Sistema predictivo de demanda y recomendación de estaciones de TransMilenio basado en datos históricos

Elaborado por:

Andrés Camilo Cruz Moncada

Adriana Valentina Suárez Cañón

Madeline Yizell Vargas Martínez

Facultad de ingeniería, Universidad EAN

Proyecto de grado

Lina María Chacón Rivera

05 de febrero de 2026

Contenido

Resumen 3
Abstract 3
Introducción 4
Marco teórico 5
Metodología 9
Estado del arte 13
Resumen del estado del arte 17
Resultados 18
Análisis de resultados 24
Propuesta de solución 25
Arquitectura del sistema web 26
Discusión 32
Plan de divulgación 32
Conclusiones 34
Referencias 35

Resumen

Este proyecto propone el análisis y predicción de la demanda en tres estaciones críticas del sistema de Transmilenio en Bogotá: Los Héroes, Calle 76 y Calle 85. Mediante registros históricos de validaciones, se busca identificar patrones de movilidad, predecir picos de demanda en distintas franjas horarias y días de la semana.

A partir de estos resultados se propone el desarrollo de una página web que permita anticipar niveles de congestión y generar recomendaciones sobre la estación menos saturada, con el fin de optimizar los tiempos de desplazamiento y mejorando la experiencia de viaje de los usuarios.

Palabras clave: Transmilenio, predicción, congestión, recomendación de estaciones, comparación de congestión, movilidad urbana

Abstract

This project proposes the analysis and prediction of demand at three critical stations of the Transmilenio transit system in Bogotá: Los Héroes, Calle 76, and Calle 85. Using historical validation records, the study aims to identify mobility patterns and predict demand peaks across different time slots and days of the week.

Based on these results, the development of a web-based platform is proposed to anticipate congestion levels and provide recommendations on the least crowded stations. The ultimate goal is to optimize commuting times and improve the overall travel experience for users through data-driven insights.

Keywords: Transmilenio, prediction, congestion, station recommendation, congestion comparison, urban mobility.

Introducción

El sistema de Transmilenio conforma el principal medio de movilidad urbana de Bogotá, movilizando diariamente a millones de usuarios a través de la red de estaciones troncales. Sin embargo, la distribución de la demanda no es uniforme, no todos van al mismo destino a la misma hora existen estaciones que están estratégicamente ubicadas como lo es la estación de héroes, calle 76 y calle 85 las cuales presentan altos niveles de movilidad, especialmente en horas pico laborales y académicas.

Dicha concentración de usuarios genera largos tiempos de espera, saturación de personas en los torniquetes, sobrecarga operativa, también reducción de la calidad del servicio recibido por parte de los usuarios. Aunque el sistema dispone de datos históricos de validaciones, esta información no es transformada en herramientas predictivas accesibles a un usuario final. Esto significa, que las decisiones de ingreso al servicio se hacen sin hacer una comparación con otros datos de saturación que se produjeran en otras estaciones que estén ubicadas en la misma zona, por el camino del usuario.

En un contexto en el que haya más de una estación en la zona del usuario que puede representar diferentes posibilidades de traslado para el desplazamiento del usuario surge la necesidad de desarrollar una solución tecnológica que permita medir, comparar y predecir la congestión entre estaciones. Lo útil de esta solución tecnológica está relacionada con la toma de decisiones y la mejora de la experiencia de transporte para poder desplazarse o el hecho de poder en algún momento determinado ir a la estación menos saturada.

Este proyecto propone que, mediante el análisis, comparación y predicción de los patrones de demanda y niveles de congestión en tres estaciones estratégicas del sistema de Transmilenio a partir de los registros de validaciones, se pueda anticipar y recomendar al usuario

la opción de la estación menos saturada clasificadas por franjas horarias y días de la semana para optimizar su desplazamiento y mejorar su experiencia de viaje. De tal forma, mediante el análisis descriptivo se compararán las tendencias de demanda para desarrollar un modelo predictivo, el cual, permita anticipar picos de afluencia en las estaciones y evaluar el desempeño de aprendizaje en algoritmos como Random Forest Classifier y Hist Gradient Boosting Regressor. Por último, los resultados obtenidos serán integrados en una página web para brindar recomendaciones al usuario sobre la estación que estará menos congestionada.

