Análisis de Patrones de Movilidad en el Sistema ECOBICI: Un Estudio de Clustering y Reducción de Dimensionalidad Durante el Mes del concierto de Taylor Swift en Ciudad de México

Juan Fernando Ramírez - 20666
Oscar Méndez - 20402
Departamento de Ingeniería
Universidad del Valle de Guatemala
Guatemala, Guatemala

***Abstract*—Este estudio examina los patrones de uso del sistema de bicicletas compartidas ECOBICI en la Ciudad de México durante el mes de agosto de 2023, un periodo marcado por numerosos eventos culturales y sociales. A través de técnicas avanzadas de análisis de datos, incluyendo clusterización y reducción de dimensionalidad, se busca entender las dinámicas de movilidad y ofrecer insights para mejorar la eficiencia y satisfacción del servicio. Se utilizan herramientas de análisis como Pandas, Matplotlib, Seaborn y Scikit-learn para procesar y visualizar los datos, aplicando métodos como PCA, t-SNE, y clustering jerárquico. Los resultados revelan diferencias significativas en los patrones de uso entre días laborables y fines de semana, y una correlación no directa entre la ubicación de las estaciones y la frecuencia de uso. Este análisis proporciona una base valiosa para la toma de decisiones estratégicas en la gestión del sistema ECOBICI, destacando la importancia de considerar múltiples factores en la planificación del transporte urbano sostenible.**

I. Introducción

El uso de sistemas de bicicletas compartidas ha crecido significativamente en muchas ciudades alrededor del mundo como una solución sostenible para el transporte urbano. En la Ciudad de México, el sistema ECOBICI representa una de las iniciativas más importantes en esta área, ofreciendo un medio de transporte alternativo que promueve la movilidad sostenible y reduce la congestión vehicular. Este estudio se centra en analizar los patrones de uso del sistema ECOBICI durante el mes de agosto de 2023, un periodo caracterizado por una serie de eventos culturales y sociales que podrían influir significativamente en la movilidad urbana.

El objetivo principal de este estudio es identificar y analizar los patrones de movilidad dentro del sistema ECOBICI, utilizando técnicas avanzadas de análisis de datos como la clusterización y la reducción de dimensionalidad. Estas técnicas permiten no solo entender mejor cómo se comportan los usuarios del sistema, sino también ofrecer insights para la optimización del servicio y la planificación de futuras expansiones o modificaciones.

Metodológicamente, este trabajo emplea varias bibliotecas y herramientas de análisis de datos en Python, incluyendo Pandas para la manipulación de datos, Matplotlib y Seaborn para la visualización, y Scikit-learn para la implementación de técnicas de machine learning. A través del uso de PCA, t-SNE, y otros métodos de reducción de dimensionalidad, junto con algoritmos de clustering como K-Means y Clustering Jerárquico, este estudio desglosa la estructura subyacente de los datos recopilados.

Los resultados revelan patrones interesantes en relación con las preferencias de los usuarios y las diferencias en la utilización del sistema durante los días laborables en comparación con los fines de semana. Además, se observó que las ubicaciones de las estaciones no están necesariamente correlacionadas con la frecuencia de uso, lo que sugiere que otros factores como la accesibilidad y las conexiones con otros modos de transporte podrían jugar un rol más significativo en la determinación de la demanda.

II. Materiales y Métodos

A. Materiales

Los datos utilizados en este estudio provienen del sistema ECOBICI de la Ciudad de México, que emplea el estándar General Bikeshare Feed Specification (GBFS). Los conjuntos de datos incluyen:

- **Datos Históricos de Viajes de ECOBICI**: Detalles de cada viaje realizado durante agosto de 2023.
- **Información de Cicloestaciones**: Datos sobre ubicación y capacidad de cada estación.

B. Métodos

El análisis fue realizado usando Python, con varias bibliotecas para procesamiento de datos, visualización y análisis estadístico y de machine learning:

- **Pandas**: Para la manipulación y limpieza de datos.
- **NumPy**: Para operaciones numéricas avanzadas.
- **Matplotlib** y **Seaborn**: Para visualización de datos.
- **SciPy**: Usada especialmente para cálculos estadísticos avanzados y pruebas.
- **Scikit-learn**: Para técnicas de aprendizaje automático incluyendo:

  - Clustering (K-Means, Spectral Clustering)
  - Reducción de dimensionalidad (PCA)
  - Métodos de transformación (StandardScaler, PowerTransformer)
- **Scikit-learn's Metrics**: Para evaluación de modelos usando métricas como Davies-Bouldin Score y Silhouette Score.
- **Scikit-learn's Manifold**: Para técnicas de reducción de dimensionalidad no lineal y visualización, incluyendo:
  - Locally Linear Embedding (LLE)
  - t-Distributed Stochastic Neighbor Embedding (TSNE)
  - Spectral Embedding
  - Isometric Mapping (Isomap)
  - Multi-Dimensional Scaling (MDS)
- **SciPy's Cluster**: Para análisis jerárquico de clustering, incluyendo la creación de dendrogramas y la asignación de clústeres.

