Facultad de Ciencias Económicas
Elements of Machine Learning

Proyecto 01
EcoBici CDMX 2019

Daniel Hidalgo
Denisse Bolaños

Introducción

ECOBICI es el sistema de bicicletas públicas de la Ciudad de México, diseñado para promover la movilidad sostenible en zonas urbanas clave. Con cientos de estaciones y miles de viajes diarios, genera datos valiosos sobre hábitos de transporte, distribución geográfica y preferencias de usuarios.

Objetivos específicos:

Aplicar técnicas de machine learning vistas en clase (como PCA para reducción dimensional y análisis de clusters) para:

- Identificar patrones de uso según edad y género.
- Evaluar el impacto de las variables climáticas en la demanda.
- Desarrollar un sistema básico de recomendación de estaciones basado en la proximidad geográfica.

Descripción de Datos y Preprocesamiento

Datos Originales

El análisis se realizó utilizando datos completos correspondientes a los 12 meses del año 2019 (enero-diciembre). Las variables clave consideradas incluyen:

- Genero_Usuario (género del usuario)
- Edad_Usuario (edad del usuario)
- Bici (identificador de bicicleta)
- Ciclo_Estacion_Retiro/Arribo (estaciones de origen/destino)
- Fecha_Retiro/Arribo (fechas de viaje)
- Hora_Retiro/Arribo (horarios de uso)

Selección del Periodo:

Se analizaron específicamente los datos de 2019 por ser el último año con patrones de movilidad previos a la pandemia. Este periodo permitía observar el comportamiento habitual del sistema ECOBICI sin las distorsiones que vendrían después.

La elección también consideró el contexto único de movilidad de ese año, cuando el tráfico en la CDMX alcanzó niveles tan altos que incluso inspiró campañas como la de Burger King atendiendo a automovilistas atrapados en el tráfico, haciendo especialmente relevante el estudio de alternativas como ECOBICI.

Para manejar eficientemente el volumen de datos, se trabajó con una muestra representativa del 20% de los viajes de cada mes, lo que permitió conservar los patrones generales manteniendo un tamaño manejable para el análisis. Este enfoque aseguró la diversidad temporal sin comprometer la capacidad de procesamiento.

Muestra

Debido al peso de cada mes de data, se optó por la construcción de una muestra que refleje un año. Se tomó una muestra equivalente al 20% de cada mes de 2019, para después unirlas y

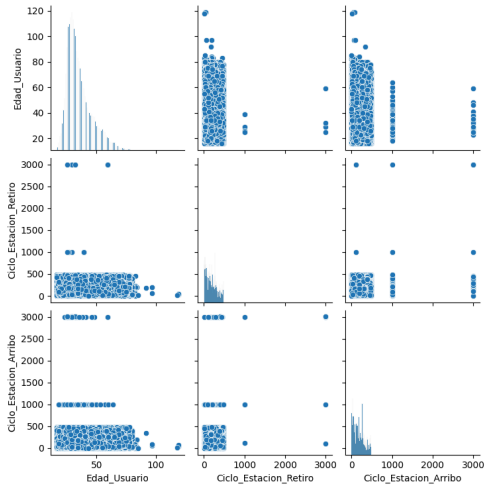
formar un “año sintético”. Esta se complementó con data geográfica de las estaciones de bici existentes.

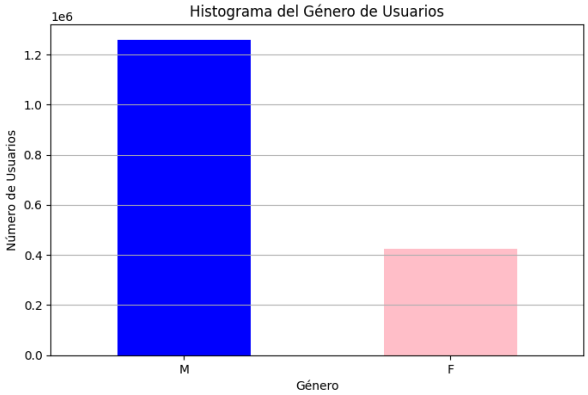
Análisis Exploratorio (EDA)

