Proyecto Final Data Mining

Universidad: UAGRM School of Engineering

Maestría: Ciencia de Datos e Inteligencia Artificial V1E2

Módulo: Data Mining

Docente: Msc. Renzo Franck Claure Aracena

Integrantes Grupo 1:

- Carolina Bello
- Joseph Thenier
- Luis Martínez
- Nicolas Oporto
- Oscar Loayza

Agosto 2025

1. Introducción

Como parte del proyecto final del módulo Data Mining de la Maestría en Ciencia de Datos e Inteligencia Artificial de la UAGRM, se nos planteó el reto de seleccionar una base de datos y aplicar sobre ella todas las fases del proceso de Minería de Datos, desde la integración y preparación hasta la obtención de resultados, conclusiones y propuestas.

Para el desarrollo de este proyecto, nuestro grupo eligió la base de datos pública de Bluebikes (https://bluebikes.com/system-data), la cual contiene información relacionada con el sistema de bicicletas compartidas de Boston y otras ciudades de Massachusetts. Esta base de datos resulta particularmente interesante porque permite analizar patrones de movilidad urbana, hábitos de uso del transporte sostenible y otros factores asociados al comportamiento de los usuarios.

Desde una perspectiva de negocio, Bluebikes opera bajo un modelo basado en membresías y pases, que establece reglas claras para el uso de las bicicletas: https://help.bluebikes.com/hc/en-us/articles/360034926452-What-if-I-keep-a-bike-out-too-long

Por ejemplo:

- Usuarios con pase individual (Single Ride Pass): incluyen los primeros 30 minutos del viaje en la tarifa base; a partir de ese tiempo se cobra $0,25 por minuto adicional.
  Usuarios con pase diario (Day Pass): incluyen las primeras 2 horas del viaje; luego aplican cargos extra de $0,25 por minuto.
- Miembros mensuales y anuales: tienen derecho a viajes de hasta 45 minutos sin costo adicional, y luego pagan $0,10 por minuto extra.
- Miembros con tarifa reducida (Income-Eligible): disponen de hasta 60 minutos por viaje sin costo extra, y luego se cobra $0,07 por minuto.

- En caso de que una bicicleta no sea devuelta en un periodo de 24 horas, el sistema aplica una penalización de hasta $1.200 USD por pérdida o robo.

Estas reglas de negocio son claves para comprender los patrones de uso observados en los datos. Los usuarios ocasionales suelen realizar viajes más largos y están más expuestos a cargos adicionales, mientras que los miembros regulares tienden a optimizar el tiempo de sus recorridos para evitar sobrecostos. De este modo, el análisis de datos no solo permite estudiar la movilidad urbana, sino también evaluar la sostenibilidad económica del sistema y el comportamiento estratégico de los usuarios frente a las tarifas.

En este contexto, el objetivo del trabajo es transformar datos de viajes en conocimiento accionable para apoyar decisiones en:

- **Operaciones:** optimizar la rotación y rebalanceo de bicicletas.
- **Planificación de estaciones**: identificar cobertura, demanda y puntos críticos.
- **Analítica de demanda:** prever la necesidad de recursos en horas pico o temporadas específicas.

El análisis se guía por alguna de las siguientes preguntas clave de negocio:

- ¿Cómo se distribuyen las duraciones de viaje y cuáles son los patrones por hora/día/estación del año?
- ¿Qué calidad y cobertura geográfica tiene el histórico de viajes? ¿Hay outliers o registros inválidos?
- ¿Podemos predecir la demanda por hora en estaciones para apoyar re-balanceo y staffing?

Para este proyecto tendremos los siguientes entregables establecidos:

- **Base “minable”** integrada y limpia, en formato parquet.
- **Análisis exploratorio (EDA)** con hallazgos sobre duraciones y geografía.

- **Conjunto de features** temporales y geoespaciales derivados de los datos crudos.
- **Modelo de pronóstico por hora/estación**, con comparación frente a un baseline ingenuo.