1) **Carga de Datos**:
   - Se cargaron dos conjuntos de datos: `stations.csv` para las estaciones y `ecobici_2023_08.csv` para los viajes.
2) **Preprocesamiento de Datos**:
   - Conversión de las columnas de fecha y hora de retiro y arribo al formato datetime.
   - Cálculo del tiempo total de uso en minutos.
   - Codificación de la variable categórica `Genero_Usuario` a valores numéricos.
   - Eliminación de registros con valores faltantes y conversión de ciertas columnas a tipos numéricos.
3) **Creación del Dataset**:
   - Combinación de los datos de viajes con la información de las estaciones correspondientes utilizando la función `merge`.
   - Agrupación de los datos por la estación de retiro para análisis a nivel de estación, calculando estadísticas agregadas como el número total de viajes y el tiempo promedio de uso por estación.
4) **Exploración de Datos**:
   - Uso de funciones como `head()` y `columns` para visualizar las primeras filas y las columnas de los datasets, respectivamente. Luego se graficaron las coordenadas de las estaciones, la demanda de bicicletas y la horas pico de cada estación
5) **Escalamiento Multidimensional de cantidad de retiros por estación**:
   - **Propósito**:Calculamos la cantidad de bicicletas retiradas por estación por día. A partir de esto calculamos una matirz de correlación entre estaciónes la cual se utilizó para hacer el escalamiento multidimencional.
6) **Clustering y Reducción de Dimensionalidad**:
   - Aplicación de algoritmos de clustering incluyendo:
     - **KMeans**: Utilizado para identificar agrupaciones basadas en las características de los viajes y las estaciones.
     - **Clustering Jerárquico**: Empleado para visualizar y determinar la estructura de clústeres a través de un dendrograma.
     - **Spectral Clustering**: Aplicado para aprovechar las propiedades del espacio generado por la matriz de similitud de los datos.
   - Uso de técnicas de reducción de dimensionalidad como PCA y diversos métodos de embedding (TSNE, Locally Linear Embedding, Spectral Embedding, Isomap, Multi-Dimensional Scaling) para facilitar la visualización y mejorar la interpretación de los datos.
7) **Evaluación de Modelos**:
   - **Análisis de las Métricas de Clustering** para evaluar la efectividad de los agrupamientos generados, incluyendo:
     - **Inercia**: Mide la suma de las distancias al cuadrado entre cada objeto del clúster y el centro de su clúster. Cuanto menor sea el valor, mejor será la agrupación.
     - **Índice de Davies-Bouldin**: Evalúa la separación entre clústeres; un valor más bajo indica una mejor separación entre los clústeres.
     - **Coeficiente de Silueta**: Mide cómo de similar es un objeto a su propio clúster comparado con otros clústeres. Los valores cercanos a +1 indican que el objeto está bien emparejado con su propio clúster y mal emparejado con los vecinos.
     - **Índice de Dunn**: Mide la relación entre las distancias mínimas interclúster y las máximas distancias intraclúster. Un valor mayor indica una mejor agrupación.
   - **Visualización de Clusters con Boxplots**:
     - Para analizar la distribución interna de los clusters y entender mejor la variabilidad dentro de cada cluster formado, se utilizaron boxplots. Estos gráficos permitieron visualizar la dispersión de los datos dentro de los clusters, identificando claramente cualquier outlier y la densidad de los datos.
     - Los boxplots fueron especialmente útiles para comparar las características de los clusters en varias dimensiones, como el tamaño del cluster, la duración promedio de los viajes en bicicleta y otras variables claves del estudio.
   - Estas técnicas de visualización y análisis permiten una evaluación profunda de la coherencia y la homogeneidad de los clusters, facilitando la interpretación de los resultados del análisis de clustering y ayudando en la toma de decisiones para futuras estrategias de análisis.

III. Resultados

A. Exploración de Datos

Se realizaron varias visualizaciones para explorar la distribución y las relaciones entre las distintas variables del conjunto de datos.