Marco teórico

El transporte público es un pilar importante para el desarrollo económico, social y ambiental de las ciudades, ya que posibilita que la comunidad acceda a empleo, educación y diversos servicios (Banco Mundial, 2021). La implementación de sistemas de movilidad masiva, en un principio se pensó como una estrategia con la finalidad de eliminar la congestión vehicular y la contaminación urbana. Sin embargo, estos sistemas hoy en día enfrentan retos hacia la saturación en horas pico, la falta de optimización de rutas y la necesidad de desarrollar tecnologías más modernas (BID, 2020).

En Bogotá, TransMilenio (en funcionamiento desde el año 2000) es consolidado como el principal transporte público de la ciudad, movilizando diariamente a millones de pasajeros. A pesar de esto, el aumento continuo de la población, las limitaciones en infraestructura y la distribución de actividades urbanas han generado altos niveles de congestión en estaciones estratégicas, especialmente en los periodos de mayor demanda, lo cual afecta directamente la eficiencia del servicio y por ende la experiencia del usuario. (TransMilenio, 2025).

La literatura especializada, ha estudiado los factores que determinan el comportamiento de la demanda de transporte. Velásquez (2009) evidenció que variables como la densidad poblacional y los precios del sistema afectan significativamente en el uso del servicio, identificando reducciones directamente en la demanda ante la subida del precio a la tarifa. Por su parte, Ortúzar y Willumsen (2001) señala que los patrones de movilidad urbana están influenciados por elementos socioeconómicos como la localización de los principales centros laborales y educativos.

Dentro del ámbito del análisis predictivo, se han empleado distintos enfoques para anticipar la afluencia de pasajeros. Los modelos tradicionales de series de tiempo, como ARIMA, permiten identificar tendencias y comportamientos estacionales en los datos históricos (Box & Jenkins, 2015). Con el tiempo, herramientas más flexibles como Prophet han facilitado la integración de componentes estacionales adicionales y variables temporales dentro del proceso de pronóstico (Taylor & Letham, 2018).

En este marco, el presente estudio busca ir más allá del análisis convencional de la demanda al comparar múltiples estaciones del sistema y generar recomendaciones basadas en la evaluación histórica y en modelos predictivos. El objetivo es estimar los niveles de saturación en diferentes puntos de la red y proponer un mecanismo que permita contrastar su nivel de ocupación, de manera que los usuarios puedan identificar la alternativa menos congestionada según la franja horaria.

El modelo predictivo Random Forest Classifier basa su entrenamiento en un algoritmo de aprendizaje supervisado, el cual está enfocado en la construcción de varios árboles de decisión para ser entrenados por medio de los subconjuntos de variables y datos aleatorios, generando un promedio mediante los resultados de predicciones. Debido a esto, se reduce la varianza de

sobreajuste del modelo y minimiza el error, lo que permite que la predicción sea estable frente a datos significativamente variables.

Por parte del modelo Hist Gradient Boosting Regressor, está fundamentado en la técnica de *ensemble learning* para su entrenamiento de construcción secuencial de árboles de decisión en donde cada árbol busca corregir los errores generados anteriormente. Mientras Hist Gradient Boosting Regressor sirve para construir modelos de forma iterativa para optimizar su proceso basado en gradientes, Random Forest Classifier utiliza arboles de forma independiente.

Para determinar la precisión del modelo predictivo se utilizaron distintas métricas de evaluación. Una de las más comunes en problemas de regresión es el Error Absoluto Medio (MAE), permite medir el promedio de las diferencias absolutas entre los valores reales y los predichos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_t|$$

Otra de las métricas es el Error Cuadrático Medio (MSE), penaliza con mayor intensidad los errores grandes al elevar las diferencias al cuadrado.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_t)^2$$

De igual forma, existe la Raiz Cuadrática del Error Medio (RMSE), interpreta el erro en la misma escala de la variable objetivo, lo que permite un análisis sencillo del desempeño predictivo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_t)^2}$$

De otro modo, el Coeficiente de Determinación (R^2), que mide la capacidad explicativa del modelo respecto a la variabilidad de los datos observados.

$$R^2 = 1 - \frac{\sum(y_i - y_t)^2}{\sum(y_i - \overline{y})^2}$$

También, el acurracy el cual es el porcentaje de instancias correctamente clasificadas Con esta medida se debe tener cuidado sobre todo si los datos están desbalanceados.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde:

- **TN / True Negative:** Caso cuando es negativo y se predice negativo.
- **TP / True Positive:** Caso cuando es positivo y se predice positivo.
- **FN / False Negative:** Caso cuando es positivo, pero se predice negativo.
- **FP / False Positive:** Caso cuando es negativo, pero se predice positivo.

