Identificación de patrones de omisión de paradas en conductores del SITP mediante técnicas de clustering no supervisado

Elaborado por:

Wilson Arley Sarmiento Hernández

Programa: Especialización en Machine Learning

William Alberto Rodríguez Cruz

Programa: Especialización en Gerencia de Proyectos

Universidad Ean

Escuela de Formación en Investigación

Seminario de Investigación de Postgrado

Bogotá

14/05/2025

Contenido

Contenido de Figuras

Contenido tablas

1. Resumen

Esta investigación nace a partir de una situación recurrente que enfrenta el Consorcio Express en Bogotá operador del sistema de transporte público. Muchos usuarios han manifestado a través de sus PQR que algunos conductores omiten paradas durante el recorrido generando afectaciones en la calidad del servicio y provocando molestia en los pasajeros. Esta realidad motivó la necesidad de analizar los datos operativos del sistema con el fin de identificar patrones que permitan entender este comportamiento y tomar decisiones informadas para mejorar el servicio.

La base del proyecto está en reconocer que, aunque existen grandes volúmenes de información sobre el funcionamiento diario de las rutas aún no se aprovechan del todo para explicar o anticipar las situaciones que los usuarios denuncian. Por eso se propone construir una herramienta analítica que facilite el reconocimiento de tendencias relacionadas con las omisiones de paradas y los momentos en que ocurren con mayor frecuencia. Para lograrlo se aplican métodos de análisis de datos con un enfoque exploratorio. Primero se utilizan técnicas de aprendizaje no supervisado como el agrupamiento o clustering que permiten descubrir comportamientos similares entre rutas horarios y puntos específicos de abordaje. Después de los resultados del clustering, se podrán incorporar modelos de regresión o predicción para evaluar el posible comportamiento que pueda llegar a tener los operadores. La intención final es aportar información clara y útil que apoye la toma de decisiones del Consorcio Express en su compromiso con un mejor servicio al usuario

2. Problema de investigación

El Sistema Integrado de Transporte (SITP) en Bogotá tiene diferentes concesiones para la prestación del servicio público de pasajeros. Consorcio Express es uno de los concesionarios y es en el que se desarrolla esta investigación. Las causas principales son la disminución de ingresos para la compañía por la disminución de pasajeros transportados por las rutas asociadas al concesionario (Transmilenio S.A., 2024). El origen principal es la percepción de los usuarios por la omisión de las paradas en los paraderos establecidos para cada ruta. Esto se identifica por medio de las reclamaciones de los mismos usuarios al concesionario directamente o por medio del ente regulador Transmilenio (Chaparro Rivera, 2019).

En algunas ocasiones, la no parada en los paraderos establecidos puede ocurrir porque el vehículo alcanza la capacidad total de usuarios a transportar por ese vehículo (Secretaría Distrital de Movilidad, 2023). En otras situaciones, es solo percepción de los usuarios, ya que no logran identificar bien los vehículos de cada concesionario y posiblemente toman datos de vehículos que no son (Rodríguez, 2019). Si la situación continúa sin una intervención con evidencias, la percepción de los usuarios y la confianza sobre el sistema de transporte público de Bogotá decaería. También podría desincentivar el uso del transporte público, y para las concesiones prestadoras del servicio, se verían afectadas con sanciones por el no cumplimiento de los indicadores de servicio establecidos por el ente gestor (TransMilenio S.A., 2000).

Viendo este panorama, se propone un análisis de los datos operativos de la actividad de cada vehículo y de quién lo opera. Esto se hará por medio de la recolección de los datos obtenidos mediante los sistemas SIRCI de cada vehículo, y con los cuales se podrán realizar diferentes análisis, como la agrupación de datos. Esto nos permitirá identificar aquellos vehículos u operadores con comportamientos diferentes a la media del

comportamiento de los demás, respecto a las validaciones o ascensos de pasajeros en cada ruta y parada (Alcaldía Mayor de Bogotá, 2021). Con esta identificación, se podrá entrar a realizar un estudio adicional del comportamiento o del por qué tiene menos pasajeros o validaciones, para tomar decisiones internas para cada uno (Alcaldía Mayor de Bogotá, 2018).

¿Qué impacto tiene la omisión de paradas de los conductores en la cantidad de pasajeros transportados y como se puede identificar patrones de comportamiento en la omisión de paradas permitiendo predecir el cumplimiento del servicio?

3. Objetivo general

Analizar cuál es el impacto que genera la omisión de las paradas de los conductores de Consorcio Express, identificando patrones con técnicas de aprendizaje automático no supervisado relacionado con la cantidad de pasajeros transportes.

4. Objetivos específicos

a. Identificar tendencias en la omisión de paradas a partir del análisis de datos operativos del SITP, considerando variables como paradero, ruta, operador, cantidad de pasajeros transportados.

b. Desarrollar un modelo descriptivo basado en técnicas de clustering no supervisado, con K-Means, para clasificar a los conductores según su comportamiento en la recolección de pasajeros.

c. Implementar un modelo de visualización en Power BI con los resultados obtenidos en el modelo.

d. Realizar un modelo de regresión predictiva con técnicas como Regresión Lineal o Random Forest Regressor para estimar la probabilidad de omisión de paradas en función de variables operativas.

5. Justificación

La movilidad en Bogotá actualmente representa un desafío sobre la calidad de vida de los ciudadanos y usuarios del sistema. En este contexto, el sistema integrado de transporte público (SITP), especialmente en el componente zonal, se convierte en un actor fundamental en la conexión entre los barrios y las principales vías troncales. Sin embargo, la problemática de la omisión de paradas por parte de los conductores afecta la experiencia de los usuarios, la accesibilidad y el cumplimiento de los objetivos del sistema.

Viéndolo desde la perspectiva de convivencia, este estudio buscar por medio de los datos generar conocimiento sobre el comportamiento operativo de los conductores de Consorcio Express, Entender porque y como se pueden generar las omisiones de paradas esto permitirá avanzar en poder tomar decisiones mucho más informadas con evidencia que puede servir de base para implementar estrategias o técnicas dentro de la organización o del sistema.

La relevancia social del estudio radica en el impacto que tendrá en los ciudadanos. Cada omisión de paradas representa la oportunidad perdida de un usuario de acceder al sistema de transporte y el poder llegar a tiempo a su lugar de destino. Estas situaciones afectan a los ciudadanos que dependen del transporte público como su único medio de movilidad. Investigar este fenómeno desde un enfoque técnico y analítico es también un compromiso con el derecho a un transporte digno, eficiente y accesible.

En cuanto a las implicaciones técnicas, aunque la investigación no está orientada a una propuesta de mejoras operativas directas, si es posible abrir la posibilidad de comprender

el problema a profundidad. El análisis de patrones mediante las técnicas de machine learning permitirá identificar posibles tendencias de omisión de paradas en ciertas rutas y paraderos por los operadores, así como operadores con perfiles específicos que afectan la calidad del servicio, este conocimiento es fundamental para las áreas de planeación, programación, operaciones y control.

El valor teórico se encuentra en la articulación de los enfoques de movilidad, comportamiento humano de los operadores y analítica de datos. Este estudio se nutre de investigaciones internacionales, integrando teorías sobre eficiencia operativa en el transporte público y técnicas de análisis de datos no supervisados, con esto también se aporta al desarrollo académico entre la tecnología, transporte y gestión urbana.

Para finalizar, la metodología de clustering aplicado y sustentado en datos reales del sistema, además de ser innovadora, fortalece las competencias en el análisis de datos, estadística aplicada y modelamiento de datos en contextos reales.

Este proyecto se enmarca en el campo de Ciencia, Tecnología e innovación. Dentro del grupo de investigación tecnológico Ontare, y en la línea de Optimización de procesos, enfocado con los lineamientos institucionales de la Universidad.

6. Marco teórico

Sistemas de Transporte Público Urbano

Características y desafíos de los sistemas de transporte masivo

Los sistemas de transporte público son pilares fundamentales en la estructura urbana, especialmente en ciudades con alta densidad poblacional como Bogotá. Gestionar estos sistemas implica enfrentar retos constantes relacionados con la eficiencia en la operación, la calidad ofrecida y, por supuesto, la satisfacción de quienes los utilizan a diario. Algunos de los desafíos más importantes incluyen ajustar la oferta a los patrones cambiantes de movilidad de los ciudadanos, administrar de manera óptima los recursos disponibles, como la flota de vehículos y el personal, asegurar que se cumplan los estándares de calidad y puntualidad, y reducir al mínimo problemas operativos, como la omisión de paradas designadas.