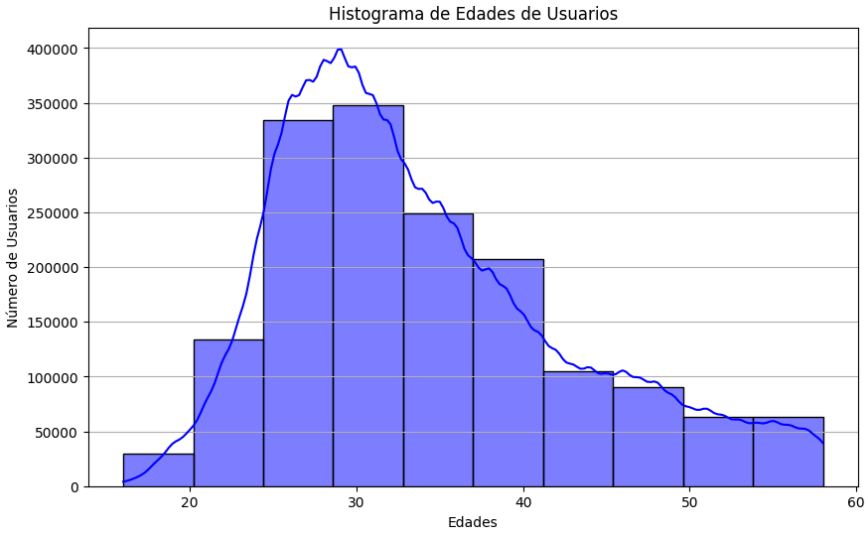
Después del proceso de muestreo, se procedió con el análisis exploratorio. Se empezó limpiando la data al buscar datos nulos, los cuales no existían en la muestra. Se procedió por la estandarización de fechas y horas en el dataset para poder manejarlas adecuadamente durante el proyecto. Finalmente, se identificaron y eliminaron los outliers de la variable edad utilizando el método intercuartil.

Visualización de Patrones

Para el entendimiento de patrones y distribuciones de la data, se realizaron distintos gráficos como histogramas para comprender mejor la base de datos con la que estaríamos trabajando. Más adelante, también realizó un heatmap pero con variables que construimos para el PCA.







Análisis de Comportamiento de Usuario

Segmentación por Edad y Género

Los usuarios se clasificaron en grupos según edad (<20, 20-29, 30-39, 40-49, 50-59, 60+) y género para identificar patrones de uso diferenciados.

Hallazgos Principales

1.Distribución de viajes

- Los usuarios masculinos predominan en todos los grupos de edad.
- Los rangos 20-29 y 30-39 concentran la mayor actividad, con una brecha notable entre géneros (hombres > mujeres).
- La frecuencia de viajes disminuye progresivamente con la edad.

2. Duración de viajes

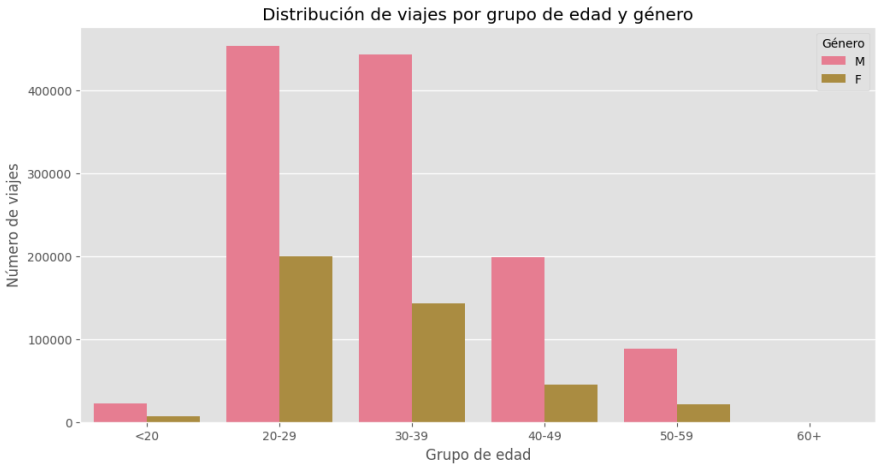
- Se realizó una prueba ANOVA para determinar si existen diferencias significativas en los tiempos promedio de viaje entre grupos etarios y géneros.
- El resultado fue de: F=0.39, p-value=0.8164, lo que indica que no hay diferencias estadísticamente significativas en los tiempos de viaje entre los segmentos analizados.