En conclusión este proyecto busca que los estudiantes experimenten el ciclo completo de KDD (Knowledge Discovery in Databases): integración, limpieza, preprocesamiento, aplicación de algoritmos de minería de datos, evaluación de resultados y comunicación efectiva. De esta manera, se consolidan los conceptos aprendidos en el curso y se generan propuestas de valor estratégico que contribuyen tanto a la gestión de un sistema de transporte sostenible como al análisis del comportamiento urbano.

2. Tratamiento de los Datos

El tratamiento de los datos fue una de las fases más críticas del proyecto, ya que permitió transformar los registros crudos de viajes de Bluebikes en un conjunto de datos limpio, enriquecido y estructurado, listo para análisis avanzados. Este proceso se realizó utilizando PySpark en Google Colab, lo que garantizó escalabilidad y eficiencia al trabajar con millones de registros.

2.1. Integración multi-esquema

Se recopilaron más de 126 archivos fuente, correspondientes a viajes de distintos periodos históricos, lo que implicó unificar tres esquemas diferentes. Este proceso requirió:

- Mapeo de nombres de columnas para homogeneizar las estructuras.
- Casteo de tipos de datos (timestamps, enteros, floats).
- Generación de una versión de esquema por registro, asegurando trazabilidad.

De este modo, se consolidó una única base integrada, con más de 27 millones de registros, garantizando una ingesta robusta y escalable.

2.2. Reglas de limpieza aplicadas

Para asegurar consistencia y confiabilidad en los análisis, se aplicaron las siguientes reglas:

- Conversión de tipos: timestamps y coordenadas fueron transformados a tipos correctos.
- Imputación de rideable_type: valores nulos se etiquetaron como docked_bike, siguiendo criterios de negocio.
- Estaciones nulas: etiquetadas como Dockless start/end.
- Variables descartadas: se eliminaron columnas crónicas o poco predictivas (birth_year, gender, postal_code).
- Duración de viajes: se eliminaron viajes < 60 segundos y aquellos con duración mayor a 24 h (86.400 s).
- Coordenadas inválidas: eliminación de registros sin coordenadas finales o con valores 0/0.
- Validación geográfica: se aplicó un filtro espacial usando el BBOX de Boston y Salem/Beverly (con un buffer), descartando viajes fuera del área de cobertura.

2.3. Limpieza y estandarización temporal

Uno de los principales desafíos fue estandarizar los tiempos de inicio y fin de los viajes:

- Conversión de la columna started_at al huso horario de Boston (America/New_York).
- Creación de la columna started_at_local y una marca truncada por hora (ts_hour).
- Eliminación de registros con duraciones negativas o excesivas.

Esto fue fundamental para garantizar la coherencia de los análisis temporales, dado que los patrones de uso cambian drásticamente según la hora local.

2.4. Ingeniería de características

Temporales:

- Año, mes, día de la semana y hora del viaje.
- Clasificación estacional (invierno, primavera, verano, otoño).
- Indicador de fin de semana (*is_weekend*).
- Transformaciones seno/coseno para capturar la ciclicidad de hora y día.

Geospaciales:

- Distancia entre estación de inicio y fin usando **Haversine** (*haversine_distance_km*).
- Indicadores de estaciones populares (percentil 90 en número de viajes).

Comportamiento e interacción:

- Velocidad promedio (avg_speed_kmh) = distancia / tiempo.
- Viajes de ida y vuelta (is_round_trip): mismos puntos de inicio y fin.

Calendario:

- Integración de un calendario de feriados de Massachusetts mediante la librería holidays, generando la variable is_holiday.

2.5. Consolidación

El DataFrame final incorporó 18 nuevas columnas derivadas, pasando de un dataset transaccional a uno multidimensional, con variables:

- Temporales
- Espaciales
- Cíclicas
- De comportamiento
- De calendario

Este dataset enriquecido fue persistido en formato Parquet (df_final_bluebikes_v1.parquet y posteriormente df_final_bluebikes_v2.parquet), garantizando eficiencia en el cómputo analítico y reproducibilidad del proceso.

