-07-27 12:00:00	False	True	True
2020-07-27 13:00:00	True	False	False
2020-07-27 15:00:00	False	False	False

Tabla 23: Elaboración del regresor Accup [Fuente propia]

La tabla anterior describe el funcionamiento del regresor, lo que permite observar que este solo tendrá efecto en las horas en las que el estado de la columna “Accup” se encuentre en true.

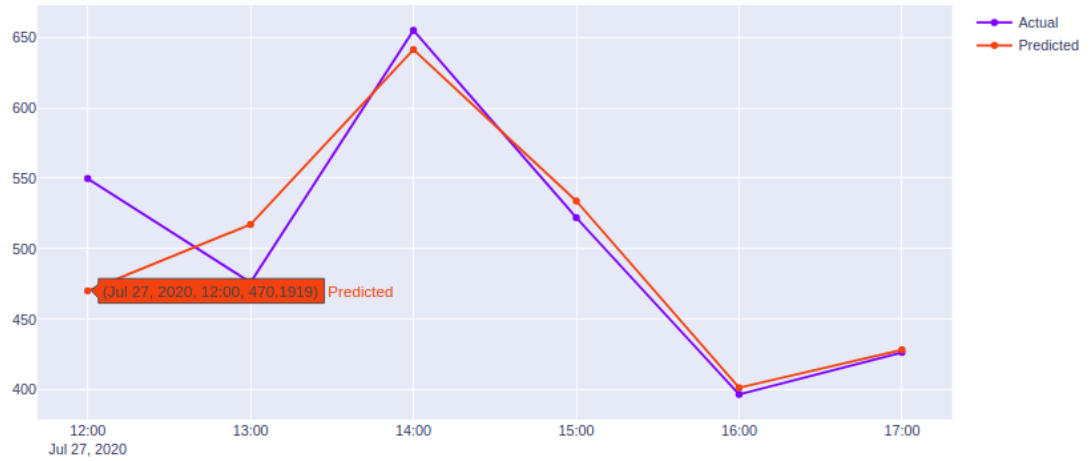

Figura 100: Predicción sin utilizar el regresor “Accup”

Al implementar el regresor se mejora la predicción en las horas 12, 14 y 17. Esto es debido a que se tiene en cuenta el incremento de los desplazamientos provocados por los accidentes de tránsito. Cabe recordar que los accidentes de tránsito que se han estudiado en esta sección tienen primeramente un efecto de reducción de los desplazamientos y posteriormente una hora después se observa un aumento en estos.

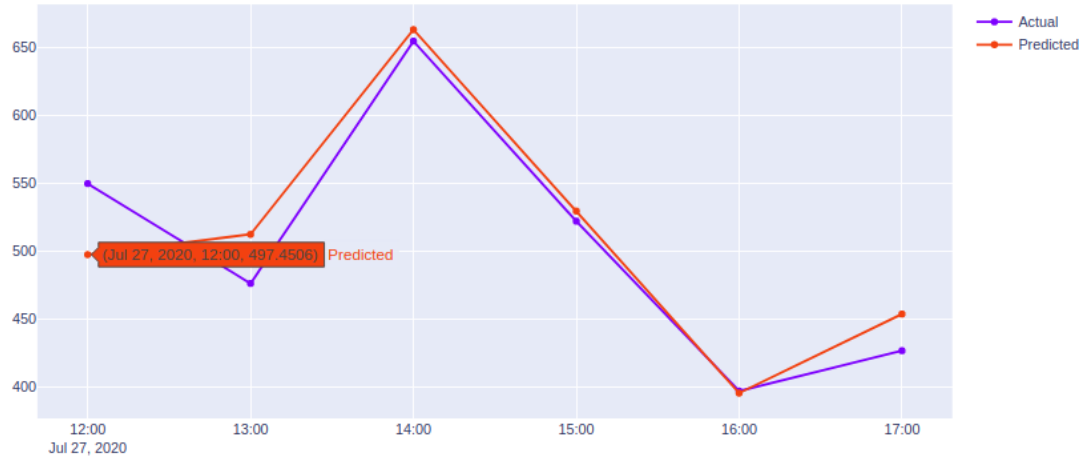

Figura 101: Predicción utilizando el regresor “Accup”

Aunque la predicción con el regresor “Accup” arrojó buenos resultados, se logra identificar un aumento no deseado en la hora 17. Este aumento en la predicción puede ser debido a que no se está teniendo en cuenta la agrupación de los datos de accidentes, en función del distrito en donde se presenta la incidencia de tráfico o el tipo de accidente que se ha registrado. Al no realizar esta agrupación, el aumento en la predicción es similar para las 3 horas identificadas.

Funcionamiento del regresor “Accdown”

El regresor “Accdown” fue diseñado de igual manera que el regresor “Accup” y ajusta las predicciones en función de los accidentes que ocurrieron en el mismo día de la predicción. Se espera que este regresor disminuya el valor de la predicción en horas específicas del día. A continuación se evidencia en una tabla de verdad el funcionamiento del regresor de acuerdo a las incidencias de tráfico reportadas en el dataset de accidentes.

Ds	Accidente1	Accidente2	Accdown
2020-07-27 11:00:00	True	True	False
2020-07-27 13:00:00	True	False	True
2020-07-27 14:00:00	False	True	False
2020-07-27 15:00:00	False	False	False

Tabla 24: Elaboración del regresor Accdown

Para que el regresor tenga un efecto en la predicción de una hora en específico la columna “Accdown” debe de tener un estado true. En las figuras 102 y 103 se presenta gráficamente la contribución del regresor.

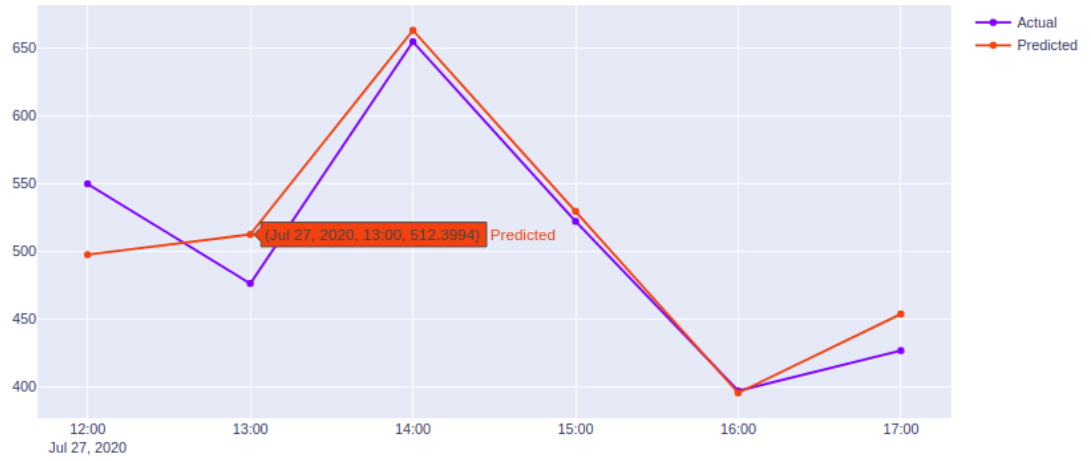

Figura 102: Predicción sin utilizar el regresor “Accdown”

Al observar la figura 103, donde se muestra el comportamiento del regresor “Accdown”, se tiene que el efecto de disminución no se presenta y por el contrario el regresor incrementa el valor de la predicción en la hora 13 y en la hora 16.

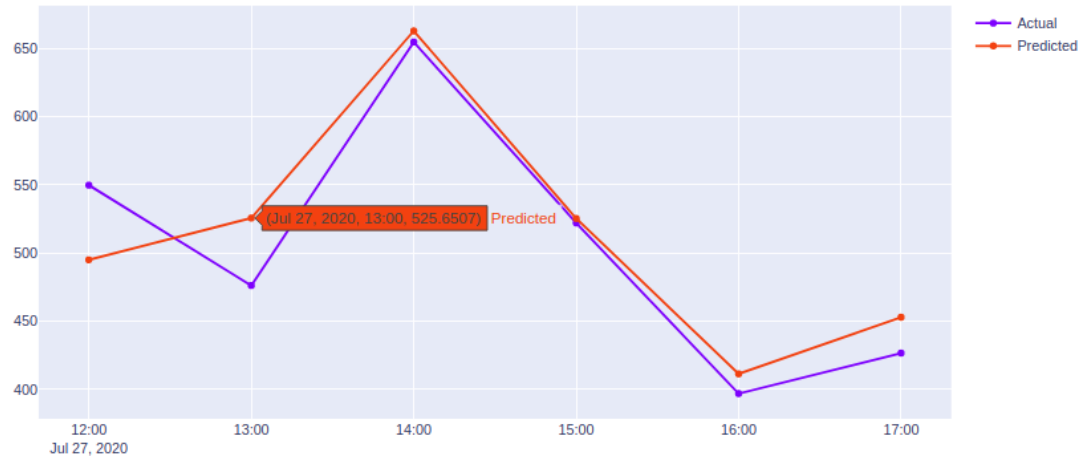

Figura 103: Predicción utilizando el regresor “Accdown”

Es probable que el incremento observado se genere por no separar en diferentes grupos los datos de incidencias de tráfico, ya que si se logra una agrupación de los accidentes de acuerdo al distrito donde ocurre o de acuerdo al tipo de accidente, el modelo podría identificar adecuadamente el efecto de reducción en los desplazamientos en el momento que se registre un incidente en la variable “Accidente1”.