Estudios realizados en ciudades con dinámicas parecidas a las de Bogotá, como es el caso de Tianjin, nos enseñan algo clave: la calidad del transporte público no depende solo de la infraestructura que vemos, los buses o las estaciones. También influyen, y mucho, cómo actúan los conductores y cómo es la ciudad misma con sus desafíos diarios (Pang et al., 2023).

Y aquí es donde entra un problema bien conocido: cuando un bus omite una parada. Esto no es solo un contratiempo; es un fallo en la operación que afecta directamente cómo la gente percibe el servicio y, desde luego, su confianza en él.

El Sistema Integrado de Transporte Público (SITP) de Bogotá

El SITP de Bogotá representa un verdadero desafío para ordenar y mejorar el funcionamiento del transporte en la ciudad. Este sistema está organizado en tres componentes principales, siendo el primero el componente Troncal, donde circulan aquellos buses articulados y biarticulados, los más grandes, que vemos moverse por carriles exclusivos en las principales avenidas. El segundo componente es el Zonal, formado por buses de menor tamaño que tienen la misión de recorrer los barrios y facilitar la conexión con las troncales, acercando el servicio a zonas más residenciales y por último el complementario que ayuda a conectar con zonas de difícil acceso con el sistema principal a través del servicio TransMiCable

La eficiencia del componente zonal es vital para la movilidad en Bogotá. Su buen funcionamiento depende de factores como minimizar los tiempos de espera, optimizar el uso de la flota y asegurar que se cumplan los estándares de calidad, lo cual incluye, de manera fundamental, la detención en todas las paradas programadas.

Comportamiento de los Conductores en Sistemas de Transporte Público

Factores que influyen en el comportamiento de los conductores

El modo en que los conductores operan los vehículos de transporte público está sujeto a una variedad de influencias. Estas se pueden agrupar en tres grandes categorías:

1. **Factores Operativos:** Aquí encontramos elementos clave como la forma en que se programan las rutas, los tiempos que se establecen para completar los recorridos, cuántos pasajeros puede llevar cada bus y, por supuesto, las

condiciones del tráfico en la ciudad. Por ejemplo, cuando hay mucha congestión vehicular, esto no solo aumenta la cantidad de emisiones que producen los buses urbanos, sino que también puede influir en las decisiones que toman los conductores durante su jornada (Rosero et al., 2023).

2. **Factores Individuales:** Estos tienen que ver directamente con cada conductor como persona: qué tanta experiencia tiene detrás del volante, qué tan buena ha sido la capacitación que ha recibido, qué lo motiva a hacer bien su trabajo y, en general, cuál es su actitud frente a sus responsabilidades laborales.
3. **Factores Contextuales:** Se refieren a todas esas condiciones externas que el conductor enfrenta día a día, como el estado en que se encuentran las calles y avenidas, qué tan congestionada está la ciudad en diferentes momentos, y las particularidades de cada zona por donde pasan las rutas que debe cubrir.

Cuando analizamos cómo estos factores se relacionan entre sí, y lo hacemos apoyándonos en técnicas modernas de aprendizaje automático, podemos descubrir patrones interesantes en la forma de conducir que realmente afectan qué tan eficiente es el servicio en general, tal como lo señalan Anil y Anudev (2022). Si nos enfocamos específicamente en el problema de los conductores que se saltan las paradas, vemos que hay varias posibles razones detrás de este comportamiento: puede ser que el bus ya esté completamente lleno y no pueda recibir más pasajeros, que el conductor sienta la presión de cumplir con unos horarios demasiado ajustados, que no haya recibido la capacitación adecuada o simplemente no le importe lo suficiente, o incluso que factores externos como los trancones o bloqueos en las vías lo estén afectando.

Patrones de conducción y su impacto en la calidad del servicio

Los estilos o patrones de conducción en el transporte público han sido un área de interés investigativo. Warren et al. (2019) resaltan cómo estos patrones se relacionan directamente con indicadores clave de la calidad del servicio, tales como la puntualidad, la seguridad y la percepción general del usuario. Por ejemplo, un estudio en Hong Kong empleó técnicas de agrupamiento para identificar distintos patrones de conducción asociados a diferentes tipos de rutas (urbanas congestionadas, interdistritales y expresas), demostrando que el comportamiento varía según el contexto operativo (Tong & Ng, 2021). Entender estas variaciones es fundamental, ya que permite analizar el comportamiento de los conductores de manera contextualizada, lo cual podría ser muy útil para comprender la omisión de paradas en las diversas condiciones urbanas de Bogotá.

Técnicas de Aprendizaje Automático para el Análisis de Datos de Transporte

Métodos de clustering no supervisado

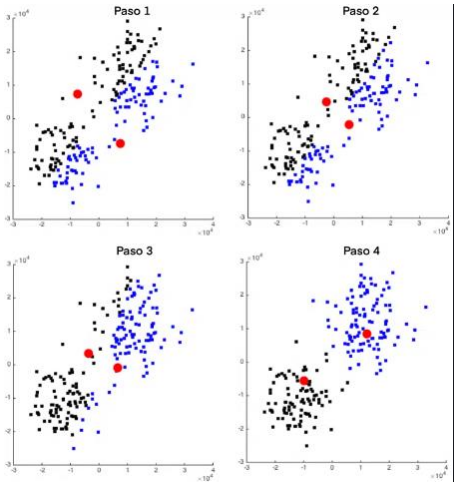
El aprendizaje no supervisado, y en particular las técnicas de *clustering* o agrupamiento, ha probado ser valioso para descubrir patrones ocultos en los datos masivos generados por los sistemas de transporte. Para esta investigación, algunos de los métodos más pertinentes son:

1. **K-Means Clustering:** Este es un algoritmo ampliamente utilizado para segmentar datos en grupos que comparten características similares. Se ha aplicado en el análisis de transporte, por ejemplo, para definir escenarios de tráfico a partir de datos de sensores y perfiles de conducción de autobuses. En este estudio, K-Means podría permitir agrupar a los conductores basándose en similitudes en su

comportamiento operativo, como la frecuencia con que omiten paradas o la cantidad de pasajeros que recogen, facilitando la identificación de perfiles de conducta. El proceso básico implica seleccionar K centroides iniciales, asignar cada dato al centroide más cercano, recalcular los centroides como el promedio de los puntos de cada grupo, y repetir hasta que los grupos se estabilicen

Figura 1

Proceso de agrupamiento con K-Means.



Nota. Imagen tomada de Gamco (2022). ¿Qué es clustering? Disponible en https://gamco.es/wp-content/uploads/2022/03/que-es-clustering-768x819.png.webp

El algoritmo K-Means funciona siguiendo estos pasos:

- Inicialización: Se seleccionan K puntos como centroides iniciales (donde K es el número de clusters deseado).
- Asignación: Cada dato se asigna al centroide más cercano, formando clusters provisionales.
- Actualización: Se recalculan los centroides como el promedio de todos los puntos asignados a cada cluster.

- Iteración: Se repiten los pasos 2 y 3 hasta que los centroides se estabilicen o se alcance un número máximo de iteraciones.

  En el marco de esta investigación, el método K-Means nos permitiría agrupar a los conductores en diferentes categorías según cómo se comportan cuando se trata de saltarse paradas y recoger pasajeros. Esto nos ayudaría a identificar tanto los patrones más comunes como aquellos que son inusuales.

2. **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise): Este método tiene una ventaja interesante sobre K-Means: no necesitamos decirle de antemano cuántos grupos queremos formar. Además, es capaz de encontrar agrupaciones con formas que no son necesariamente regulares e incluso puede identificar puntos que no encajan bien en ningún grupo (lo que llamamos ruido u outliers). Estas características lo hacen especialmente útil cuando trabajamos con datos de transporte, donde los patrones de comportamiento no siempre siguen líneas rectas o son fáciles de detectar.
   DBSCAN funciona estableciendo una especie de vecindario alrededor de cada punto de datos (basándose en un radio ε y un número mínimo de puntos llamado MinPts) para ir formando los grupos. Entre sus ventajas está que puede detectar comportamientos que se salen de lo normal y que es flexible en cuanto a la forma que pueden tener los grupos que identifica. Investigadores como Wang y sus colegas (2022) han combinado DBSCAN con K-Means++ para analizar cómo conducen las personas, lo que demuestra lo útiles que pueden ser estos enfoques basados en la densidad de los datos.

