SAMSUNG
2024
Innovation
Campus AI

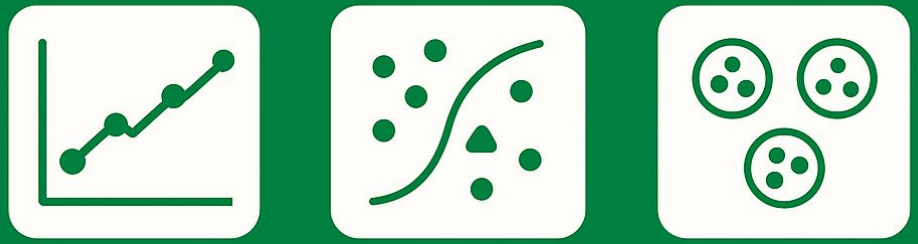
MiBici

BikeShareNetwork

Said Baruqui Ramirez

Salvador Arana Mercado

Ramón Parra Galindo

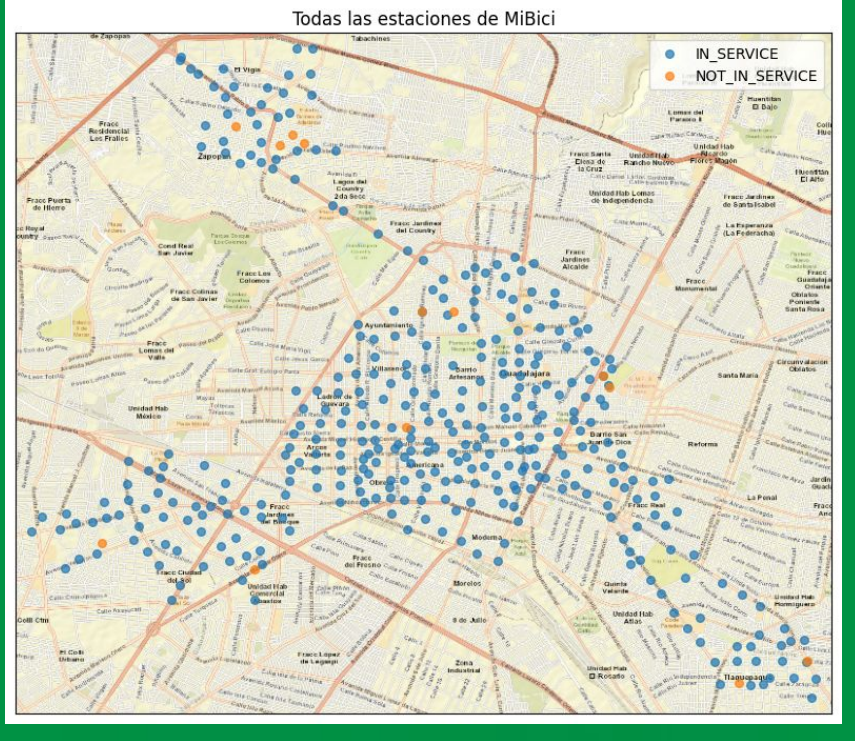


ML PROJECT

UNIVERSIDAD DE GUADALAJARA
PIENSA Y TRABAJA

UNIVERSIDAD DE MONTERREY
HOMO HOMINIS IN MINISTERIO PERFICITUR

¿Por qué?



- Crecimiento Rápido: A partir de su inauguración en 2014, actualmente MiBici cuenta con alrededor de 372 estaciones.
- Gracias al flujo asimétrico de bicicletas, la calidad del servicio depende directamente de predecir las acciones del usuario.
- Ciudades con Sistemas de Bicicletas: Ejemplos incluyen Nueva York (Citi Bike), Londres (Santander Cycle) y Ciudad de México (Ecobici).

Objetivo general

Utilizar modelos de Machine Learning para analizar el uso del sistema de bicicletas públicas MiBici en Guadalajara, para identificar y predecir patrones de uso, con el objetivo de encontrar algoritmos que permitan optimizar la movilidad.

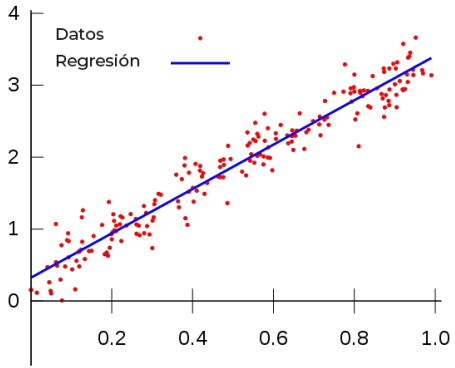
Objetivo específicos

- Aplicar modelos de regresión para predecir la demanda de bicicletas por estación.
- Usar clasificación para categorizar viajes y usuarios con base en su comportamiento.
- Implementar técnicas de agrupamiento para detectar estaciones con patrones similares y segmentar usuarios.

Que se espera de cada modelo

Regresión

- ¿Cuántos viajes se esperan mañana en una estación? (Red LSTM)
- ¿Cuántos viajes se iniciarán en una estación en un horario específico?



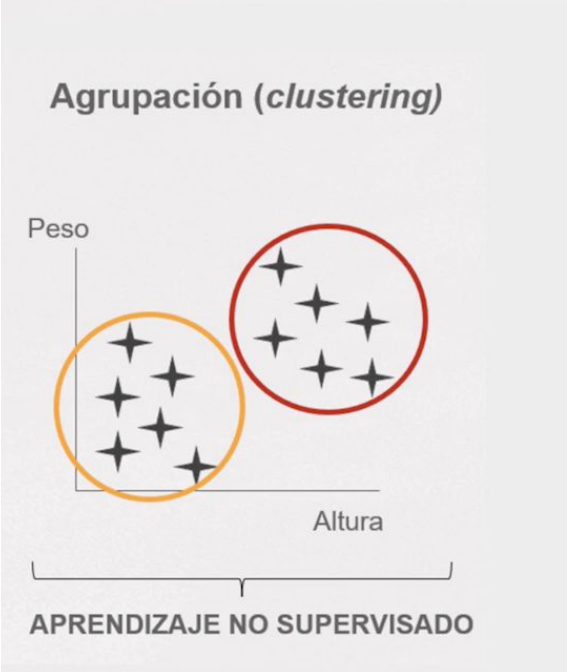
Métricas de evaluación para cada modelo

- SARIMA (pronóstico de demanda) *Error Medio Absoluto (MAE)* y *Error Cuadrático Medio (RMSE)*.

Que se espera de cada modelo

Agrupamiento

- Agrupar estaciones en 3 niveles de demanda (baja/media/alta) basado en su popularidad. (*K-Means)*
- Calcular densidad espacial de estaciones (*BallTree (Análisis de Densidad Geográfica)*)



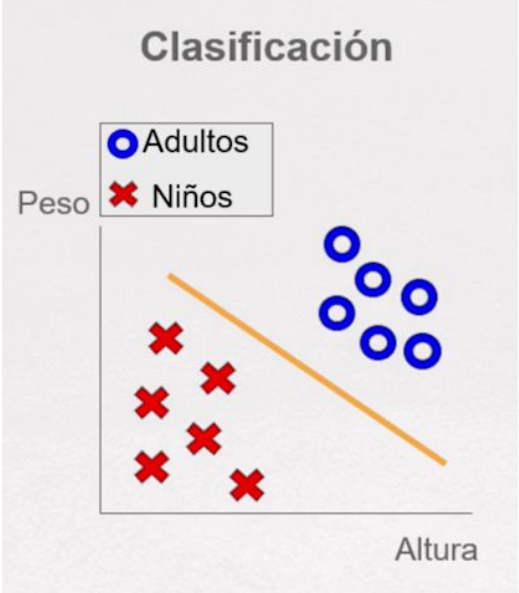
Métricas de evaluación para cada modelo

- Método del Codo, *Silhouette Score*.

Que se espera de cada modelo

Clasificación

- ¿Un usuario renovará su membresía o no? (*Arbol de desicion*)



Métricas de evaluación para cada modelo

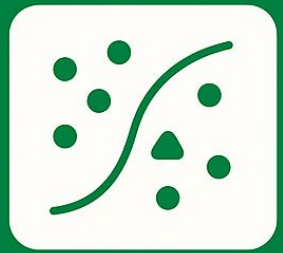
- *Precisión, Recall, F1-Score, Accuracy.*

Acerca del Dataset

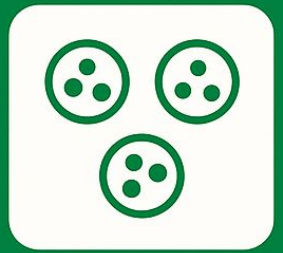
MiBici

BikeShareNetwork

Regression



Classification

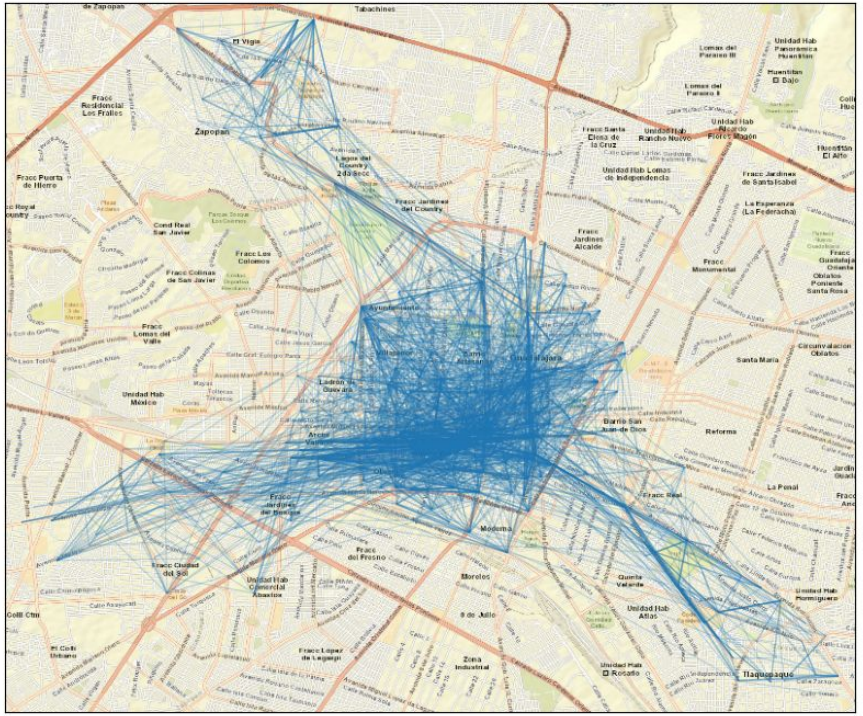


Clustering

ML PROJECT

Acerca Dataset

- Fuente: Datos abiertos del sistema de bicicletas compartidas MiBici en Guadalajara, México.
- Periodo de Datos: Diciembre de 2014 - Enero de 2024 (110 meses).
- Volumen de Viajes: 25,863,690 viajes registrados en el sistema.
- Estaciones: 372 estaciones de bicicletas públicas en la ZMG.
- Codificación de Datos: Archivos separados por comas (.csv)

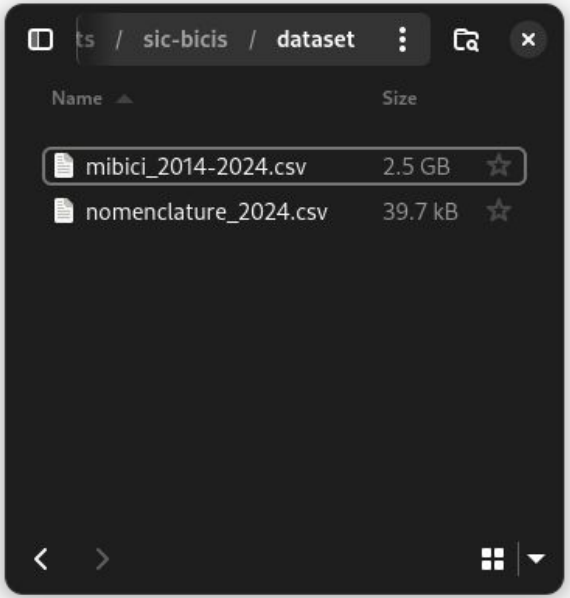


3000 viajes visualizados (aprox. 10 horas de servicio)

Acerca Dataset

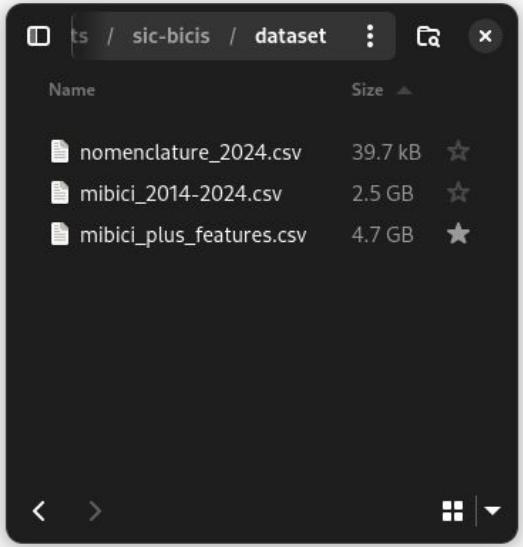
Los archivos y sus variables mas relevantes:

- **nomenclature_2024.csv: (39.7kB)**
  - ID
  - coordenadas de las 372 estaciones
- **mibici_2014-2024.csv: (2.5 GB)**
  - Fecha y hora de inicio y fin de los viajes
  - ID de inicio y fin del viaje
  - Edad del usuario
  - ID del usuario (anonimizado)

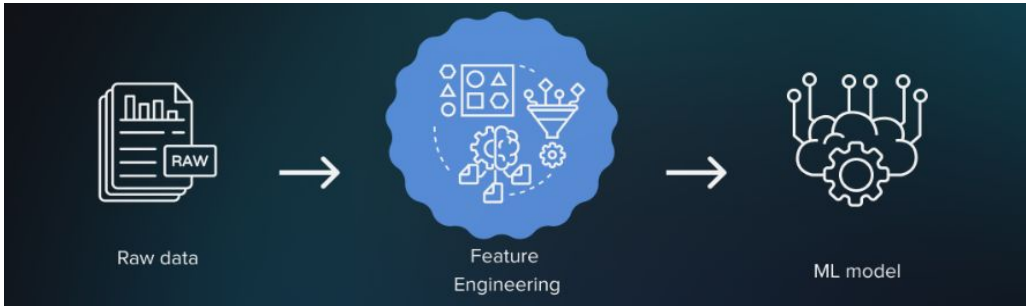


Feature Engineering

Se aplicaron diversas técnicas de Feature Engineering para convertir los datos en información útil.