Según los resultados obtenidos en el análisis del regresor “Accdown”, se puede concluir que éste no aporta de manera positiva al modelo. Por lo tanto, se ha decidido no considerar este regresor en modelos posteriores.

E. Anexo: Modelo con estacionalidad personalizada por día de la semana

Este modelo esta diseñado para realizar una predicción de Lunes a Domingos y para cualquier hora deseada. Por lo tanto se entrena este modelo añadiendo una estacionalidad personalizada para cada día de la semana aplicando la función `add_seasonality` de Prophet, la cual requiere como entrada un nombre, un periodo en días, el orden de fourier que mejor represente el comportamiento de un día y una condición que debe de estar presente dentro del dataframe que se está entrenando. En adición a las estacionalidades ajustadas en la figura 104, Prophet también ajusta por defecto una estacionalidad diaria, semanal y ajusta una tendencia mediante una detección automática de puntos de cambio (changepoints) en la serie temporal.

Modelo de estacionalidad personalizada por día de la semana

```
m = Prophet()
m.add_seasonality(name='monthly', period=30.5, fourier_order=5)
m.add_seasonality(name='Lunes_season', period=1, fourier_order=5, condition_name='Lunes')
m.add_seasonality(name='Martes_season', period=1, fourier_order=5, condition_name='Martes')
m.add_seasonality(name='Miércoles_season', period=1, fourier_order=5, condition_name='Miércoles')
m.add_seasonality(name='Jueves_season', period=1, fourier_order=5, condition_name='Jueves')
m.add_seasonality(name='Viernes_season', period=1, fourier_order=5, condition_name='Viernes')
m.add_seasonality(name='Sábado_season', period=1, fourier_order=5, condition_name='Sábado')
m.add_seasonality(name='Domingo_season', period=1, fourier_order=5, condition_name='Domingo')
```

Figura 104: Configuración de parámetros del Modelo F

En la figura 105 aparece la gráfica de componentes donde se puede apreciar la tendencia y la variación estacional de los datos de estudio. De igual manera se hace evidente la similitud que existe entre la estacionalidad de los días lunes, martes y miércoles, así como la similitud entre los días jueves y viernes. Los días sábados y domingos presentan una estacionalidad muy diferente a las demás.

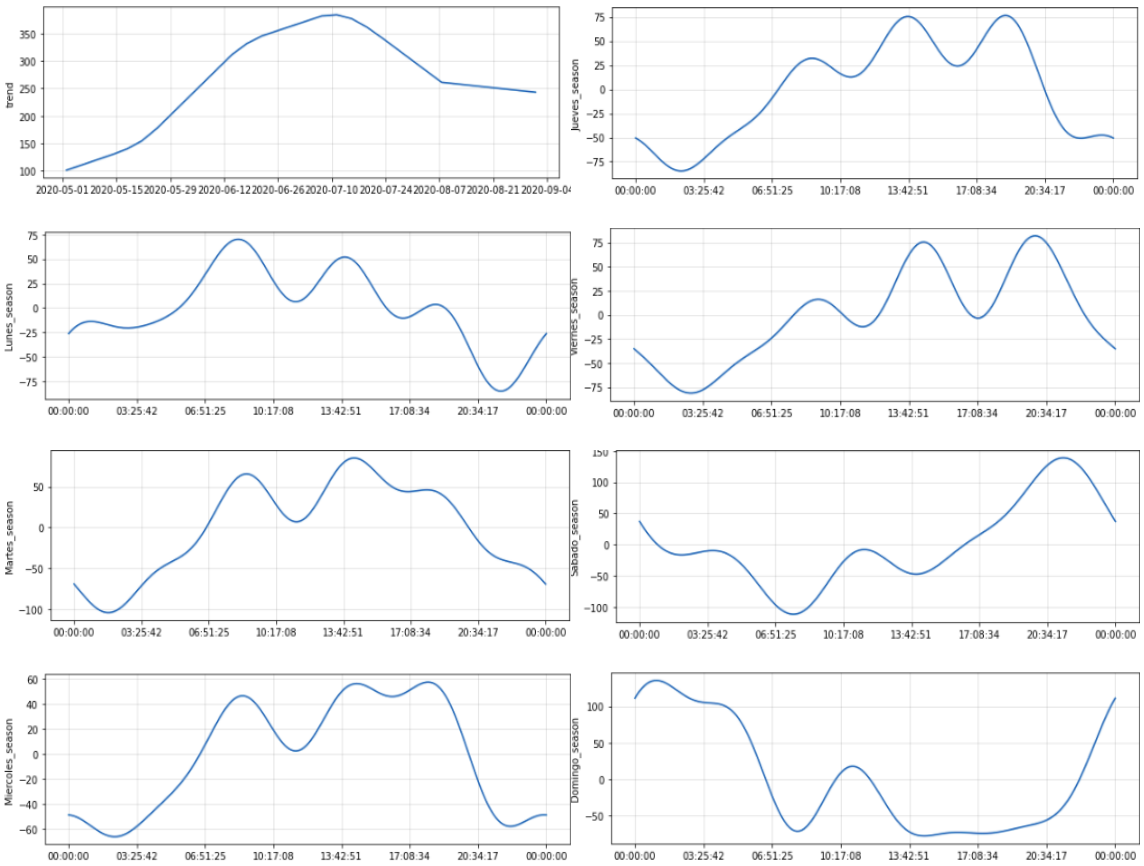

Figura 105: Gráfica de componentes de tendencia y estacionalidades personalizadas

Con el propósito de evaluar el modelo es preciso definir unos datos de entrenamiento y unos datos prueba, para este caso en particular se considera un 90% de datos de entrenamiento y un 10% de datos de prueba. Teniendo en cuenta esta proporción, el tiempo de ejecución del programa fue de 50 minutos para al final obtener el gráfico de la figura 106 y un valor RMSE de 57.8 que refleja la precisión del modelo. Es importante destacar que el modelo tiene un tiempo de ejecución prolongado debido a que se utiliza la técnica de evaluación de modelos denominada `Forward Cross-Validation` y la técnica de predicción continua `Rolling Forecasting`, dichas técnicas son computacionalmente costosas ya que realizan una predicción para cada punto en el conjunto de prueba utilizando los valores anteriores como conjunto de entrenamiento.

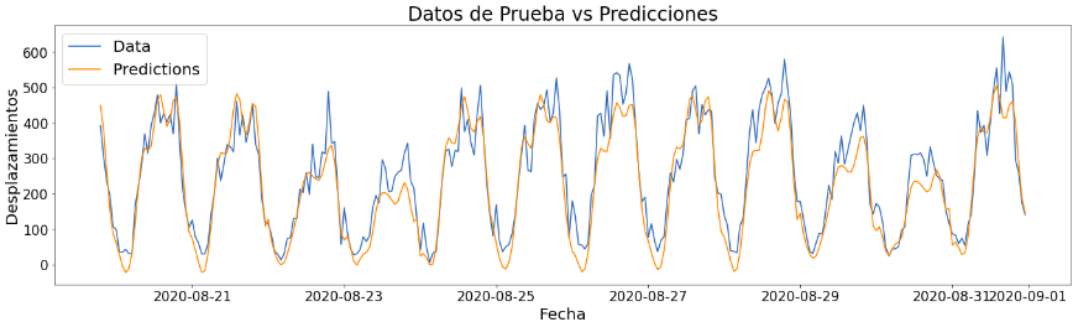

Figura 106: Gráfica de comparación entre datos reales y predicciones del modelo ajustado por día de la semana

El notebook que generó los resultados de la figura 106 se encuentra disponible en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `Modelo_Prophet_EstacionalidadDiadelaSemana.ipynb`

En la gráfica de la figura 107 se puede apreciar los resultados de un modelo ajustado con sus parámetros por defecto (Modelo Default). Este modelo es evaluado con el fin de tener una referencia en el desempeño de los modelos, la métrica RMSE que se obtuvo para este caso fue de 76.9.