De igual manera, la precisión es intuitivamente la capacidad del clasificador de no etiquetar como positiva una muestra que es negativa y el recall es intuitivamente la capacidad del clasificador de encontrar todas las muestras positivas.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Por último, el F1 Score es una combinación de precisión y recall, útil cuando tienes un equilibrio entre estos dos aspectos.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

La implementación de Random Forest Classifier y Hist Gradient Boosting Regressor permiten desarrollar un modelo predictivo robusto para la predicción a la demanda de estaciones de TransMilenio, aprovechando el análisis de datos históricos para proporcionar una mejor planificación operativa, optimización de recurso y aportar a la toma de decisiones en espacios de transporte inteligente.

Metodología

Esta investigación, se realiza bajo un enfoque cuantitativo, centrado en analizar y modelar los datos históricos de validaciones registradas en las estaciones Los Héroes, Calle 76 y Calle 85 del sistema de Transmilenio. El diseño metodológico es no experimenta y longitudinal retrospectivo, ya que se utilizan registros históricos sin manipular las variables, además de analizar patrones temporales a lo largo de cierto tiempo.

La investigación se desarrolla por 3 fases articuladas. En primer lugar, re realiza la recopilación y preparación de los datos provenientes del sistema "TuLlave", incluyendo procesos de limpieza, estandarización y agregación por franjas horarias. Esta etapa permite transformar los registros brutos en una estructura temporal idónea para análisis de tipos comparativo y de modelo predictivo. En segundo lugar, se lleva a cabo el análisis descriptivo y comparativo de la demanda entre estaciones con el propósito de identificar patrones recurrentes, horarios pico y diferencia en los niveles de saturación. A partir de esos resultados, se establece un índice de

congestión normalizado en porcentaje (0-100), el cual permite comparar de manera uniforme los niveles de ocupación entre estaciones, no obstante, su volumen absoluto de validaciones.

Después se implementan los modelos de aprendizaje automático, que se orientan a predecir la demanda futura, por estación. El modelo se entrena con variables temporales y características derivadas de los datos históricos, procediendo a evaluar el comportamiento del modelo a partir de distintas métricas de error, tales como la MSE o la MAE, para de esta manera permitir prever situaciones de congestionamiento en franjas horarias determinadas.

Finalmente, se diseña un motor de recomendación que integra los valores predichos del Índice de Congestión para comparar estaciones cercanas y sugerir al usuario la alternativa menos saturada. Dicho motor se incorpora en una aplicacion web que visualiza los niveles actuales y proyectados de congestión mediante indicadores gráficos, facilitando la interpretación y la toma de decisiones informadas.

Por lo tanto, la metodología propuesta para la integración de análisis descriptivo, modelado predictivo y generación de recomendaciones da lugar a la construcción de un sistema en funcionamiento que tiene la finalidad de optimizar la experiencia del usuario en el transporte público urbano.

El proyecto se divide en seis fases principales, empezando desde la definición inicial del problema hasta la validación final y la preparación de la sustentación. Cada fase incluye actividades específicas que permiten avanzar progresiva y organizacionalmente en el desarrollo del modelo predictivo y del prototipo web para la recomendación de estaciones con menor congestión.

Fase 1. Iniciación

En esta etapa, se establecen los objetivos del proyecto y se detallan los requisitos necesarios para su desarrollo. Se define el alcance del sistema, abordando el problema de la congestión en las estaciones de transporte público y la necesidad de crear predicciones que ayuden a los usuarios a encontrar estaciones con menos afluencia. Luego, se lleva a cabo una revisión de artículos centrados en estudios sobre la predicción de la demanda en el transporte público, utilizando técnicas de aprendizaje automático. Esta revisión es clave para identificar metodologías, algoritmos y enfoques que se han utilizado en investigaciones anteriores.

Fase 2. Preparación de datos

En esta etapa, nos enfocamos en recopilar los datos necesarios para desarrollar el modelo predictivo. Estos datos provienen de la información histórica sobre la demanda de pasajeros en las estaciones del sistema de transporte TransMilenio.

Luego, se lleva a cabo un proceso de limpieza de datos para eliminar registros inconsistentes, valores faltantes o errores que podrían afectar el rendimiento del modelo. Por último, normalizamos los datos para asegurarnos de que las variables que utilizamos en los algoritmos de aprendizaje automático estén en escalas comparables, lo que facilita el entrenamiento de los modelos.

Fase 3. Análisis exploratorio de datos

Durante esta etapa, se llevó a cabo un análisis exploratorio de datos (EDA) con el fin de entender cómo se comporta la demanda de pasajeros en las estaciones que estamos estudiando.