- ➔ Conversión de tipos de datos
- ➔ Obtención de variables
- ➔ Uso de variables geográficas
- ➔ Cálculo de distancias
- ➔ Combinación con información climática histórica
- ➔ Popularidad de estaciones
- ➔ Segmentación del usuario
- ➔ Codificación de variables categóricas

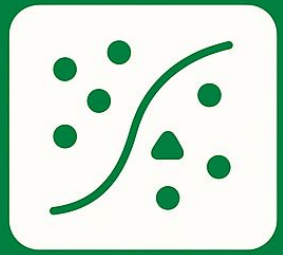


Análisis exploratorio de datos (EDA)

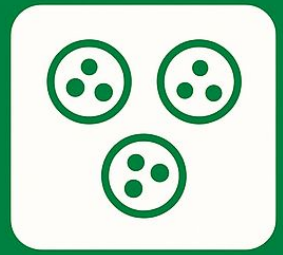
MiBici

BikeShareNetwork

Regression



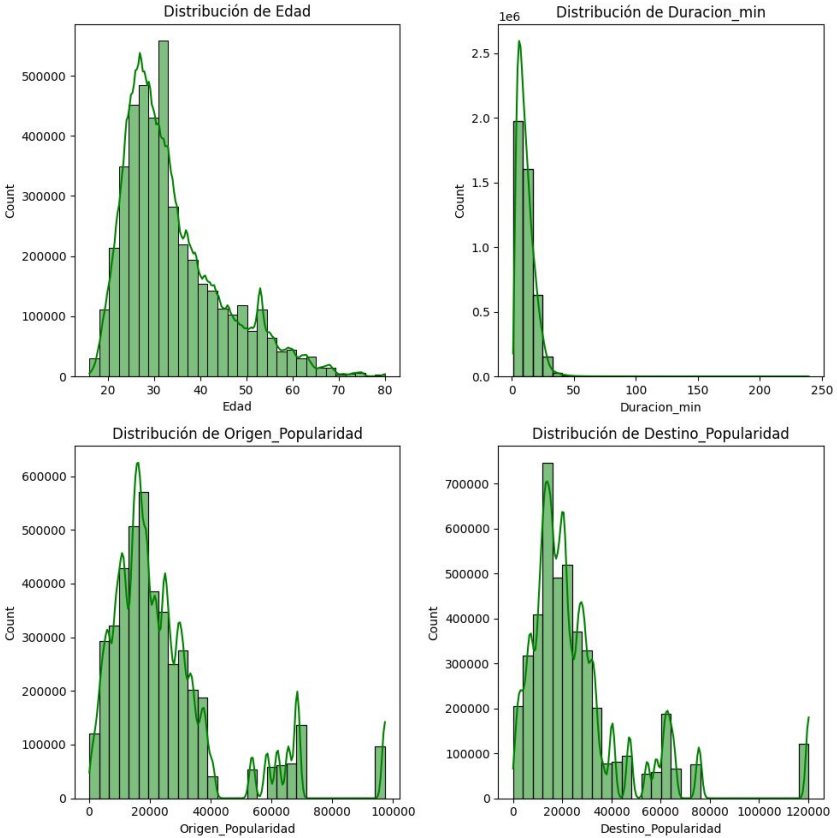
Classification



Clustering

ML PROJECT

Variables numéricas



Distribución de Edad

Indica que hay alrededor de 500,000 registros en el grupo de edad más frecuente (probablemente entre 25-35 años)

Distribución de Origen Popularidad

La mayoría de viajes salen de estaciones poco populares (probablemente muchas estaciones con pocos viajes)

Duración min

La mayoría de viajes son cortos (<30 min), con pocos viajes muy largos.

Destino Popularidad

Muestra distribución similar a origen pero con ligeras diferencias

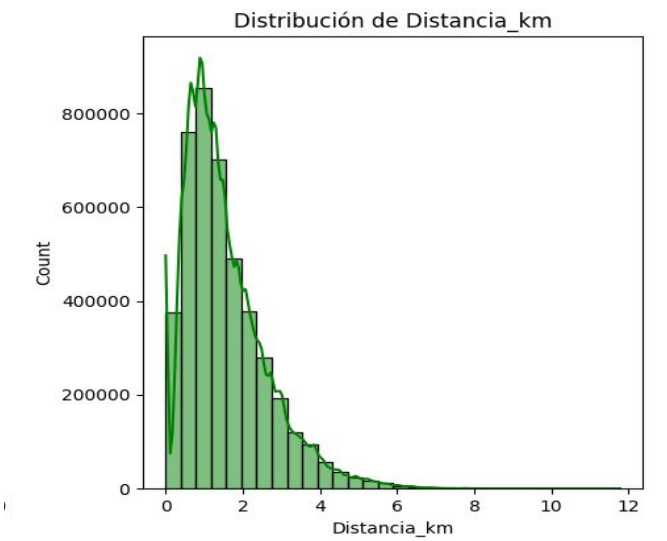
Variables numéricas

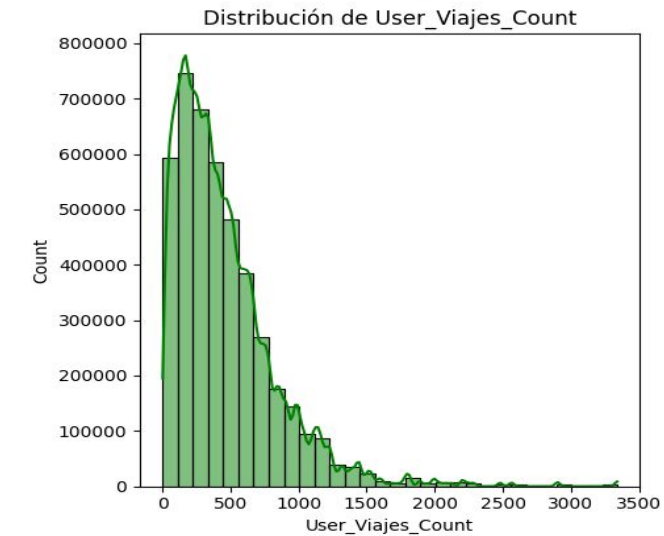
Distancia km

Alrededor de 800,000 viajes en el rango más común (probablemente <2 km)

User Viajes Count

Cuántos usuarios tienen X cantidad de viajes

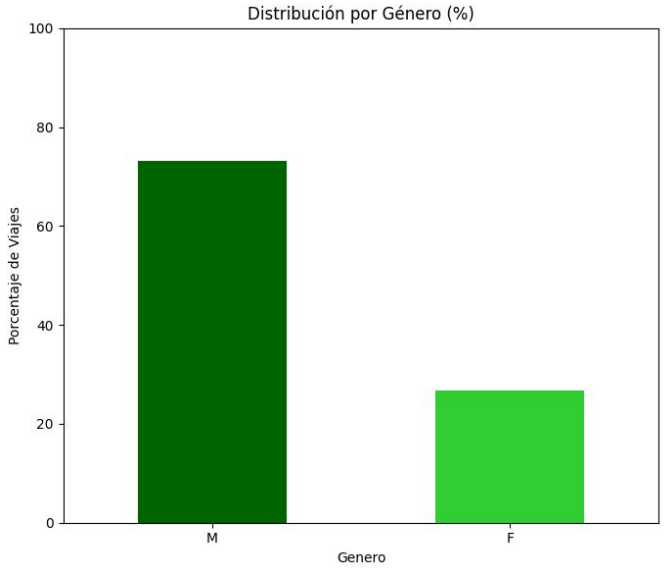




Variables categóricas

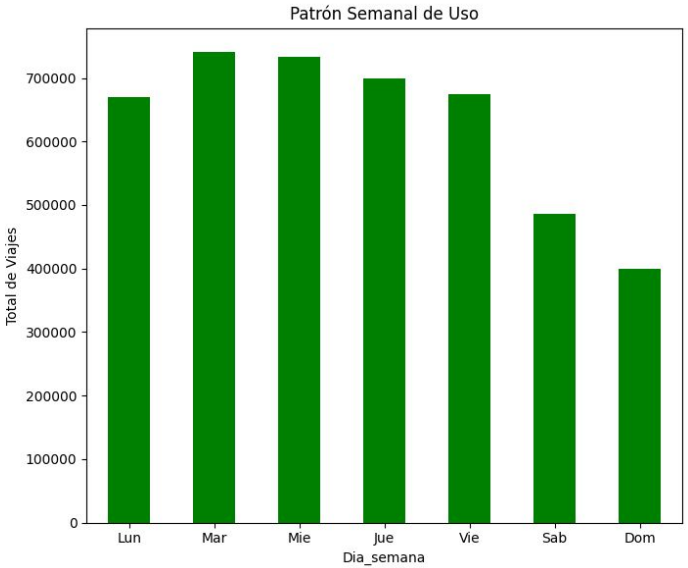
Distribución por género

Podemos visualizar que los hombres usan más este tipo de transporte que las de género femenino.



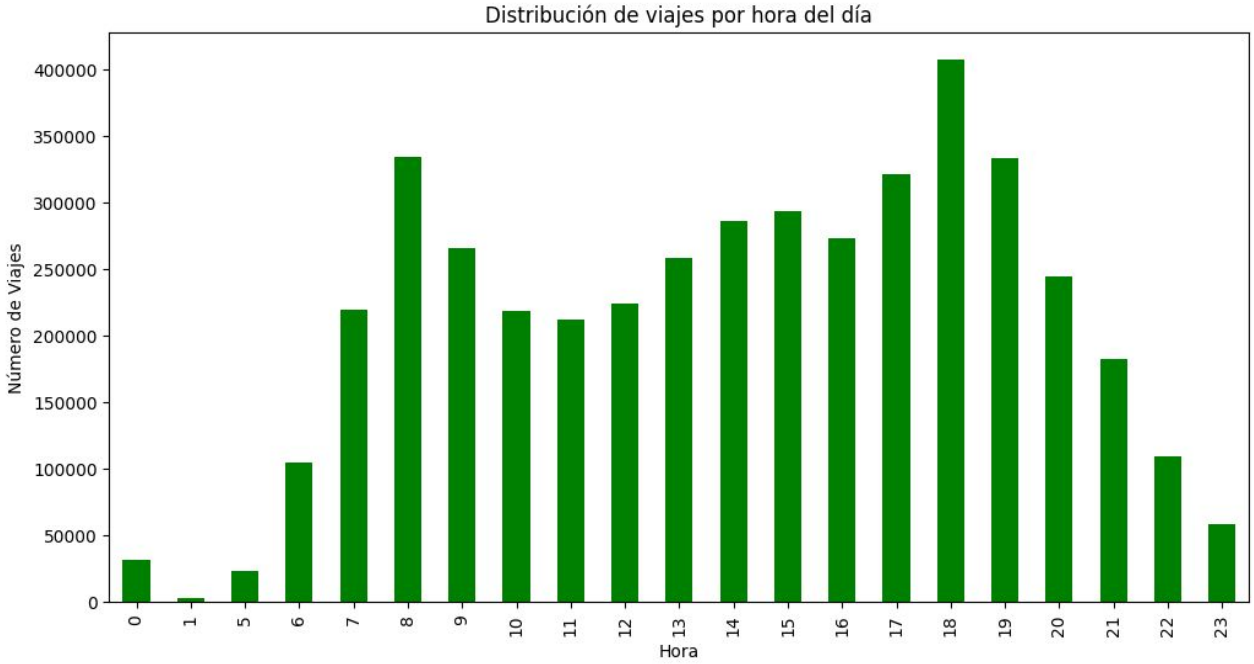
Patrón semanal de Uso

Visualizamos que entre semana es más usado el sistema de mibici.