Figura 2

Funcionamiento del algoritmo DBSCAN

Funcionamiento del algoritmo DBSCAN.

Nota. Guía del algoritmo de agrupación DBSCAN tomado de https://www.datacamp.com/es/tutorial/dbscan-clustering-algorithm

El algoritmo DBSCAN se basa en dos parámetros principales:

ε (epsilon): Define el radio de vecindad alrededor de cada punto.

MinPts: Número mínimo de puntos requeridos dentro del radio ε para formar un cluster.

Las principales ventajas de DBSCAN para esta investigación son:

Detección de outliers: Identifica conductores con comportamientos atípicos.

Flexibilidad en la forma de los clusters: No asume que los grupos deben tener forma esférica.

No requiere especificar el número de clusters: Útil cuando no se conoce a priori cuántos patrones de comportamiento existen.

Modelos de regresión predictiva

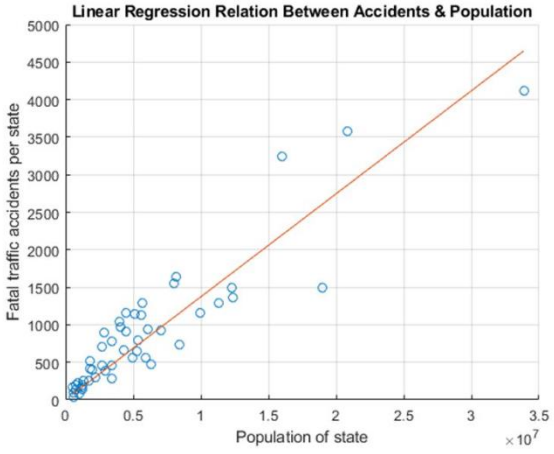
Los modelos de regresión son un complemento perfecto para nuestro análisis de agrupamiento, ya que nos permiten predecir valores continuos, como cuántos pasajeros puede transportar un bus o qué tan probable es que un conductor se salte una parada. Entre los métodos que más nos pueden ayudar están:

Regresión Lineal

La regresión lineal nos ayuda a entender cómo se relacionan ciertas variables predictoras (pensemos en características específicas de la ruta, la hora del día en que circula el bus, o incluso el historial del conductor) con lo que queremos predecir (como la frecuencia con que se omiten paradas). Lo bueno de este método es que, además de ser sencillo, nos permite interpretar fácilmente los resultados, lo que es clave para establecer conexiones de causa-efecto entre cómo opera el sistema y cómo se comportan los conductores.

Figura 3

Representación Gráfica de Regresión Lineal



Nota. Imagen tomada de Dzenan Hamzic (Big Data Science & Software Engineering), disponible en https://dzenanhamzic.com/2016/07/25/linear-regression-with-one-variable-in-matlab/

En términos matemáticos, la regresión lineal se expresa con esta ecuación:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$

Donde:

(**y**) es lo que queremos predecir (por ejemplo, cuántos pasajeros llevará el bus)

β_0 es el punto de partida o intercepto

$\beta_1, \beta_2, ..., \beta_n$ son los valores que nos dicen cuánto influye cada variable predictora

$x_1, x_2, ..., x_n$ son nuestras variables predictoras (como la hora del día o qué día de la semana es)

(**ε**) representa el margen de error que siempre existe

¿Por qué elegir la regresión lineal? Tiene varias ventajas:

Podemos entenderla fácilmente: Los coeficientes nos muestran claramente qué variables son más importantes y si su efecto es positivo o negativo.

Es sencilla: No hace falta ser un experto para implementarla y comprender sus resultados.

No consume muchos recursos: A diferencia de modelos más complejos, funciona bien incluso con computadoras modestas.

7. Metodología

a. Primer nivel

i. Enfoque y alcance de la investigación

En esta investigación se adopta un enfoque cuantitativo, ya que se utilizan y se trabaja con datos provenientes de los sistemas SIRCI de los vehículos los cuales obtienen registros operativos del componente zonal del SITP. El análisis está centrado en el procesamiento de grandes volúmenes de datos para poder extraer los patrones mediante algoritmos de aprendizaje automático no supervisado esto sin intervención directa sobre las variables, el objetivo es observar y realizar una caracterización del comportamiento de los operadores a partir de la información histórica lo cual refuerza el objetivo de la investigación.

ii. Diseño de la investigación

Para la investigación se adopta un diseño no experimental, transversal y de un tipo descriptivo correlacional.

- **No Experimental:** porque no se manipulan las variables, sino que se observan y analizan comportamientos de los registros con datos reales.
- **Transversal:** esto porque los datos corresponden a un periodo especifico que comprende de los días hábiles del mes de febrero y marzo del 2025.
- **Descriptivo-Correlacional:** porque se busca identificar patrones de comportamiento por medio de clústeres y poder relacionarlos con variables como cantidad de pasajeros por ruta, paradero y operadores, aunque no se

establece efectos directos si se identifican relaciones consistentes entre las variables operativas.

iii. Definición de variables

A continuación, se presenta la descripción conceptual y operacional de las principales variables utilizadas en el modelo.

Tabla 1

Descripción Conceptual y Operacional de las Variables del Modelo

Variable	Definición Conceptual	Definición Operacional	Dimensiones Principales
Pasajeros	Número total de validaciones (ascensos) registradas por el sistema SIRCI en cada parada	Medida directa obtenida del archivo de validaciones este valor se suma por operador, paradero, ruta y fecha	Cantidad
Ruta	Recorrido operativo asignado a un vehículo para el servicio de transporte de pasajeros	Variable categórica extraída del registro se usa para agrupar datos por contexto operativo	Código de línea
Código de Paradero	Identificador único asignado a cada punto de ascenso y descenso de pasajeros del sistema SITP	Se utiliza como variable de agrupación para calcular el desempeño y frecuencia por ubicación	Ubicación
Conductor	Persona encargada de operar el vehículo en cada viaje y para cada ruta	Se usa como variable para agrupar datos y asignar puntuación relativa según el clúster	Identificador

Frecuencia Total	Número de veces que un conductor pasa por un mismo paradero durante el periodo analizado	Se calcula por agrupación conductor-ruta-paradero y se utiliza como variable de entrada al modelo de clustering	Repetición
Índice Relativo	Medida estandarizada del desempeño de un conductor frente al promedio del contexto operativo	Se calcula con una fórmula tipo z-score diferencia entre su valor y la media del grupo dividido por la desviación estándar.	Desempeño por ruta y paradero
Clúster (grupo)	Clasificación automática generada por el algoritmo K-Means según similitud de variables operativas.	Resultado del modelo de aprendizaje automático donde se etiquetan 3 grupos, bajo, medio y alto desempeño relativo.	Segmento de desempeño
Latitud y Longitud	Coordenadas geográficas del paradero	Se agrega por agrupación de datos para análisis con clustering geoespacial	Ubicación geografica

Nota. Esta tabla presenta las variables clave utilizadas en el modelo, sus definiciones y cómo fueron medidas u obtenidas.

iv. Población y muestra

La población objeto de esta investigación está compuesta por los registros operativos de validaciones de ascensos de pasajeros registrados por cada uno de los vehículos de Consorcio Express empresa concesionaria del componente zonal del SITP en Bogotá con dos contratos asociados, San Cristóbal y Usaquén, cuenta con 12 patios operacionales y un total de 94 rutas operativas y un aproximado de 23.912.322 registro durante 1 año estos registros son las agrupaciones por fecha, paradero, ruta y operador porque los

registros de las validaciones únicamente en Consorcio Express para el año 2024 fue de un total de 123.346.805 validaciones.

Tamaño de la población: 3.985.387 registros únicos correspondientes al periodo entre el 1 de febrero y el 31 de marzo de 2025 únicamente para los días hábiles.

Cobertura: Rutas, paraderos y operadores activos durante ese periodo.