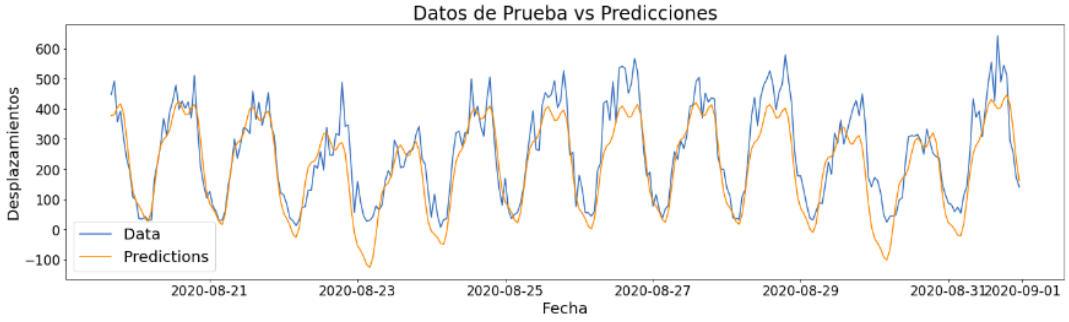

Figura 107: Gráfica de comparación entre datos reales y predicción del modelo Default

El notebook que generó los resultados de la figura 107 se encuentra disponible en el siguiente reposi- torio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `1Modelo_Prophet_Default.ipynb`

F. Anexo: Modelos ajustados para la etapa de desescalada y Nueva normalidad

Al examinar los modelos previos se percibe que la ventana de predicción es demasiado extensa para cumplir con el objetivo del trabajo de investigación, que es apoyar la toma de decisiones en la planificación de la oferta del servicio de transporte público masivo. Para lograr alcanzar este objetivo es necesario tener una ventana de predicción por horas para que el sistema pueda proveer información valiosa.

En esta sección se fija la ventana de predicción en un solo día. Además, se separan los datos en dos periodos que serán identificados como “Desescalada” y “Nueva normalidad”. A continuación en las figuras 108 y 109 se grafican los modelos por defecto (Modelos Default) para tener una referencia que permita evaluar el desempeño de los demás modelos.

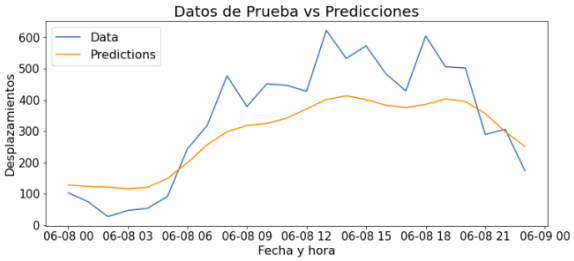

Figura 108: Comparación entre datos reales y predicciones del Modelo Default - Desescalada

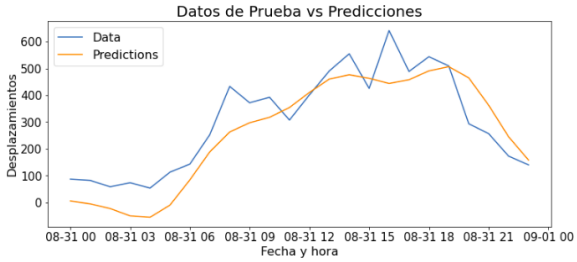

Figura 109: Comparación entre datos reales y predicciones del Modelo Default - Nueva Normalidad

Las gráficas de las figuras 108 y 109 mostradas previamente, son el resultado de ajustar el modelo por defecto, el cual ajusta automáticamente una tendencia, una estacionalidad semanal y una estacionalidad diaria. Al evaluar los modelos empleando la métrica RMSE se obtienen los siguientes resultados.

Modelo	Etapa	RMSE
Default	Desescalada	108.3
Default	Nueva normalidad	93.6

Tabla 25: Evaluación comparativa de los modelos Default en diferentes etapas de la pandemia

El notebook que generó los resultados de las figuras 108 y 109 se encuentran disponibles en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `.ipynb`

Periodo de desescalada 02/05/2020 - 08/06/2020

El objetivo al realizar la separación de los datos en dos periodos es encontrar un patrón de desplazamientos de acuerdo a la etapa de pandemia que pueda ser representado mediante un modelo predictivo. Para ello en este apartado se muestran los resultados encontrados para el periodo de desescalada ajustando el modelo con una estacionalidad

por días y una estacionalidad por horas. En la figura 110 se puede apreciar el ajuste de las estacionalidades de acuerdo al día de la semana.

Modelo de estacionalidad personalizada por día de la semana - Desescalada

```
m = Prophet(weekly_seasonality=True,daily_seasonality=True)
m.add_seasonality(name='Lunes_season', period=1, fourier_order=5, condition_name='Lunes')
m.add_seasonality(name='Martes_season', period=1, fourier_order=5, condition_name='Martes')
m.add_seasonality(name='Miércoles_season', period=1, fourier_order=5, condition_name='Miércoles')
m.add_seasonality(name='Jueves_season', period=1, fourier_order=5, condition_name='Jueves')
m.add_seasonality(name='Viernes_season', period=1, fourier_order=5, condition_name='Viernes')
m.add_seasonality(name='Sábado_season', period=1, fourier_order=5, condition_name='Sábado')
m.add_seasonality(name='Domingo_season', period=1, fourier_order=5, condition_name='Domingo')
```

Figura 110: Configuración de parámetros del Modelo G - Desescalada

Una vez se ha realizado el entrenamiento del modelo ajustando las estacionalidades por día, se procede a comparar los datos de prueba con las predicciones del modelo tal como se aprecia en la figura 111. En este gráfico se está llevando a cabo una predicción para el último día lunes dentro del periodo desescalada y se observa en la línea amarilla como el modelo representa la estacionalidad del lunes. Aparentemente el modelo identifica unos picos a las 8AM, 14PM y 19PM pero no son suficientes para representar de manera exacta lo que está sucediendo con los desplazamientos en este día en específico. Además, se puede notar en la figura 112 la discrepancia entre el modelo y los datos reales, siendo deseable que las predicciones se acerquen lo más posible a la línea roja.

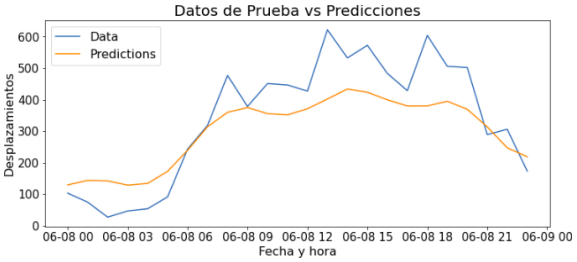

Figura 111: Comparación entre datos reales y predicciones del Modelo ajustado por día de la semana - Desescalada

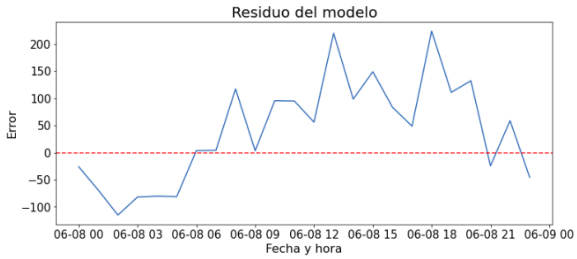

Figura 112: Residuo entre datos reales y predicciones del Modelo ajustado por día de la semana - Desescalada

Modelo	Etapa	RMSE
G	Desescalada	101.9

Tabla 26: Evaluación del modelo G

Periodo de nueva normalidad 09/06/2020 - 31/08/2020

El ajuste de los parámetros para el modelo de nueva normalidad es igual al modelo previo ajustado por día de la semana en el periodo de desescalada.

Modelo de estacionalidad personalizada por día de la semana - Nueva normalidad

```
m = Prophet(weekly_seasonality=True,daily_seasonality=True)
m.add_seasonality(name='Lunes_season', period=1, fourier_order=5, condition_name='Lunes')
m.add_seasonality(name='Martes_season', period=1, fourier_order=5, condition_name='Martes')
m.add_seasonality(name='Miércoles_season', period=1, fourier_order=5, condition_name='Miércoles')
m.add_seasonality(name='Jueves_season', period=1, fourier_order=5, condition_name='Jueves')
m.add_seasonality(name='Viernes_season', period=1, fourier_order=5, condition_name='Viernes')
m.add_seasonality(name='Sábado_season', period=1, fourier_order=5, condition_name='Sábado')
m.add_seasonality(name='Domingo_season', period=1, fourier_order=5, condition_name='Domingo')
```

Figura 113: Configuración de parámetros del Modelo H - Nueva normalidad

En la gráfica de la figura 114 se puede apreciar un aumento en el valor de las predicciones realizadas después de la hora 5:00 y una disminución en el valor de las predicciones en horas de la madrugada. Esto es debido a las características propias del periodo de nueva normalidad. A pesar de que el modelo predictivo se aproxima más a los valores reales, en algunas horas todavía hay un margen de error significativo, tal como se nota en la hora 16:00.