Análisis temporal

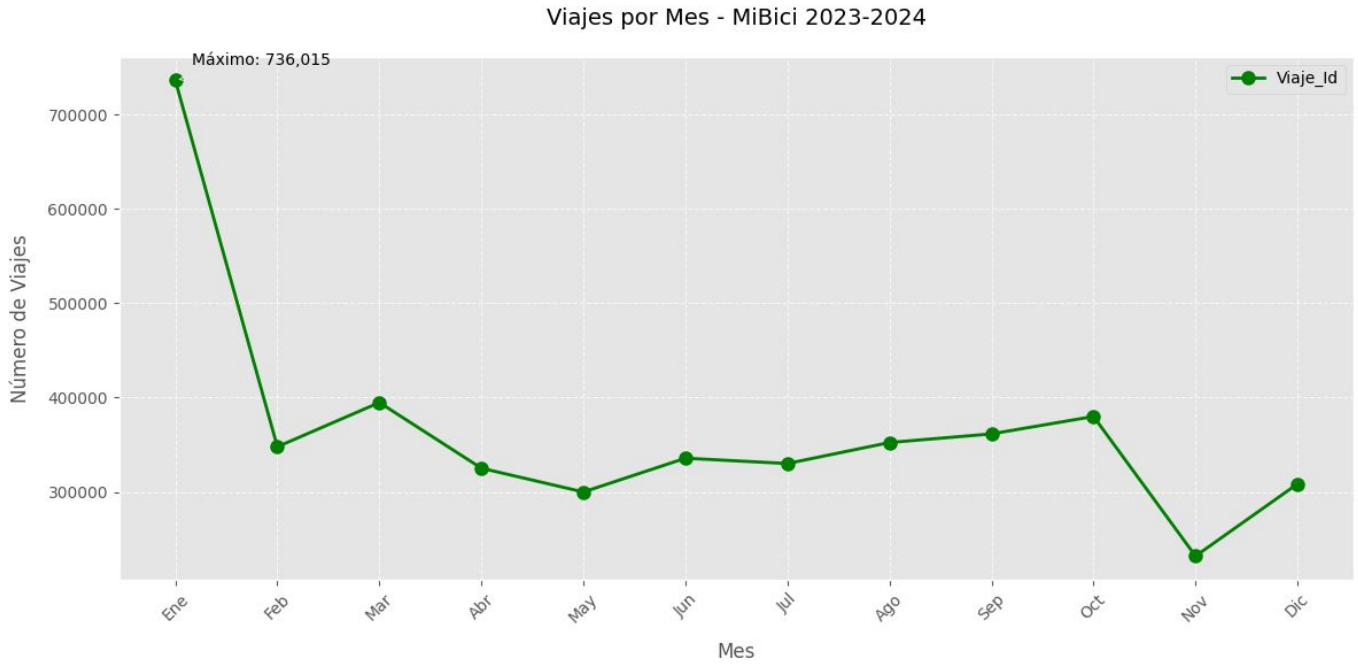
Vemos que a las 8:00 y a las 18:00 son las horas que más se usan las bicis.



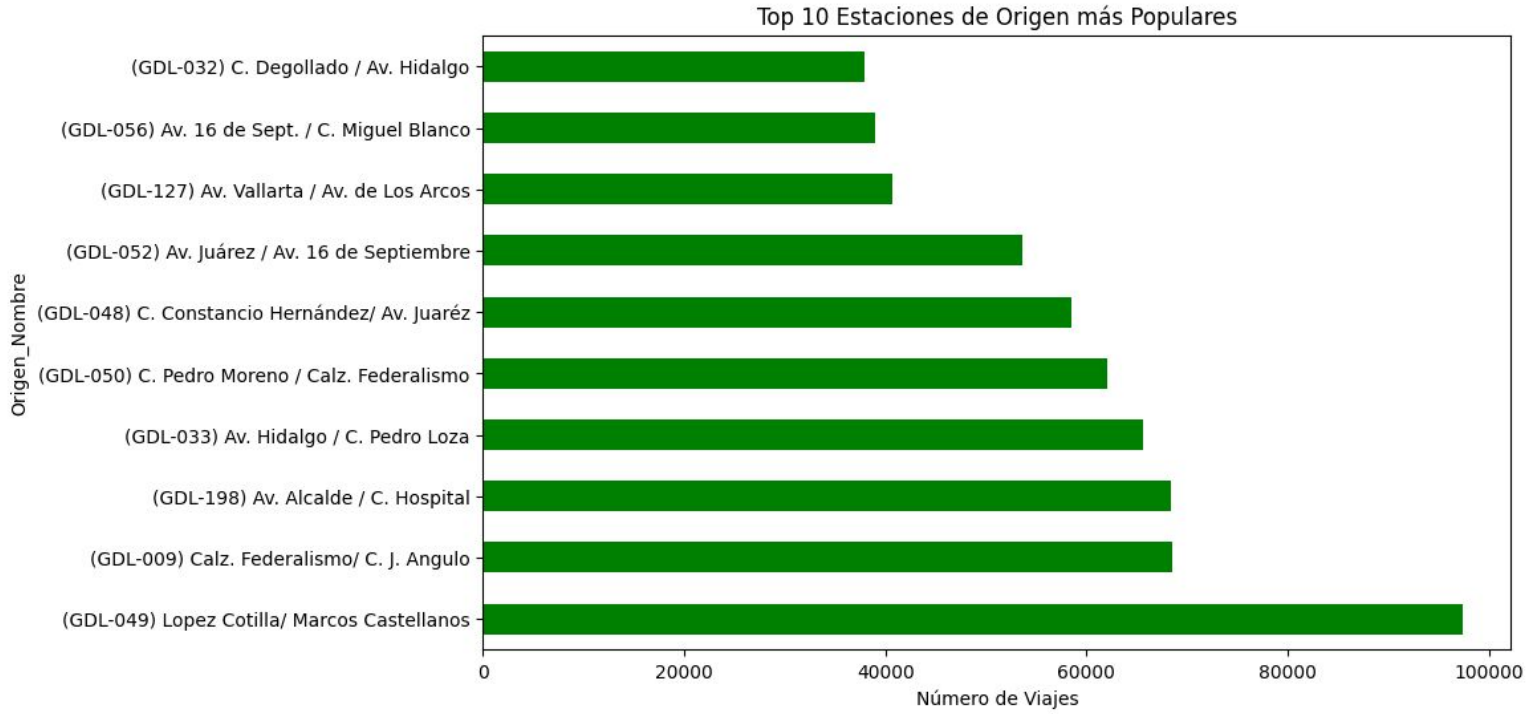
Análisis temporal

Viajes por meses

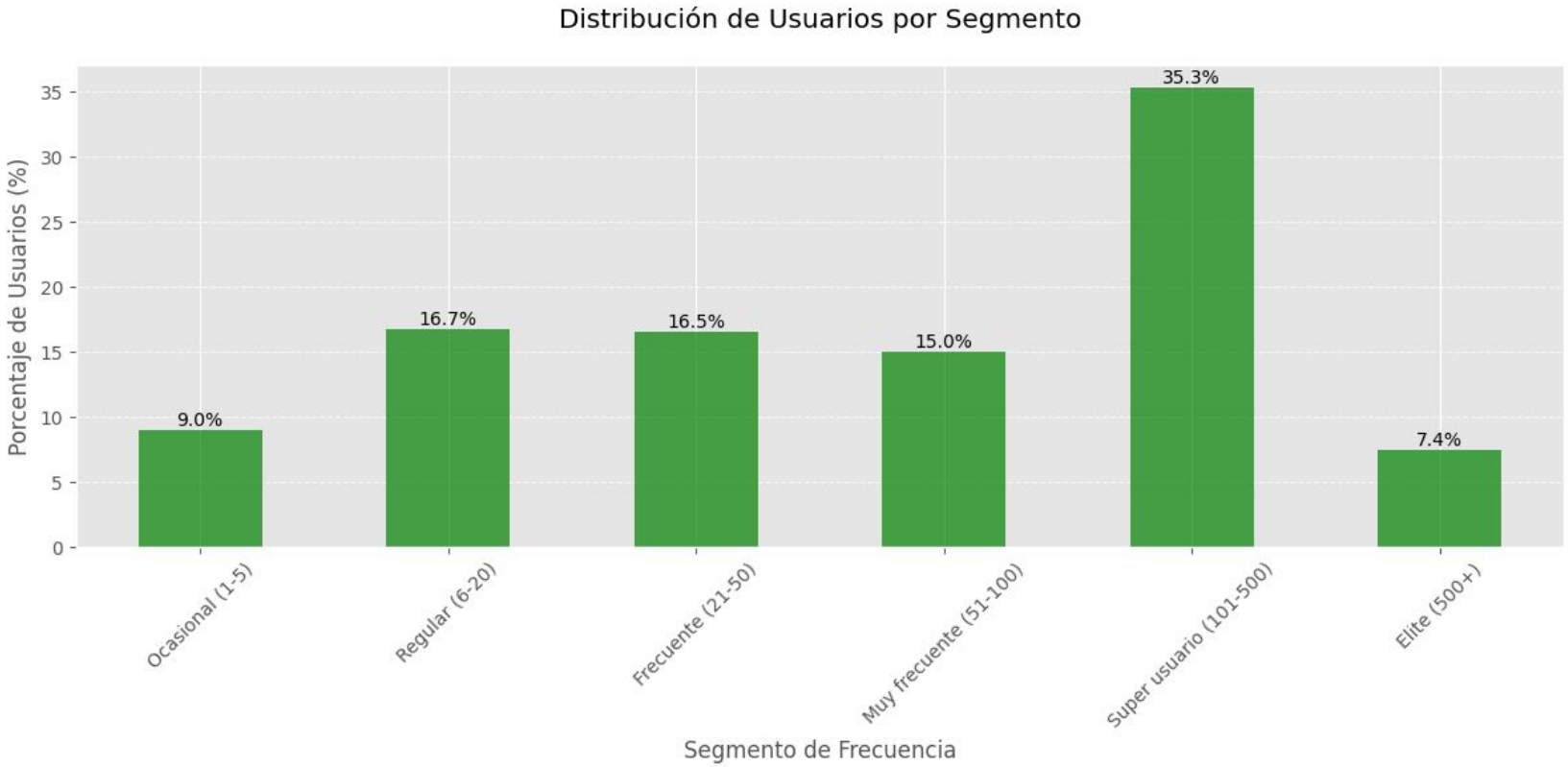
Podemos ver que en enero es donde mas se usan las bicis del 2023 al 2024