3. Relación edad-género

- Se realizó una prueba de chi-cuadrado para evaluar la relación entre género y grupo etario en la cantidad de viajes.
- El resultado fue de: chi2=17699.97, p-value=0.0000, lo que indica que existe una asociación significativa entre la edad, el género y la cantidad de viajes.

Visualización de Datos

Se generó un gráfico de barras que muestra la distribución de viajes por grupo de edad y género.



Se observa que los hombres tienen una mayor participación en todos los segmentos, con una diferencia más pronunciada en los grupos 20-29 y 30-39.

Análisis de Estaciones con PCA

El Análisis de Componentes Principales (PCA) se emplea en este estudio para reducir la dimensionalidad de los datos relacionados con las estaciones de Ecobici en la Ciudad de México. Dado que los datos originales pueden contener redundancias o correlaciones entre variables, el uso de PCA permite encontrar combinaciones lineales de las variables originales que explican la mayor variabilidad posible. Esto facilita la interpretación de patrones espaciales y la agrupación de estaciones con características similares.

Componentes principales identificados

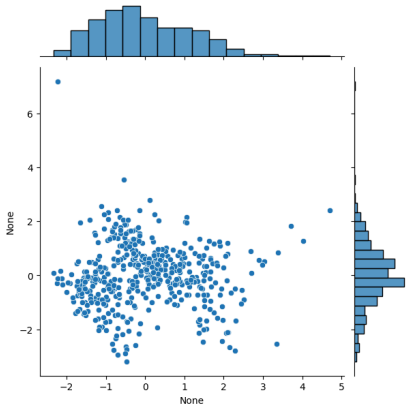
Tras aplicar PCA a los datos de estaciones, se identificaron las siguientes tendencias principales:

- **Primer Componente Principal (PC1):** Explica la mayor parte de la varianza y está fuertemente relacionado con la ubicación geográfica de las estaciones y la cantidad de usos.
- **Segundo Componente Principal (PC2):** Captura diferencias en la distribución de estaciones en función de su frecuencia de uso y proximidad a zonas de alta densidad.

Interpretación espacial de resultados

El gráfico PCA generado muestra una dispersión clara de estaciones en un espacio bidimensional, donde las estaciones con características similares tienden a agruparse. Se identifican las siguientes observaciones clave:

- **Agrupaciones:** Se observan clusters de estaciones con alta demanda, posiblemente ubicadas en zonas como Reforma, Condesa y Polanco.
- **Estaciones atípicas:** Algunas estaciones aparecen como valores atípicos en el espacio PCA, lo que podría indicar ubicaciones con baja demanda o en áreas de expansión del sistema.
- **Distribución geográfica:** La representación bidimensional permite inferir que las estaciones siguen un patrón alineado con la estructura urbana de la ciudad, favoreciendo ciertas direcciones como el eje norte-sur.



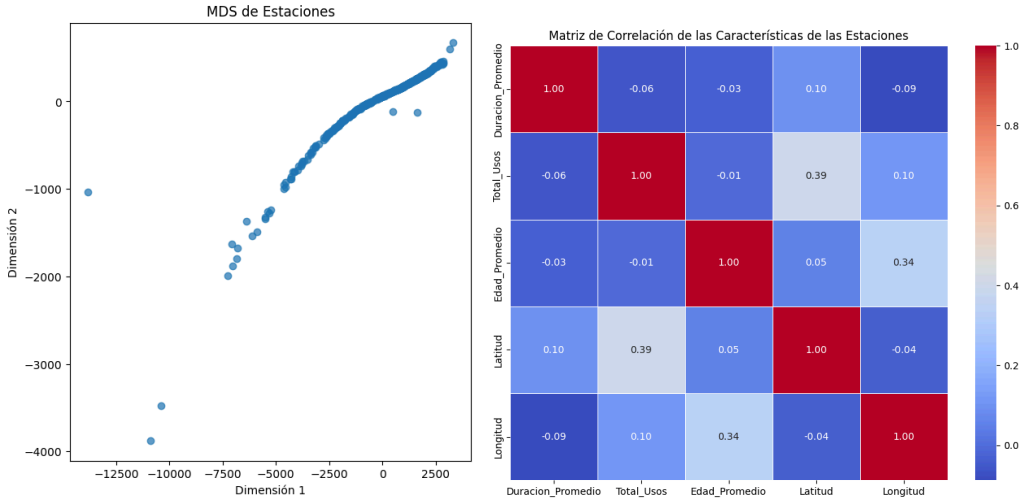
Análisis complementario: MDS y otras métricas