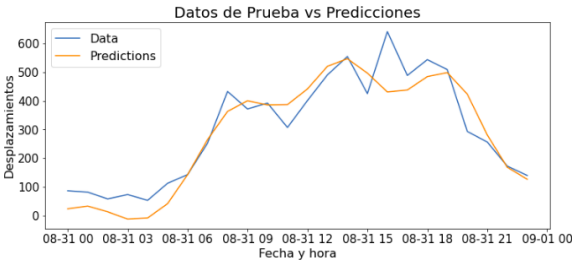

Figura 114: Comparación entre datos reales y predicciones del Modelo ajustado por día de la semana - Nueva normalidad

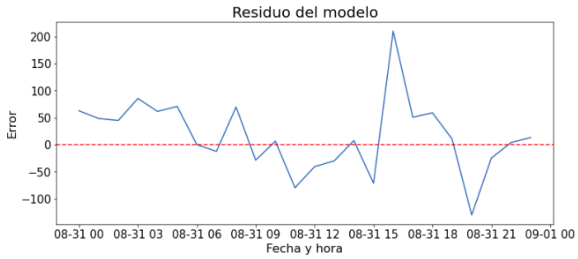

Figura 115: Residuo entre datos de reales y predicciones del Modelo ajustado por día de la semana - Nueva normalidad

Modelo	Etapa	RMSE
H	Nueva normalidad	68.5

Tabla 27: Evaluación del modelo H

El notebook que generó los resultados de la figuras 111, 112, 114 y 115 se encuentra disponible en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `Modelo_Prophet_Des_Nnor-Dia.ipynb`

Los Modelos ajustados por día de la semana reflejan una mejora con respecto al modelo Default tanto en la etapa de Desescalada como en la de Nueva normalidad. Aunque existe una mejora al comparar el valor RMSE de los modelos, la estacionalidad de los días lunes sigue sin ser suficiente para representar todas las variaciones que se presentan en el día. También se ha encontrado hasta este punto que algunas mejoras en la predicción obedecen a factores aleatorios y no a patrones de movilidad.

G. Anexo: Ajuste de regresores adicionales para los modelos de desescalada y Nueva normalidad

Después de haber evaluado el rendimiento de los modelos anteriormente diseñados en base a la estacionalidad, se incluye en esta nueva sección el ajuste de un nuevo parámetro definido como “extra regressor” o regresor adicional. Los regresores incorporados en el modelo son variables complementarias adicionales que actúan como predictores para ayudar a predecir de manera más precisa la variable de interés.

Periodo de desescalada 02/05/2020 - 08/06/2020

El modelo I ajusta cuatro estacionalidades, mensual, semanal, diaria y la estacionalidad personalizada para los días lunes. Adicionalmente se ajusta una estacionalidad multiplicativa que representa adecuadamente el incremento progresivo de los desplazamientos en el periodo de desescalada. Para ajustar el Regresor se calcula un promedio para cada patrón de hora presente en los días lunes y se obtiene un patrón promedio con el que se ajusta el modelo.

Modelo #: Estacionalidad personalizada días Lunes y regresor 24H - Desescalada

```
m = Prophet(seasonality_mode='multiplicative',weekly_seasonality=True,daily_seasonality=True)
m.add_seasonality(name='monthly', period=30.5, fourier_order=5)
m.add_seasonality(name='Lunes_season', period=1, fourier_order=5, condition_name='Lunes')
m.add_regressor('Patron1AM_D'), m.add_regressor('Patron2AM_D')
m.add_regressor('Patron3AM_D'), m.add_regressor('Patron4AM_D')
m.add_regressor('Patron5AM_D'), m.add_regressor('Patron6AM_D')
m.add_regressor('Patron7AM_D'), m.add_regressor('Patron8AM_D')
m.add_regressor('Patron9AM_D'), m.add_regressor('Patron10AM_D')
m.add_regressor('Patron11AM_D'), m.add_regressor('Patron12PM_D')
m.add_regressor('Patron13PM_D'), m.add_regressor('Patron14PM_D')
m.add_regressor('Patron15PM_D'), m.add_regressor('Patron16PM_D')
m.add_regressor('Patron17PM_D'), m.add_regressor('Patron18PM_D')
m.add_regressor('Patron19PM_D'), m.add_regressor('Patron20PM_D')
m.add_regressor('Patron21PM_D'), m.add_regressor('Patron22PM_D')
m.add_regressor('Patron23PM_D'), m.add_regressor('Patron24AM_D')
```

Figura 116: Configuración de parámetros del Modelo I - Desescalada

En la figura 117 se presenta la gráfica de componentes del modelo I, en ella se muestran las estacionalidades del modelo, la tendencia en el periodo desescalada y el regresor adicional. En la representación visual del regresor adicional, se logra apreciar el patrón promedio para cada día lunes dentro del periodo de estudio.

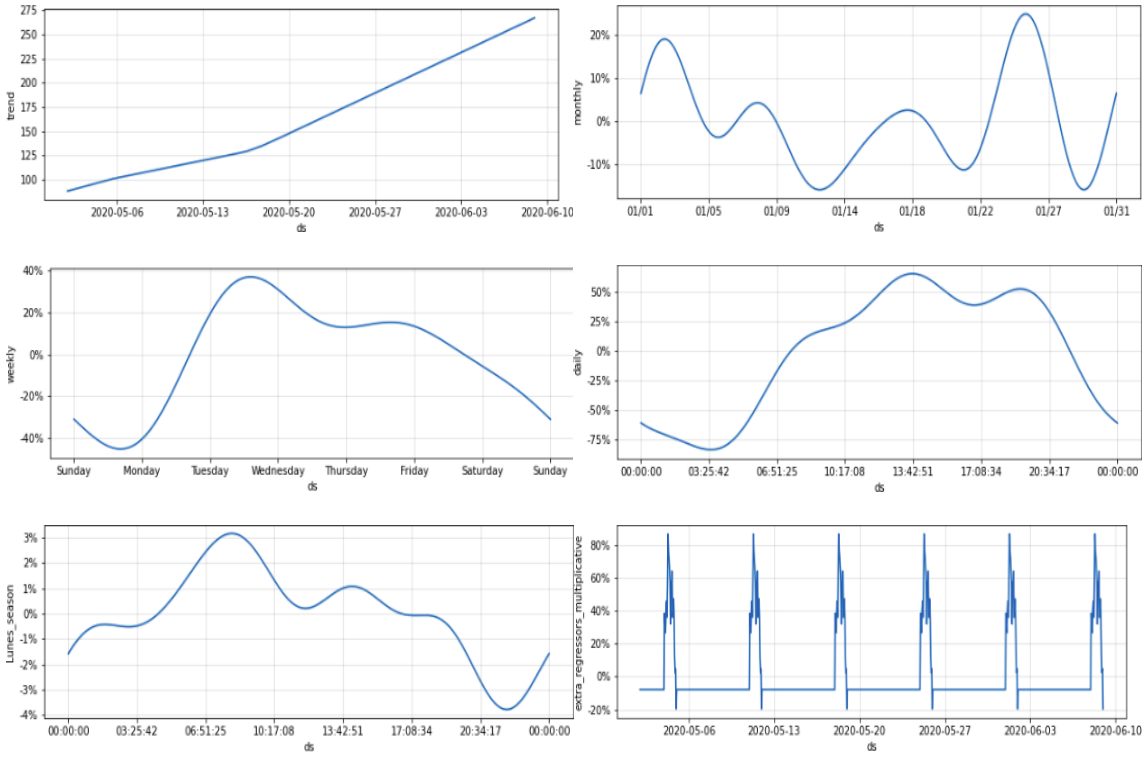

Figura 117: Gráfica de componentes del Modelo I

Al realizar la gráfica de la figura 118, que muestra la comparación entre los datos reales y la predicción del modelo, se logra apreciar un aumento en la exactitud de la predicción. Además en la tabla 28, se logra apreciar una disminución considerable en el valor RMSE del modelo. Estos resultados sugieren que el ajuste del patrón promedio es efectivo.

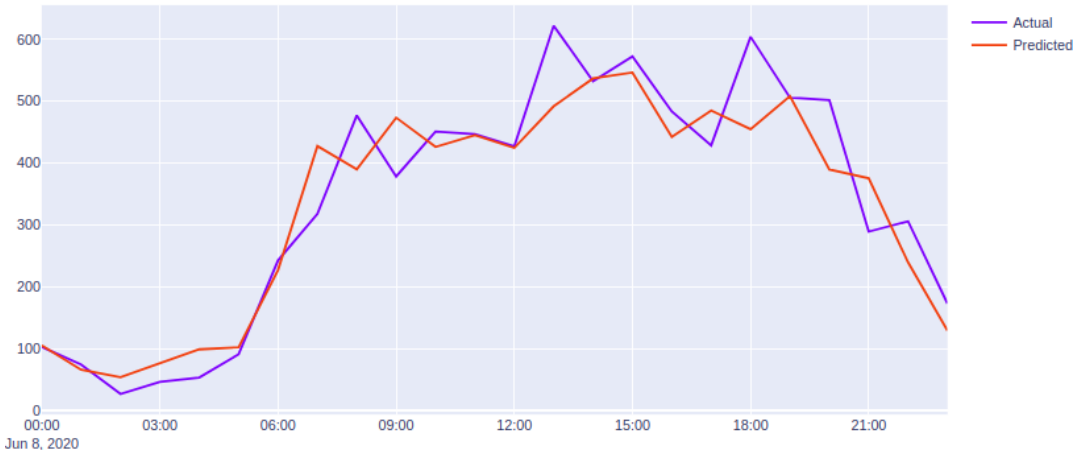

Figura 118: Comparación entre datos reales y predicciones del Modelo I

Modelo	Etapa	RMSE
I	Desescalada	65.9

Tabla 28: Evaluación del modelo I

Sin embargo en la gráfica de la figura 118 se sigue observando un error alto en la predicción para las horas 13:00, 18:00, 19:00. Las cuales presentan una diferencia de más de 100 entre el valor real y la predicción del número de desplazamientos.