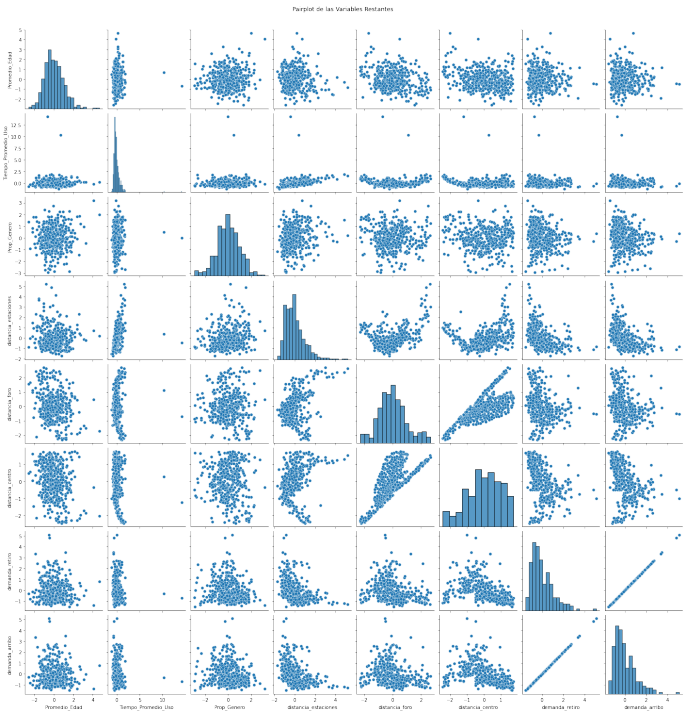

Fig. 1. Relaciones entre diferentes variables representadas a través de un pairplot.

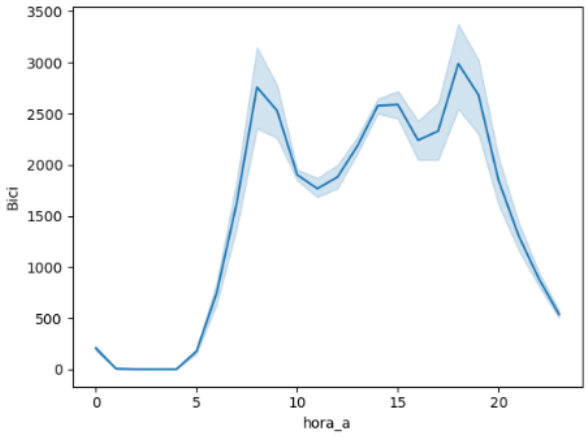

Fig. 2. Demanda de bicicletas durante las horas pico.

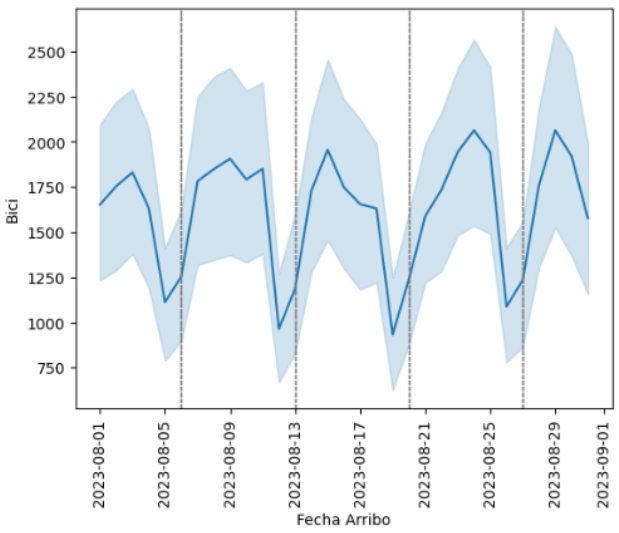

Fig. 3. Visualización de la demanda general de bicicletas.

B. Escalamiento Multidimensional de cantidad de retiros por estación

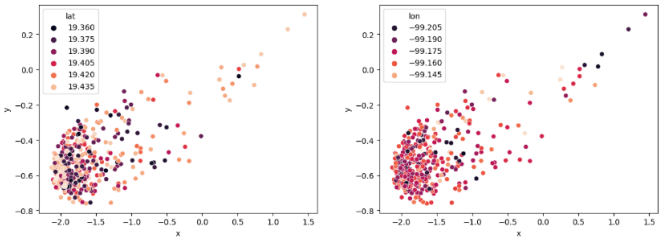

Fig. 4. Escalamiento Multidimensional .

C. Reducción de Dimensionalidad y Proyecciones

Aplicación de técnicas de reducción de dimensionalidad para la proyección de los datos en espacios de menor dimensión, facilitando la interpretación de las estructuras y patrones complejos.

