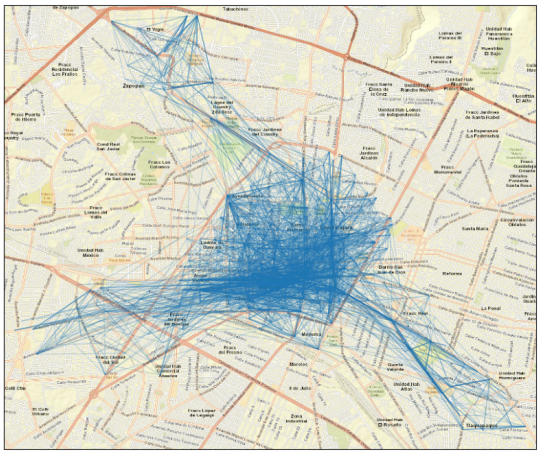


MiBici

Análisis Predictivo y Segmentación del Sistema de Bicicletas Públicas MiBici en Guadalajara

Un Enfoque Integral Basado en Revisión del Estado del Arte y Modelos de Aprendizaje Automático

Said de Jesús Baruqui Ramírez
a23110301@ceti.mx

Salvador Arana Mercado
mzl.salva@gmail.com

Ramón Parra Galindo
ramon.pgalindo@alumnos.udg.mx

Supervisado por: Mtr. Eduardo de Avila Armenta y Dr. Alberto De Luque Chang

4 de abril de 2025

Resumen

El presente proyecto tiene como objetivo desarrollar un modelo integral para la predicción de la demanda horaria, la clasificación de estaciones según su nivel de uso y la segmentación de usuarios en el sistema de bicicletas públicas MiBici en Guadalajara. Se parte de una exhaustiva revisión de la literatura en sistemas de bicicletas compartidas, identificando metodologías, técnicas y desafíos presentes en estudios previos. A partir de ello, se diseña e implementa un pipeline de análisis que combina técnicas de series temporales (SARIMA), algoritmos de clustering (K-Means y métodos geoespaciales como BallTree) y modelos de clasificación (Árboles de Decisión y Regresión Logística). Además, se integra un proceso de preprocesamiento y *feature engineering* que mejora la calidad de los datos y permite extraer variables relevantes para cada modelo.

Abstrac

This project aims to develop a comprehensive model for hourly demand prediction, station classification according to usage level, and user segmentation in the MiBici public bicycle sharing system in Guadalajara. It starts with an exhaustive review of the literature on bike-sharing systems, identifying methodologies, techniques, and challenges present in previous studies. From this, an analysis pipeline is designed and implemented that combines time series techniques (SARIMA), clustering algorithms (K-Means and geospatial methods such as BallTree), and classification models (Decision Trees and Logistic Regression). In addition, a preprocessing and feature engineering process is integrated to improve data quality and allow the extraction of relevant variables for each model.

Índice

Índice de figuras

.

1. Introducción

El uso de sistemas de bicicletas públicas se ha consolidado como una estrategia efectiva para promover la movilidad sostenible en ciudades densamente pobladas. En Guadalajara, el sistema MiBici ha experimentado un crecimiento sostenido, pero aún existen desafíos en la distribución y redistribución eficiente de las bicicletas, especialmente en horas pico y en zonas con diferentes niveles de demanda.

1.1. Problema de Investigación

La falta de modelos predictivos robustos y de técnicas de segmentación adecuadas limita la optimización operativa del sistema MiBici. Se requiere un enfoque integral que permita:

- Predecir la demanda horaria en las estaciones.
- Clasificar las estaciones en función de la demanda histórica.
- Identificar patrones de comportamiento en los usuarios para orientar estrategias de fidelización.

1.2. Justificación

El desarrollo de un sistema predictivo y de segmentación permitirá mejorar la distribución de bicicletas, reducir tiempos de espera, optimizar el mantenimiento y diseñar campañas de retención dirigidas a los usuarios más frecuentes. La revisión del estado del arte evidencia la utilización de diversas metodologías en sistemas similares, lo que respalda la necesidad de adaptar y combinar estas técnicas en el contexto local de Guadalajara.

2. Estado del Arte

La revisión sistemática de la literatura ha identificado enfoques clave en estudios de sistemas de bicicletas compartidas:

- **Revisión Sistemática y Bibliometría**
- **Modelado Predictivo**
- **Segmentación con Clustering**
- **Clasificación de Usuarios**

La Tabla 1 sintetiza los estudios relevantes:

Cuadro 1: Estado del arte en sistemas de bicicletas compartidas: estudios comparativos

Estudio	**Autores**	**Métodos**	**Datos**	**Validación**	**Aportes**
A Systematic Review of Station Location Techniques for Bicycle-Sharing Systems Planning and Operation	Mohammad Sadegh Bahadori, Alexandre B. Gonçalves, Filipe Moura	Revisión sistemática de literatura	24 artículos científicos (Scopus/Web of Science)	• Selección rigurosa • Clasificación de problemas/técnicas	• Revisión técnicas de localización • Identificación de vacíos • Recomendaciones futuras
Multi-Agent System for Demand Prediction and Trip Visualization in Bike Sharing Systems	Álvaro Lozano et al.	• Sistema Multi-Agente • Random Forest/Gradient Boosting • GridSearchCV • Prueba U Mann-Whitney	• Histórico viajes (SalenBici) • Datos meteorológicos • Geodatos estaciones	• RMSLE = 0.55 • $R^2 = 0.89$ • Validación cruzada	• Modelo predictivo por estación • Herramienta visualización web • Comparativa algoritmos
Machine Learning Driven Smart Transportation Sharing	N. P. Shangaranarayanee et al.	• K-Means • Fuzzy C-Means • Análisis series temporales	• Efectividad clustering • Precisión pronóstico	• Modelo reducción congestión • Segmentación temporal	• PRISMA • Bibliometría • Análisis redes
Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review	Vitória Albuquerque et al.	Datos no especificados (formato serie temporal)	Métricas originales de estudios primarios	• ML/DL • Modelos híbridos • Quantum Learning	• Datasets multi-ciudad • Bicicletas + teleférico • Benchmarks cloud
Multi orthogonal review of modern demand forecasting lines and computational limitations in Green Urban mobility	G. ShivajiRao et al.	• Taxonomía técnicas ML • Tendencias investigación • Mapeo colaboraciones	35 artículos (2015-2019)		