Muestreo: Se utilizó un censo de aproximadamente el 16% del total de los registros, dado que el análisis requiere observar el comportamiento global de todos los operadores en las distintas condiciones operativas ya sea rutas o frecuencias diferentes.

Justificación: Según la disponibilidad del archivo consolidado (pasajeros_total.parquet) escogió utilizar los registros de estos dos meses y de los días hábiles ya que la estacionalidad de frecuencia de pasajeros es estable por el retorno de vacaciones de fin de año de la gran mayoría de la población como estudiantes y empleados con el fin de no sesgar el análisis de patrones especialmente en zonas o rutas con baja frecuencia.

b. Segundo nivel

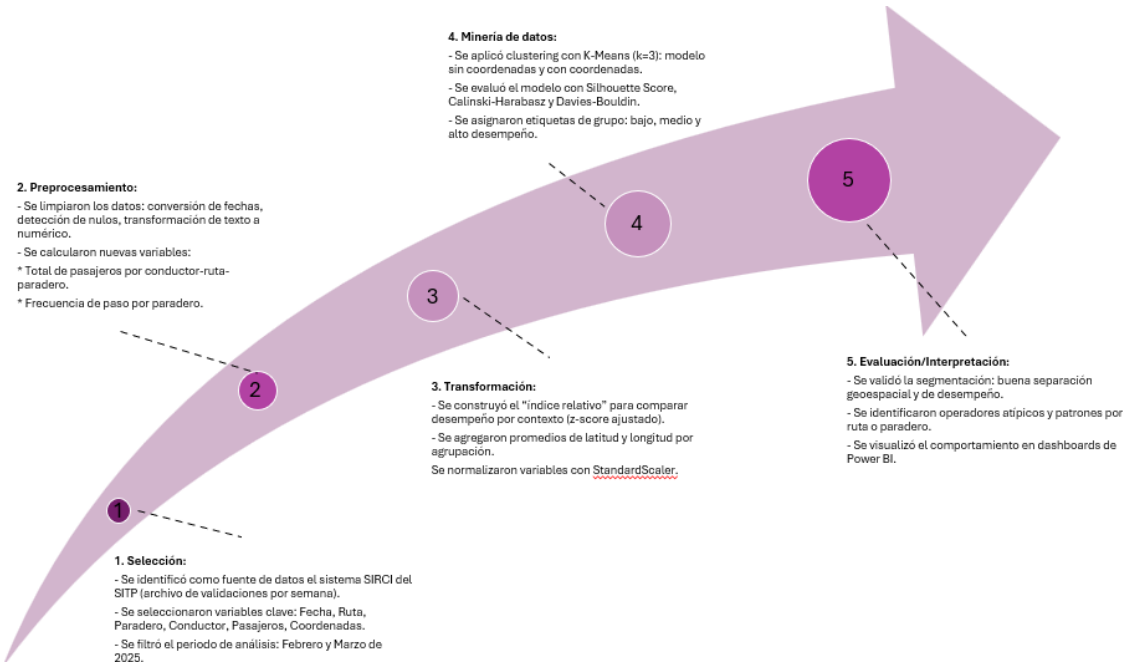
i. Enfoque metodológico en CRISP-DM

Con el fin de estructurar el proceso para realizar el análisis de los datos se adoptó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) como guía principal, el cuál es utilizado en proyectos de minería de datos y ciencia de datos ya que proporciona una ruta clara desde comprender el problema hasta la implementación de soluciones basada en datos.

Las fases de la metodología fueron contextualizadas en el análisis de la de la investigación sobre la omisión de paradas.

Figura 4

Fases desarrolladas para el tratamiento de los datos basado en la metodología CRISP-DM



Nota: Elaboración propia

ii. Selección de instrumentos y modelos

Este estudio de enfoque cuantitativo se sustenta en el uso de instrumentos digitales de recolección masiva de datos, sistema de información del SITP y las bases exportadas desde estos sistemas.

Se utilizaron archivos procesados con anterioridad por la gran cantidad de registro de las validaciones de ascensos de pasajeros y actividad de los vehículos que contienen variables como:

- Fecha de la validación del pasaje
- Ruta y Paradero
- Código del conductor
- Cantidad de Pasajeros
- Coordenadas del Paradero

Posteriormente se consolidaron en un único archivo maestro (pasajeros_total.parquet) el cual sirvió de insumo para las etapas de limpieza, transformación y modelado.

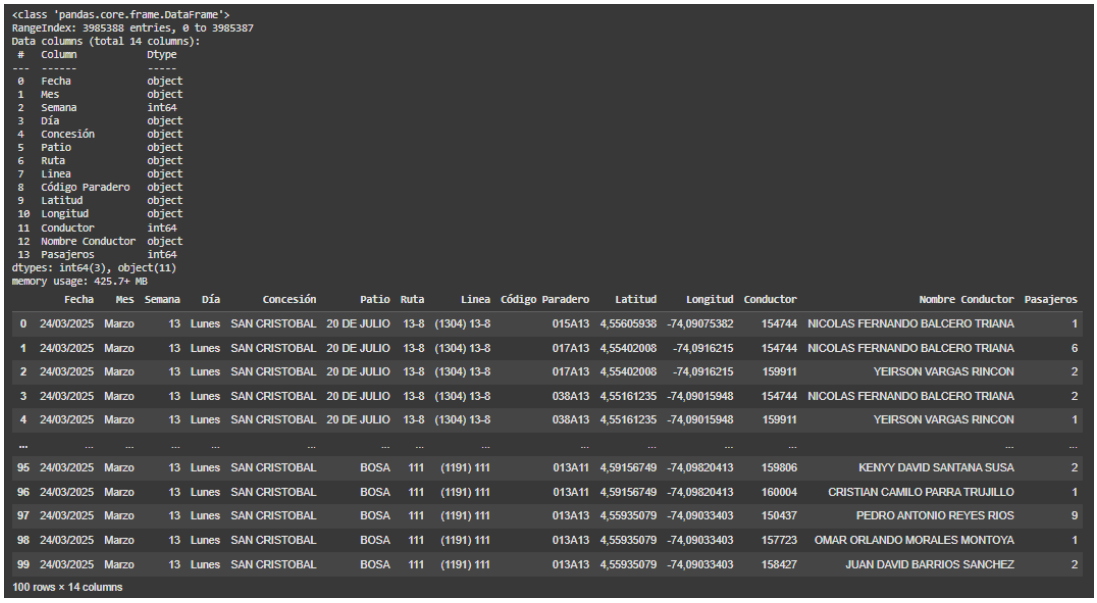
Como instrumento de modelado se utilizó el algoritmo K-means, el cual fue validado para clasificar registros en grupos con comportamiento similares sin necesidad de etiquetas dadas con anterioridad.

Técnicas de análisis de datos:

Una vez se realizó la preparación de los datos se aplicaron las siguientes técnicas con instrumentos o modelos.

Figura 5

Flujo del Sistema SIRCI para la Recolección Automática de Datos de Ascensos por Paradero, Ruta y Fecha