El notebook que generó los resultados de la figura 118 se encuentra disponible en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `RegressorDes_PatronPromedio.ipynb`

Periodo de nueva normalidad 09/06/2020 - 31/08/2020

El modelo J a diferencia del modelo anterior ajusta dos estacionalidades, la estacionalidad personalizada para el día lunes y la estacionalidad diaria. En este modelo en particular se ajusta el parámetro “changepoint_prior_scale”, el cual adapta la tendencia teniendo en cuenta más puntos de cambio dentro de la serie temporal. Por último se ajustan los regresores adicionales que ajustan un patrón promedio.

Modelo #: Estacionalidad personalizada días Lunes y regresor 24H - Nueva normalidad

```
m = Prophet(changepoint_prior_scale=0.5,weekly_seasonality=False,daily_seasonality=True,
            seasonality_mode='multiplicative')
m.add_seasonality(name='Lunes_season', period=1, fourier_order=5, condition_name='Lunes')
m.add_regressor('Patron1AM_N',mode='multiplicative'), m.add_regressor('Patron2AM_N',mode='multiplicative')
m.add_regressor('Patron3AM_N',mode='multiplicative'), m.add_regressor('Patron4AM_N',mode='multiplicative')
m.add_regressor('Patron5AM_N',mode='multiplicative'), m.add_regressor('Patron6AM_N',mode='multiplicative')
m.add_regressor('Patron7AM_N',mode='multiplicative'), m.add_regressor('Patron8AM_N',mode='multiplicative')
m.add_regressor('Patron9AM_N',mode='multiplicative'), m.add_regressor('Patron10AM_N',mode='multiplicative')
m.add_regressor('Patron11AM_N',mode='multiplicative'), m.add_regressor('Patron12PM_N',mode='multiplicative')
m.add_regressor('Patron13PM_N',mode='multiplicative'), m.add_regressor('Patron14PM_N',mode='multiplicative')
m.add_regressor('Patron15PM_N',mode='multiplicative'), m.add_regressor('Patron16PM_N',mode='multiplicative')
m.add_regressor('Patron17PM_N',mode='multiplicative'), m.add_regressor('Patron18PM_N',mode='multiplicative')
m.add_regressor('Patron19PM_N',mode='multiplicative'), m.add_regressor('Patron20PM_N',mode='multiplicative')
m.add_regressor('Patron21PM_N',mode='multiplicative'), m.add_regressor('Patron22PM_N',mode='multiplicative')
m.add_regressor('Patron23PM_N',mode='multiplicative'), m.add_regressor('Patron24AM_N',mode='multiplicative')
```

Figura 119: Configuración de parámetros del Modelo J - Nueva normalidad

El parámetro “changepoint_prior_scale” es probablemente el parámetro que tiene más impacto en el modelo predictivo por lo que determina la flexibilidad de la tendencia. Este parámetro por defecto tiene un ajuste de 0.05, el cual funciona para muchas series temporales. Sin embargo en este modelo se modifica este parámetro con el propósito de identificar cambios abruptos en la serie temporal. Se selecciona un valor moderado que le permita al modelo entender los datos de forma más precisa ya que la predicción que se está buscando es por horas, de igual forma al justar el “changepoint” existe la posibilidad de sobre-ajustar el modelo. En la figura 120 se puede apreciar la gráfica de tendencia, la cual refleja una mayor variación en el periodo de datos de entrenamiento. Además se observa las estacionalidades y el patrón promedio para los días lunes.

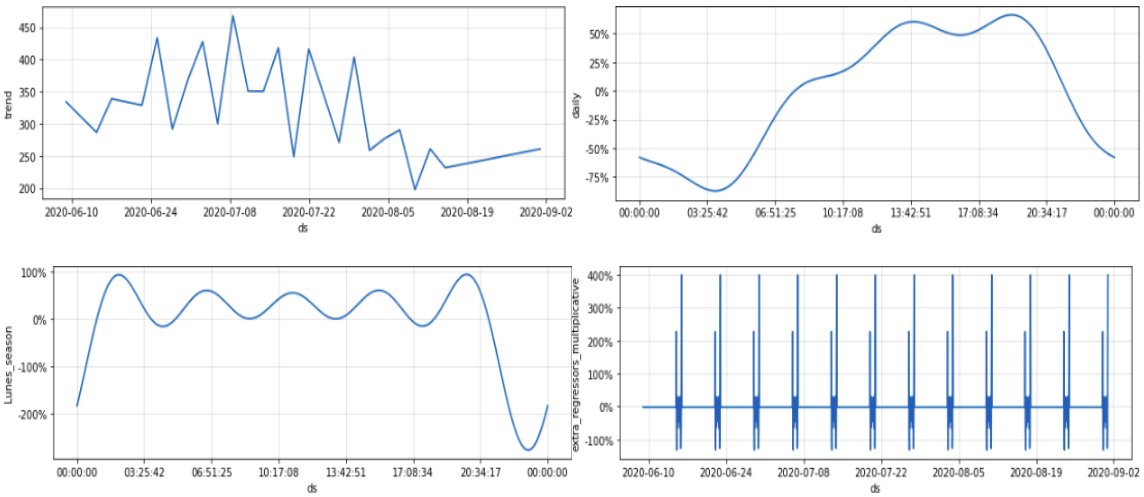

Figura 120: Gráfica de componentes del Modelo J

En la figura 121 se observa como el modelo logra representar adecuadamente el intervalo de horas 00:00 - 07:00 y el intervalo 19:00 - 23:00, que corresponden a horarios con muy pocos cambios en los patrones de desplazamientos. Caso contrario ocurre en horas de la tarde donde ocurren cambios inesperados, tal como se ve reflejado en el pico de desplazamientos en la hora 16:00.

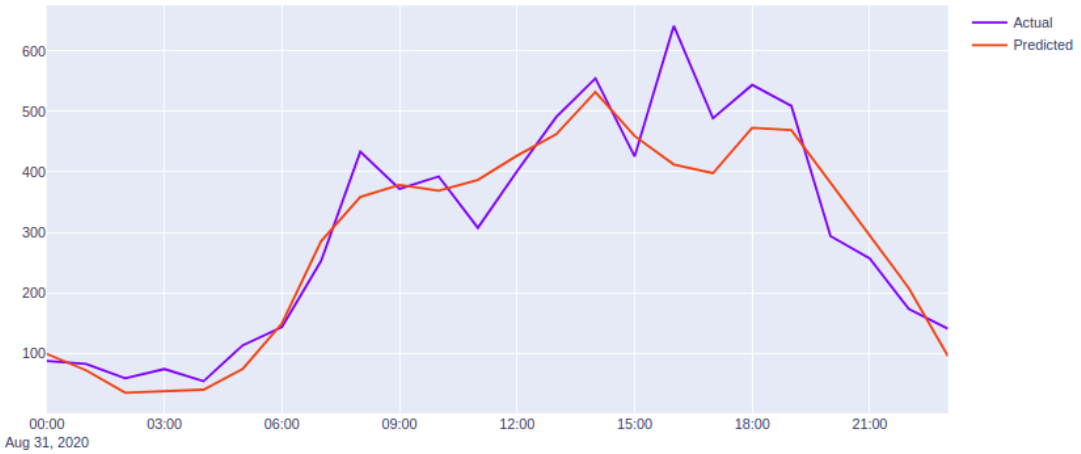

Figura 121: Comparación entre datos reales y predicciones del Modelo J

Modelo	Etapa	RMSE
J	Nueva normalidad	65.1

Tabla 29: Evaluación del modelo J

El notebook que generó los resultados de la figura 121 se encuentra disponible en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `RegressorNnor_PatronPromedio.ipynb`

En conclusión, los ajustes de parámetros realizados en esta sección logran representar un patrón general de los desplazamientos en los días lunes en el periodo de estudio pero no logra representar detalladamente los cambios observados en determinadas horas donde se presentan picos de desplazamientos inesperados.

H. Anexo: Modelo Random Forest

Existen múltiples técnicas de aprendizaje automático para realizar análisis y pronósticos de datos. Prophet, una técnica ampliamente empleada para el pronóstico de series temporales, es solo una de ellas. En esta nueva sección se examinará el desempeño del algoritmo de aprendizaje automático Random Forest. Esta técnica combina múltiples arboles de decisión para construir un modelo predictivo preciso.