Además del PCA, se ha realizado un análisis de Multidimensional Scaling (MDS) para representar las estaciones en un espacio de menor dimensión basado en sus similitudes. Se identificaron los siguientes puntos clave:

- **Patrón lineal:** Se observa que las estaciones tienden a distribuirse en una estructura alineada, posiblemente reflejando la disposición geográfica de la red de Ecobici a lo largo de ejes principales de la ciudad.
- **Clusters diferenciados:** Al igual que en el PCA, algunas estaciones se agrupan en zonas con alta demanda, mientras que otras aparecen más dispersas.

Además, la matriz de correlaciones mostró una relación positiva moderada (~0.39) entre la latitud de las estaciones y la cantidad de usos, lo que sugiere que ciertas áreas tienen una mayor demanda de bicicletas. En cambio, la edad promedio de los usuarios no mostró una correlación significativa con otras variables del sistema.

Estos hallazgos complementan el análisis PCA y refuerzan la importancia de considerar factores espaciales y de demanda en la planificación y expansión del sistema Ecobici.

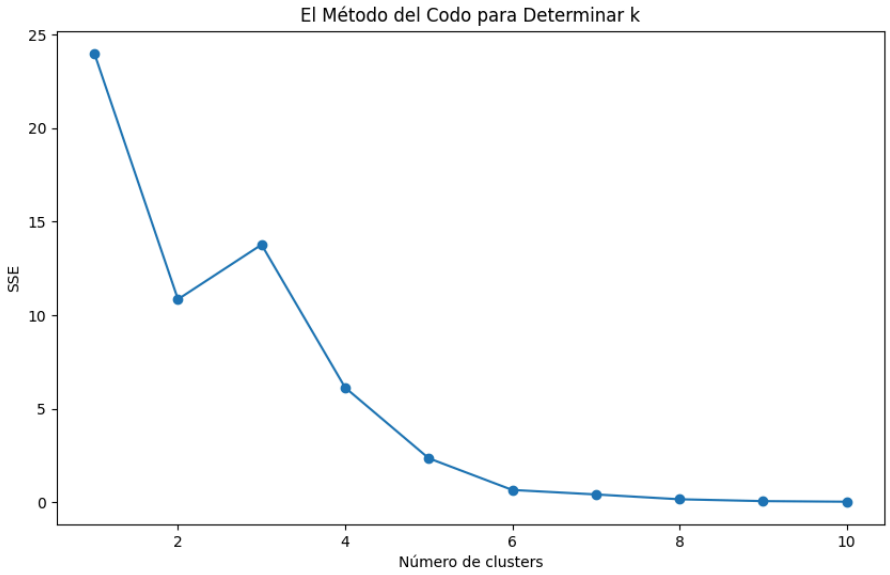


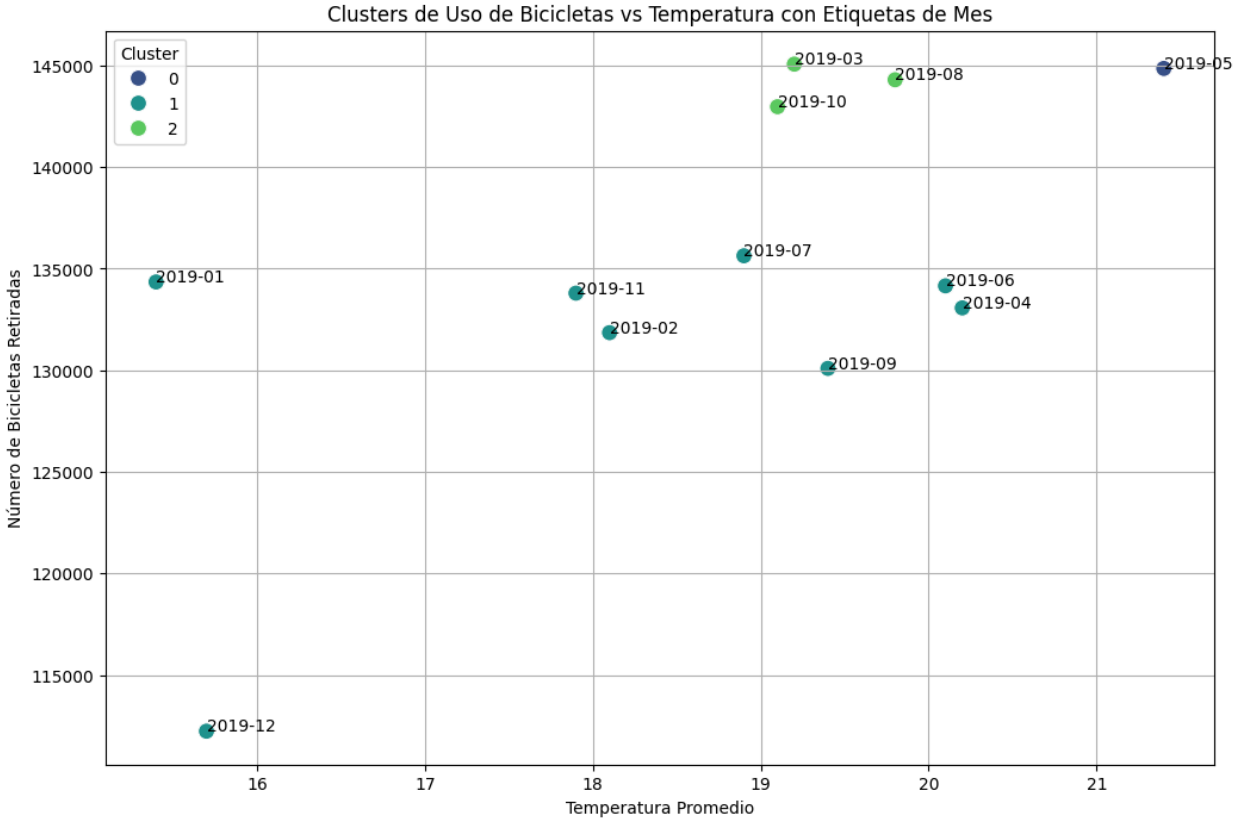
Integración con Variables Climáticas