Identificamos patrones de comportamiento, los horarios con mayor afluencia y los posibles picos de demanda. Con base en este análisis, establecemos un umbral que nos ayudará a clasificar el nivel de ocupación de las estaciones en diferentes categorías, como baja, media o

alta congestión. Esta información será muy útil más adelante en nuestro sistema de recomendación.

Fase 4. Modelado predictivo

En esta etapa, se pone en marcha varios modelos de aprendizaje automático con el propósito de predecir la demanda de pasajeros. Primero, se utiliza un modelo basado en Random Forest Regressor, que es excelente para captar relaciones no lineales entre las variables del conjunto de datos. Luego, se implementa un segundo modelo que emplea Gradient Boosting Regressor, famoso por su habilidad para aumentar la precisión al combinar varios modelos más simples. Ambos modelos se entrenan y validan con los datos que ya han sido procesados. Al final, se evalúan ambos modelos usando métricas como MAE (Error Absoluto Medio), MSE (Error Cuadrático Medio) y R^2, con el objetivo de medir el rendimiento de cada modelo. Para la clasificación del nivel de congestión se entrena el Random Forest Classifier el cual traduce las validaciones predichas en una categoría interpretable para el usuario final, permitiendo que el sistema recomiende la estación menos saturada.

Fase 5. Desarrollo del prototipo web

En esta etapa, se crea un prototipo web que permite ver las predicciones que genera el modelo. El sistema cuenta con un backend que se encarga de procesar estas predicciones en tiempo real, utilizando el modelo que ha sido entrenado.

Además, se desarrolla un frontend web que facilita al usuario consultar el nivel de congestión en diferentes estaciones. La interfaz mostrará el estado de las estaciones a través de un sistema de colores (verde, amarillo y rojo) que indica niveles de ocupación bajos, medios y altos, respectivamente, ayudando al usuario a elegir la estación con menos congestión.

Por último, se llevan a cabo pruebas de usabilidad para asegurarse de que el sistema funcione correctamente y para evaluar lo fácil que es usar la interfaz.

Fase 6. Cierre y validación

En la fase final se realiza la validación del sistema desarrollado, incluyendo pruebas de funcionamiento y revisión de requisitos no funcionales.

Estado del arte

Proceso de selección de artículos

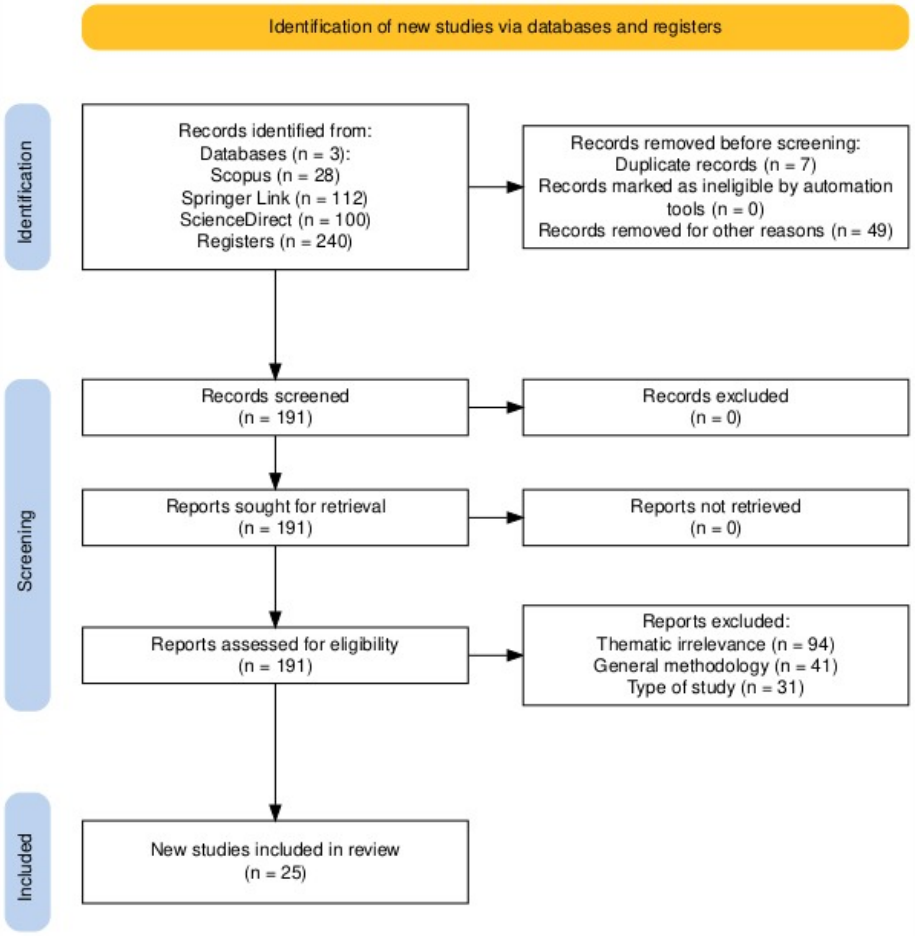
El proceso de selección de artículos se realizó siguiendo la metodología PRISMA. El objetivo fue identificar estudios relevantes para el análisis y predicción de demanda en el sistema masivo de transporte público en Bogotá.