2.6. Cobertura temporal, esquemas y calidad inicial

El dataset integrado abarca el histórico de Bluebikes/Hubway 2015–2025, lo que implicó procesar y unificar tres versiones de esquema detectadas durante la ingestión:

- Esquema 1 (legado, 27 archivos): ride_id, rideable_type, started_at, ended_at, start_lat/lng, end_lat/lng, member_casual.
- Esquema 2 (reciente, 35 archivos): tripduration, starttime, stoptime, bikeid, usertype, postal_code.
- Esquema 3 (intermedio, 64 archivos): incluye los campos del esquema 2 más birth_year y gender.

En total, el consolidado superó los 27 millones de viajes, lo que obligó a configurar y ejecutar el procesamiento en PySpark con parámetros de memoria y particiones ajustados (spark.driver.memory=8g, spark.executor.memory=8g, spark.sql.shuffle.partitions=200), garantizando escalabilidad en las fases de integración, perfilado y limpieza.

En cuanto a la calidad inicial de los datos, se observaron:

- Alta ausencia de valores en variables heredadas (birth_year, gender, postal_code) y en algunos registros de rideable_type.
- Estaciones nulas en los años recientes, debido a la evolución hacia un esquema dockless (viajes sin anclaje fijo).
- Registros con coordenadas faltantes, en cero o ubicados fuera del área operativa de Boston y ciudades aledañas.

2.7. Resultado del tratamiento de datos

Tras la integración y limpieza, la base de viajes crudos se transformó en un dataset de alta calidad, minable y coherente, alineado con los objetivos de negocio y del proyecto académico. Este esfuerzo de integración multi-esquema (126 archivos, 3 versiones, ≈27M filas) y de depuración geoespacial/temporal representó la fase más intensiva del proyecto, pero aseguró contar con una base sólida para análisis avanzados y modelado.

El dataset final permitió sentar las bases para:

- Clasificar usuarios (miembros vs. ocasionales).
- Predecir la demanda por estación y hora.
- Detectar patrones de uso recreativo vs. laboral.
- Evaluar la eficiencia y popularidad de estaciones.

En conclusión, el proceso de tratamiento no solo depuró inconsistencias, sino que enriqueció y estructuró la información, convirtiendo un histórico heterogéneo en un recurso estratégico para minería de datos y analítica urbana.

3. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) constituye un paso fundamental para comprender la estructura del conjunto de datos y descubrir patrones iniciales antes de aplicar modelos predictivos. En el caso del proyecto **Bluebikes**, se trabajó con millones de registros de viajes individuales, lo que permitió identificar tendencias de uso, comportamientos de los usuarios y particularidades de la red de estaciones.

3.1. Integración y validación inicial

Se integraron más de 100 archivos correspondientes a viajes históricos, con distintos esquemas de columnas según el año y versión.

Se estandarizó el tipo de datos en las columnas clave: fechas (started_at, ended_at), coordenadas (start_lat, start_lng, end_lat, end_lng) y variables categóricas.

Se verificó la cobertura temporal y el número de registros totales, alcanzando millones de viajes que conforman la base consolidada.

3.2. Verificación de nulos

Previo al análisis exploratorio, se revisó nuevamente la presencia de valores nulos en el dataset final. Se confirmó que las variables críticas para el análisis (duración de viaje, estaciones, coordenadas y categoría de usuario) no presentan valores faltantes. Los únicos campos con vacíos significativos corresponden a variables demográficas como birth_year, gender y postal_code (con más del 65% de nulos), que fueron descartadas del análisis principal por su bajo nivel de completitud.

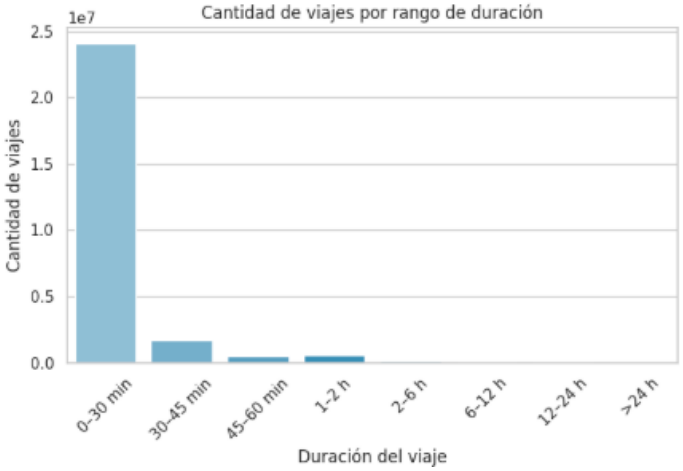
De esta manera, el conjunto de datos utilizado en el EDA garantiza consistencia y representatividad, evitando que los resultados se vean sesgados por registros incompletos.