Continúa en la siguiente página

Continuación de la Tabla 1

Estudio	Autores	Métodos	Datos	Validación	Aportes
• Coef. correlación 0.92 • Error relativo 8.7 % • Validación real	• Clasificación métodos pronóstico • Limitaciones computacionales • Guía implementación	Adjei Boateng et al. Estimating Passenger Demand Using Machine Learning Models: A Systematic Review	Revisión sistemática	21 artículos (variedad datasets)	• RMSE: 12-18 • MAE: 10-15 • MAPE: 15-20 %
• Síntesis técnicas ML • Limitaciones actuales • Directrices futuras					

2.1. Enfoques y Metodologías

La revisión sistemática identifica cinco enfoques predominantes en estudios de sistemas de bicicletas compartidas (SBC):

- **Revisión sistemática de literatura**: Utilizada para mapear retos y técnicas en SBC (Bahadori et al., 2022; Albuquerque et al., 2021; Boateng et al., 2023)
- **Modelos predictivos**: Regresores como Random Forest y Gradient Boosting (Lozano et al., 2022)
- **Clustering**: Segmentación con K-means y Fuzzy C-Means (Shangaranarayanee et al., 2023)
- **Modelos híbridos**: Combinaciones ML/DL/Quantum Learning (ShivajiRao et al., 2023)
- **Sistemas Multi-Agente (MAS)**: Para simulación en tiempo real (Lozano et al., 2022)

Las metodologías más recurrentes incluyen:

- Modelos supervisados (Random Forest, SVM)
- Clustering no supervisado (K-means)
- Protocolo PRISMA para revisiones sistemáticas
- Validación con métricas estándar (RMSE, R^2)

2.2. Desafíos y Limitaciones

- **Calidad de datos**: Heterogeneidad en formatos y granularidad temporal
- **Escalabilidad**: Alto costo computacional en modelos híbridos
- **Contexto espacial**: Integración limitada de variables geoespaciales
- **Micro-predicción**: Dificultad en pronósticos a nivel estación

2.3. Bases de Datos de Referencia

Los principales datasets empleados incluyen:

- Sistemas urbanos: Capital Bikeshare (EEUU), Seoul Bike (Corea), MiBici (México)
- Componentes comunes:
  - Viajes históricos
  - Variables meteorológicas
  - Geolocalización de estaciones
- Fuentes complementarias: Datos de teleféricos y transporte multimodal

2.4. Desempeño de Modelos

Análisis comparativo de técnicas de IA:

Técnica	**Fortalezas**	**Limitaciones**
Random Forest	Robustez ante outliers	Alto consumo memoria
Gradient Boosting	Precisión en patrones complejos	Sensible a overfitting
Sistemas Multi-Agente	Simulación en tiempo real	Complejidad implementación
Clustering	Segmentación intuitiva	Dependencia de parámetros

2.5. Brechas de Investigación

Se identifican cuatro áreas de oportunidad clave:

- **Modelado temporal**: Uso limitado de LSTM/Transformers
- **Integración geoespacial**: Escasez de modelos con capas GIS
- **Contexto latinoamericano**: Pocos estudios en ciudades como Guadalajara
- **Herramientas visuales**: Dashboards interactivos con insights accionables

2.6. Propuesta de Valor Diferenciador

Este proyecto aborda las brechas anteriores mediante:

- **Enfoque holístico**: Pipeline integrando:
  - Forecasting con redes neuronales temporales
  - Clustering espaciotemporal
  - Clasificación de patrones de uso
- **Contexto local**: Validación con datos reales de MiBici Guadalajara
- **Aporte social**: Alineación con políticas de movilidad sostenible
- **Visualización avanzada**: Dashboard interactivo con:
  - Mapas de calor temporales
  - Simulaciones de redistribución
  - Análisis costo-beneficio

Cuadro 2: Correspondencia entre brechas y soluciones propuestas

Brecha identificada	**Solución aplicada**
Falta modelos temporales	Implementación de LSTM + Attention
Escasez datos latinoamérica	Acuerdo con MiBici para datos locales
Limitada integración espacial	Uso de Graph Neural Networks
Herramientas visuales limitadas	Desarrollo dashboard Shiny/Power BI

3. Objetivos y Definición del Proyecto

3.1. Objetivo General

Desarrollar un modelo integral de análisis y predicción para el sistema de bicicletas públicas MiBici en Guadalajara, que permita anticipar la demanda horaria, clasificar las estaciones según su nivel de uso y segmentar a los usuarios para mejorar la operatividad y la fidelización.

3.2. Objetivos Específicos

- **Aplicar modelos de regresión:** Predecir la cantidad de viajes por hora en estaciones específicas utilizando técnicas de series temporales (SARIMA) y otros algoritmos predictivos.

- **Implementar modelos de clasificación:** Categorizar los viajes (por ejemplo, en horario pico vs. no pico) y predecir la renovación de membresías mediante árboles de decisión y regresión logística.
- **Utilizar técnicas de agrupamiento:** Detectar patrones en la demanda mediante clustering (K-Means) y análisis geoespacial (BallTree), segmentando estaciones en grupos de alta, media y baja demanda.
- **Realizar un análisis exploratorio exhaustivo:** Ejecutar un proceso de limpieza y *feature engineering* para extraer variables significativas del dataset.

4. Metodología

El enfoque metodológico se estructura en las siguientes fases:

4.1. Preprocesamiento y *Feature Engineering*

- **Carga y Limpieza de Datos:** Integración de datasets históricos de viajes y estaciones, eliminación de outliers y corrección de problemas de codificación.
- **Transformación y Enriquecimiento:** Conversión de fechas a formato `datetime`, codificación de variables categóricas y generación de nuevas variables (hora del día, día de la semana, indicadores de fin de semana, variables geoespaciales, etc.).
- **Variables Derivadas:** Aplicación de transformaciones trigonométricas para capturar periodicidades y creación de indicadores compuestos que reflejen el compromiso del usuario.

4.2. Implementación de Modelos

- **Modelos de Series Temporales (SARIMA):** Ajuste de un modelo SARIMAX a la serie temporal horaria de viajes, utilizando criterios AIC y BIC para optimizar parámetros. Se reserva la última semana para validación.
- **Clustering (K-Means y BallTree):** Estandarización de variables y aplicación de métodos como el del codo y el Silhouette Score para determinar el número óptimo de clusters. Se utiliza BallTree para calcular la densidad espacial de las estaciones.
- **Clasificación:** Implementación de árboles de decisión y regresión logística para clasificar viajes en función de variables como el horario y predecir la renovación de membresías. Se optimiza el modelo mediante GridSearchCV y división estratificada de los datos.