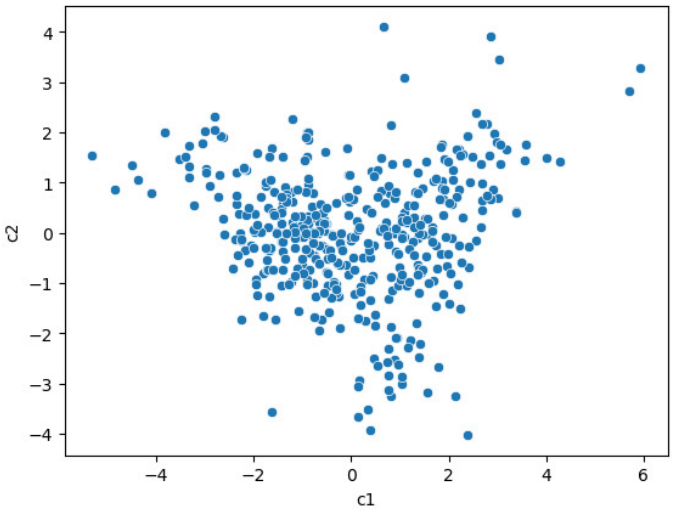

Fig. 5. Proyección de los datos utilizando Análisis de Componentes Principales (PCA).

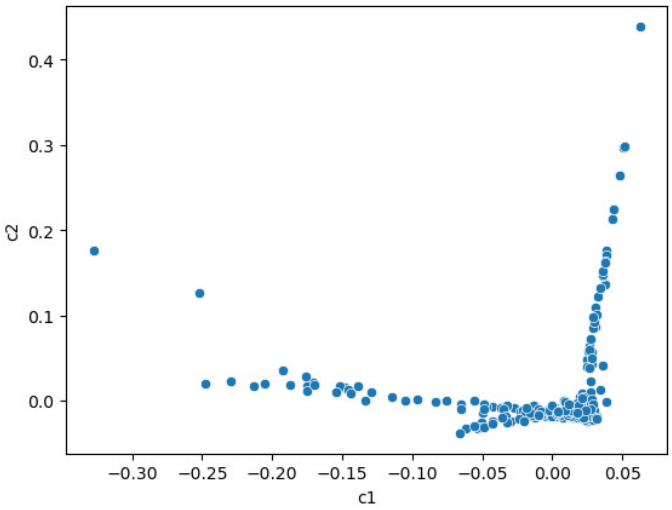

Fig. 6. Embedding localmente lineal de los datos.

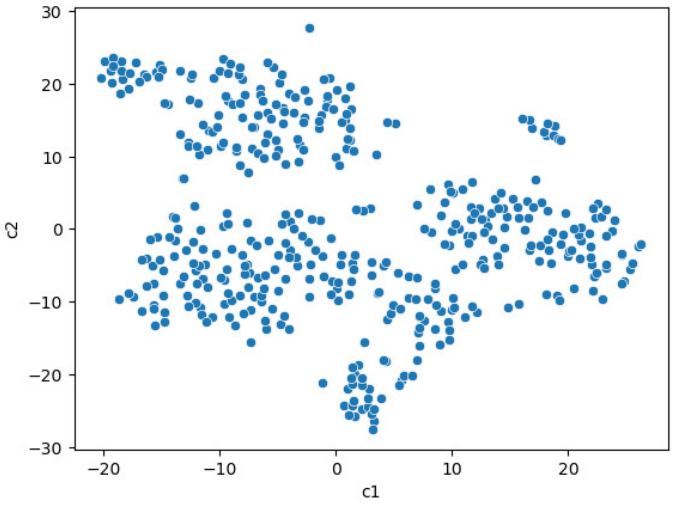

Fig. 9. Técnica de t-SNE para la visualización de los datos de alta dimensión.

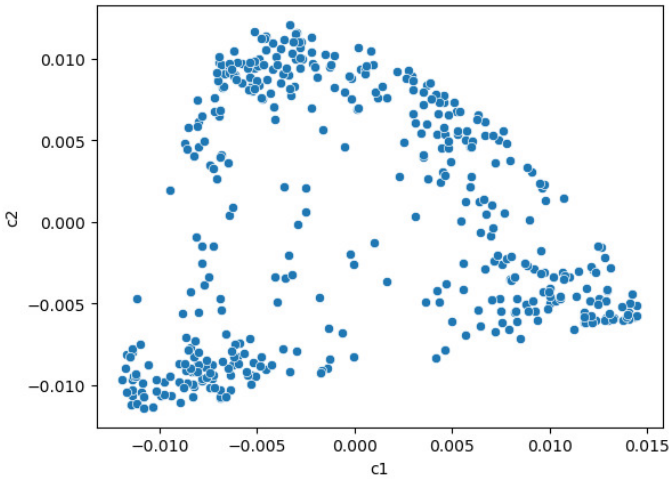

Fig. 7. Embedding espectral de los datos.