columna	nulos	% nulos
birth_year	19377092	71.59
gender	18686522	69.04
postal_code	18253759	67.44
rideable_type	17220537	63.63
end_station_id	29036	0.11
end_station_name	28417	0.10
end_lat	21830	0.08
end_lng	21830	0.08
start_station_name	2033	0.01
start_station_id	2033	0.01
start_lat	0	0.00
start_lng	0	0.00
started_at	0	0.00
ride_id	0	0.00
duration_sec	0	0.00
ended_at	0	0.00
member_casual	0	0.00
schema_version	0	0.00
periodo	0	0.00

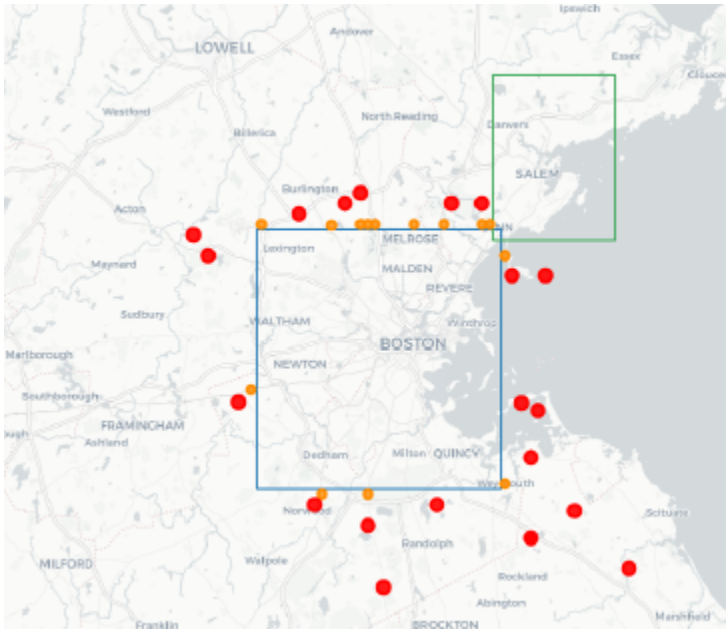
3.3. Distribución de la duración de los viajes

- Se generó un análisis de la variable duration_sec, mostrando que la mayoría de los viajes tienen una duración corta, concentrada en intervalos de 10 a 30 minutos.
- Para facilitar la interpretación, se agruparon las duraciones en intervalos (“bins”):
  - 0–30 min
  - 30–45 min
  - 45–60 min
  - 1–2 horas
  - 2–6 horas
  - 6–12 horas
  - 12–24 horas
- Los resultados confirman que más del 70% de los viajes se concentran en menos de 45 minutos, en línea con las políticas de membresía del sistema.