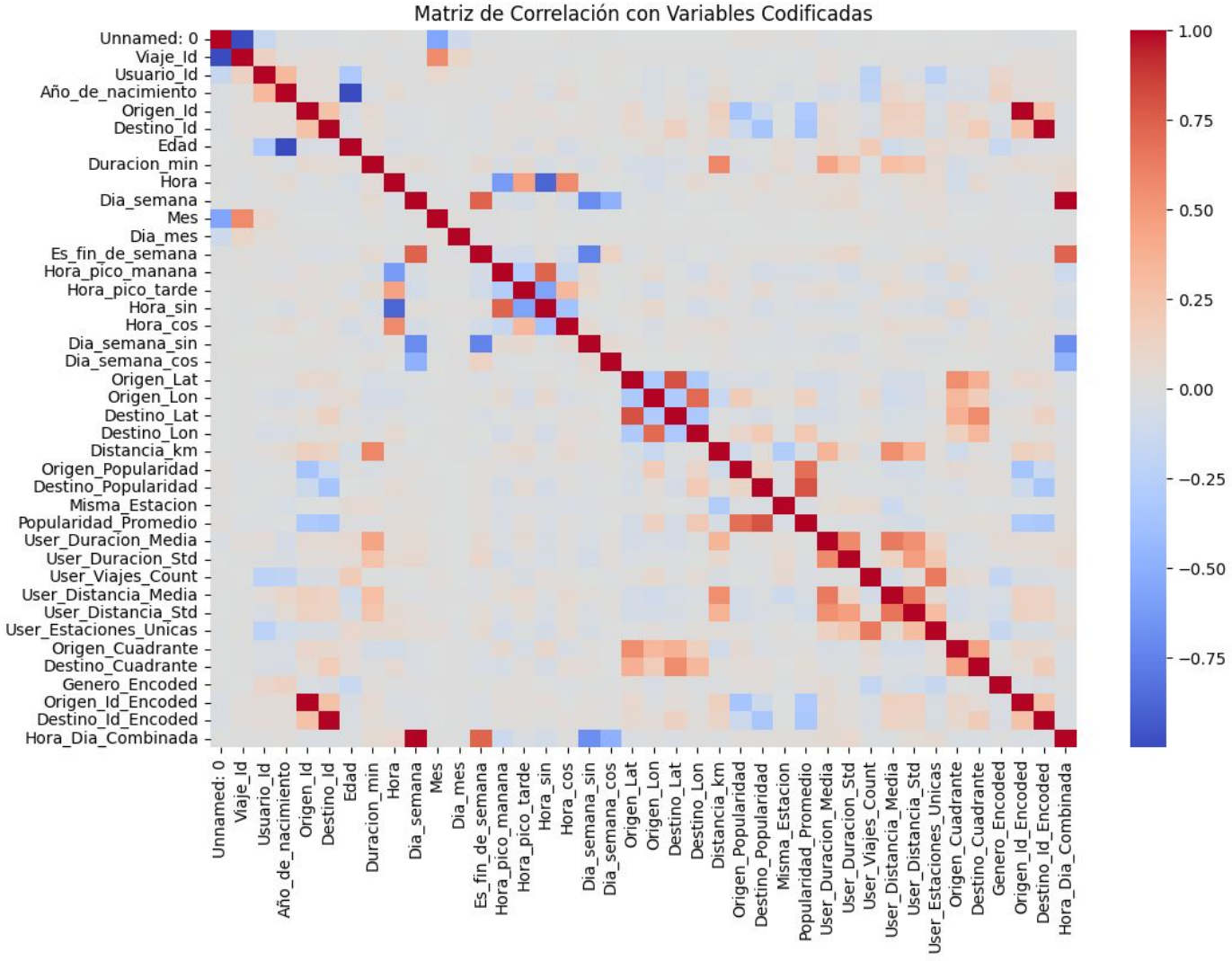
Análisis geográfico



Análisis Usuarios



Correlaciones

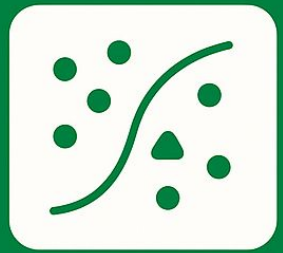


Implementación de modelos

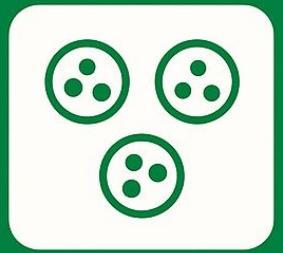
MiBici

BikeShareNetwork

Regression



Classification



Clustering

ML PROJECT

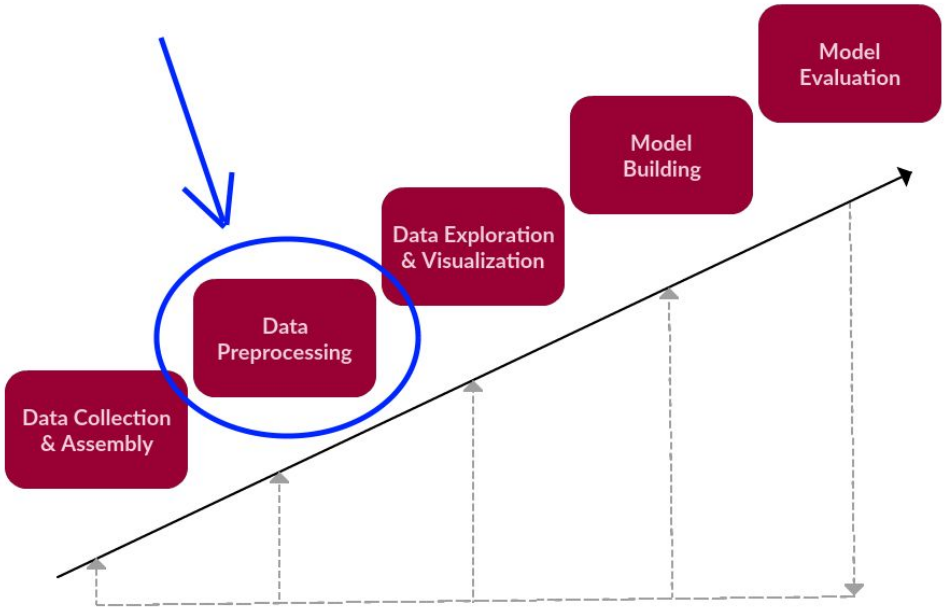
Regresión

Predicción de Demanda

Predecir la cantidad de viajes por hora en una estación específica del sistema MiBici, utilizando un modelo de series temporales.

Regresión

Regresión (SARIMA)



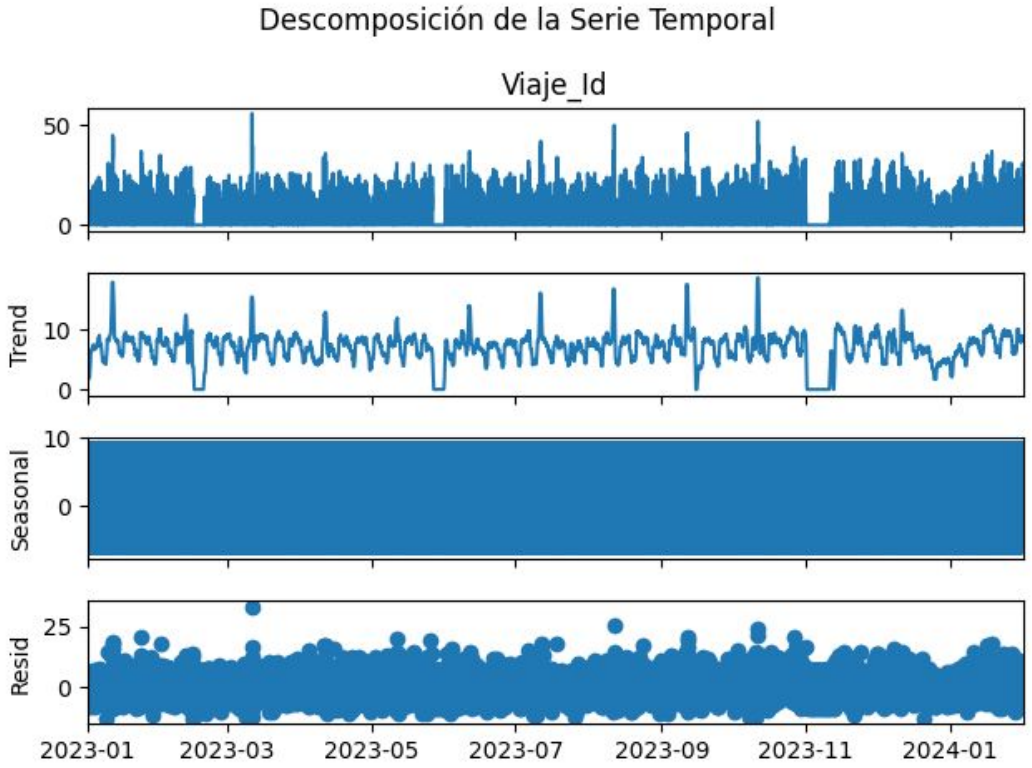
Metodología

Preprocesamiento

- Conversión de timestamps a formato datetime.
- Agrupación de viajes por hora
- Relleno de horas sin viajes con 0.

Regresión

Regresión (SARIMA)

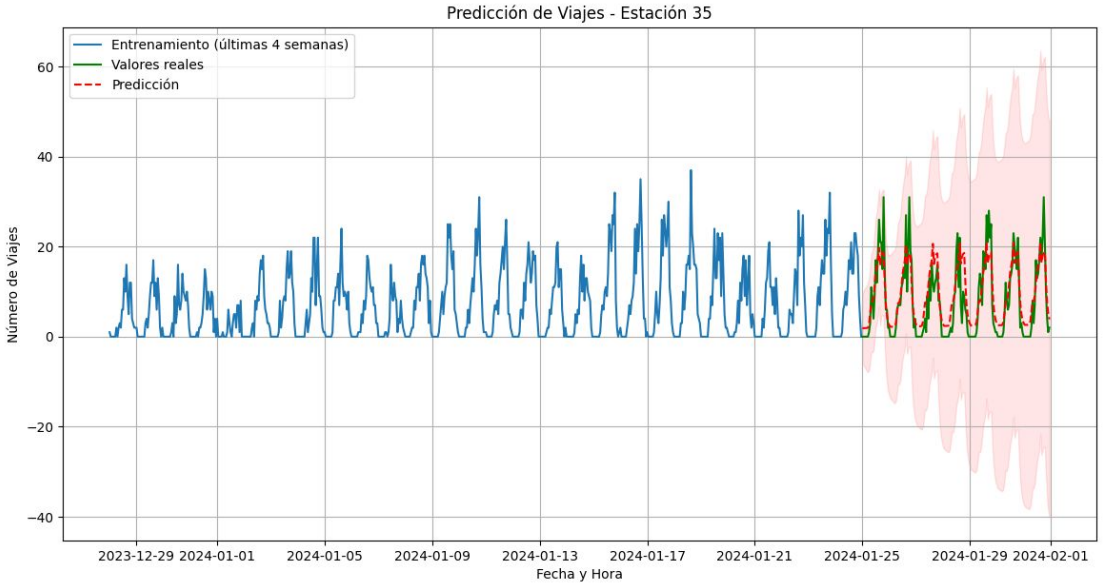


Descomposición de series temporales

- **Serie Original:** Muestra la dirección general del uso a lo largo de
- **Tendencia (Trend):** Muestra la dirección general del uso a lo largo del tiempo.
- **Estacionalidad (Seasonal):** Refleja patrones que se repiten en intervalos regulares.
- **Residuo (Resid):** Captura anomalías o variaciones no explicadas por la tendencia o estacionalidad.

Regresión

Regresión (SARIMA)



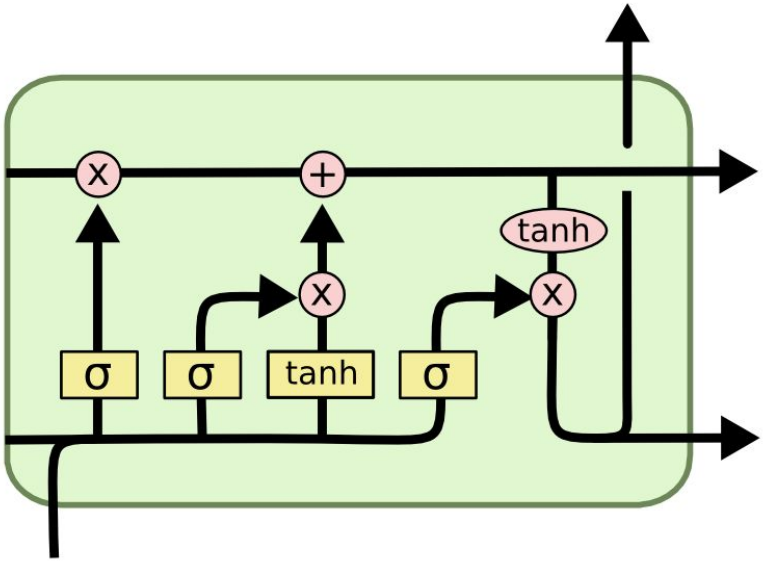
Modelo SARIMA

La gráfica permite visualizar qué tan cerca estuvieron las predicciones (rojo) de los valores reales (verde), evaluando así el rendimiento del modelo. Los ejes muestran fechas/horas (eje X) y cantidad de viajes (eje Y).

.

Regresión

(LSTM)



LSTM: Long Short-Term Memory

Predicción de Demanda

A partir de la demanda histórica de la estación de bicicletas y variables climáticas que sirven de ancla en la estacionalidad de los datos.

Información climática obtenida de: OpenMeteo

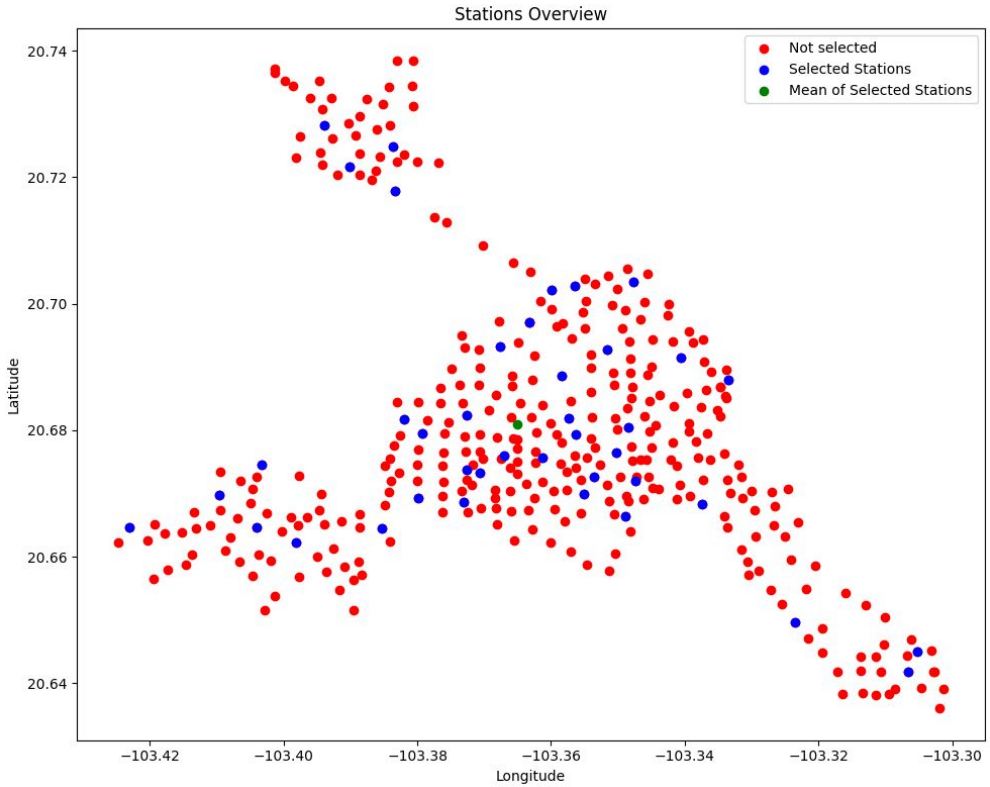
Regresión

(LSTM)

Metodología

Preprocesamiento

- Conversión de timestamps a formato datetime.
- Filtrado de salidas en falso (viajes duración < 1min)
- Agrupación de viajes por día y estación.
- Relleno de días sin viajes con 0.
- Cálculo de la distancia manhattan.
- Muestreo de las estaciones.
- Obtención y unión de variables históricas de clima.