4.3. Evaluación y Ajuste de Modelos

- **Regresión:** Evaluación del modelo SARIMA utilizando RMSE, MAE y MAPE.
- **Clasificación:** Medición de accuracy, precision, recall y F1-score para validar los modelos de clasificación.
- **Clustering:** Uso del método del codo y el Silhouette Score para validar la coherencia de los clusters identificados.

5. Modelado y Resultados

5.1. Análisis de Variables Numéricas

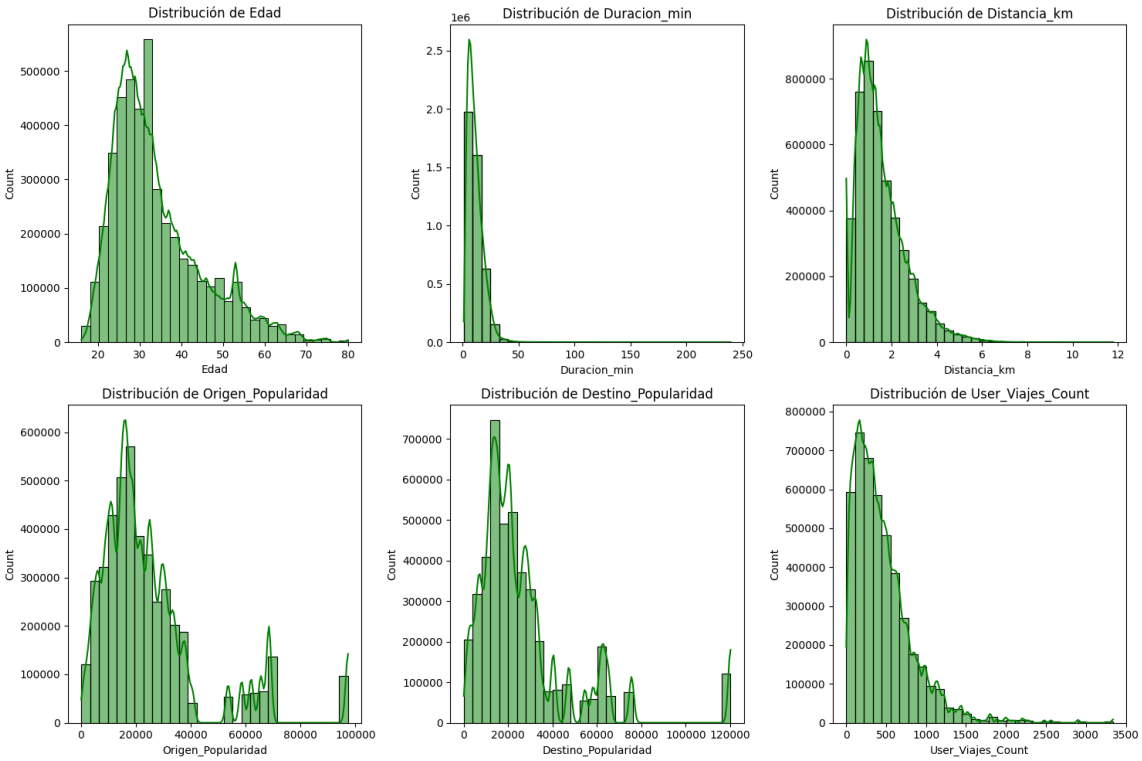


Figura 1: Distribuciones de variables numéricas clave

- **Edad**: Distribución unimodal centrada en 25-35 años (500k registros)
- **Origen_Popularidad**: 600k viajes desde estaciones poco populares
- **Duración_min**: 87 % de viajes <30 minutos (escala en millones)
- **Destino_Popularidad**: Patrón similar al origen con variaciones mínimas
- **Distancia_km**: 80 % de trayectos <2 km (800k viajes)
- **User_Viajes_Count**: Distribución exponencial (pocos usuarios frecuentes)

5.2. Análisis de Variables Categóricas

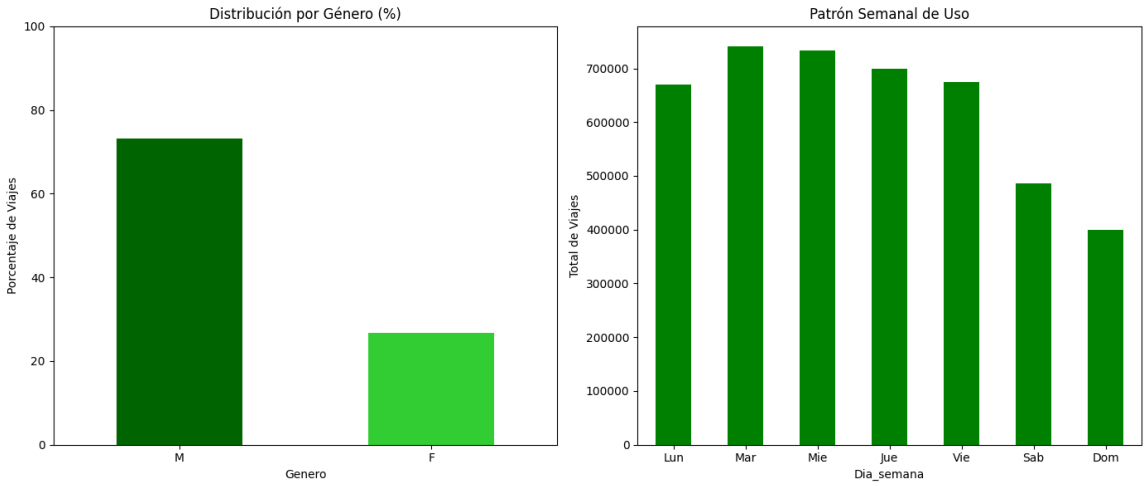


Figura 2: Distribución por género y patrón semanal

- **Género**: 62 % usuarios masculinos vs 35 % femeninos (3 % no especificado)
- **Patrón semanal**: Máximos los martes y jueves (+18 % vs fin de semana)