Con la finalidad de enriquecer el análisis de la data, se buscó implementar variables climáticas con la finalidad de extraer más información de valor. Al principio, se optó por el uso de un API climático. Con la finalidad de eficientizar la cantidad de llamadas al API se empezó con una estrategia donde se guardaría tres temperaturas por cada día del 2019; una para la mañana, tarde y noche. Está después agregaría por rangos a la data antes mencionada. Sin embargo, seguían siendo muchas llamadas. Otra estrategía utilizada fue la incluir solo una temperatura por día. A pesar de esto, el uso del API nos fue imposibilitada por temas de límites financieros en el proyecto.

Finalmente, se optó por un csv del gobierno que contaba la temperatura promedio por mes durante el año 2019. Esto nos ayudó a entender un poco mejor el comportamiento del uso de bicis en CDMX por época del año y temperatura. Este análisis se hizo aplicando el modelo

KMeans. A pesar de ser una solución viable para el proyecto, se recomienda la búsqueda de información más enriquecedora para un análisis climático más exhaustivo.





Sistema de Recomendación

Evolución del Enfoque

El desarrollo del sistema de recomendación pasó por varias iteraciones, adaptándose a los hallazgos obtenidos en cada fase:

Primer Intento:

Se implementó un modelo de Factorización Matricial (NNMF) que consideraba edad, género y bicicleta utilizada. A pesar de los esfuerzos por personalizar las recomendaciones, el modelo no logró capturar patrones significativos, resultando en una precisión de 0.