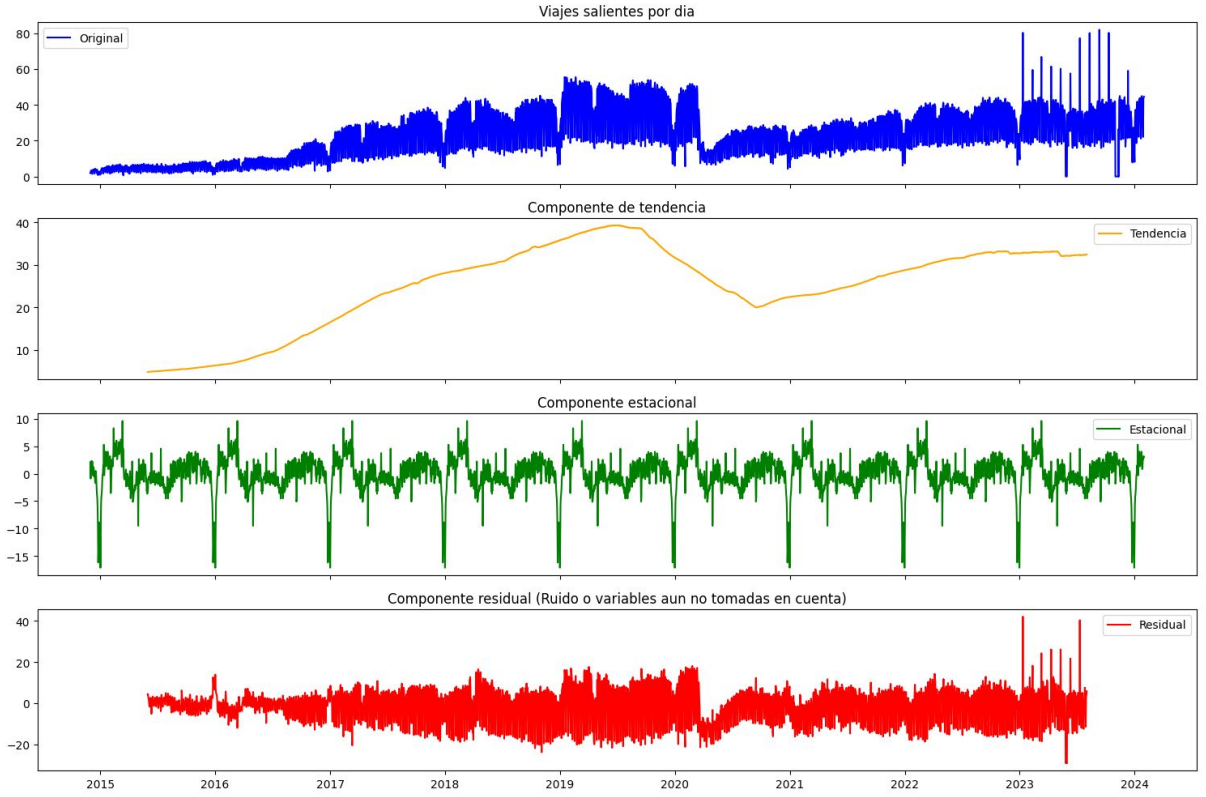
Regresión

(LSTM)

Metodología

Exploración de técnicas

Se observó que los datos tienen un componente estacional bastante fuerte con periodicidad de 1 año, siguen una tendencia propia y en el componente residual aún quedan variables por ser consideradas.



Regresión

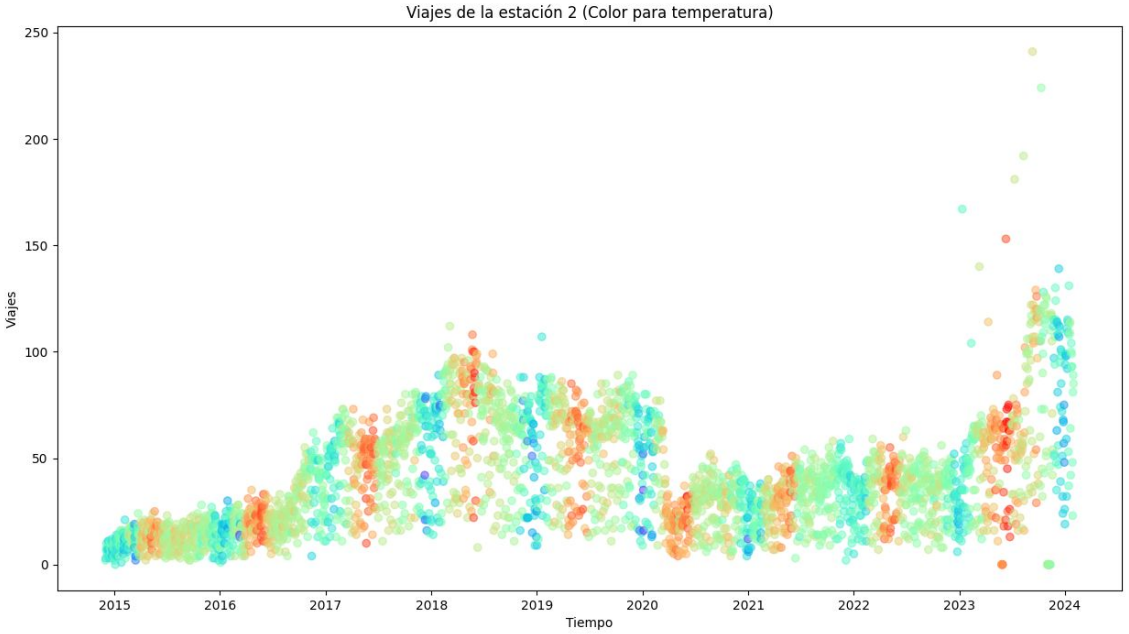
(LSTM)

Metodología

Implementación

Como heurísticas, se optó por otorgar información estacional al modelo con la temperatura media del día y los mm de lluvia total del dia.

Y el componente de tendencia lo obtiene a partir de la sumatoria de viajes salientes de la estación y su distancia manhattan media.



Regresión

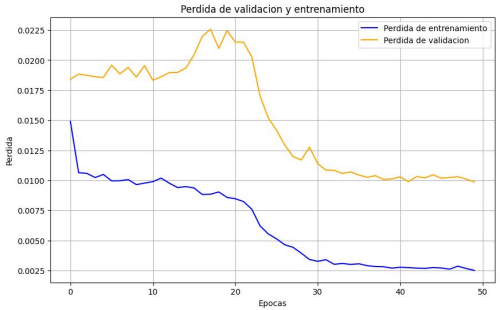
(LSTM)

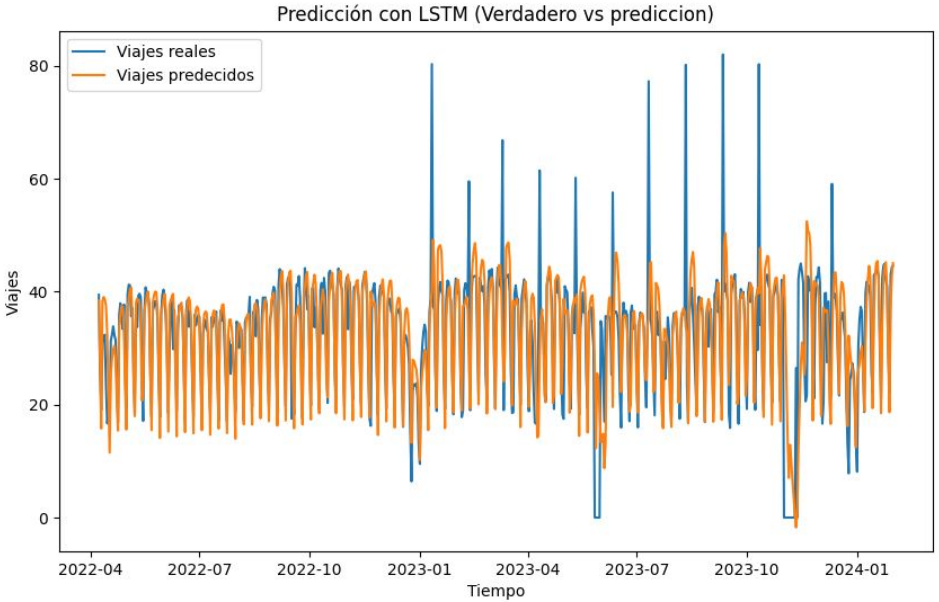
Metodología

Resultados

Características:

- Viajes salientes
- Distancia media de viajes
- Temperatura media
- Lluvia total del dia





Regresión

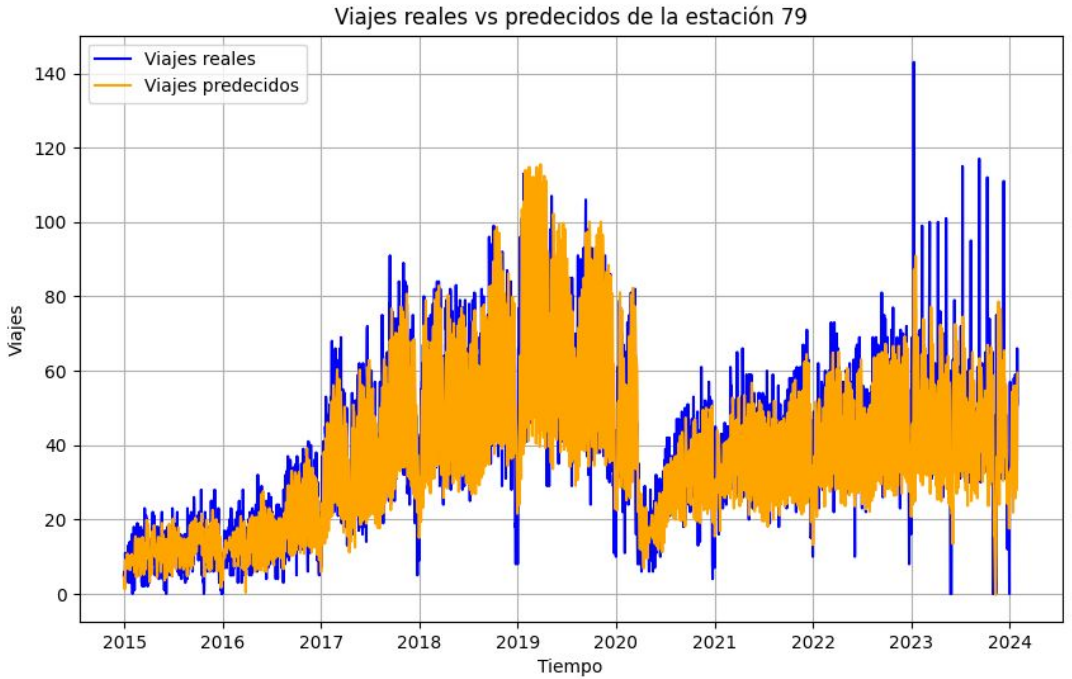
(LSTM)

Metodología

Resultados

Características:

- Viajes salientes
- Distancia media de viajes
- Temperatura media
- Lluvia total del dia



Regresión

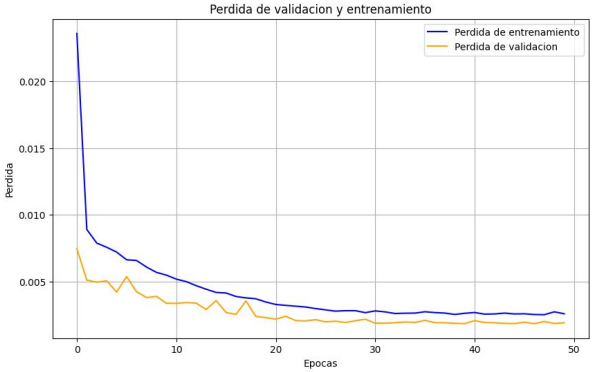
(LSTM)