Inicialmente se encontraron 240 registros de tres bases de datos académicas: Scopus (28), Springer Link (112) y ScienceDirect (100) mediante la implementacion de la siguiente combinacion de terminos booleanos: ("public transportation" OR "public transit" OR "urban transport" OR "mass transit") AND ("passenger demand" OR “ridership” OR "passenger flow" OR "travel demand") AND (“prediction” OR “forecasting” OR “estimation”) AND ("machine learning" OR "deep learning" OR "artificial intelligence") AND (“congestion” OR “crowding” OR "passenger congestion"). Luego, se eliminaron 7 registros duplicados y 49 registros por no pertenecer a artículos de revistas indexadas, dando como resultado 191 artículos únicos para el proceso de revisión.

En la fase de evaluación, se revisaron los títulos, resúmenes y contenidos de los artículos para determinar su relevancia con el tema de investigación. En esta etapa se excluyeron 166 estudios. De estos, 94 no estaban directamente relacionados con el problema de investigación, 41

abordaban tecnologías de forma demasiado general sin una aplicación específica al transporte, y 31 eran de tipos de estudio no relevantes para el objetivo del trabajo.

Finalmente, se seleccionaron 25 artículos científicos por su relación directa con la predicción de demanda, el flujo de pasajeros y la congestión en el transporte público. Estos artículos constituyen la base del estado del arte de la investigación.



Nota. Diagrama de flujo del proceso de selección de estudios. Adaptado de *PRISMA Flow Diagram*, por Page et al. (2021), PRISMA (http://www.prisma-statement.org/).

Autor(es) /Año	Título del estudio	Objetivo principal	Metodología empleada	Variables analizadas	Principales hallazgos	Limitaciones	Relación con el proyecto
Zhang et al., 2017	Topology augmented dynamic spatial-temporal network for passenger flow forecasting	Predecir flujo en metro	Deep Learning (ST-ResNet)	Tiempo, ubicación	Alta precisión espacio-temporal	Alto requerimiento de datos	Base para modelo predictivo
Liu & Chen, 2017	Short-term passenger flow prediction	Predicción a corto plazo	Deep Learning	Flujo histórico	Mejora frente a métodos tradicionales	No incluye variables externas	Similar al enfoque del proyecto
Wang et al., 2022	TMS-GNN for bus passenger flow prediction	Predecir flujo en buses	Graph Neural Networks	Red, tiempo	Captura relaciones espaciales	Complejidad alta	Aplicable a estaciones
Sun et al., 2021	Station-wise passenger flow prediction	Flujo por estación	ML (RF, SVM)	Boarding data	Random Forest destaca	Datos limitados	Similar a TransMilenio
Li et al., 2022	OD demand prediction in rail transit	Predicción origen-destino	Modelos probabilísticos	OD matrices	Alta precisión	Difícil obtención datos	Útil para rutas
Zhao et al., 2017	Passenger demand prediction metro	Predicción demanda	Probabilístico	Demanda histórica	Buen desempeño	No dinámico	Aplicable
Yu et al., 2020	MTDTSN bus flow prediction	Predicción buses	Deep Learning	Tiempo, espacio	Mejora multiescala	Alto costo computacional	Similar
Liu et al., 2019	Public transit flow prediction	Predicción flujo	Deep Learning	Datos históricos	Alta precisión	Generalización limitada	Directamente relacionado
Chen et al., 2021	Bus demand prediction crowdsourcing	Predicción demanda	DL + crowdsourcing	Datos usuarios	Mejora precisión	Dependencia externa	Innovación aplicable
Wang et al., 2020	Daily bus demand prediction	Predicción diaria	Machine Learning	Tiempo, clima	Buen desempeño	Limitado a buses	Similar
van Oort et al., 2015	Ridership forecasting	Predecir demanda	ML	Históricos	Mejora planificación	Datos limitados	Igual objetivo