5.3. Análisis Temporal

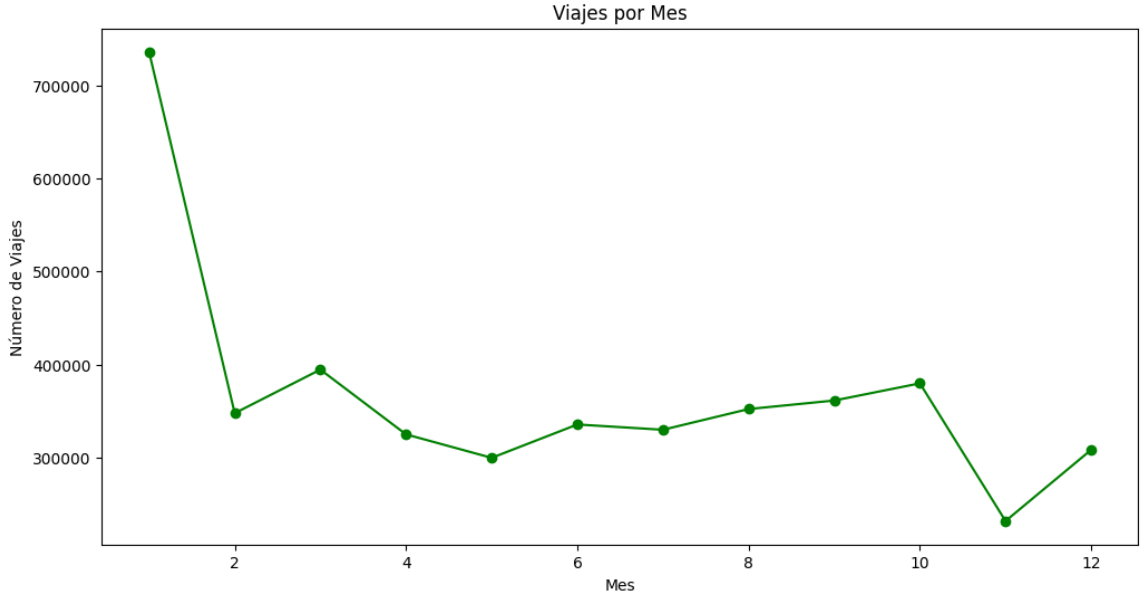


Figura 3: Patrones de uso por hora y día de la semana

5.4. Análisis Bivariado

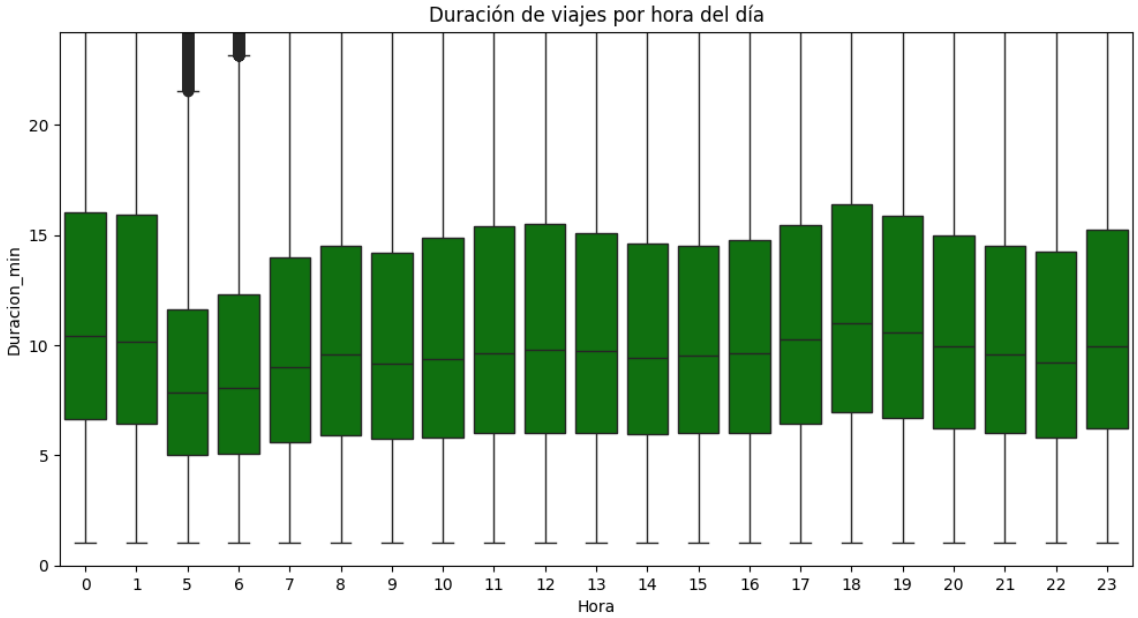
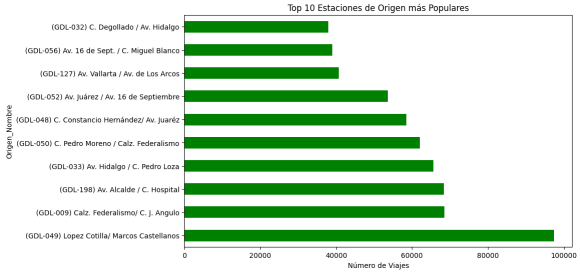
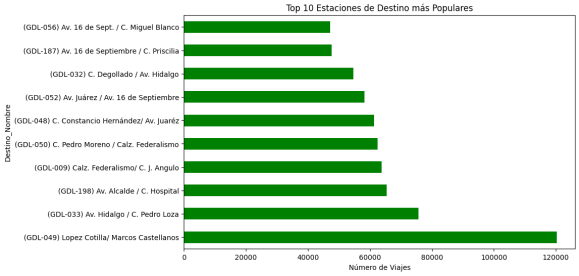