Metodología

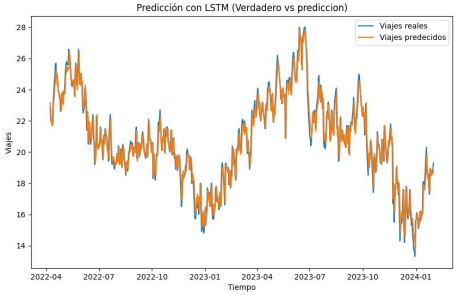
Resultados

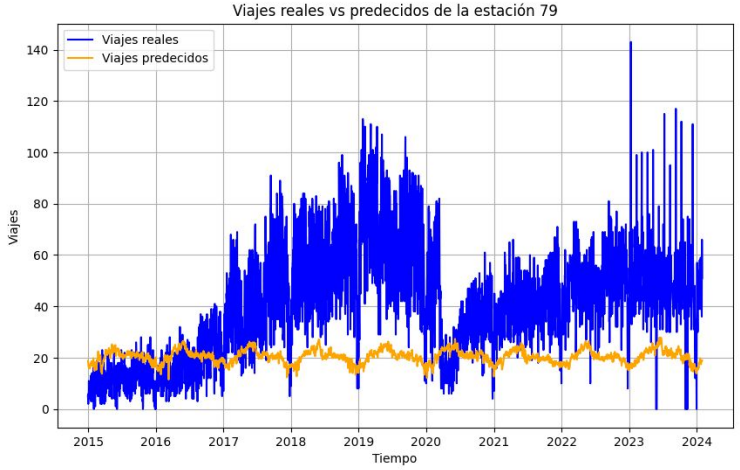
Características:

- *SIN Viajes salientes*
- *SIN Distancia media de viajes*
- Temperatura media
- Lluvia total del dia



Validación ->





<- Prueba real

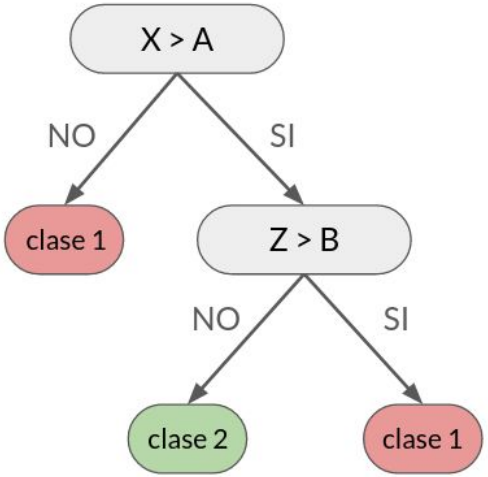
Clasificación

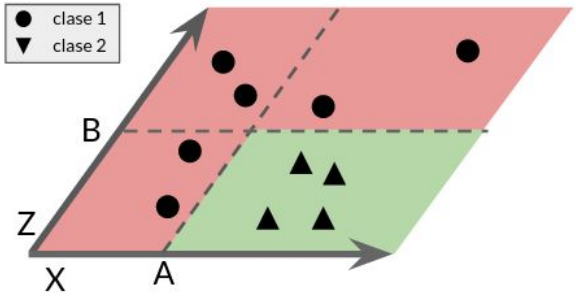
Árbol de decisión

Renovación de Usuarios

Implementación

Implementa un modelo de árbol de decisión para predecir si un usuario renovará su membresía en un sistema de Mibici.



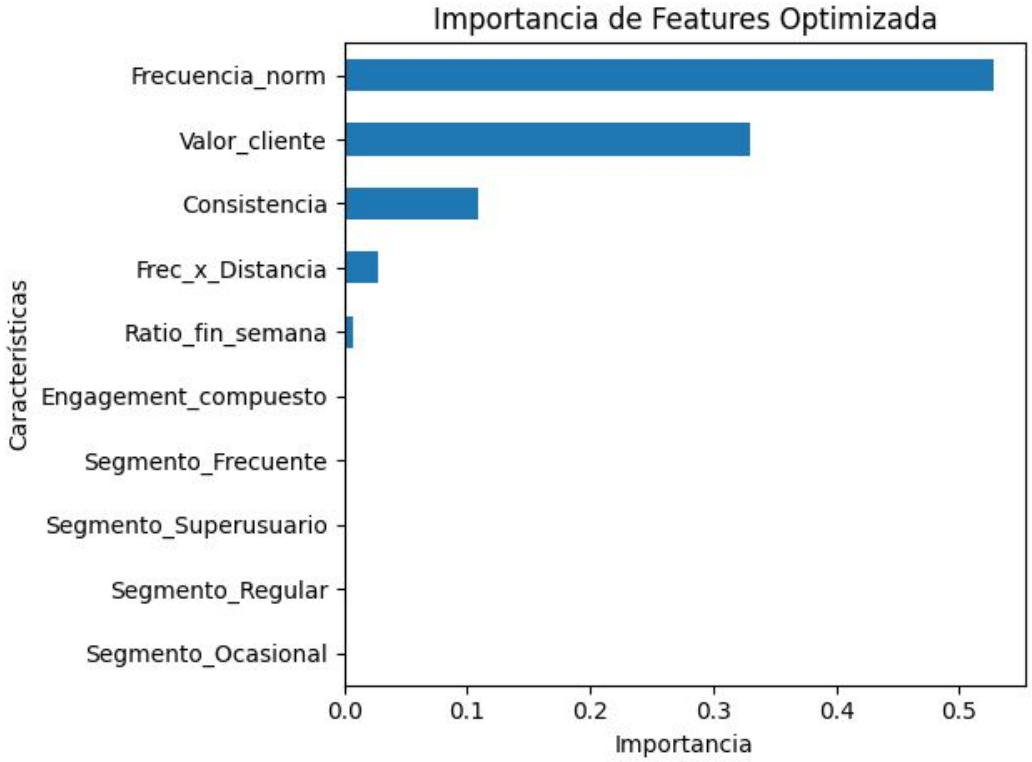


Clasificación

Arbol de desicion

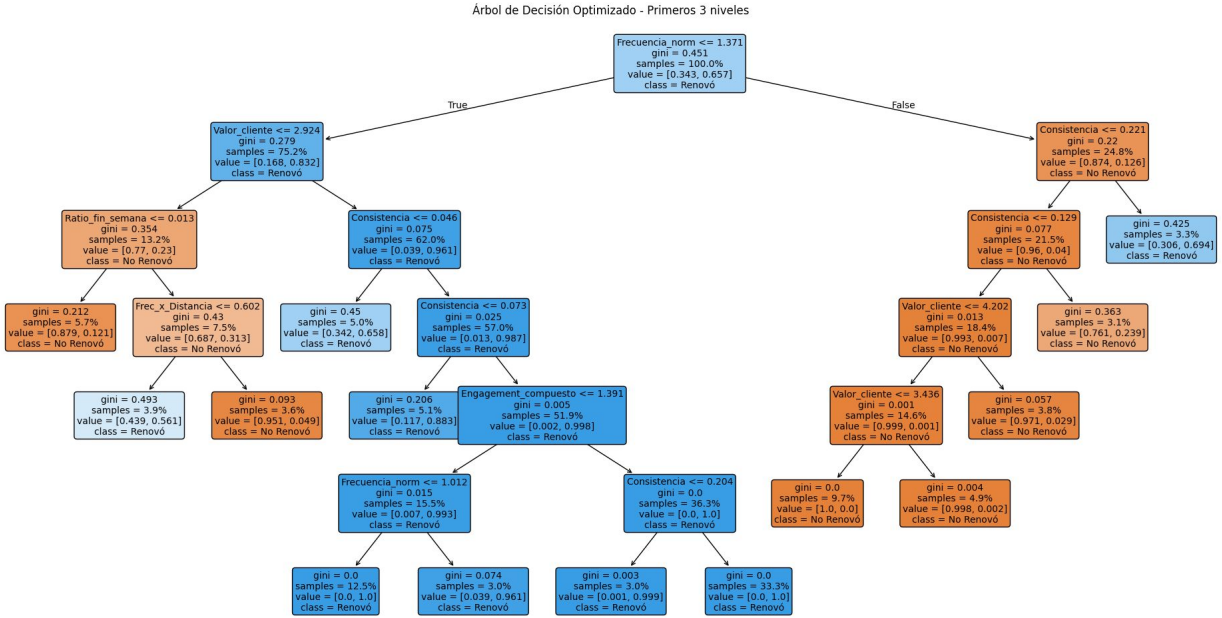
Importancia de Features

Las variables más relevantes para predecir la renovación.



Clasificación

Arbol de desicion



Árbol de Decisión

Azul Oscuro -> Alta probabilidad de no renovar (90%).

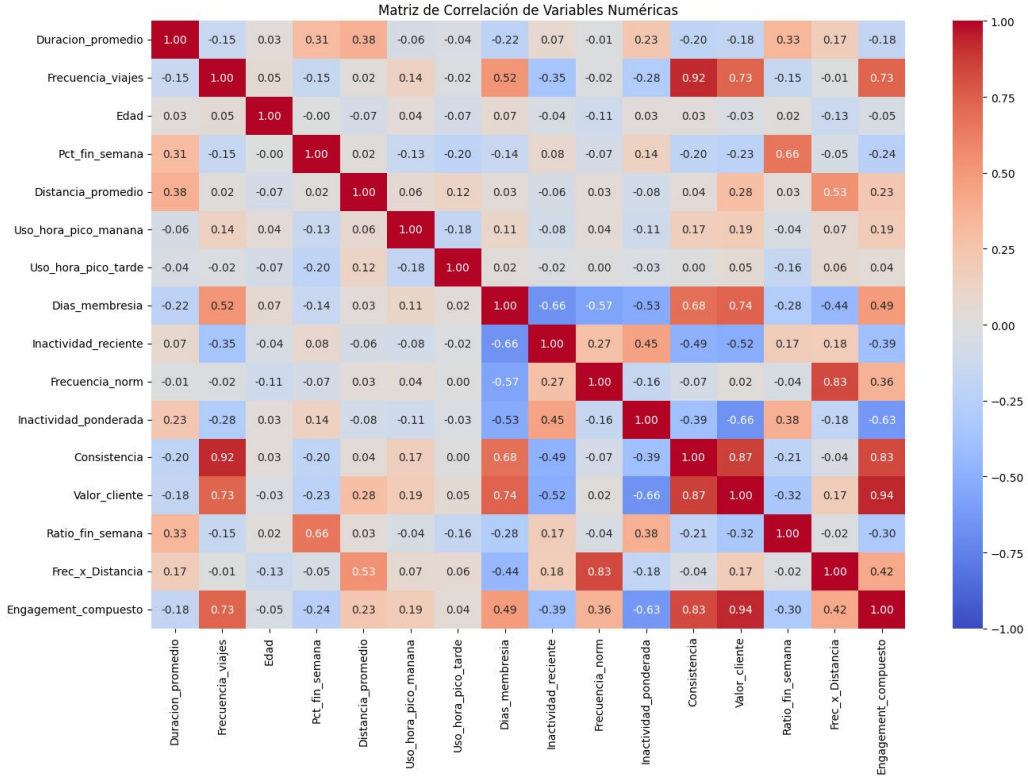
Azul Claro -> Riesgo moderado de no renovar (60%).

Naranja Oscuro -> Alta probabilidad de renovación (90% en este caso).

Naranja Claro -> Probabilidad moderada de renovación (60-80%).

Clasificación

Arbol de desicion

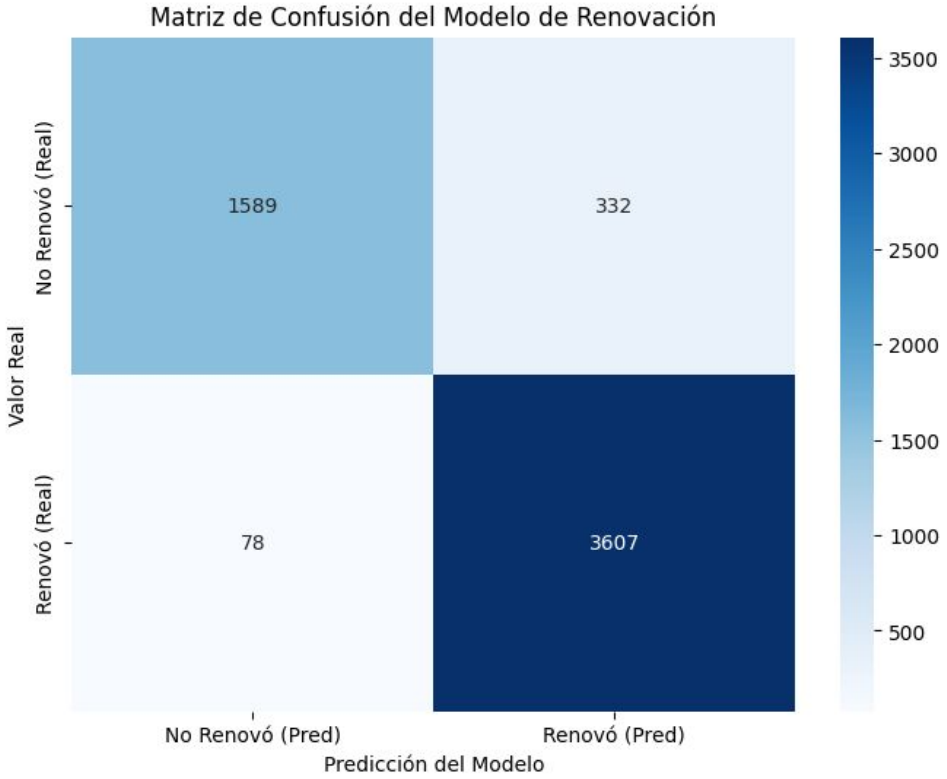


Matriz de Correlación

- **Frecuencia de viajes vs Consistencia (0.92)**
  - Los usuarios que realizan más viajes tienden a usar el sistema de manera más constante.
- **Consistencia vs Valor del cliente (0.87):**
  - Cuanto más consistente es el uso, mayor es el valor que representa el cliente para el sistema.
- **Valor del cliente vs Engagement compuesto (0.94)**
  - El nivel de engagement refleja casi perfectamente el valor del cliente.
- **Días de membresía vs Frecuencia de viajes (0.52) y Consistencia (0.68)**
  - Cuantos más días tiene un usuario con membresía activa, más frecuente y consistente es su uso del sistema.