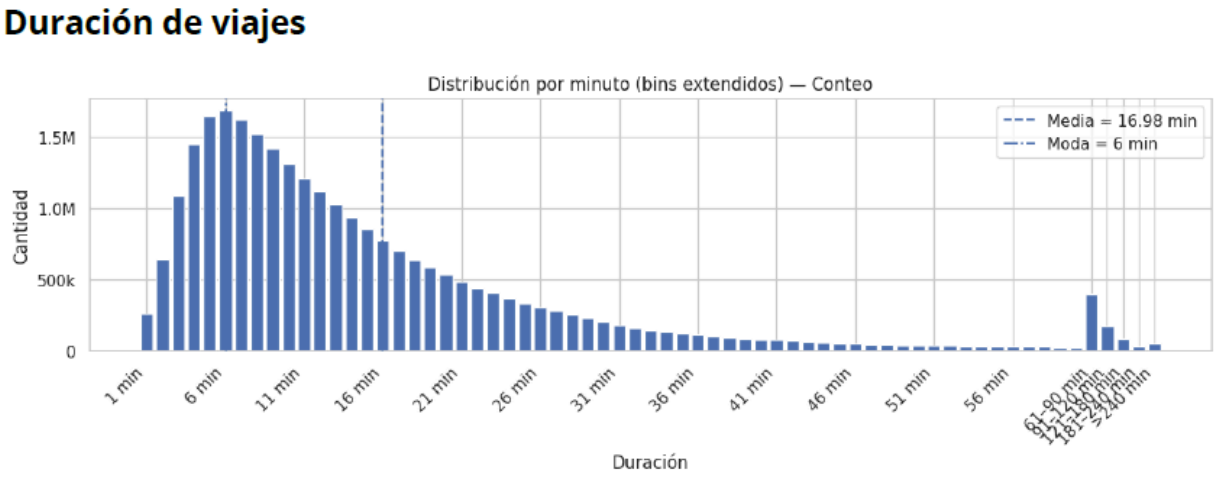
	bin_label	count	pct
0	0–30 min	24072068	89.14
1	30–45 min	1669216	6.18
2	45–60 min	490552	1.82
3	1–2 h	583348	2.16
4	2–6 h	135336	0.50
5	6–12 h	16728	0.06
6	12–24 h	16875	0.06
7	>24 h	19299	0.07



Se ajustaron los bins mayores a una hora



3.4. Visualizaciones exploratorias



Se elaboraron histogramas y gráficos de barras para representar la distribución de las duraciones. Estas visualizaciones permitieron identificar:

- La alta concentración en viajes cortos (uso laboral/diario).

- La menor, pero existente, proporción de viajes largos, más vinculados a recreación o turismo.
- La presencia de valores atípicos en viajes superiores a 24 horas, que se consideran anomalías o posibles errores de registro.

3.5. Análisis preliminar de usuario

- Se comenzó a contrastar el comportamiento de miembros (usuarios regulares) frente a ocasionales (casual riders).
- Los miembros tienden a ajustarse más a viajes cortos (<45 minutos), optimizando su uso para evitar costos adicionales.
- Los usuarios ocasionales muestran mayor dispersión en la duración, con más probabilidad de viajes largos.

3.6. Conclusiones EDA

El análisis exploratorio permitió confirmar que el dataset de Bluebikes contiene patrones claros en las duraciones de viaje y diferencias en el comportamiento de usuarios. Estas observaciones iniciales orientan las fases posteriores de modelado, en las que se podrán aplicar técnicas de clasificación (miembro vs. ocasional), clustering (agrupación de patrones de viajes) y predicción de demanda por estación y periodo.