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3985388 entries, 0 to 3985387
Data columns (total 14 columns):
 #   Column            Dtype
---  ------            -----
 0   Fecha             object
 1   Mes               object
 2   Semana            int64
 3   Día               object
 4   Concesión         object
 5   Patio             object
 6   Ruta              object
 7   Linea             object
 8   Código Paradero   object
 9   Latitud           object
 10  Longitud          object
 11  Conductor         int64
 12  Nombre Conductor  object
 13  Pasajeros         int64
dtypes: int64(3), object(11)
memory usage: 425.7+ MB
```

	Fecha	Mes	Semana	Día	Concesión	Patio	Ruta	Linea	Código Paradero	Latitud	Longitud	Conductor	Nombre Conductor	Pasajeros
0	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	20 DE JULIO	13-8	(1304) 13-8	015A13	4,55605938	-74,09075382	154744	NICOLAS FERNANDO BALCERO TRIANA	1
1	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	20 DE JULIO	13-8	(1304) 13-8	017A13	4,55402008	-74,0916215	154744	NICOLAS FERNANDO BALCERO TRIANA	6
2	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	20 DE JULIO	13-8	(1304) 13-8	017A13	4,55402008	-74,0916215	159911	YEIRSON VARGAS RINCON	2
3	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	20 DE JULIO	13-8	(1304) 13-8	038A13	4,55161235	-74,09015948	154744	NICOLAS FERNANDO BALCERO TRIANA	2
4	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	20 DE JULIO	13-8	(1304) 13-8	038A13	4,55161235	-74,09015948	159911	YEIRSON VARGAS RINCON	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	BOSA	111	(1191) 111	013A11	4,59156749	-74,09820413	159806	KENYY DAVID SANTANA SUSA	2
96	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	BOSA	111	(1191) 111	013A11	4,59156749	-74,09820413	160004	CRISTIAN CAMILO PARRA TRUJILLO	1
97	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	BOSA	111	(1191) 111	013A13	4,55935079	-74,09033403	150437	PEDRO ANTONIO REYES RIOS	9
98	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	BOSA	111	(1191) 111	013A13	4,55935079	-74,09033403	157723	OMAR ORLANDO MORALES MONTOYA	1
99	24/03/2025	Marzo	13	Lunes	SAN CRISTOBAL	BOSA	111	(1191) 111	013A13	4,55935079	-74,09033403	158427	JUAN DAVID BARRIOS SANCHEZ	2

100 rows × 14 columns

Nota: Elaboración propia.

Figura 6

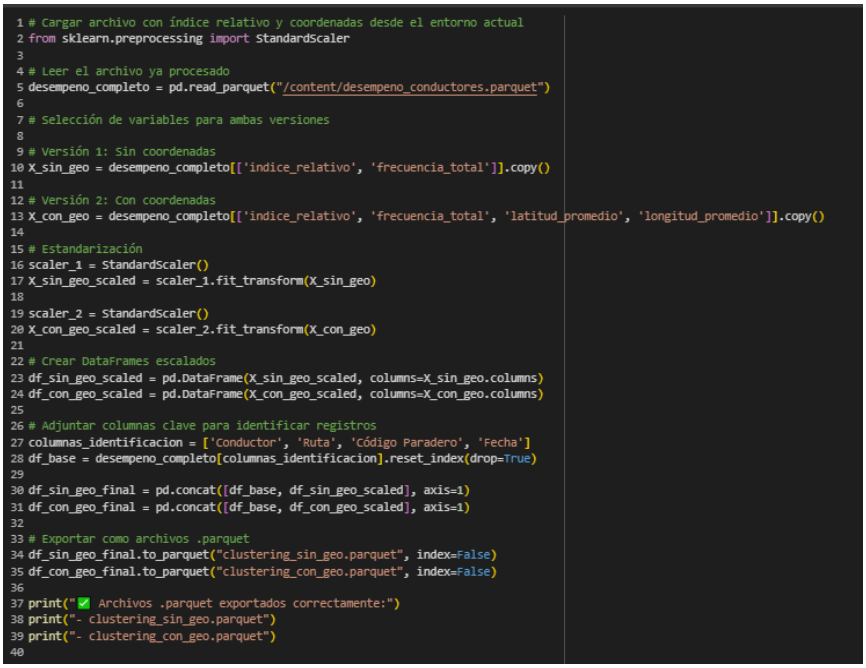
Proceso de Limpieza, Agregación de Variables y Creación del Índice Relativo a Partir del Archivo Pasajeros_total.parquet

	Conductor	Ruta	Código Paradero	Fecha	pasajeros_conductor	promedio_general	desviacion_general	indice_relativo	latitud_promedio	longitud_promedio
0	100120	BA900	021A00	2025-02-06	5	4.777778	3.032234	0.073287	4.656973	-74.070889
1	100120	BA900	022A00	2025-02-06	3	1.714286	0.951190	1.351691	4.645869	-74.072408
2	100120	BA900	023A00	2025-02-06	5	2.750000	1.488048	1.512048	4.644132	-74.072861
3	100120	BA900	025A01	2025-02-06	1	1.875000	1.125992	-0.777093	4.748458	-74.032414
4	100120	BA900	028A01	2025-02-06	11	4.000000	3.214550	2.177599	4.750562	-74.044063
5	100120	BA900	030A01	2025-02-06	4	3.000000	1.732051	0.577350	4.713825	-74.047507
6	100120	BA900	031A01	2025-02-06	1	1.200000	0.447214	-0.447214	4.707342	-74.048771
7	100120	BA900	032A01	2025-02-06	11	3.375000	3.420004	2.229529	4.707672	-74.048379
8	100120	BA900	037A00	2025-02-06	2	2.000000	0.816497	0.000000	4.642057	-74.075190
9	100120	BA900	058A01	2025-02-06	1	4.125000	3.044316	-1.026503	4.729159	-74.045299
10	100120	BA900	077A01	2025-02-06	9	4.400000	3.062316	1.502131	4.740739	-74.043363
11	100120	BA900	116A01	2025-02-06	10	4.200000	3.224903	1.798504	4.724606	-74.046225
12	100120	BA900	117A00	2025-02-06	6	4.875000	3.440826	0.326956	4.658608	-74.069805
13	100120	BA900	117A01	2025-02-06	10	5.000000	3.944053	1.267731	4.732175	-74.044950
14	100120	BA900	118A00	2025-02-06	9	5.000000	3.605551	1.109400	4.667838	-74.063871
15	100120	BA900	151A01	2025-02-06	11	4.666667	3.614784	1.752064	4.747236	-74.024679
16	100120	BA900	169A00	2025-02-06	11	3.200000	3.326660	2.344694	4.669771	-74.064294
17	100120	BA900	170A00	2025-02-06	1	1.200000	0.447214	-0.447214	4.649209	-74.072989
18	100120	BA900	190B01	2025-02-06	4	2.285714	0.951190	1.802254	4.754761	-74.024647
19	100120	BA900	212A03	2025-02-06	6	2.375000	1.846812	1.962842	4.677365	-74.066779

Nota: Elaboración propia.

Figura 7

Aplicación de StandardScaler para la Normalización y Escalado de Variables Numéricas Previo al Modelado

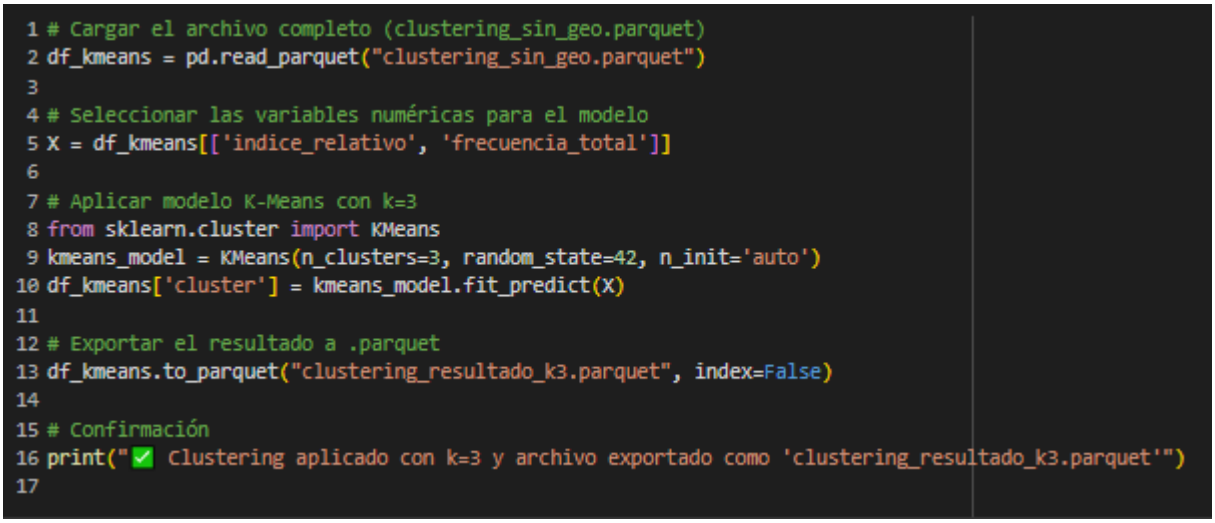


```
# Cargar archivo con índice relativo y coordenadas desde el entorno actual
from sklearn.preprocessing import StandardScaler