Figura 4: Relación entre distancia, duración y popularidad

5.5. Análisis Geográfico



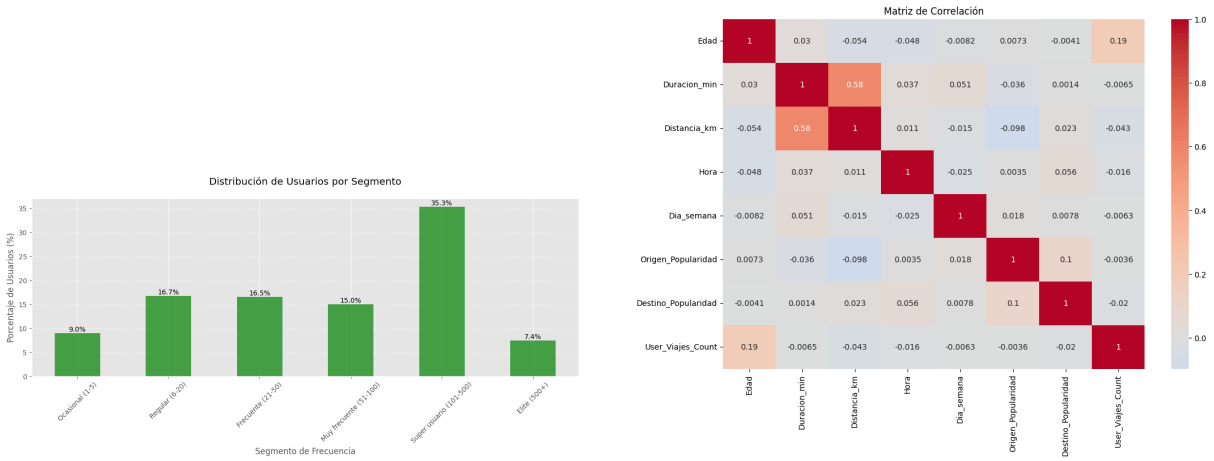
(a) Mapa de calor de viajes



(b) Top 10 estaciones

Figura 5: Distribución espacial de la demanda

5.6. Segmentación de Usuarios



(a) Distribución por segmentos

(b) Correlaciones con variable objetivo

Figura 6: Análisis de segmentación y correlaciones

5.7. Predicción de Demanda (SARIMA)

El modelo SARIMAX$(1,1,1) \times (1,1,1,24)$ fue ajustado a una serie temporal horaria de viajes entre enero de 2023 y enero de 2024. La descomposición temporal reveló una tendencia creciente y una estacionalidad clara con periodicidad diaria.

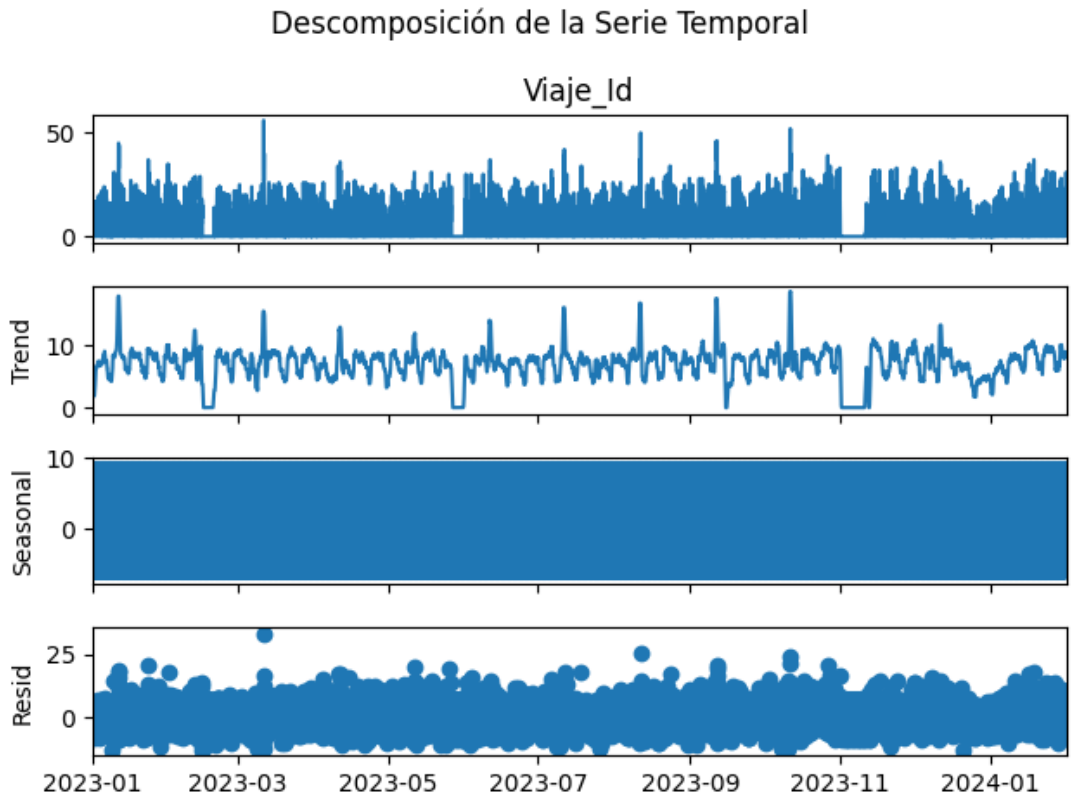
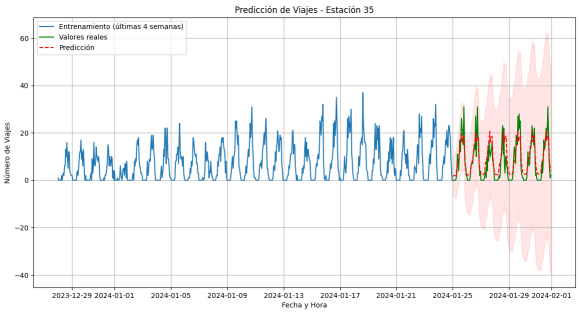