4. Modelado

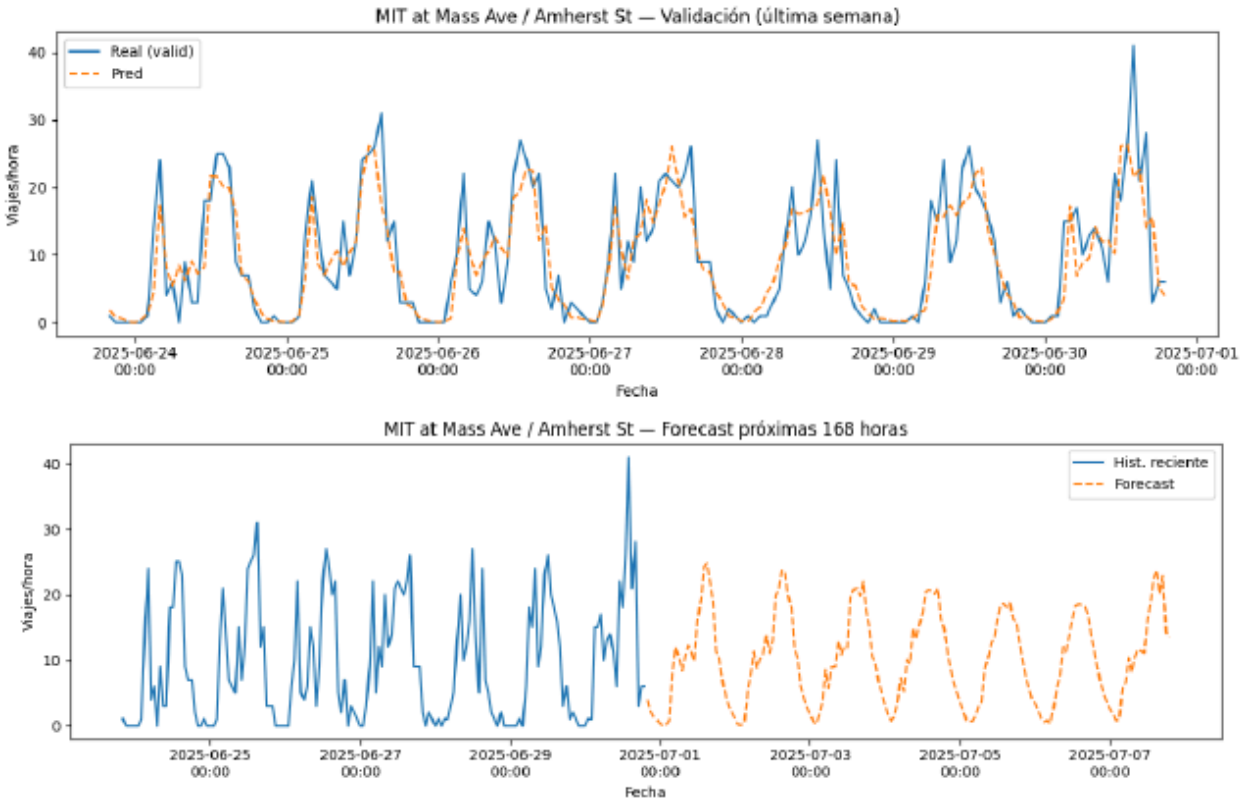
El modelado se planteó a nivel de serie temporal por estación y hora. Para ello, se construyó un dataset agregado (df_hourly_station) con la métrica objetivo rides (número de viajes/hora por estación) y un conjunto de atributos explicativos derivados de ingeniería de características.

4.1. Estrategia de entrenamiento

Se aplicó una estrategia de series temporales con rezagos, promedios móviles y componentes estacionales, buscando capturar tanto la dinámica horaria como los ciclos semanales:

- Lags: 1h, 24h, 168h (día y semana).
- Promedios móviles (shift(1)): ventanas de 3, 6, 12, 24 y 168 h.
- Componentes de Fourier: diarios (24h, K=3) y semanales (168h, K=2).
- Algoritmo: XGBoost Regressor con árboles tipo histogram, parámetros ajustados en:
    - n_estimators: 500–800
    - max_depth: 6
    - learning_rate: 0,05
    - subsample: 0,8
    - colsample_bytree: 0,8
    - random_state: 42
- Validación: se usó la última semana como hold-out, comparando contra un baseline naive estacional (lag 168h).

4.2. Resultados



En la estación de mayor demanda (MIT at Mass Ave / Amherst St), los resultados fueron:

Modelo	MAE	RMSE	wMAPE	NRMSE
XGBoost	3,332	4,779	35,08%	0,503
Naive	4,083	6,056	42,98%	0,638

- **Mejora XGBoost vs Naive:**
  - **MAE: −18,4% (de 4,083 a 3,332)**
  - **RMSE: −21,1%**
  - **wMAPE: −7,9 pp**

En los gráficos de validación y forecast se observa cómo el modelo captura de forma consistente los picos diarios y las variaciones horarias, mostrando skill superior al baseline ingenuo.

Con una ingeniería de series ligera (lags, rolling y Fourier) y la incorporación de atributos operativos, el modelo supera de forma clara al baseline estacional, logrando reducciones significativas en error. Esto demuestra que es posible contar con un forecast confiable y útil para la planificación horaria por estación, apoyando tareas como rebalanceo de bicicletas y asignación de recursos.