# Leer el archivo ya procesado
desempeno_completo = pd.read_parquet("/content/desempeno_conductores.parquet")

# Selección de variables para ambas versiones

# Versión 1: Sin coordenadas
X_sin_geo = desempeno_completo[['indice_relativo', 'frecuencia_total']].copy()

# Versión 2: Con coordenadas
X_con_geo = desempeno_completo[['indice_relativo', 'frecuencia_total', 'latitud_promedio', 'longitud_promedio']].copy()

# Estandarización
scaler_1 = StandardScaler()
X_sin_geo_scaled = scaler_1.fit_transform(X_sin_geo)

scaler_2 = StandardScaler()
X_con_geo_scaled = scaler_2.fit_transform(X_con_geo)

# Crear DataFrames escalados
df_sin_geo_scaled = pd.DataFrame(X_sin_geo_scaled, columns=X_sin_geo.columns)
df_con_geo_scaled = pd.DataFrame(X_con_geo_scaled, columns=X_con_geo.columns)

# Adjuntar columnas clave para identificar registros
columnas_identificacion = ['Conductor', 'Ruta', 'Código Paradero', 'Fecha']
df_base = desempeno_completo[columnas_identificacion].reset_index(drop=True)

df_sin_geo_final = pd.concat([df_base, df_sin_geo_scaled], axis=1)
df_con_geo_final = pd.concat([df_base, df_con_geo_scaled], axis=1)

# Exportar como archivos .parquet
df_sin_geo_final.to_parquet("clustering_sin_geo.parquet", index=False)
df_con_geo_final.to_parquet("clustering_con_geo.parquet", index=False)

print("✅ Archivos .parquet exportados correctamente:")
print("- clustering_sin_geo.parquet")
print("- clustering_con_geo.parquet")

```

Nota: Elaboración propia.

Figura 8

Resultados del Clustering No Supervisado con K-Means para la Agrupación de Conductores en Tres Clústeres de Desempeño Relativo

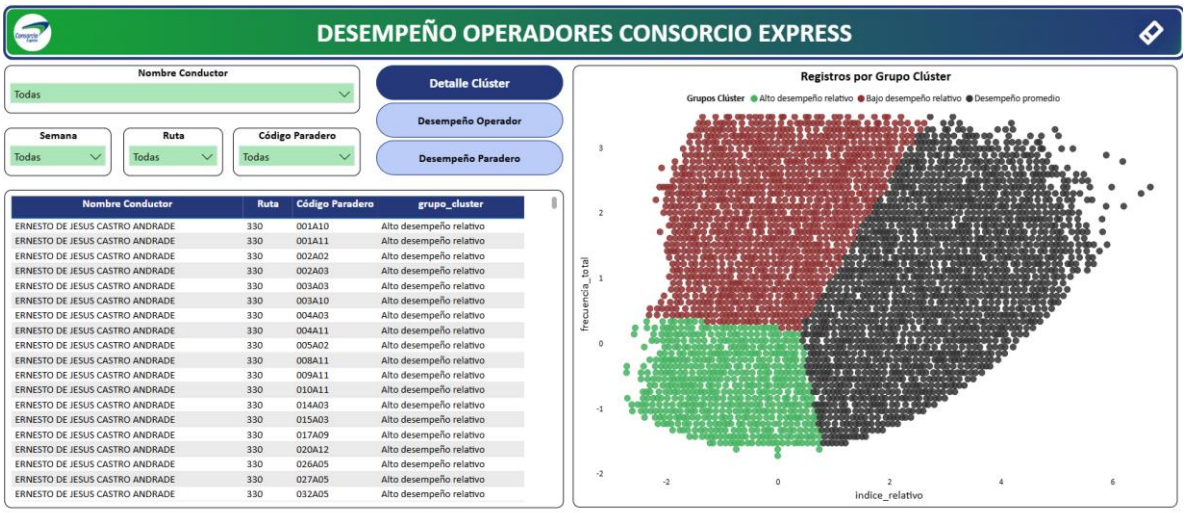


```
# Cargar el archivo completo (clustering_sin_geo.parquet)
df_kmeans = pd.read_parquet("clustering_sin_geo.parquet")

# Seleccionar las variables numéricas para el modelo
X = df_kmeans[['indice_relativo', 'frecuencia_total']]

# Aplicar modelo K-Means con k=3
from sklearn.cluster import KMeans
kmeans_model = KMeans(n_clusters=3, random_state=42, n_init='auto')
df_kmeans['cluster'] = kmeans_model.fit_predict(X)