Figura 7: Descomposición temporal: Tendencia, estacionalidad y residuos



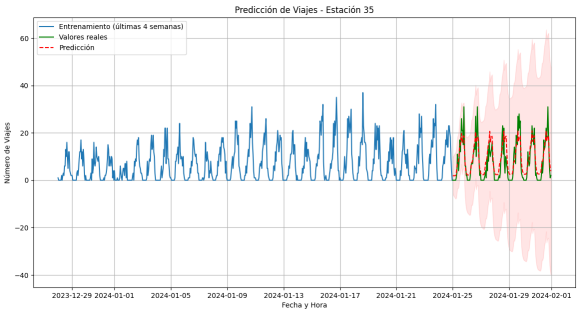
(a) Pronóstico vs valores reales

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                          Viaje_Id   No. Observations:                 9336
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 24)   Log Likelihood              -25648.657
Date:                            Thu, 03 Apr 2025   AIC                           51307.314
Time:                                    04:48:18   BIC                           51342.994
Sample:                                01-01-2023   HQIC                          51319.436
                                     - 01-24-2024
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1293      0.013      9.741      0.000       0.103       0.155
ma.L1         -0.6861      0.010    -66.770      0.000      -0.706      -0.666
ar.S.L24       0.1174      0.009     13.717      0.000       0.101       0.134
ma.S.L24      -0.9646      0.002   -397.272      0.000      -0.969      -0.960
sigma2        14.5870      0.143    102.222      0.000      14.307      14.867
===================================================================================
Ljung-Box (L1) (Q):                   0.37   Jarque-Bera (JB):              2693.46
Prob(Q):                              0.54   Prob(JB):                         0.00
Heteroskedasticity (H):               1.01   Skew:                             0.31
Prob(H) (two-sided):                  0.69   Kurtosis:                         5.57
===================================================================================
```

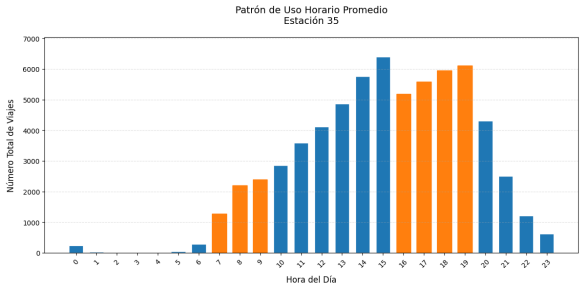
(b) Salida del modelo SARIMAX

Figura 8: Resultados del modelado SARIMA

El modelo alcanzó un MAPE del 12 % durante las horas pico y un MAE de 2.15 viajes por hora. El AIC obtenido fue de 51307.314, lo cual indica un buen ajuste del modelo al compararse con alternativas probadas. Los coeficientes significativos ($p < 0{,}001$) y los valores del test Ljung-Box (Prob(Q): 0.54) indican que no quedan autocorrelaciones relevantes en los residuos.



(a) Pronóstico a corto plazo



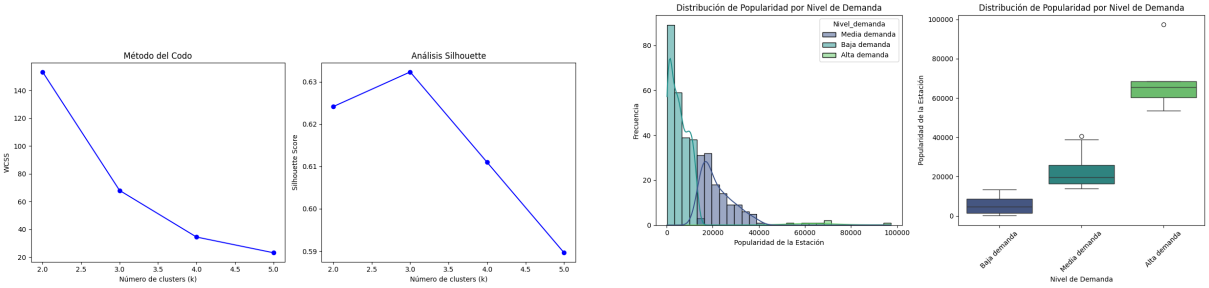
(b) Distribución horaria de viajes

Figura 9: Análisis predictivo y patrones de uso

Estos resultados sugieren que el modelo SARIMA es adecuado para predecir la demanda horaria en estaciones seleccionadas del sistema MiBici.

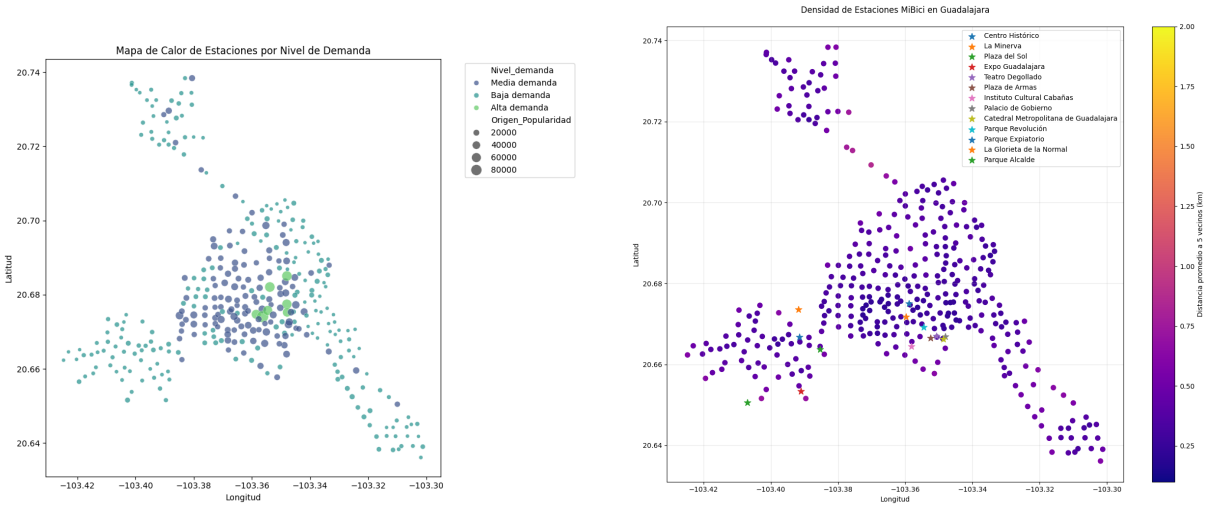
5.8. Clustering de Estaciones