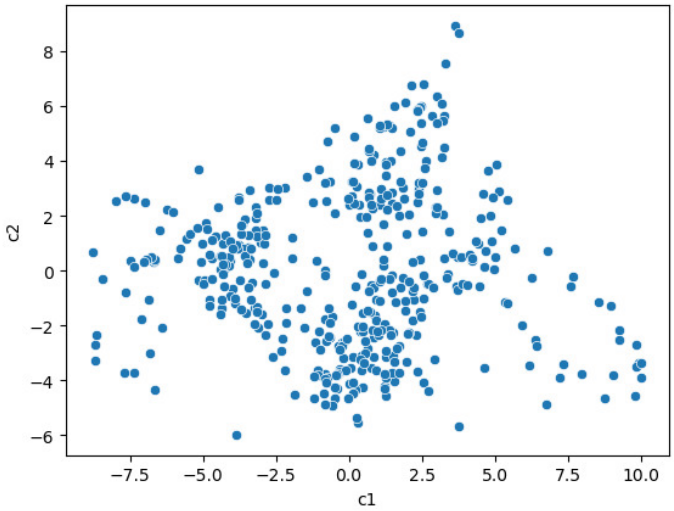

Fig. 8. Reducción de dimensionalidad utilizando Isomap.

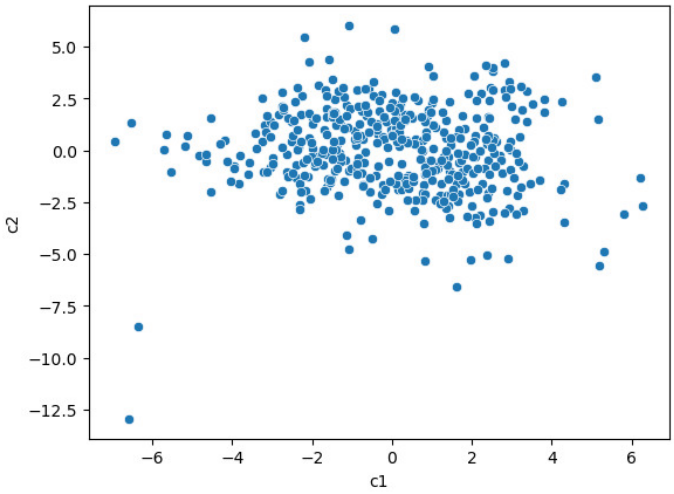

Fig. 10. Escalamiento Multidimensional (MDS) aplicado al conjunto de datos.

D. Clusters

Visualización de los clústeres identificados a través de diferentes técnicas de clustering y proyecciones.

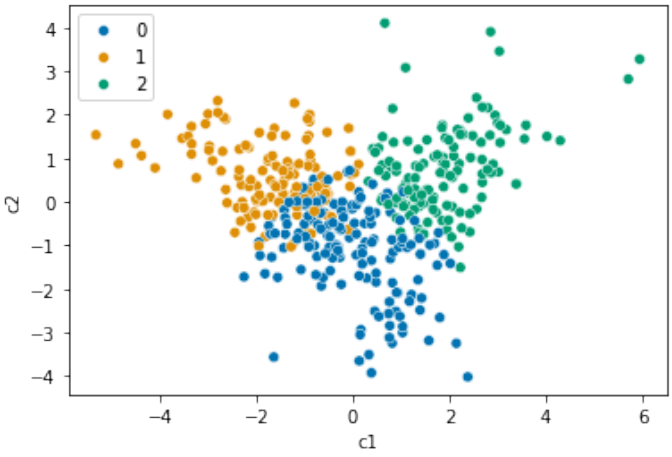

Fig. 11. Clústeres identificados por el algoritmo KMeans y visualizados con PCA.

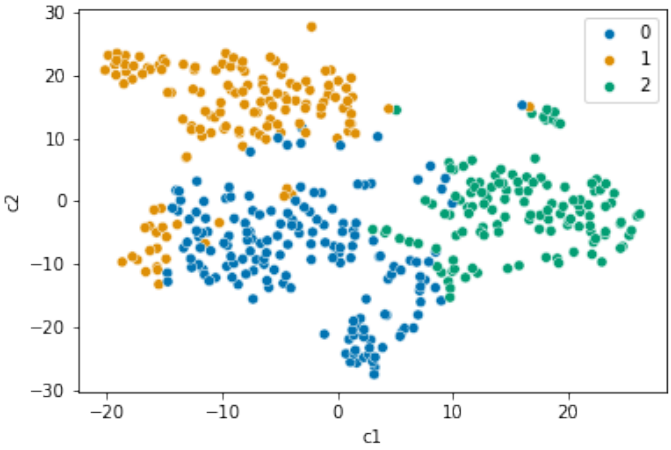

Fig. 12. Clústeres identificados por KMeans visualizados utilizando t-SNE.