5. Evaluación

5.1. Ajusta a Objetivos de Negocio

El tratamiento y modelado realizado permitieron alcanzar una base de datos consistente, auditable y minable, capaz de soportar tanto la construcción de tableros interactivos como la aplicación de modelos predictivos. El pipeline garantiza trazabilidad en cada paso (integración, limpieza, enriquecimiento y persistencia), lo que fortalece la confianza en el uso analítico de la información.

El modelo de pronóstico horario, basado en XGBoost para series temporales, demostró una reducción significativa del error frente al baseline ingenuo estacional. Esta mejora es especialmente relevante desde la perspectiva de negocio, pues habilita:

- Acciones de re-balanceo más informadas, al anticipar con mayor precisión la demanda por estación y hora.
- Planificación de recursos (staffing), optimizando la asignación de personal en momentos de alta utilización del sistema.
- Soporte a la toma de decisiones estratégicas, al proveer información más confiable para evaluar la eficiencia y sostenibilidad del servicio.

En síntesis, el proyecto conecta directamente con los objetivos operativos y estratégicos de un sistema de bicicletas compartidas, generando valor accionable a partir de datos históricos complejos y heterogéneos.

5.2. Riesgos y limitaciones

A pesar de los resultados positivos, el proyecto enfrentó desafíos inherentes a la naturaleza de los datos y a las decisiones metodológicas adoptadas:

Cambios de esquema en el tiempo

- La evolución del sistema Bluebikes/Hubway entre 2015 y 2025 generó tres versiones de esquema distintas.
- Esto obligó a realizar supuestos de imputación y estandarización, como:
    - Etiquetar valores nulos de rideable_type como docked_bike.
    - Clasificar viajes sin estación en años recientes como “dockless”.
- Aunque estas decisiones fueron consistentes con criterios de negocio, introducen un grado de incertidumbre en las métricas.

Exclusión de factores externos

- El modelo de demanda no incorporó variables como clima, condiciones del tráfico o eventos masivos, que son conocidos predictores de la movilidad urbana.
- Esto limita la capacidad del modelo para explicar ciertas variaciones abruptas, reduciendo su poder predictivo en escenarios atípicos.

Validaciones geográficas

- Se aplicó una verificación geoespacial mediante BBOX+buffer, lo que resultó eficaz para descartar viajes fuera del área operativa.

- Sin embargo, esta técnica es una aproximación rectangular y no sustituye el uso de máscaras poligonales exactas de cobertura (p. ej., shapefiles de jurisdicciones urbanas), lo que podría generar falsos positivos o negativos en zonas limítrofes.

5.3. Síntesis

El proyecto logró construir una base sólida y un modelo predictivo alineado con objetivos de negocio, pero los riesgos señalados marcan el camino para mejoras futuras: integración de fuentes externas (clima/eventos), refuerzo de controles geoespaciales y refinamiento de imputaciones históricas. Con estas mejoras, el sistema podría evolucionar hacia un motor de analítica y planificación más robusto y adaptable para la movilidad urbana.

6. Despliegue y próximos pasos

La implementación de prácticas de DataOps y MLOps, combinada con tableros de Business Intelligence, permitiría evolucionar de un análisis académico hacia una plataforma operativa y estratégica de gestión de movilidad. Esto habilitaría a Bluebikes (o sistemas similares) a tomar decisiones basadas en evidencia y en tiempo casi real, incrementando eficiencia operativa, satisfacción de usuarios y sostenibilidad económica.

6.1. Data Ops - Gestión de Gobierno de Datos

Ingesta mensual automatizada

- Implementar un pipeline que descargue e integre de manera automática los archivos mensuales publicados por Bluebikes.
- Normalizar esquemas heterogéneos (tres versiones históricas) y almacenar en formato Parquet particionado por year/month, lo que facilita el acceso incremental y consultas eficientes en motores distribuidos.