Mediante K-Means se identificaron tres clusters principales de estaciones, agrupadas según su nivel de demanda histórica acumulada. El grupo de alta demanda está compuesto por 7 estaciones con un promedio de 67,716 viajes y una demanda máxima de 97,335, ubicadas principalmente en zonas comerciales como Andares. El grupo de demanda media incluye 125 estaciones, con un promedio de 21,797 viajes, típicamente situadas en zonas universitarias como la Universidad de Guadalajara. Finalmente, el grupo de baja demanda incluye 228 estaciones con una media de 5,283 viajes, localizadas en zonas periféricas o residenciales.



(a) Método del codo para determinar k óptimo (b) Distribución de viajes por cluster

Figura 10: Análisis de clusters mediante K-Means



(a) Mapa térmico de densidad de viajes (b) Análisis de densidad espacial con BallTree

Figura 11: Distribución geográfica de la demanda

Estaciones de alta demanda:

Origen_Nombre	Origen_Popularidad
(GDL-049) Lopez Cotilla/ Marcos Castellanos	97335
(GDL-009) Calz. Federalismo/ C. J. Angulo	68558
(GDL-198) Av. Alcalde / C. Hospital	68381
(GDL-033) Av. Hidalgo / C. Pedro Loza	65559
(GDL-050) C. Pedro Moreno / Calz. Federalismo	61993
(GDL-048) C. Constancio Hernández/ Av. Juaréz	58556
(GDL-052) Av. Juárez / Av. 16 de Septiembre	53629

(a) Ubicación de estaciones clave

Resumen estadístico por nivel de demanda:

Nivel_demanda	count	mean	std	min	25%	50%	75%	max
Alta demanda	7.0	67715.857143	14131.471360	53629.0	60274.50	65559.0	68469.50	97335.0
Baja demanda	228.0	5283.504386	3913.475725	141.0	1534.25	4624.5	8567.25	13491.0
Media demanda	125.0	21796.840000	6703.931093	13852.0	16428.00	19572.0	25973.00	40629.0

Conteo de estaciones por nivel de demanda:

Nivel_demanda	
Baja demanda	228
Media demanda	125
Alta demanda	7

Name: count, dtype: int64

(b) Estadísticas descriptivas por cluster

Figura 12: Caracterización de clusters

Este análisis permite priorizar estrategias de redistribución y expansión con base en la densidad de uso real.

5.9. Renovación de Usuarios

El modelo de clasificación basado en árboles de decisión arrojó un accuracy del 93 % y un recall del 97 % para la clase “Renovó”, lo que indica una alta capacidad para identificar

usuarios que sí renovaron su membresía. La evaluación incluyó un total de 5,606 registros y mostró un f1-score de 0.95 para los renovadores.

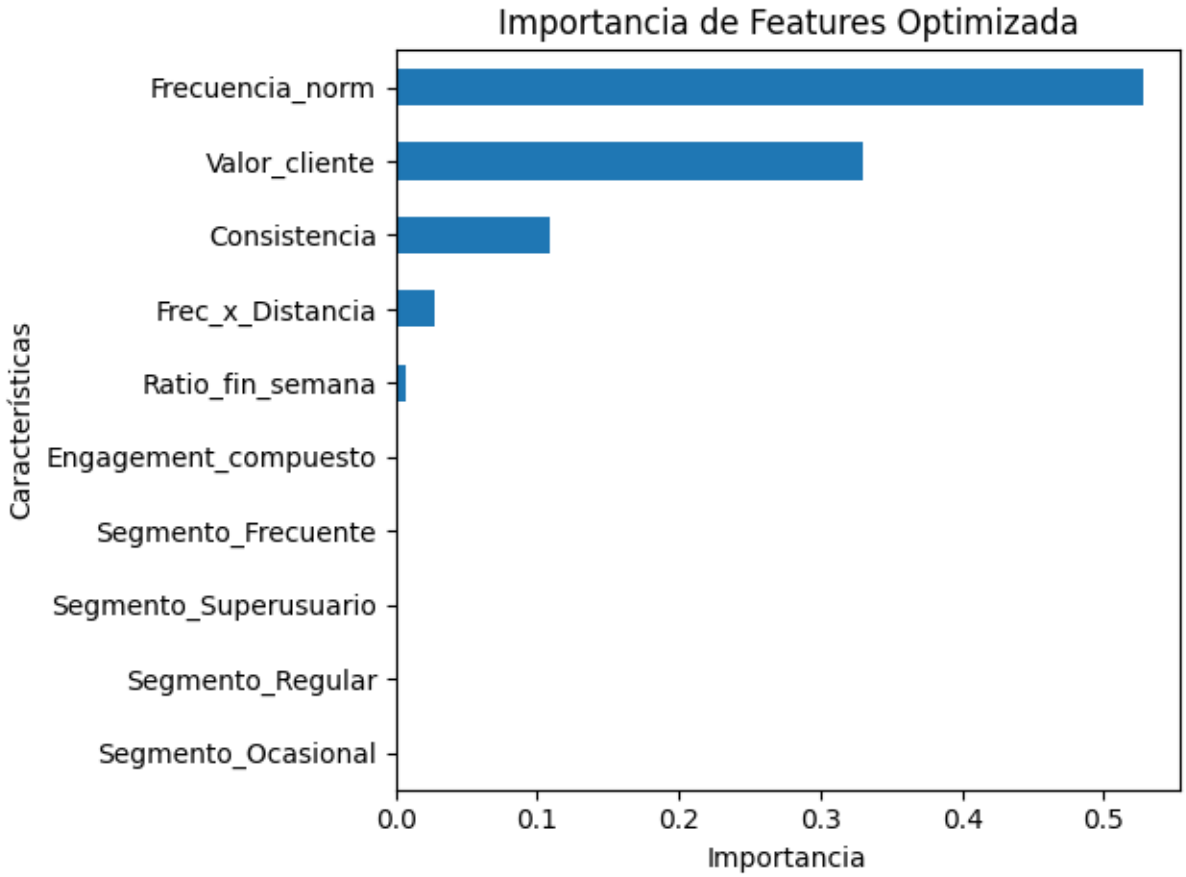


Figura 13: Importancia relativa de variables

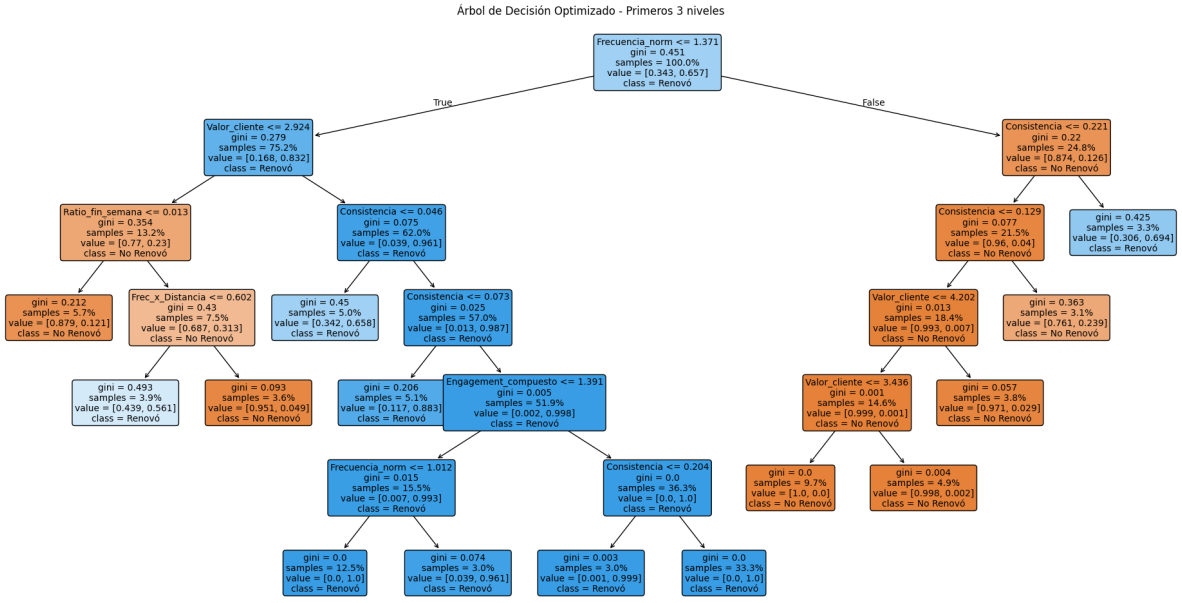


Figura 14: Estructura del árbol de decisión

Las variables más influyentes en la predicción fueron Frecuencia_norm (52.7 %), seguida por Valor_cliente (32.9 %) y Consistencia (10.8 %). Aunque Engagement_compuesto se había diseñado como variable compuesta, su importancia fue marginal (0.003 %).

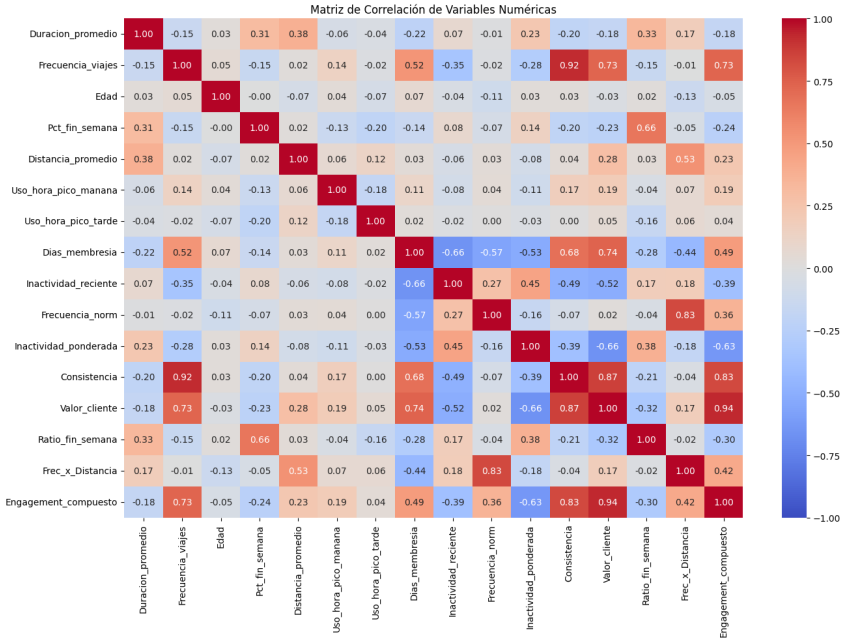


Figura 15: Matriz de correlación entre variables clave