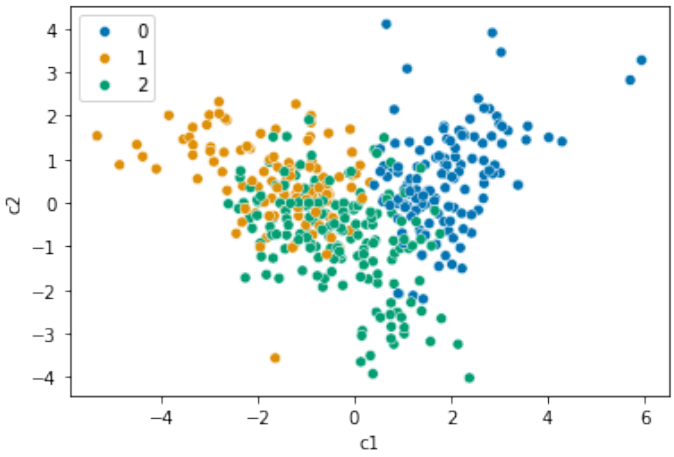

Fig. 13. Clústeres identificados mediante clustering espectral y visualizados con PCA.

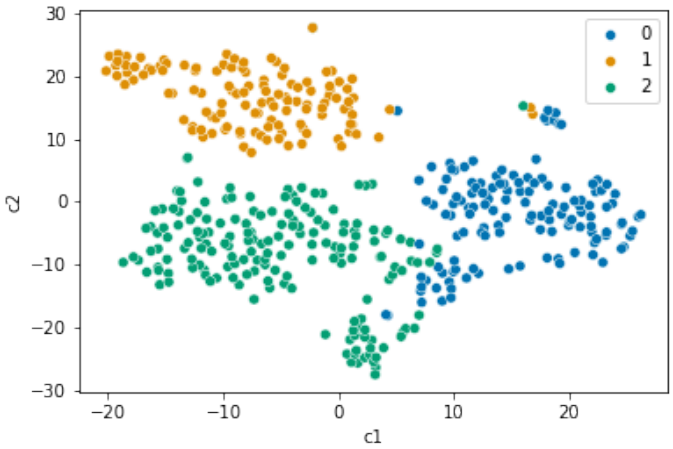

Fig. 14. Clústeres identificados mediante clustering espectral y proyectados con t-SNE.

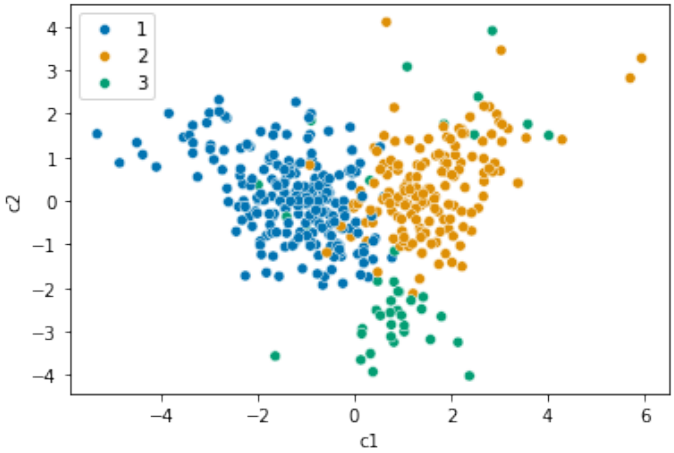

Fig. 15. Clústeres resultantes del clustering jerárquico proyectados con PCA.

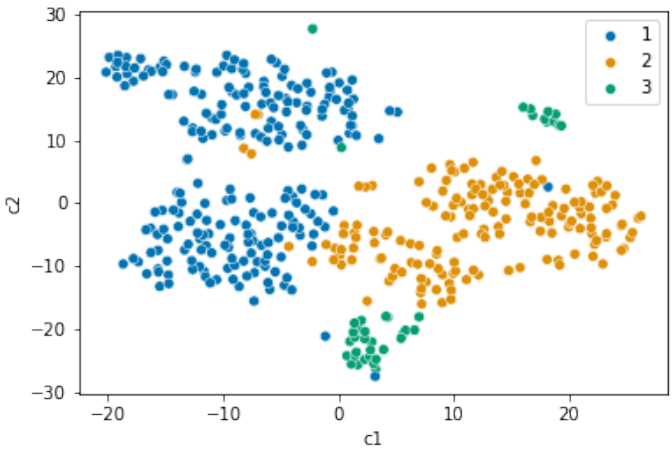

Fig. 16. Clústeres resultantes del clustering jerárquico proyectados con t-SNE.