# Exportar el resultado a .parquet
df_kmeans.to_parquet("clustering_resultado_k3.parquet", index=False)

# Confirmación
print("✅ Clustering aplicado con k=3 y archivo exportado como 'clustering_resultado_k3.parquet'")

```

Nota: Elaboración propia.

Figura 9

Dashboard Interactivo en Power BI para la Visualización e Interpretación de Clústeres y el Monitoreo del Desempeño por Zona, Ruta y Operador



Nota. Elaboración propia.

Análisis y discusión de los resultados

El análisis de los datos operativos de la actividad de los vehículos y el total de pasajeros datos que fueron recolectados de los meses de febrero y marzo de 2025 sobre el comportamiento de los conductores de Consorcio Express permitió identificar patrones mediante la técnica de clustering (K-Means). Tomando como referencia el índice relativo el cual se calculó por medio de una puntuación estandarizada por ruta y paradero, todos los registros se clasificaron en tres grupos de desempeño alto, medio y bajo. Esta calificación permitió identificar un comportamiento operativo de los conductores con base en la frecuencia de paso por las paradas y la cantidad de pasajeros recogidos.

1. Segmentación por Clúster

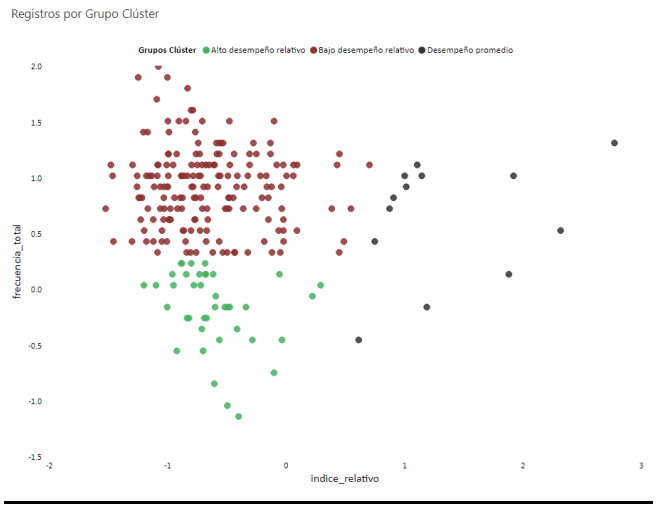
En la figura 9 se observa cómo se distribuyen los clústeres en función del índice relativo que esta ubicado en el (eje X) y la frecuencia total ubicado en el (eje Y), se puede

observar una división clara de los registros donde se pueden identificar operadores con diferentes comportamientos bajo desempeño relativo, Desempeño promedio y Alto desempeño relativo, esta segmentación es importante para realiza acciones, monitoreos poder tomar decisiones o generar intervenciones.

Ahora podemos observar la figura 10 donde se realiza un filtro por el operador Adith Ruidiaz Beleño filtrando por la ruta T21 identificando un comportamiento bajo siendo este mas relevante y teniendo en cuenta que, aunque tiene otros comportamientos esto indica que el comportamiento solo puede estar en ciertas paradas.

Figura 10

BI Investigación Omisión de Paradas SITP



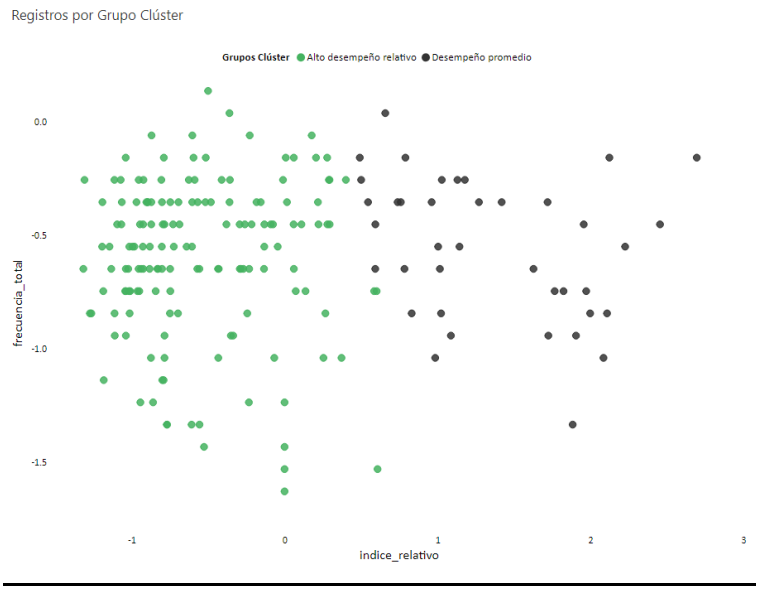
Abrir en Power BI
Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **Ruta** (es T21), **Nombre Conductor** (es ADITH RUIDIAZ BELEÑO)

Por otro lado en la (figura11), se realiza un filtro para la ruta E26B del mismo conductor observando que el comportamiento es diferente ya que esta con un desempeño promedio

y alto lo que nos puede indicar que puede ser las rutas las que pueden ocasionar este desempeño o se pueden encontrar otro tipo de circunstancias si este comportamiento es frecuente sobre estas rutas.

Figura 11

BI Investigación Omisión de Paradas SITP



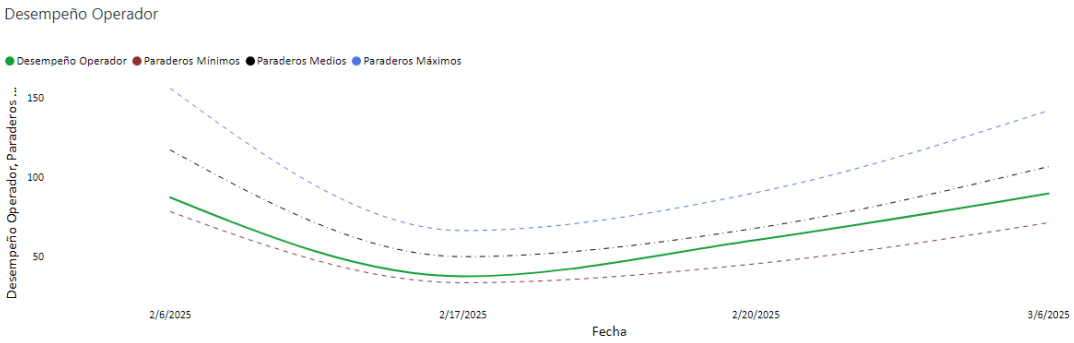
Abrir en Power BI
Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **Ruta** (es E26B), **Nombre Conductor** (es ADITH RUIDIAZ BELEÑO)

2. Análisis temporal del desempeño

El gráfico de líneas de la (figura 12) evidencia la evolución diaria del desempeño del conductor frente a las líneas de referencias de paraderos mínimos, medios y máximos. Donde se evidencia que este comportamiento del conductor es frecuente para la ruta T21 ya que esta en los días que opero la ruta por debajo de los paraderos medios a parar y recoger la media de pasajeros que normalmente generan validaciones del pasaje.

Figura 12

BI Investigación Omisión de Paradas SITP



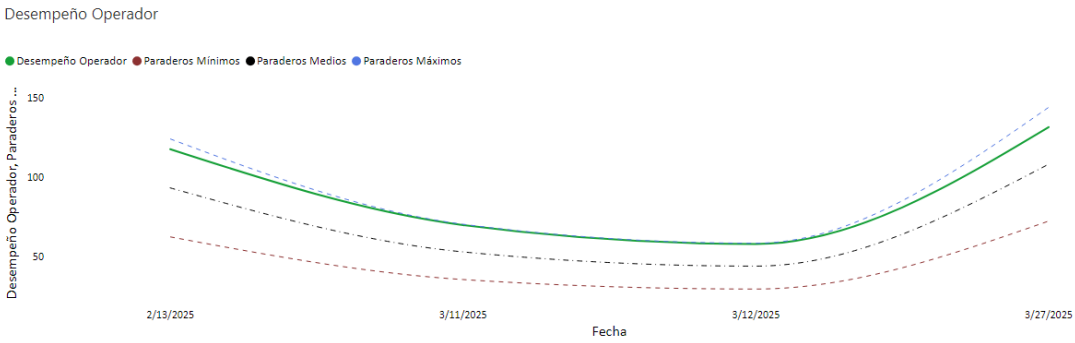
[Abrir en Power BI]

Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **Ruta** (es T21), **Nombre Conductor** (es ADITH RUIDIAZ BELEÑO)

Luego observamos la (figura 13) sobre el comportamiento obtenido en la ruta E26B y el cambio es evidente ya que esta sobre los paraderos máximos lo que indica que sobre esta ruta y esos días operativos el conductor mantuvo ascensos por arriba del promedio de pasajeros que se transportan en esta ruta.

Figura 13

BI Investigación Omisión de Paradas SITP



Abrir en Power BI
Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **Ruta** (es E26B), **Nombre Conductor** (es ADITH RUIDIAZ BELEÑO)

3. Comportamiento geográfico por paradas

La geolocalización de los clústeres por paradero según la (figura 14 y 15) muestra una concentración de desempeño bajo en la mayoría de paraderos aunque este recorrido de la ruta T21 tiende a generar un recorrido tomando vías como la calle 80 con alto flujo vehicular puede ser un factor ya que los tiempos de recorridos pueden varias pero aunque las dos rutas están sobre la zonas de recorridos con inicio del punto nor-occidental al punto nor-oriental vemos que el comportamiento de la ruta E26B que no toma la vía calle 80 y que no tiene la misma cantidad de paraderos puede dar indicios a que rutas más largas y con paso por vías principales y con alto flujo vehicular puedan ser una de las causas del comportamiento.

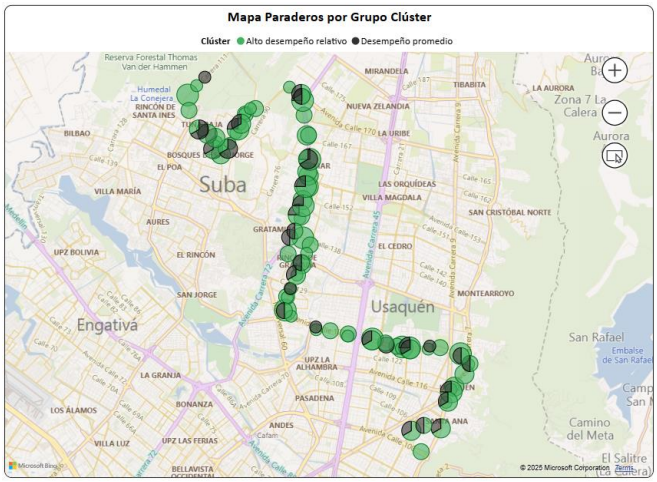
Figura 14

BI Investigación Omisión de Paradas SITP

Abrir en Power BI
Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **Ruta** (es T21), **Nombre Conductor** (es ADITH RUIDIAZ BELEÑO)

Figura 15

BI Investigación Omisión de Paradas SITP



Abrir en Power BI
Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **Ruta** (es E26B), **Nombre Conductor** (es ADITH RUIDIAZ BELEÑO)

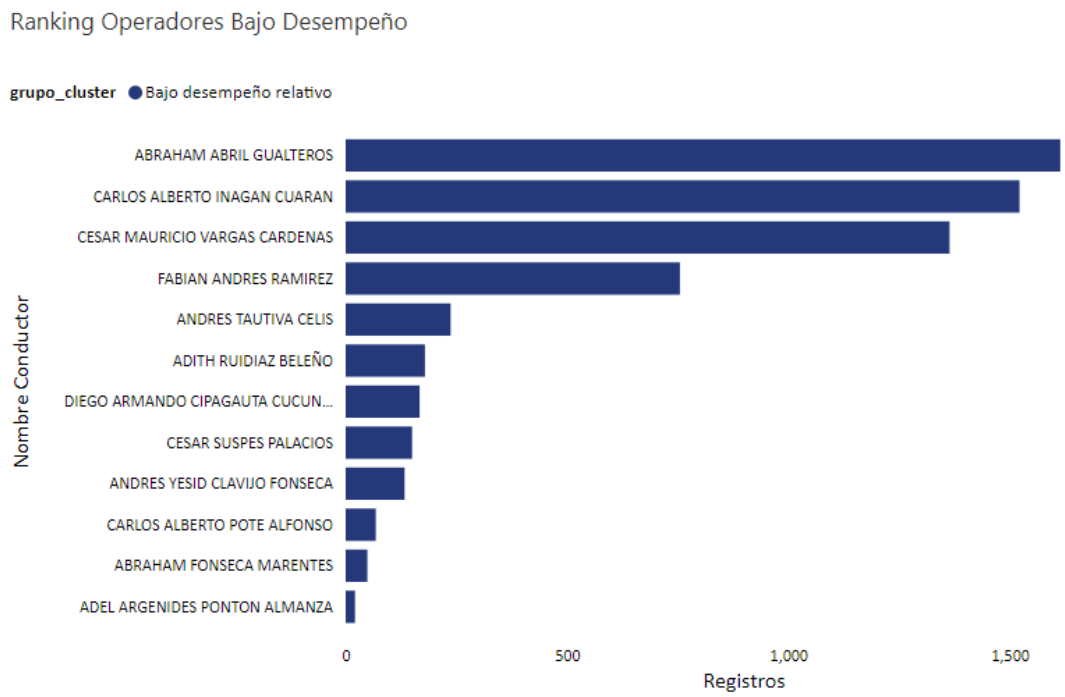
Estos comportamientos pueden tener diferentes explicaciones ya sea por condiciones de la infraestructura de la malla vial, demanda de pasajeros o congestión vehicular, pero si es claro que representa una oportunidad para estudios mas focalizados por patio, ruta, paradero y conductor.

4. Comparación entre operadores

En el panel de ranking operadores bajo desempeño de la (figura 16) y tomando como referencia al conductor Adith Ruidiaz Beleño frente a otros conductores se identifican operadores con una alta frecuencia de registros en clústeres de bajo desempeño muchos mas altos que el operador que tenemos como referencia y esta información es de gran ayuda para la gestión del talento humano operativo del concesionario priorizando procesos de retroalimentación, formación o auditorías.