Periodo de desescalada 02/05/2020 - 08/06/2020

Para obtener el mejor rendimiento del nuevo modelo es fundamental proporcionar un dataframe que contenga las variables necesarias para que el modelo pueda representar adecuadamente los datos. En la figura 122 se ven representadas dos variables, la variable “t-168” que hace referencia a los datos del mismo día y hora de la semana anterior (en este caso corresponde a un día lunes de la semana previa). La otra variable “RollingMean” contiene el cálculo de la media móvil de los datos de estudio. Por lo que para cada valor de la serie temporal, se estaría calculando la media de los 168 valores anteriores, en otras palabras se estaría haciendo una estimación de la estacionalidad semanal (que equivale a una ventana de tamaño 168). Por otra parte también es necesario indicar al modelo el número de árboles de decisión que se deben incluir. En este caso, se están utilizando 1000 árboles para construir el modelo. Un número elevado de árboles pueden mejorar el rendimiento del modelo pero a medida que se aumenta este valor puede requerir más tiempo de procesamiento y recursos computacionales.

Modelo Random Forest #: Lunes anterior y rolling mean - Desescalada

```
PeriodoDesescalada['t-168'] = PeriodoDesescalada["y"].shift(i)
PeriodoDesescalada["RollingMean"] = PeriodoDesescalada["y"].rolling(window = 168).mean()
model = RandomForestRegressor(n_estimators=1000)
```

Figura 122: Configuración de parámetros del Modelo K - Desescalada

De igual manera que en el modelo Prophet, se realiza la gráfica comparativa de la figura 123, en ella se muestran los datos reales y las predicciones de las 24 horas del último día lunes del periodo de estudio. La predicción del modelo no está evidenciando mejora con respecto a los modelos ajustados en Prophet y el modelo continua siendo incapaz de representar las variaciones del día.

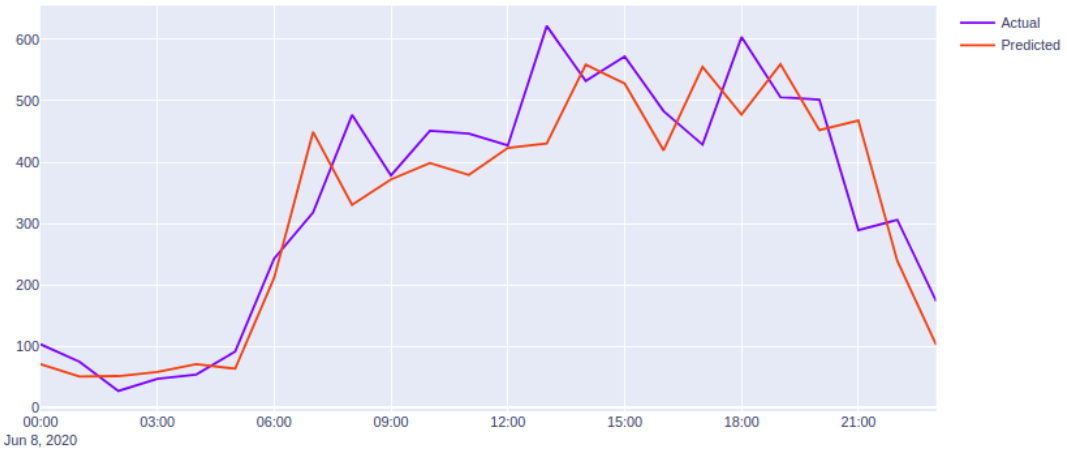

Figura 123: Comparación entre datos reales y predicciones del Modelo K

Modelo	Etapa	RMSE
K	Desescalada	84.8

Tabla 30: Evaluación del modelo K

El notebook que generó los resultados de la figura 123 se encuentra disponible en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `RandomForest_Des.ipynb`

Periodo de nueva normalidad 09/06/2020 - 31/08/2020

Para el modelo en el periodo de nueva normalidad se ajustarán dos variables al igual que el modelo en Desescalada. La variable “t-168” que corresponde a un día lunes de la semana previa y la variable “t-72” que corresponde al viernes anterior si se toma como punto de partida el día lunes. La variable “t-168” toma este nombre debido a que se tienen en cuenta los datos localizados 168 posiciones hacia atrás, de igual manera ocurre con la variable “t-72”. Finalmente y al igual que en el modelo en desescalada se ajusta el parámetro ***“n_estimators”*** con un valor de 1000.

Modelo Random Forest #: Lunes y viernes anterior - Nueva normalidad

```
PeriodoNnormalidad['t-168')] = PeriodoNnormalidad["y"].shift(i)
PeriodoNnormalidad['t-72'] = PeriodoNnormalidad["y"].shift(i)
model = RandomForestRegressor(n_estimators=1000)
```

Figura 124: Configuración de parámetros del Modelo L - Nueva normalidad

En la predicción realizada por el modelo L tampoco se evidencia una mejora en comparación con su equivalente ajustado con Prophet. Este modelo se configuró para representar los días lunes presentes en el periodo de nueva normalidad, aún así se decide tener en cuenta los datos del viernes de la semana previa dado que se observa un patrón similar en ambos días. Los resultados de la predicción son presentados en la figura 125.

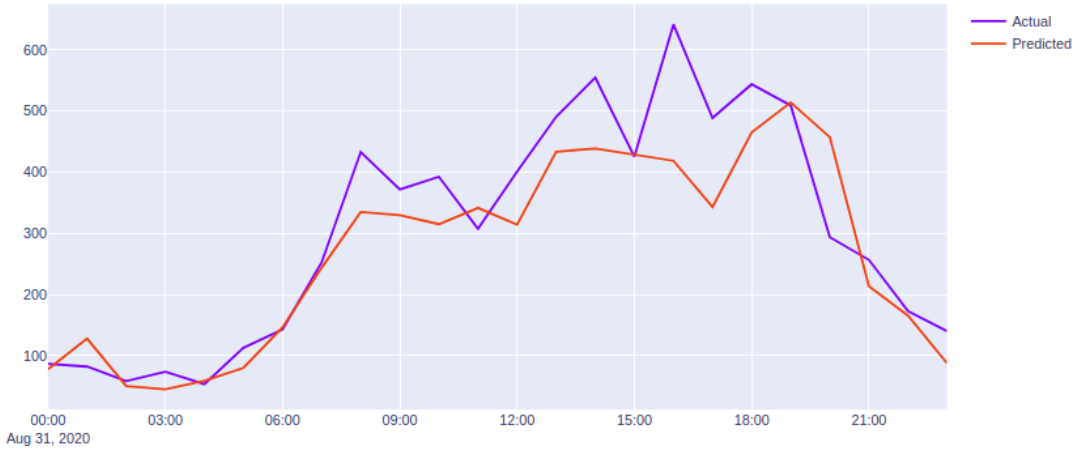

Figura 125: Comparación entre datos reales y predicciones del Modelo L

Modelo	Etapa	RMSE
L	Nueva normalidad	80.5

Tabla 31: Evaluación del modelo L

El notebook que generó los resultados de la figura 125 se encuentra disponible en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `RandomForest_Nnor.ipynb`

Examinando los resultados de la sección actual se puede concluir que aunque Random Forest es una técnica de aprendizaje automático ampliamente utilizada, existen otros métodos más adecuados y precisos para la predicción de series temporales.

Un modelo más adecuado para la predicción de series temporales es el modelo Prophet, el cual permite ajustar múltiples parámetros, entre ellos la tendencia y la estacionalidad de los datos, así como variables complementarias mediante regresores. De esta manera el ajuste de múltiples parámetros aumentan la capacidad del modelo para adaptarse a diferentes patrones.

I. Anexo: Análisis de patrones en la serie temporal

La búsqueda de patrones significativos dentro del conjunto de datos representa uno de los desafíos principales en el trabajo de investigación, por consiguiente es conveniente utilizar todas las herramientas que estén al alcance. En el contexto de análisis de series temporales es posible utilizar “Matrix Profile” y/o “Time series snippets”, a través de estas herramientas se obtienen algunos resultados que se mostrarán en este apartado. A continuación, en la figura 126 se observa el conjunto de datos que serán analizados mediante esta herramienta.