Cats et al., 2016	Ridership prediction special events	Demanda en eventos	Modelos estadísticos	Eventos	Impacto significativo	Casos específicos	Útil para picos
Liu et al., 2018	Metro ridership prediction	Demanda estaciones	Regresión	Entorno urbano	Influencia urbana	No dinámico	Relacionado
Wei et al., 2020	Ridership loss estimation	Estimar pérdida demanda	ML	Factores externos	Predice caídas	No tiempo real	Complementario
Yang et al., 2021	Rail passenger forecasting	Predicción pasajeros	Deep Learning	Series temporales	Alta precisión	Datos grandes	Igual enfoque
Duan et al., 2020	Deep learning demand prediction	Evaluar DL	Deep Learning	Tiempo	DL superior	Coste computacional	Justificación teórica
Shrivastava et al., 2024	Crowding prediction transit	Predecir congestión	LSTM	Ocupación	Predicción precisa	Datos específicos	Directamente clave
Zhang et al., 2019	Passenger flow methods	Comparar métodos	Comparativo	Series temporales	DL mejor	Generalización	Marco teórico
Li et al., 2022	Intermodal hub demand forecasting	Predicción demanda	Modelos híbridos	Riesgo, flujo	Mejora decisiones	Complejo	Aplicable
Li et al., 2023	Privacy demand forecasting	Predicción privada	ML	Datos anonimizados	Protege datos	Menor precisión	Avanzado
Chen et al., 2021	XGBoost train prediction	Predicción tren	XGBoost	Tiempo	Alta precisión	Overfitting	Alternativa ML
Sun et al., 2020	Holiday demand prediction	Demanda festivos	ML	Calendario	Cambios de demanda	Contexto limitado	Útil
Yu et al., 2021	Graph CNN traffic prediction	Flujo tráfico	GCN	Red vial	Buen rendimiento	No específico PT	Parcial
Cats et al., 2017	Metro vulnerability	Vulnerabilidad sistema	Estadístico	Disrupciones	Impacto alto	No predictivo	Contexto
Zhao et al., 2018	Congestion control AI	Control congestión	Inteligencia artificial	Flujo	Mejora control	No específico PT	Complementario

Resumen del estado del arte

El análisis realizado sobre la predicción del flujo y demanda de pasajeros en sistemas de transporte público, coinciden que es un problema fundamental para la mejora de la planificación, reducción de la congestión y optimización del tiempo. Por lo que la mayoría de los estudios demuestran que los enfoques estadísticos y probabilísticos tracciónales, han sido progresivamente reemplazados por técnicas de Machine Learning y Deep Learning. Las cuales, han permitido capturar mejor los patrones complejos del comportamiento de los pasajeros.

Se ha demostrado que los modelos más efectivos son aquellos que combinan dimensiones temporales y espaciales, como lo son las variables de tiempo (horas, días, temporadas) y ubicación (estaciones, rutas o redes). De modo que, estos modelos resultan ser los más avanzados, como las redes neuronales profundas, LSTM y redes de grafos GNN, logrando representar con mayor precisión la estructura real de los de sistemas de transporte. Por otro lado, existen las variables externas, como el clima, eventos especiales o datos provenientes de usuarios, lo que mejora significativamente la predicción, permitiendo modelar situaciones reales como umbrales de demanda o congestión.

A pesar de esto, se han logrado identificar limitaciones claves: este tipo de modelos suelen requerir grandes volúmenes de datos, alto poder computacional y, a veces se suelen presentar problemas para generalizar a otros contextos o dependen de información difícil de obtener. En consecuencia, estas herramientas no solo permiten predecir la demanda, sino también anticipar la congestión, analizar perdidas de usuarios y mejorar la toma de decisiones. Lo que da paso a una base fundamental para el continuo desarrollo en los sistemas de transporte inteligentes.

Resultados

Para el entrenamiento de los modelos Random Forest Classifier y Hist Gradient Boosting Regressor, se utilizaron datos históricos de validación en las estaciones de TransMilenio. Se tuvieron en cuenta variables temporales y operacionales como hora, tipo de día y patrones de movilidad.

En la Figura 1 se muestra el flujo metodológico utilizado para procesar los datos, entrenamiento de modelos y estimación de predicciones de demanda.