Figura 16

BI Investigación Omisión de Paradas SITP



Abrir en Power BI

Nota: Fecha de los datos: 19/5/25, 21:20. Filtrado por **grupo_cluster** (es Bajo desempeño relativo), **Nombre Conductor** (es, ADITH RUIDIAZ BELEÑO, OTROS)

5. Relación con los objetivos

- Para el objetivo (a) se dio cumplimiento mediante la identificación de patrones de omisión a partir de los datos de las validaciones en los paraderos.
- Para el objetivo (b) se evidencia con la aplicación efectiva del modelo de clustering realizado en un cuadernillo de Google Colab.
- El objetivo (c) se logró con la implementación del dashboard interactivo en Power BI.

- El objetivo (d) se desarrollo parcialmente en la etapa de análisis y evolución, dejando como paso inicial a modelos de regresión lineal o predicción futura.

Diversos estudios (Pang et al., 2023; Warren et al., 2019; Tong & Ng, 2021) destacan como el comportamiento operativo de los conductores tienen una influencia en la calidad del servicio. En este estudio, la omisión de paradas no se aborda como un fenómeno con poca importancia sino como un patrón detectable y cuantificable un poco alineado con lo que se ha observado en sistemas de transporte en ciudades similares a Bogotá, adicional el uso del método clustering como herramienta de segmentación y/o agrupación ayuda con la apertura a intervenciones focalizadas para generar efectividad en el servicio.

Conclusiones

- La aplicación de técnicas como la de clustering no supervisado permitió realizar una segmentación del comportamiento de los conductores de una forma objetiva, revelando patrones de desempeño bajo, medio y alto.
- Se identificaron zonas geográficas especificas donde la omisión de paradas o bajo desempeño es más frecuente lo que hace de este insumo una representación valiosa para realizar planificaciones operativas enfocadas en minimizar estos comportamientos y tener un aumento en el desempeño de los conductores.
- La creación del dashboard en Power BI facilito la visualización e interpretación de los resultados obtenidos después de aplicar el modelo no supervisado, este informe visual puede ser utilizado directamente por el concesionario.

- El estudio realizado confirma que la minería de datos aplicada al transporte publico permite transformar los datos operativos en insumo y conocimiento para la toma de decisiones en pro de la mejora del servicio
- Se hace la recomendación de extender este análisis a más periodos del año y la preparación y puesta en marcha de modelos predictivos con el fin de poder anticipar ciertos patrones de comportamientos.

8. Lista de referencias

- Anil, A. R., & Anudev, J. (2022). Driver behavior analysis using K-means algorithm. *Proceedings of the 2022 3rd International Conference on Intelligent Computing, Instrumentation and Control Technologies: Computational Intelligence for Smart Systems, ICICICT 2022*. https://doi.org/10.1109/ICICICT54557.2022.9917899
- Kim, M. K., Kim, S. P., Heo, J., & Sohn, H. G. (2017). Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area. *KSCE Journal of Civil Engineering*, *21*(3). https://doi.org/10.1007/s12205-016-1099-8
- Pang, L., Jiang, Y., Wang, J., Qiu, N., Xu, X., Ren, L., & Han, X. (2023). Research of Metro Stations with Varying Patterns of Ridership and Their Relationship with Built Environment, on the Example of Tianjin, China. *Sustainability (Switzerland)*, *15*(12). https://doi.org/10.3390/su15129533
- Warren, J., Lipkowitz, J., & Sokolov, V. (2019). Clusters of Driving Behavior from Observational Smartphone Data. *IEEE Intelligent Transportation Systems Magazine*, *11*(3). https://doi.org/10.1109/MITS.2019.2919516
- Alcaldía Mayor de Bogotá. (2018, diciembre 14). *Acuerdo entre TransMilenio y SITP para mejorar el servicio del sistema*. https://bogota.gov.co/mi-ciudad/movilidad/acuerdo-entre-transmilenio-y-sitp-para-mejorar-el-servicio-del-sistema
- Alcaldía Mayor de Bogotá. (2021). *Manual de operaciones del componente zonal del SITP (rutas urbanas, complementarias, especiales, alimentadoras)* (Versión 3). https://intranet.odt.gov.co/wp-content/uploads/2022/09/M-DB-003-Manual-de-Operaciones-Componente-Zonal-V.3.-1.pdf
- Chaparro Rivera, S. (2019). *Análisis de aspectos de percepción de calidad en TransMilenio* [Trabajo de grado, Universidad de los Andes]. Repositorio Uniandes.

https://repositorio.uniandes.edu.co/entities/publication/6aac2d26-dcaf-4b27-a05d-7234f123880a

- El Tiempo. (2018, octubre 9). *Trancones y omisión de paradas, detrás de las demoras en rutas zonales del SITP*. https://www.eltiempo.com/bogota/trancones-y-omision-de-paradas-detras-de-demoras-en-rutas-zonales-del-sitp-295800
- Rodríguez, F. (2019). *Análisis del impacto de las condiciones del tránsito en Bogotá en el diseño operacional inicial del Sistema Integrado de Transporte Público SITP* [Trabajo de grado, Universidad Nacional de Colombia]. Repositorio UNAL. https://repositorio.unal.edu.co/handle/unal/77547
- Secretaría Distrital de Movilidad. (2023). *Informe de gestión y resultados 2023*. https://www.movilidadbogota.gov.co/web/sites/default/files/Paginas/22-01-2024/informe_de_gestion_y_resultados_2023.pdf
- TransMilenio S.A. (2000). *Contrato de concesión para la prestación del servicio público de transporte terrestre masivo urbano de pasajeros en el sistema TransMilenio*. https://ppp.worldbank.org/public-private-partnership/sites/default/files/2024-07/Transmilenio%20Fase%20I.pdf
- TransMilenio S.A. (2024). *Estadísticas de oferta y demanda del Sistema Integrado de Transporte Público - SITP - marzo 2024*. https://www.transmilenio.gov.co/publicaciones/154050/estadisticas-de-oferta-y-demanda-del-sistema-integrado-de-transporte-publico-sitp-marzo-2024/