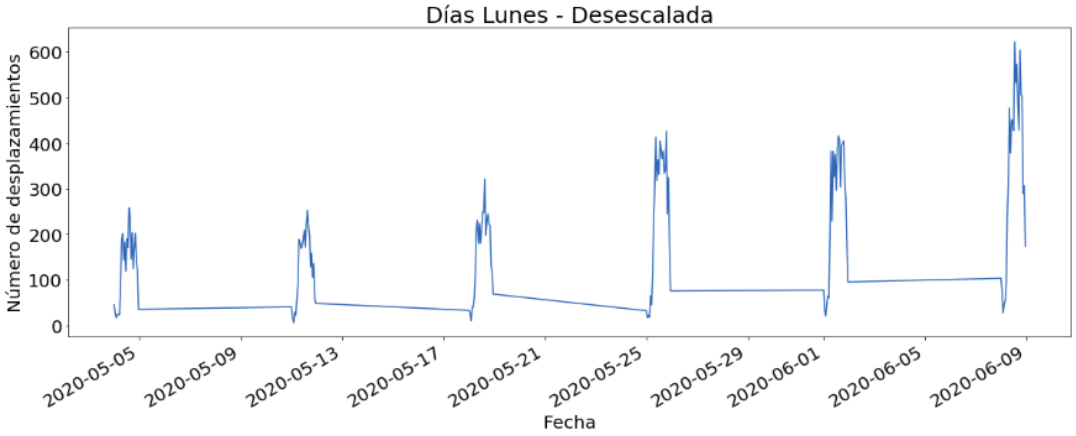

Figura 126: Conjunto de datos correspondiente a los días lunes del periodo de desescalada

La función principal de estas herramientas es identificar patrones y anomalías. Por lo tanto, resultan esenciales para proporcionar información sobre el comportamiento de la serie temporal. Para poder hacer uso de estas herramientas, se requiere la instalación de la librería “Matrixprofile” y el paquete “Stumpy”. En la siguiente figura se muestra el perfil de matriz para diferentes tamaños de ventana, se puede notar que el tamaño de la ventana cambia drásticamente el perfil de matriz.

La figura 127 está dando a conocer 4 perfiles de la serie temporal, donde se destaca especialmente el perfil de 12 horas que representa claramente la estacionalidad de los 6 lunes presentes en el periodo de desescalada. Sin embargo no logra detectar detectar un patrón reiterativo (motif) o alguna anomalía (discord) dentro del conjunto de datos. Los “motifs” dentro del perfil de matriz estarían representados por una línea, por el contrarío los “discords” estarían representadas dentro del perfil de matriz mediante una elevación repentina del pico [46].

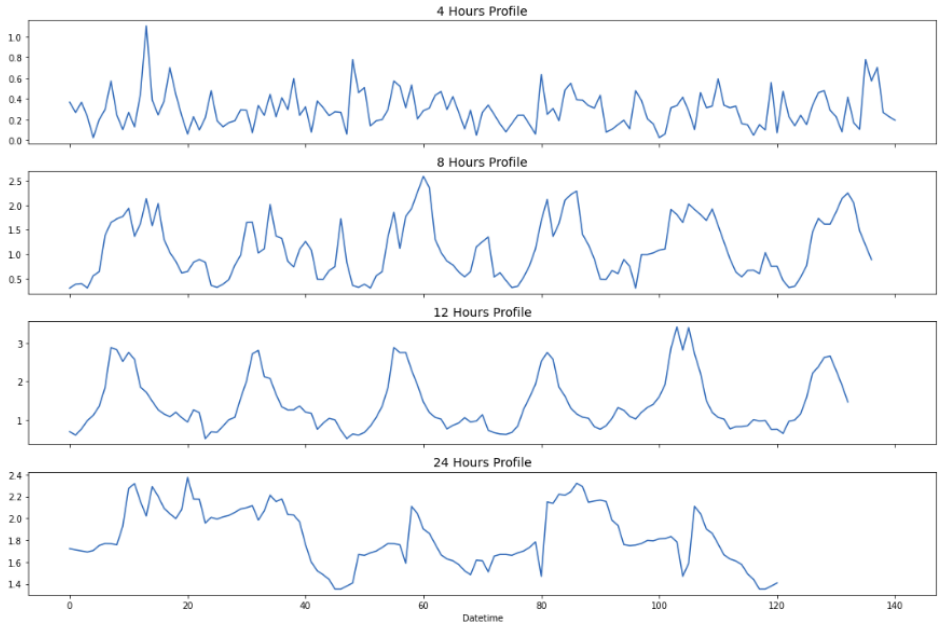

Figura 127: Perfiles de matriz con diferentes tamaños de ventana

Utilizando el paquete Stumpy se realiza una nueva prueba para encontrar los patrones presentes de los días lunes en el periodo de desescalada. La intención de esta nueva exploración es identificar un patrón reiterativo que este presente en cada lunes de la serie temporal. En la gráfica de la figura 128 se puede observar que el “motif” ocurre en diferentes horas. Sin embargo, aunque el algoritmo identifique que las áreas sombreadas tienen una similitud entre ellas, en realidad, el resultado obtenido en el contexto de la movilidad carece de sentido y el patrón encontrado se debe a una coincidencia. La Figura 128 se basa en un ejemplo similar presentado en la siguiente fuente [47].

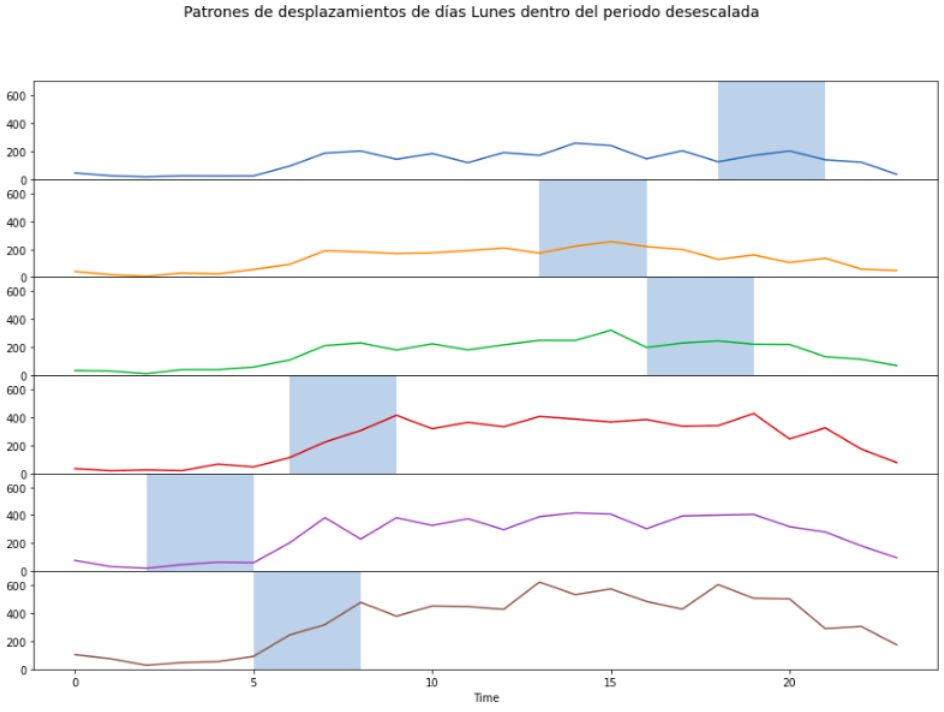

Figura 128: Patrones reiterativos identificados dentro de la serie temporal

De la exploración de esta herramienta se puede concluir que aunque no se obtuvieron

los resultados esperados, la herramienta podría ser de gran utilidad para el análisis de patrones de movilidad si se estudian exhaustivamente todas sus características.

El notebook que generó los resultados de las figuras 127 y 128 se encuentran disponibles en el siguiente repositorio de github `https://github.com/Jhonvalencia77/TFG_Unicauca.git` con el nombre de `Patrones_MatrixProfile_Desescalada.ipynb`

Referencias

[1] D. Muley, M. Shahin, C. Dias y M. Abdullah, "Role of transport during outbreak of infectious diseases: Evidence from the past," *Sustainability (Switzerland)*, vol. 12, págs. 1-22, 18 2020, ISSN: 20711050. DOI: 10.3390/SU12187367.

[2] K. Gkiotsalitis y O. Cats, "Public transport planning adaption under the COVID-19 pandemic crisis: literature review of research needs and directions," *Transport Reviews*, vol. 41, págs. 374-392, 3 2021, ISSN: 14645327. DOI: 10.1080/01441647.2020.1857886. dirección: https://doi.org/10.1080/01441647.2020.1857886.

[3] "Reportes Portal Nacional Datos Abiertos | Datos Abiertos Colombia." (), dirección: https://www.datos.gov.co/stories/s/Reportes-Portal-Nacional-Datos-Abiertos/pvyw-9yqs (visitado 11-03-2022).

[4] "Portal de Datos Abiertos de TransMilenio." (), dirección: https://datosabiertos-transmilenio.hub.arcgis.com/search?collection=Dataset (visitado 11-03-2022).

[5] "Datos Abiertos - Metro de Medellín." (), dirección: https://datosabiertos-metrodemedellin.opendata.arcgis.com/search?categories=movilidad (visitado 11-03-2022).

[6] "Datos Abiertos - METROCALI." (), dirección: https://www.metrocali.gov.co/wp/datos-abiertos/ (visitado 11-03-2022).

[7] "Open Data Movilidad | Ministerio de Transportes, Movilidad y Agenda Urbana." (), dirección: https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data/opendata-movilidad (visitado 11-03-2022).

[8] P. d. t. d. A. Madrid. "Estado de la movilidad de la ciudad de Madrid 2020." (), dirección: https://transparencia.madrid.es/FWProjects/transparencia/Movilidad/Trafico/InformesMovilidad/Ficheros/InformeMovilidad2020.pdf (visitado 05-09-2023).