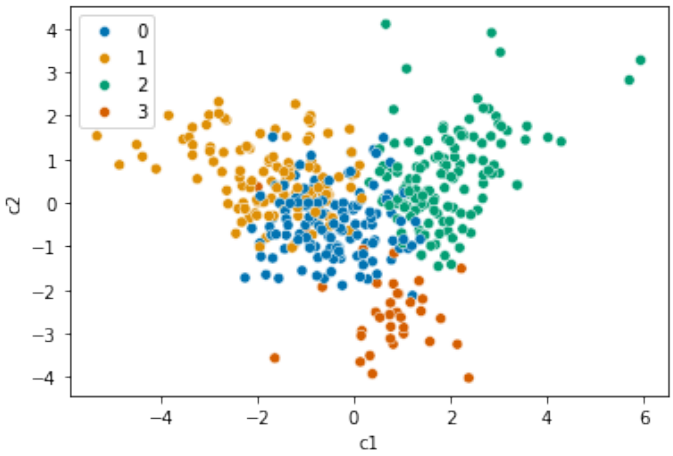

Fig. 17. Una segunda versión de clústeres KMeans proyectada con PCA.

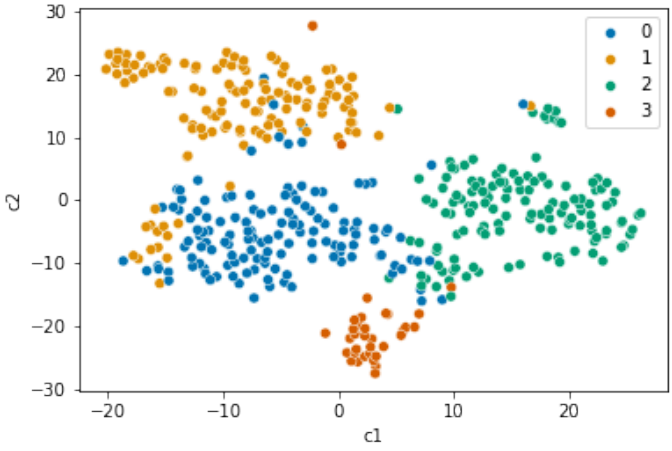

Fig. 18. Una segunda versión de clústeres KMeans visualizada con t-SNE.

IV. Discusión

A. Análisis Exploratorio de Datos

Durante la exploración de los datos, se destacaron patrones significativos. En la Figura 2, se muestra la cantidad de bicicletas utilizadas diariamente a lo largo del mes de agosto de 2023. Se observa que el servicio de ECOBICI se utiliza más como medio de transporte regular durante los días laborables en comparación con los fines de semana, lo que sugiere una orientación más funcional que recreativa del sistema. La Figura 3 revela las horas pico que coinciden con los horarios convencionales de entrada y salida laboral, con un tercer pico notable alrededor de las 3 p.m., posiblemente atribuible a la salida de escuelas y universidades, o a la hora de almuerzo. Además, las visualizaciones de pares de variables del dataset de estaciones de bicicletas revelan una correlación directa entre la cantidad de retiros y arribos, así como una relación no lineal entre la distancia de las estaciones al foro y al centro de la ciudad, lo cual es coherente con el patrón de uso observado.

B. Escalamiento Multidimensional

El Escalamiento Multidimensional (MDS) se realizó utilizando como base las correlaciones entre la cantidad de retiros de bicicletas por día en cada estación. La técnica de MDS ilustra la similitud en los patrones de demanda de retiros entre estaciones; es decir, estaciones con comportamientos diarios similares se representan cercanas entre sí en la visualización. Adicionalmente, se asignaron colores según la latitud y longitud en los gráficos para evaluar la posible correlación de la ubicación geográfica con el volumen de retiros. Los resultados indican que la ubicación no está directamente relacionada con la demanda diaria de bicicletas, ya que estaciones con comportamientos similares pueden estar geográficamente distantes, o viceversa.

C. Reducción de Dimensiones y Proyecciones

1) PCA: A través del Análisis de Componentes Principales (PCA), se evidenciaron relaciones lineales con las variables. La primera componente principal se correlaciona negativamente con las distancias al centro y al foro, así como con la distancia entre las estaciones de retiro y arribo, y positivamente con la demanda de retiros y arribos. La segunda componente principal muestra una correlación positiva con las distancias de retiro y con la demanda.

2) MDS: El MDS no logró diferenciar claramente los grupos, presentando una nube de puntos. Sin embargo, la primera componente está correlacionada positivamente con la distancia entre estaciones y negativamente con las distancias al centro y al foro, así como con las demandas de retiro y arribo. La segunda componente no muestra una correlación significativa con ninguna variable.

3) t-SNE: El algoritmo t-SNE sí logró diferenciar grupos en el espacio de proyección. Ambas componentes muestran relaciones con las variables de estudio, aunque estas no son lineales, lo que sugiere la presencia de patrones más complejos en los datos.

4) Isomap: Isomap no consiguió agrupar los datos en clústeres distintos. Sin embargo, se observa que la primera componente principal tiene relaciones significativas con varias variables, similar a las reveladas por PCA.