Clasificación

Arbol de desicion

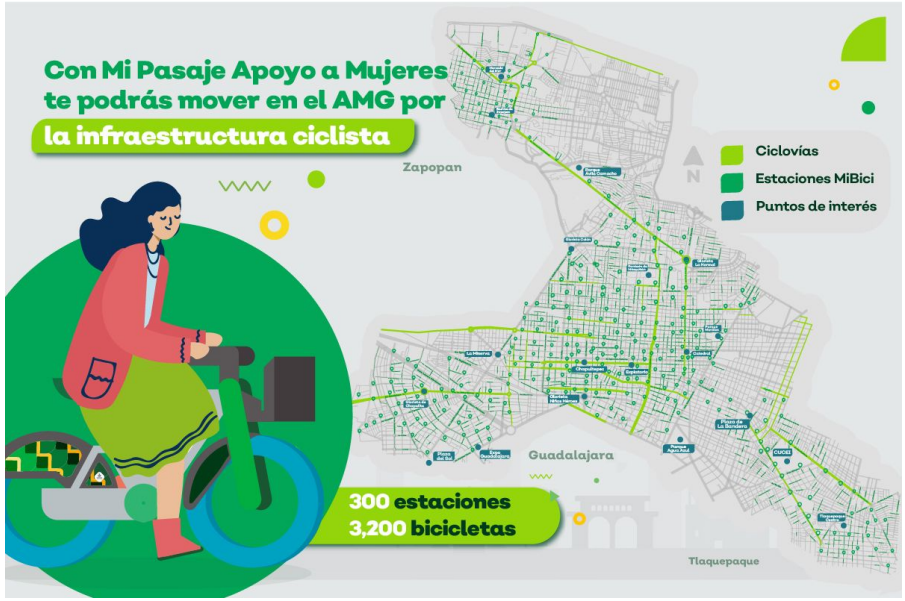


Matriz de Confusión

- El modelo tiene una alta capacidad de predicción (93% de accuracy)
- Predice muy bien quién renovará (98% de recall en la clase positiva).
- Tiene un pequeño margen de error en quienes no renovarán, con un 83% de especificidad.

Agrupamiento

Clasificación las según su popularidad



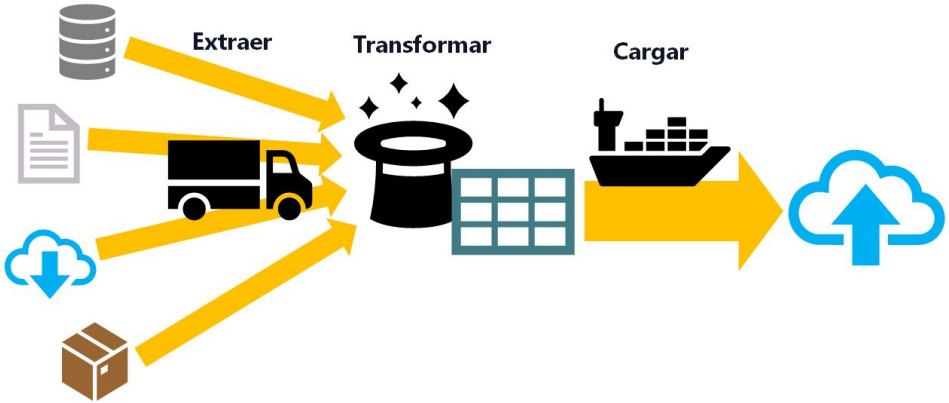
Clasificación las según su popularidad

Realizamos un análisis de clustering (agrupamiento) sobre las estaciones del sistema de bicicletas públicas MiBici, utilizando el algoritmo K-Means para clasificarlas según su popularidad (cantidad de viajes que inician en ellas).

Agrupamiento

Clasificación las según su popularidad

Métricas del modelo



Obtener métricas clave de cada estación.

Se agrupan los datos por Origen_Id (ID de la estación) y se extraen:

Origen_Popularidad: Número de viajes que parten de la estación.

Origen_Lat y Origen_Lon: Coordenadas geográficas.

Origen_Nombre: Nombre de la estación.

Agrupamiento

Clasificación las según su popularidad

Determinación del Número Óptimo de Clusters (k)

Método del Codo

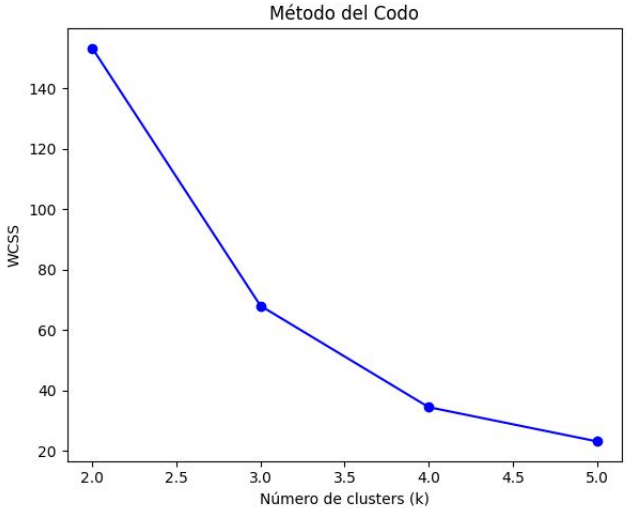
Evalúa la suma de cuadrados intra-cluster (WCSS) para diferentes valores de k.

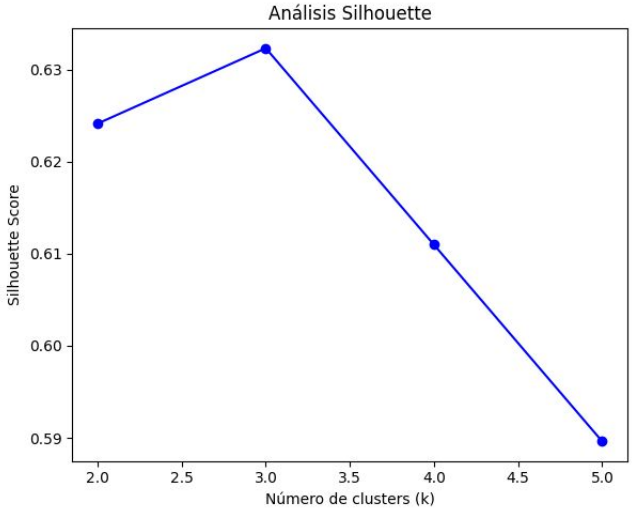
Se busca el punto donde añadir más clusters no mejora significativamente el modelo.

Análisis Silhouette

Mide qué tan bien definidos están los clusters.

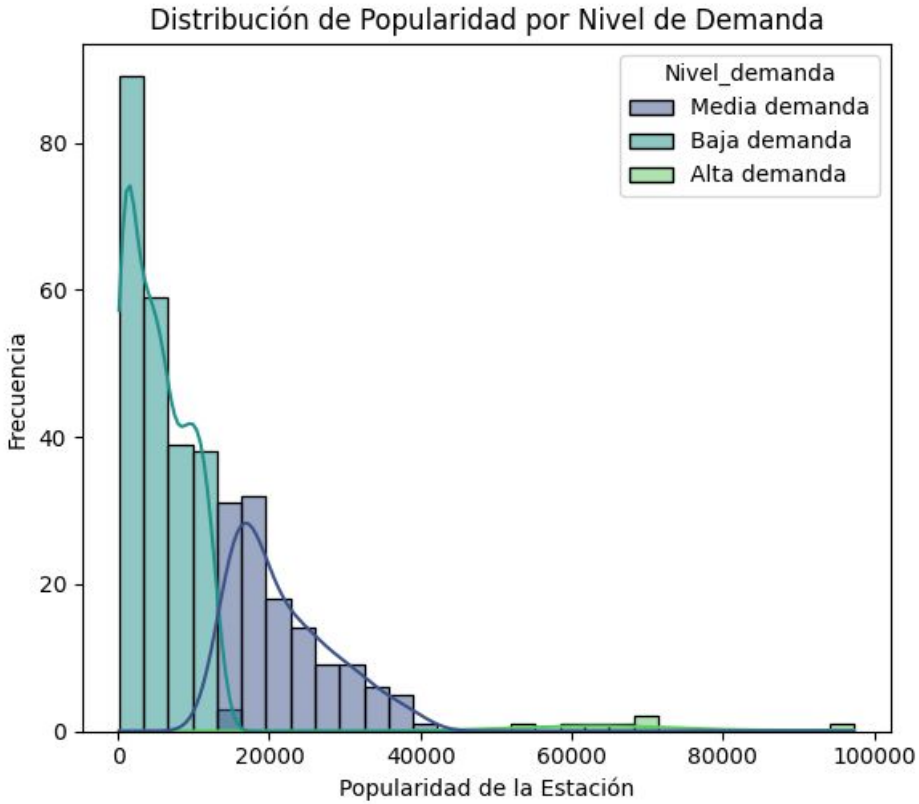
Un score alto (cercano a 1) indica clusters bien separados





Agrupamiento

Clasificación las según su popularidad



Distribución de popularidad por nivel de demanda

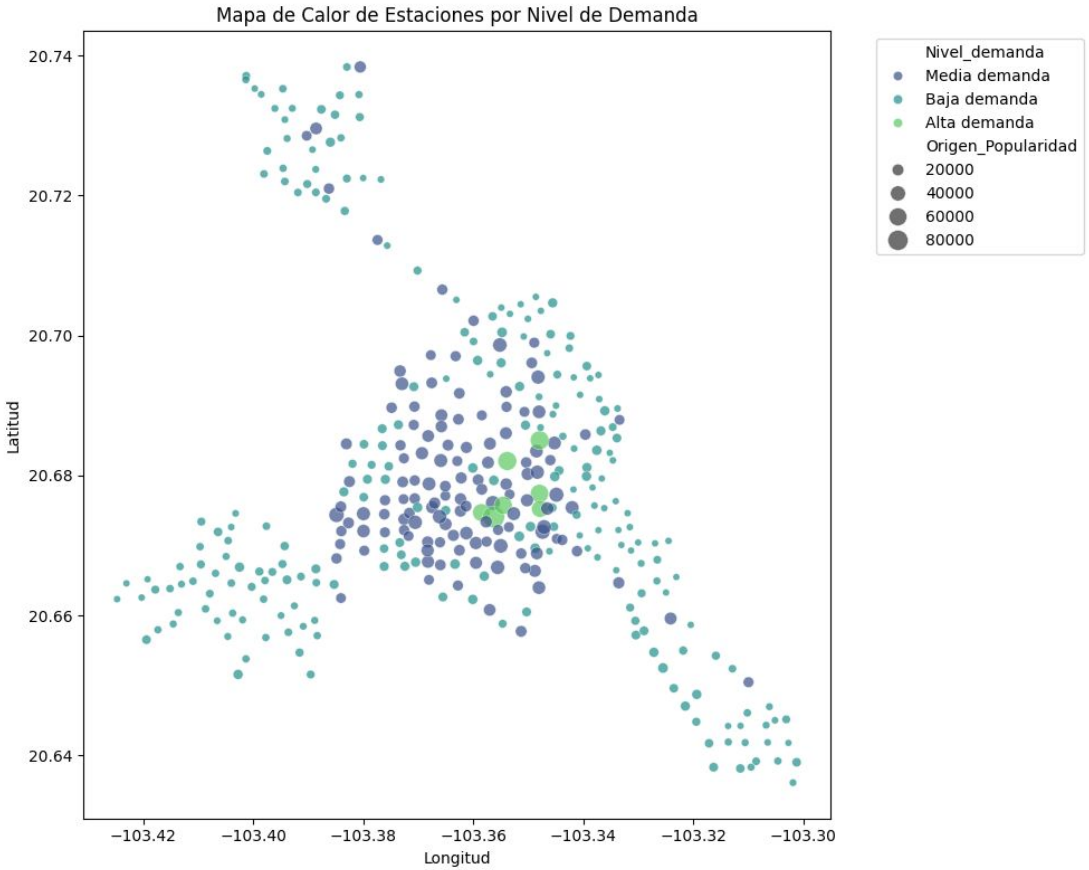
Aplicación de K-Means (k=3)