[9] A. Aloi, B. Alonso, J. Benavente et al., *Effects of the COVID-19 lockdown on urban mobility: Empirical evidence from the city of Santander (Spain)*, 2020. DOI: 10.3390/su12093870.

[10] A. Tirachini y O. Cats, "COVID-19 and public transportation: Current assessment, prospects, and research needs," *Journal of Public Transportation*, vol. 22, págs. 1-34, 1 2020, ISSN: 23750901. DOI: 10.5038/2375-0901.22.1.1.

[11] F. A. Ramirez-Buitrago, N. A. Correal-Huertas, L. D. Ramirez-Leuro, D. A. Sandoval-Pedreros y L. A. Rubio-Caballero, "Tools for the monitoring, user characterization, and their applications to the Public Integrated Transport System due to the COVID-19 disease effects: A case study in Bogota, TRANSMILENIO company," *Transportation Research Procedia*, vol. 58, págs. 431-438, 2019 2021, ISSN: 23521465. DOI: 10.1016/j.trpro.2021.11.058. dirección: https://doi.org/10.1016/j.trpro.2021.11.058.

[12] N. Shinohara, J. Sakaguchi, H. Kim et al., "Survey of air exchange rates and evaluation of airborne infection risk of COVID-19 on commuter trains," *Environment International*, vol. 157, pág. 106 774, July 2021, ISSN: 18736750. DOI: 10.1016/j.envint.2021.106774. dirección: https://doi.org/10.1016/j.envint.2021.106774.

[13] M. B. Tentor, "Sistema Basado en Analítica Predictiva para Ofrecer Recomendación sobre Uso de Transporte Público en Madrid Durante la Pandemia," Universidad Politécnica de Madrid, 2021, págs. 1-75.

[14] D. M. Fernández, "Análisis y Diseño de un Sistema para Apoyar un Modelo de Transporte Público Seguro Basado en los Datos de Movilidad Durante la Pandemia en Madrid," Universidad Politécnica de Madrid, 2021, págs. 1-76.

[15] J. C. Arias-Chicaiza, C. d. R. Arias-Chicaiza, C. A. Oñate-Haro y S. A. Diaz-Pazmiño, "La planificación como herramienta en la movilidad del transporte Urbano," *Dominio de las ciencias*, vol. 8, págs. 61-80, 2022. dirección: http://dx.doi.org/10.23857/dc.v8i2.2633.

[16] J. A. Quirós-Alés, "Determinación de frecuencias y capacidades óptimas en redes densas de ferrocarril de tránsito rápido," Universidad de Sevilla, 2015, págs. 1-164.

[17] N. Tyler, "La toma de decisiones en la formulación de un modelo de transporte público sostenible," *Revista de Tecnología ¦ Journal of Technology*, vol. 13, n.º 3, págs. 109-114, 2014.

[18] Google_Transit. "Descripción general de GTFS estáticas | Transporte público estático | Google for Developers." (), dirección: https://developers.google.com/transit/gtfs?hl=es.

[19] "Dash Documentation User Guide | Plotly." (), dirección: https://dash.plotly.com/ (visitado 05-06-2023).

[20] "Overview — Ray 2.5.1." (), dirección: https://docs.ray.io/en/latest/ray-overview/index.html (visitado 05-06-2023).

[21] S. J. Taylor y B. Letham, "Forecasting at Scale," *American Statistician*, vol. 72, págs. 37-45, 1 ene. de 2018, ISSN: 15372731. DOI: 10.1080/00031305.2017.1380080.

[22] Sitiobigdata. "Aprendizaje automatico y las Metricas de regresión." (), dirección: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/# (visitado 05-06-2023).

[23] Pronosticoexperto. "Introducción a las Métricas para la Medición de la Asertividad y Errores de los Pronósticos." (), dirección: https://www.pronosticoexperto.com/cpto-metricas-medicion-asertividad (visitado 05-06-2023).

[24] IBM. "Conceptos básicos de ayuda de CRISP-DM." (2021), dirección: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview (visitado 05-09-2023).

[25] ComunidaddeMadrid. "Municipios de la Comunidad de Madrid." (), dirección: https://www.comunidad.madrid/servicios/municipios/municipios-comunidad-madrid.

[26] AyuntamientodeMadrid. "Organización municipal." (), dirección: https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Organizacion-municipal?vgnextfmt=default&vgnextchannel=2ef308a90a1e9410VgnVCM100000171f5a0aRCRD#distritos.

[27] BarriosdeMadrid. "Distritos de madrid." (), dirección: https://barriosdemadrid.net/mapas/distritos/.

[28] MetroMadrid. "Informe corporativo 2021." (), dirección: https://www.metromadrid.es/sites/default/files/documentos/Informecorporativo2021ESP.pdf.

[29] Renfe. "Cercanías de Madrid (Mapa de Rutas y Líneas)." (), dirección: https://www.renfe.com/es/es/cercanias/cercanias-madrid/mapas.

[30] ConsorcioRegionaldeTransportesdeMadrid. "La Red de Metro Ligero/Tranvía." (), dirección: https://www.crtm.es/tu-transporte-publico/metro-ligero.aspx.

[31] ComunidaddeMadrid. "Muévete en transporte público." (), dirección: https://www.comunidad.madrid/servicios/transporte/muevete-transporte-publico.

[32] Elpais. "Coronavirus: Los datos de una pandemia en tres olas | Sociedad | EL PAÍS." (), dirección: https://elpais.com/sociedad/2021-03-10/los-datos-de-una-pandemia-en-tres-olas.html.

[33] WikimediaCommons. "Diagrama de tiempo estados alarma COVID-19." (), dirección: https://commons.wikimedia.org/wiki/File:Diagrama_de_tiempo_estados_alarma_COVID-19.jpg.

[34] Wikipedia. “Pandemia de COVID-19 en España.” (), dirección: https://es.wikipedia.org/wiki/Pandemia_de_COVID-19_en_Espa%C3%B1a.

[35] J. Valencia Bolaños. “Repositorio Github- TFG_Unicauca.” (2023), dirección: https://github.com/Jhonvalencia77/TFG_Unicauca.

[36] D. C. Corrales, J. C. Corrales y A. Ledezma, “How to address the data quality issues in regression models: A guided process for data cleaning,” *Symmetry*, vol. 10, n.º 4, págs. 1-20, 2018, ISSN: 20738994. DOI: 10.3390/sym10040099.

[37] O. P. de la Salud, *Indicadores de salud. Aspectos conceptuales y operativos*. 2018, ISBN: 978-92-75-32005-1. dirección: www.paho.org/permissions.

[38] “Covid 19 -TIA por Municipios y Distritos de Madrid - Conjuntos de datos - Datos Abiertos Comunidad de Madrid.” (), dirección: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos (visitado 18-05-2023).

[39] S. O. Blog. “Stop aggregating away the signal in your data.” (), dirección: https://stackoverflow.blog/2022/03/03/stop-aggregating-away-the-signal-in-your-data/ (visitado 01-04-2022).

[40] B. Kumar Jha y S. Pande, “Time Series Forecasting Model for Supermarket Sales using FB-Prophet,” *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, n.º Iccmc, págs. 547-554, 2021. DOI: 10.1109/ICCMC51019.2021.9418033.

[41] C. B. Aditya Satrio, W. Darmawan, B. U. Nadia y N. Hanafiah, “Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET,” *Procedia Computer Science*, vol. 179, n.º 2020, págs. 524-532, 2021, ISSN: 18770509. DOI: 10.1016/j.procs.2021.01.036. dirección: https://doi.org/10.1016/j.procs.2021.01.036.

[42] “GTFS Red de Metro | Datos Abiertos del Consorcio Regional de Transportes de Madrid.” (), dirección: https://data-crtm.opendata.arcgis.com/datasets/crtm::gtfs-red-de-metro/about.

[43] “Boletín estadístico online - Información estadística - Ministerio de Fomento.” (), dirección: https://apps.fomento.gob.es/BoletinOnline/?nivel=2&orden=07000000 (visitado 01-06-2023).

[44] Metromadrid. “Proyectos y datos | Metro de Madrid.” (), dirección: https://www.metromadrid.es/es/transparencia/proyectos-y-datos#panel1 (visitado 01-06-2023).

[45] DatosMadrid. “Accidentes de tráfico de la Ciudad de Madrid - Portal de datos abiertos del Ayuntamiento de Madrid.” (), dirección: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=7c2843010d9c3610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM10000017
vgnextfmt=default (visitado 20-12-2022).

[46] “Introduction to Matrix Profiles. A Novel Data Structure for Mining Time... | by Tyler Marrs | Towards Data Science.” (), dirección: https://towardsdatascience.com/introduction-to-matrix-profiles-5568f3375d90 (visitado 01-06-2023).

[47] “stumpy/docs/Tutorial_Consensus_Motif.ipynb at main · TDAmeritrade/stumpy · GitHub.” (), dirección: https://github.com/TDAmeritrade/stumpy/blob/main/docs/Tutorial_Consensus_Motif.ipynb (visitado 01-06-2023).