Figura 1

Flujo metodológico del sistema predictivo

Nota. Elaboración propia. Flujo metodológico del sistema predictivo para la estimación de demanda y clasificación del sistema de transporte TransMilenio.

El proceso observado en la Figuro 1 inicia con la consolidación de los datos históricos, seguido por la limpieza y normalización de los datos para después identificar variables históricas y Lags, con las cuales se realiza el entrenamiento a los modelos de Machine Learning. Finalmente ver en una aplicación web la predicción resultante del modelo para el usuario final.

Como se observa en la Tabla 1, el modelo alcanzó un R^2 superior a 0.97 en las tres estaciones, lo que evidencia una alta capacidad de la demanda. La accuracy superó el 89% en todos los casos la estación Los Héroes con mayor precisión en la categorización del nivel de congestión con 93.18%.

Tabla 1

Métricas de desempeño del modelo

	Los Héroes	Calle 76	Calle 85
MAE	10.40	9.02	7.03
RMSE	20.25	17.76	11.97
R^2	0.9836	0.9824	0.9780
Accuracy	93.18%	92.59%	89.78%

Nota. Elaboración propia. De igual forma, la Tabla 1 contiene el desempeño del modelo en las tres estaciones por separado. Esto justifica que el modelo funciona bien en todos los casos, no solo en uno.

Tabla 2

Random Forest Classifier - Estación Los Héroes

	Precisión	**Recall**	**F1-score**
Bajo	0.965	0.925	0.944
Medio	0.887	0.914	0.900

Alto	0.945	0.960	0.953

Nota. Elaboración propia. La clase ALTO presenta el mejor F1-score de 0.953, mientras que MEDIO obtiene el menor desempeño.

Tabla 3

Random Forest Classifier – Estación calle 76

	Precisión	**Recall**	**F1-score**
Bajo	0.927	0.912	0.920
Medio	0.894	0.827	0.859
Alto	0.868	0.978	0.920

Nota. Elaboración propia. La clase MEDIO presenta el recall más bajo de 0.827, indicando mayor dificultad para detectar congestión intermedia.

Tabla 4

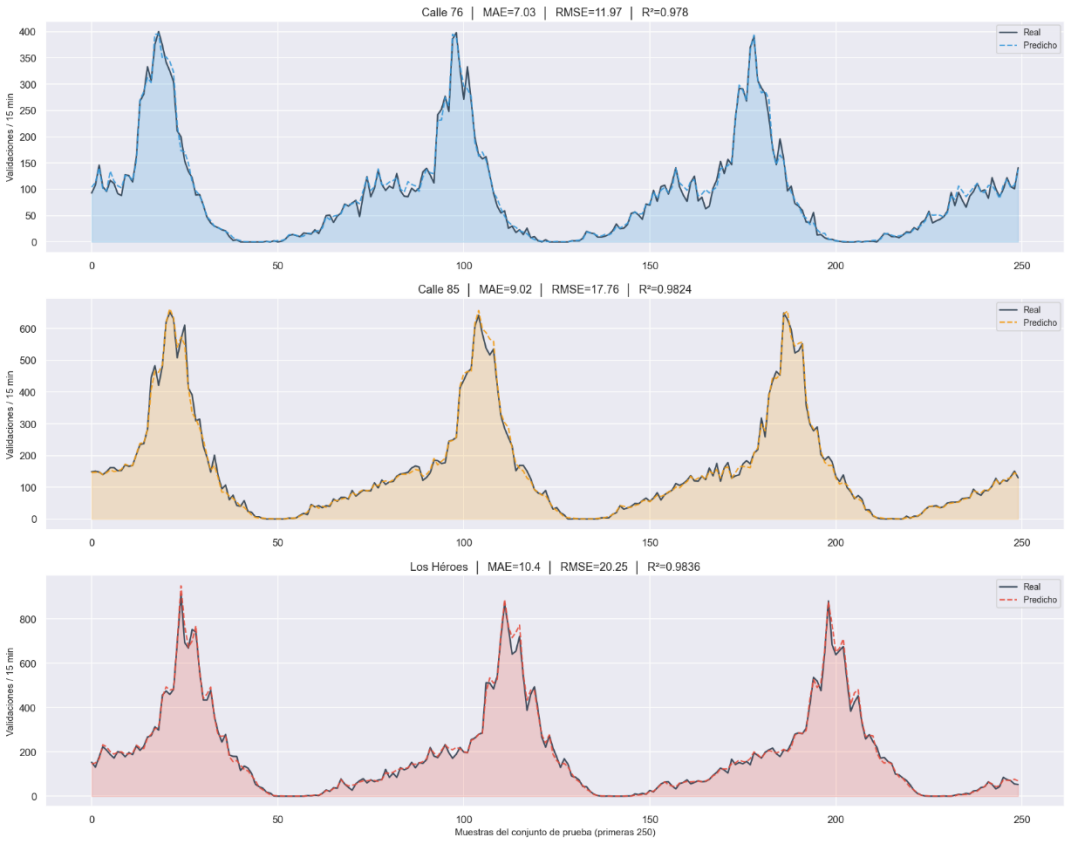
Random Forest Classifier – Estación calle 85