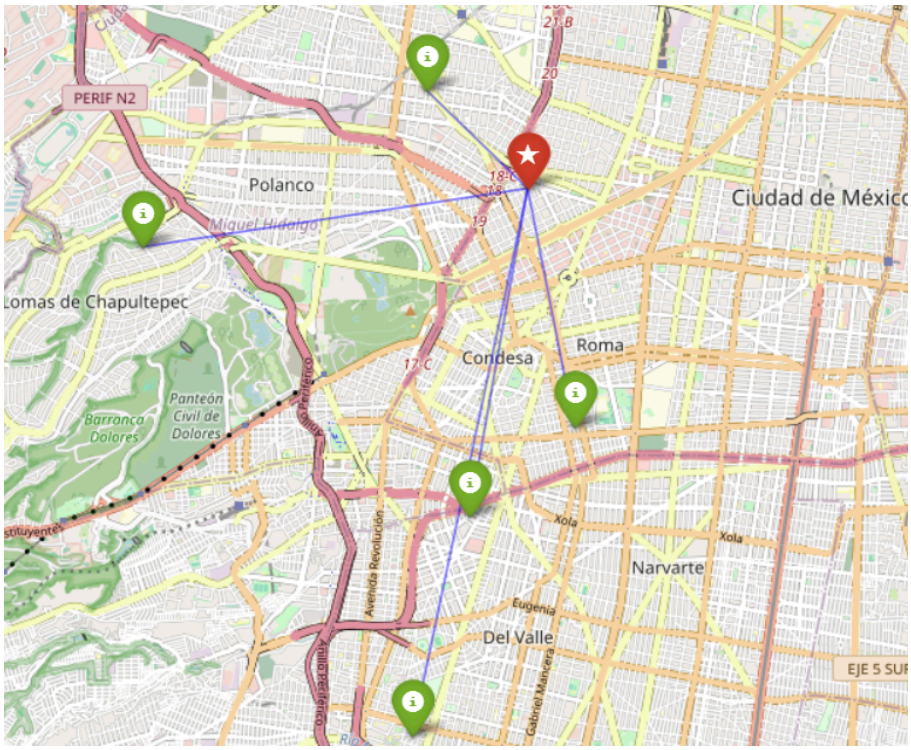
Segundo Intento:

Al eliminar la variable de bicicleta y enfocarse solo en edad y género, se añadieron mejoras como umbrales de uso mínimo y un sistema de respaldo con estaciones populares. Sin embargo, los resultados siguieron siendo insatisfactorios, manteniendo una precisión nula.

Solución Final:

Se reorientó la estrategia hacia un sistema basado en proximidad, implementando dos funciones clave:

- **recomendar_estaciones_cercanas:** Utiliza distancias geográficas para sugerir opciones relevantes.
- **recomendar_por_hora:** Combina proximidad con patrones horarios de uso.



Implementación y Limitaciones
El sistema final se alimentó de datos de usuarios y estaciones, priorizando la utilidad práctica sobre la complejidad algorítmica. Las principales limitaciones incluyeron la inefectividad de los modelos basados en perfiles y la necesidad de adoptar enfoques más directos.

Conclusiones y Recomendaciones

Optimización de Recomendaciones Demográficas
El sistema actual podría mejorarse significativamente mediante:

1.Implementación de un sistema de identificación único:
- Generación de IDs anónimos para cada usuario
- Permite seguimiento individualizado sin comprometer privacidad
- Facilita la identificación de patrones de uso personalizados

2. Incorporación de variables adicionales:
- Zona de residencia y trabajo (geolocalización)
- Frecuencia de uso (usuarios ocasionales vs. habituales)
- Preferencias de rutas basadas en historial

3. Investigación de features clave:
- Análisis de qué variables demográficas impactan más en las preferencias
- Identificación de correlaciones significativas entre características de usuarios y elección de estaciones

Integración de Datos Climáticos y de Infraestructura
Para recomendaciones más precisas:

1.Datos meteorológicos:
- Priorización de estaciones cercanas a transporte público en días lluviosos
- Sugerencia de rutas más largas en condiciones climáticas favorables
- **Limitación actual:** No se encontró una API climática con la precisión y cobertura necesarias

Información de infraestructura:
- Incorporación de datos sobre calidad de ciclovías
- Niveles de seguridad por zona
- Disponibilidad de estaciones