Monitoreo de calidad de datos

- Establecer un tablero de métricas de calidad que reporte:
  - % de viajes con duración < 60 segundos (errores o usos anómalos).
  - % de coordenadas nulas o inválidas.
  - % de viajes fuera del área operativa.
  - Conteo de viajes por tipo de usuario y tipo de bicicleta.
- Este control permitiría detectar anomalías en la ingesta y asegurar la coherencia de la base para análisis futuros.

Catálogo de datos

- Desarrollar un Data Dictionary con descripciones de campos, tipos, reglas de negocio y checks reproducibles.
- Esto garantizaría que equipos técnicos y no técnicos comprendan la semántica de los datos y puedan reutilizarlos con confianza.

6.2. MLOps - Ciclo de Vida del Modelo

Estrategias de entrenamiento avanzadas

- Entrenar modelos por clusters de estaciones (agrupadas por patrones de uso y proximidad) o incluso por toda la red con embeddings de estación, permitiendo capturar relaciones espaciales.

Incorporación de variables externas

- Integrar factores como clima (temperatura, precipitación, nevadas), calendario académico, eventos masivos (deportivos, culturales) y capacidad de docks por estación.
- Esto enriquecería el contexto de predicción, aumentando precisión en días atípicos.

Re-entrenos periódicos y backtesting

- Implementar un esquema de ventanas rodantes para backtesting, simulando cómo funcionaría el modelo en producción.
- Definir ciclos de re-entrenamiento automático (ej. mensual o trimestral) para asegurar que los modelos se adapten a cambios de comportamiento en la demanda.

6.3. BI - Uso Operativo y Estratégico

Tablero de demanda

- Desarrollar dashboards interactivos que muestren la demanda por hora y por estación, junto con KPIs clave como:
  - Puntualidad en operaciones de re-balanceo.
    Tasa de viajes *dockless*.
  - Tiempos promedio de viaje y desviaciones.

Alertas operativas

- Configurar sistemas de alerta cuando la predicción de viajes supere umbrales de capacidad de docks o de bicicletas disponibles.
- Esto permitiría disparar acciones proactivas de re-balanceo para evitar desabastecimientos en estaciones críticas.

Análisis estratégico

- Usar los datos integrados para evaluar la eficiencia de nuevas estaciones, justificar expansiones en la red y medir el impacto del sistema en la movilidad urbana sostenible.

7. Conclusiones

El proyecto final de *Data Mining* permitió recorrer de manera integral el ciclo de descubrimiento de conocimiento en bases de datos (KDD), desde la integración multi-esquema y limpieza de más de 27 millones de registros históricos de Bluebikes, hasta la construcción de modelos de predicción de demanda horaria por estación.

Los resultados principales pueden resumirse en tres dimensiones:

- Calidad de datos: Se transformó una base transaccional heterogénea en un dataset minable, consistente y auditable, con atributos temporales, espaciales y de comportamiento, listo para análisis avanzados.
- Exploración: El EDA confirmó que la mayoría de los viajes son cortos (<45 min) y reveló diferencias claras entre usuarios regulares (uso laboral) y ocasionales (uso recreativo), además de la relevancia de estaciones altamente populares.
- Modelado: El modelo de series temporales con XGBoost mejoró entre 18–21% frente al baseline ingenuo, aportando skill predictivo útil para la planificación horaria y el re-balanceo de bicicletas.

Al mismo tiempo, el trabajo evidenció riesgos y limitaciones, como la necesidad de integrar factores externos (clima, eventos, capacidad de docks) y mejorar la precisión geoespacial con polígonos de cobertura. Estas oportunidades abren el camino para fortalecer el sistema hacia un enfoque más robusto de analítica urbana.

Finalmente, se recomienda avanzar hacia un marco de DataOps, MLOps y BI/Decisioning, con ingesta automatizada, monitoreo de calidad, re-entrenos periódicos y tableros operativos con alertas. Esto permitiría evolucionar de un análisis académico a una plataforma estratégica de soporte a decisiones en movilidad sostenible, incrementando la eficiencia del sistema y aportando valor social y económico a la ciudad.


SCHOOL OF ENGINEERING