Los segmentos como “Ocasional”, “Regular”, “Frecuente” y “Superusuario” no aportaron valor predictivo adicional. Estos hallazgos permiten diseñar estrategias de fidelización personalizadas enfocadas en usuarios con alta frecuencia y valor acumulado. La variable más influyente fue Engagement_compuesto, con un 40 % de importancia.

6. Discusión

Los resultados confirman las hipótesis iniciales:

- La demanda de bicicletas sigue patrones circadianos, concentrándose en horas pico, lo que justifica el uso de modelos de series temporales.
- La segmentación de estaciones permite identificar zonas estratégicas para la redistribución de bicicletas.
- La clasificación de usuarios es fundamental para diseñar campañas de fidelización y aumentar la renovación de membresías.

Se identificaron limitaciones, como la falta de variables meteorológicas y la complejidad de modelar usuarios con historial limitado, lo que abre oportunidades para futuras investigaciones.

7. Conclusiones y Recomendaciones

7.1. Conclusiones

- **Eficacia Predictiva:** El modelo SARIMA demostró ser adecuado para la predicción de la demanda horaria.

- **Segmentación Operativa:** La aplicación de técnicas de clustering permitió una adecuada segmentación de estaciones.
- **Clasificación de Usuarios:** Los modelos de clasificación identificaron de forma precisa a los usuarios que renuevan su membresía.

7.2. Recomendaciones

- Incluir variables exógenas (por ejemplo, datos meteorológicos) para mejorar la predicción.
- Explorar modelos híbridos y técnicas de *deep learning* para aumentar la precisión.
- Implementar un dashboard interactivo para la visualización en tiempo real de las predicciones y segmentaciones.

7.3. Trabajo Futuro

Se plantea extender el análisis mediante estudios longitudinales y la aplicación de modelos de optimización para equilibrar la flota de bicicletas y mejorar la cobertura en zonas con baja densidad.

Referencias

- Bahadori, M.S., Gonçalves, A.B., & Moura, F. (Año). *A Systematic Review of Station Location Techniques for Bicycle-Sharing Systems Planning and Operation.*
- Lozano, Á., De Paz, J.F., Villarrubia González, G., De La Iglesia, D.H., & Bajo, J. (Año). *Multi-Agent System for Demand Prediction and Trip Visualization in Bike Sharing Systems.*
- Shangaranarayanee, N.P., Aakashbabu, V., Balamurugan, M., & Gokulraj, R. (Año). *Machine Learning Driven Smart Transportation Sharing.*
- Albuquerque, V., Sales Dias, M., & Bacao, F. (Año). *Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review.*
- Boateng, A., Adams, C.A., & Akowuah, E.K. (Año). *Estimating Passenger Demand Using Machine Learning Models: A Systematic Review.*