5) Spectral Embedding: El Spectral Embedding logró separar los datos en grupos distintos, acumulando ciertas estaciones en una región específica del gráfico y dispersando las demás a lo largo de dos ejes. Las relaciones con las variables son complejas pero discernibles, y hay una conexión con las coordenadas geográficas.

6) Locally Linear Embedding: El Locally Linear Embedding dividió los datos en tres líneas distintas, lo que puede indicar una separación visual útil para el análisis. Las relaciones con las variables son no lineales y siguen patrones similares a los de la proyección bidimensional.

D. Clusterización

1) K-Means: Usando el método K-Means, se determinó que el número óptimo de clústeres es tres. Los boxplots por variable, que se pueden encontrar en los anexos, permiten entender mejor las características de cada clúster. El clúster 0 está compuesto por estaciones con una distancia media entre puntos de retiro y arribo y con una menor cantidad de bicicletas utilizadas. El clúster 1 se caracteriza por su distancia del centro y del foro y una baja utilización de bicicletas, similar al clúster 0. Por último, el clúster 2 agrupa estaciones cercanas al centro y al foro, con una alta frecuencia de uso.

2) Spectral Clustering: El Spectral Clustering asignó clústeres de manera eficaz en concordancia con las agrupaciones observadas en la visualización t-SNE. La correspondencia de los clústeres con las proyecciones t-SNE y PCA indica una división coherente de los datos. Adicionalmente podemos ver que la características de cada grupo son básicament las mismas que en las de K-Means.

3) Clustering Jerárquico: El Clustering Jerárquico presentó resultados distintos a los obtenidos con K-Means y Spectral Clustering, sugiriendo la existencia de dos grupos principales y un tercer grupo menos definido. Los boxplots revelan que el clúster 1 agrupa estaciones distantes tanto del centro como del foro y con una baja utilización, mientras que el clúster 2 reúne estaciones con una alta frecuencia de uso y que no estan tan lejos del centro y el foro. Finalmente el clúster 3 agrupa los que están más cerca del centro y del foro pero que no se usan tanto.

4) K-Means con Transformaciones: Se aplicó una variante de K-Means con transformaciones previas, utilizando Box-Cox para normalizar las distribuciones y PCA para reducir la dimensionalidad, considerando componentes que acumulaban el 95% de la varianza explicada. Los resultados fueron similares a los obtenidos sin transformaciones, lo que indica que estas no tuvieron un impacto significativo en la efectividad del método dado que las distribuciones originales ya eran similares a las normales.

V. Conclusiones

1) **Utilización del Servicio ECOBICI**: Los resultados del análisis exploratorio indican claramente que el servicio ECOBICI se utiliza predominantemente como medio de transporte durante los días laborables, con una disminución notable en su uso durante los fines de semana. Esta tendencia sugiere que las políticas y estrategias de promoción del servicio podrían enfocarse en mejorar y facilitar aún más su uso durante los días laborables, considerando también estrategias específicas para incrementar la utilización recreativa los fines de semana.
2) **Impacto de la Ubicación Geográfica**: A través del Escalamiento Multidimensional (MDS), se demostró que la ubicación geográfica de las estaciones no influye directamente en los patrones de demanda diaria, lo que destaca la importancia de otros factores, como la accesibilidad y las conexiones con otros modos de transporte, sobre la simple proximidad a puntos de interés como el centro o el foro.
3) **Eficiencia de las Técnicas de Reducción de Dimensiones**: Las técnicas como PCA y t-SNE fueron efectivas para identificar y visualizar las relaciones y agrupaciones subyacentes en los datos. En particular, t-SNE ayudó a diferenciar claramente los grupos basados en la proximidad de las características de las estaciones, lo que puede ser útil para entender mejor las dinámicas de uso y mejorar la gestión del servicio.
4) **Relevancia del Clustering en la Operativa del Servicio**: La aplicación de diferentes técnicas de clustering reveló patrones significativos de uso, destacando cómo las características de las estaciones influyen en su desempeño. El clustering jerárquico, por ejemplo, sugirió la existencia de grupos con características operativas distintas, que podrían requerir enfoques de gestión diferenciados. Esto subraya la utilidad de estas técnicas para la toma de decisiones estratégicas en la planificación y mejora del sistema de bicicletas compartidas.

Estas conclusiones no solo subrayan la utilidad de las técnicas de análisis de datos en el contexto de sistemas de transporte público compartido, sino que también proporcionan una base para futuras investigaciones y mejoras en la gestión de ECOBICI, orientadas a maximizar su eficacia y satisfacción del usuario.

References

[1] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Disponible en: https://scikit-learn.org/stable/modules/clustering.html

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Disponible en: https://scikit-learn.org/stable/modules/manifold.html

[3] P. Virtanen et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020. [Online]. Disponible en: https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[6] F. Pedregosa et al., ”Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Disponible en: https://scikit-learn.org/stable/modules/model_evaluation.html