Se elige k=3 para clasificar las estaciones en:

- Baja demanda
- Media demanda
- Alta demanda

Agrupamiento

Clasificación las según su popularidad



Mapa Geográfico de Estaciones

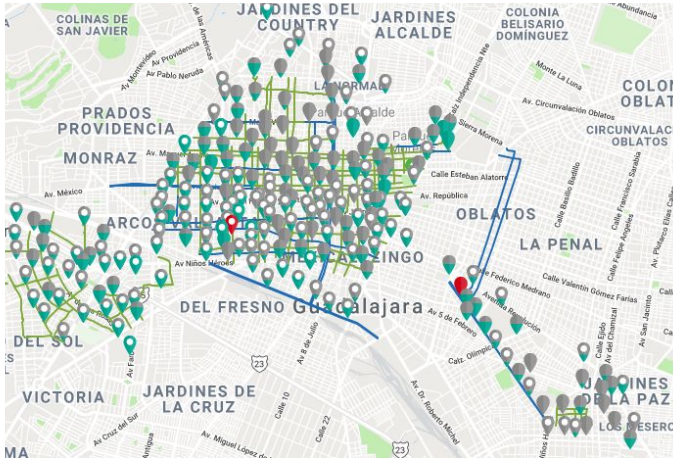
Agrupamiento

Clasificación las según su popularidad

Conteo de estaciones por nivel de demanda:

Nivel demanda

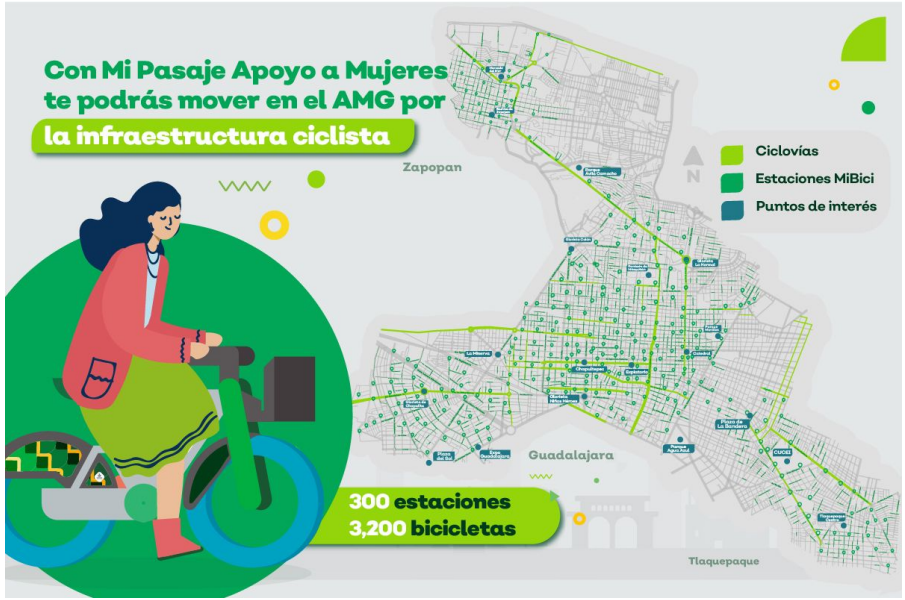
Baja demanda	228
Media demanda	125
Alta demanda	7



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

Agrupamiento

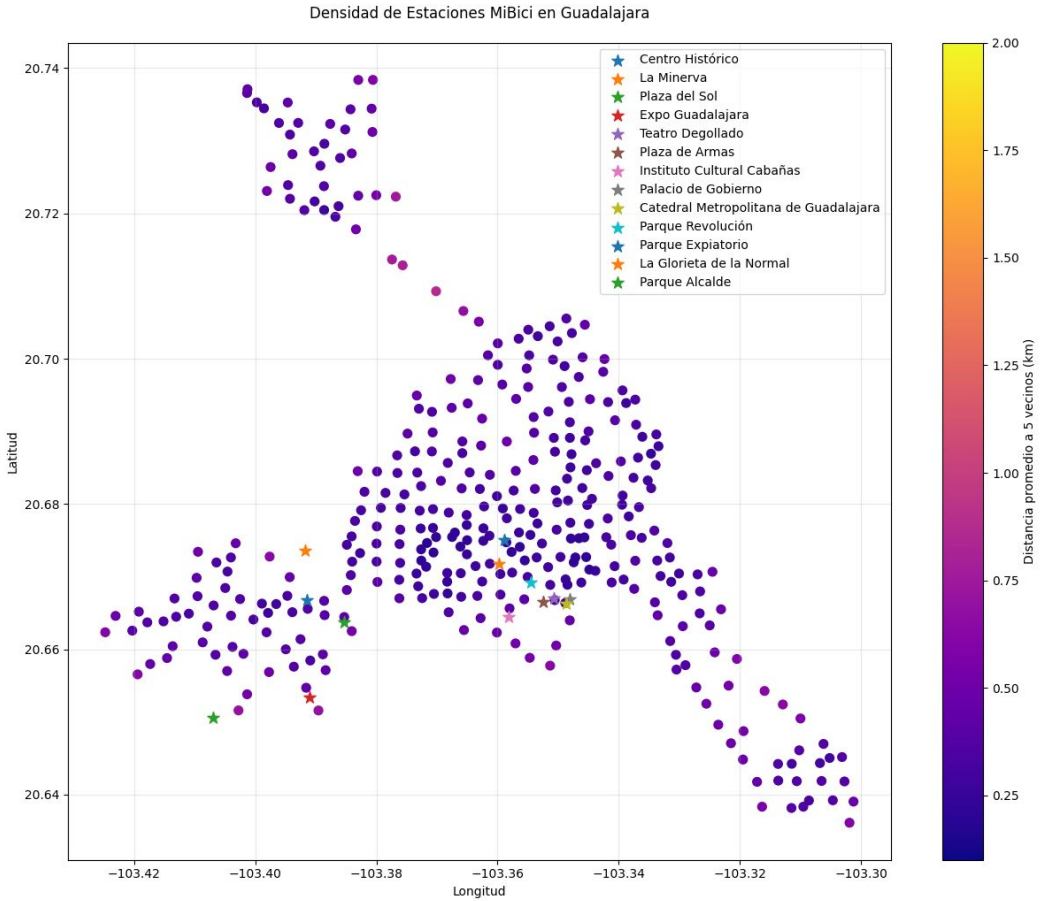


Densidad espacial de estaciones

Calculamos las distancias promedio entre estaciones y visualizando densidad en un mapa interactivo

Agrupamiento

Densidad espacial de estaciones



Mapa de calor con gradiente

- Amarillo: Alta densidad
- Azul: Baja densidad
- Basado en distancias promedio a 5 estaciones cercanas.
- Puntos de referencia clave (Centro Histórico) marcados con estrellas.

- BallTree se usa en algoritmos de agrupamiento
- BallTree no es un modelo de aprendizaje

Agrupamiento

Densidad espacial de estaciones

Conteo de estaciones por nivel de demanda

Top 10 estaciones más aisladas

Origen_Nombre	Densidad_Estaciones
(GDL-188) Av. de la Presa / Av. Manuel A. Cam	0.869742
(ZPN-070)Av. Avila Camacho / Av. Patria	0.819053
(ZPN-056) Av. M. Ávila Camacho / C. San Jorge	0.771751
(ZPN-069)Jesús María Romo /Av. Aurelio Ortega	0.757365
(ZPN-067) C. Chimalhuacán / Av. López Mateos	0.703386
(GDL-158) Av. Faro / Av. Las Rosas	0.693868
(GDL-192) Calle A1 / Av. Manuel A. Camacho	0.690288
(GDL-226) C. Luis G. Cuevas /Av. Revolución	0.641076
(GDL-176) C. San Bonifacio/ Paseo Benedictino	0.631031
(GDL-210) Av. Washington / Calz. Independenci	0.629025

Top 10 estaciones más centrales

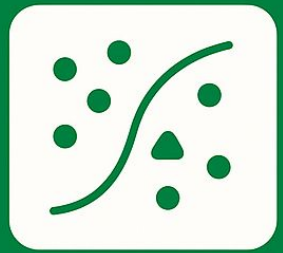
Origen_Nombre	Densidad_Estaciones
(GDL-044) C. Gral. San Martín /Av. Vallarta	0.200568
(GDL-065) C. Simón Bolivar / Av. La Paz	0.208693
(GDL-195) C. Ramón Corona / Av. Juárez	0.218261
(GDL-063) C. Colonias / C. López Cotilla	0.220653
(GDL-046) C. Emerson / Av. Vallarta	0.220797
(GDL-049) Lopez Cotilla/ Marcos Castellanos	0.221375
(GDL-040) C. Morelos / C. Progreso	0.223138
(GDL-187) Av. 16 de Septiembre / C. Priscilia	0.224474
(GDL-043) C. Simón Bolivar / C. López Cotilla	0.225699
(GDL-030) C. Pedro Moreno/ C. Progreso	0.227051

Conclusiones

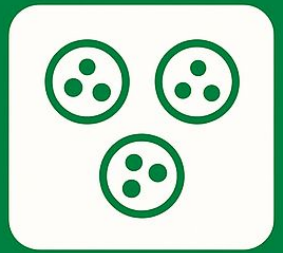
MiBici

BikeShareNetwork

Regression



Classification

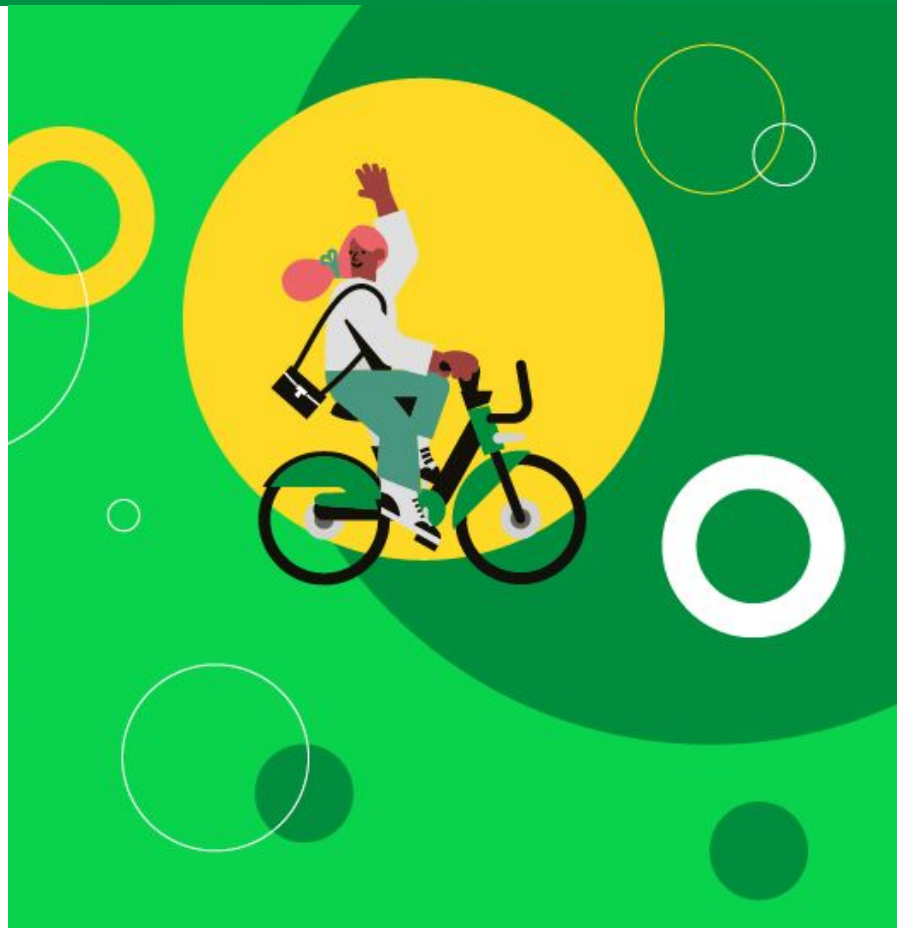


Clustering

ML PROJECT

Conclusiones

- El modelo SARIMA resultó efectivo para predecir la demanda horaria de bicicletas públicas.
- Se identificaron patrones temporales y estacionales consistentes en los datos.
- Esta segmentación facilita la planificación y expansión del sistema en zonas clave.



Recomendaciones

- Implementar redistribución dinámica de bicicletas en horas pico.
- Priorizar estaciones de alta demanda cercanas a centros comerciales y oficinas.
- Diseñar campañas de fidelización para usuarios ocasionales.