	Precisión	**Recall**	**F1-score**
Bajo	0.953	0.933	0.943
Medio	0.873	0.905	0.889
Alto	0.954	0.939	0.946

Nota. Elaboración propia. La clase ALTO obtiene el F1-score más alto de 0.946, mientras que MEDIO presenta el menor desempeño de 0.889.

Figura 2

Real vs. Predicho (HistGradientBoosting Regressor)



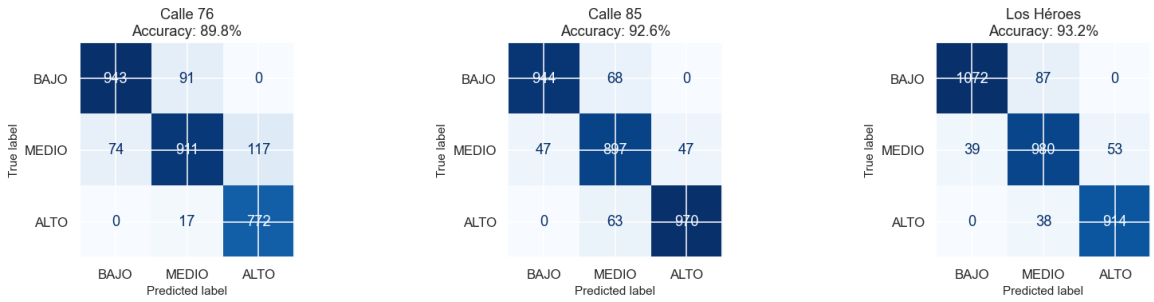
Nota. Elaboración propia. Figura generada mediante el modelo HistGradientBoosting Regressor implementado en Python con la librería scikit-learn en Visual Studio Code.

La Figura 2 muestra qué tan bien el modelo de regresión predice el número real de validaciones en las tres estaciones. La línea sólida es el valor real y la línea punteada es la predicción del modelo.

Figura 3

Matrices de Confusión (Random Forest Classifier)

Nota. Elaboración propia. Figura generada mediante el modelo Random Forest Classifier

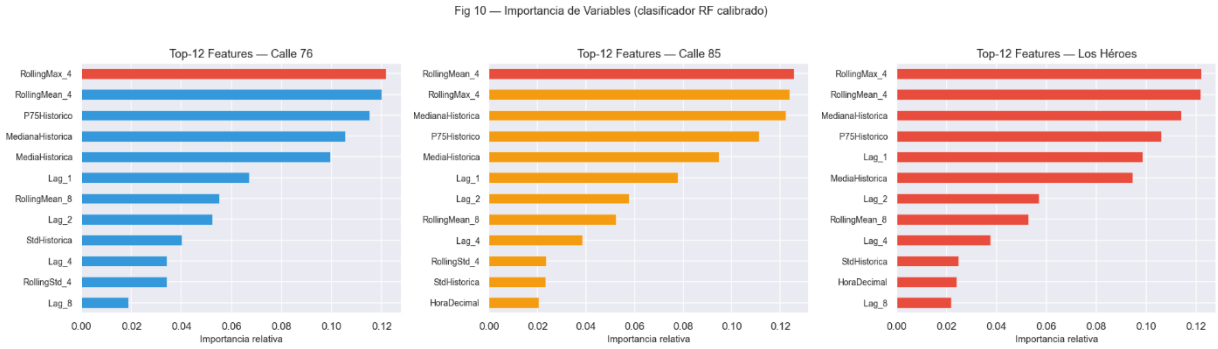


implementado en Python con la librería scikit-learn en Visual Studio Code.

En la Figura 3 se visualizan las matrices de confusión que interpretan cómo el clasificador asigna cada observación real a una categoría predicha. Cada fila es la clase real (BAJO, MEDIO, ALTO) y cada columna es la clase predicha, los valores en la diagonal principal son las predicciones correctas.

Figura 4

Importancia de Variables



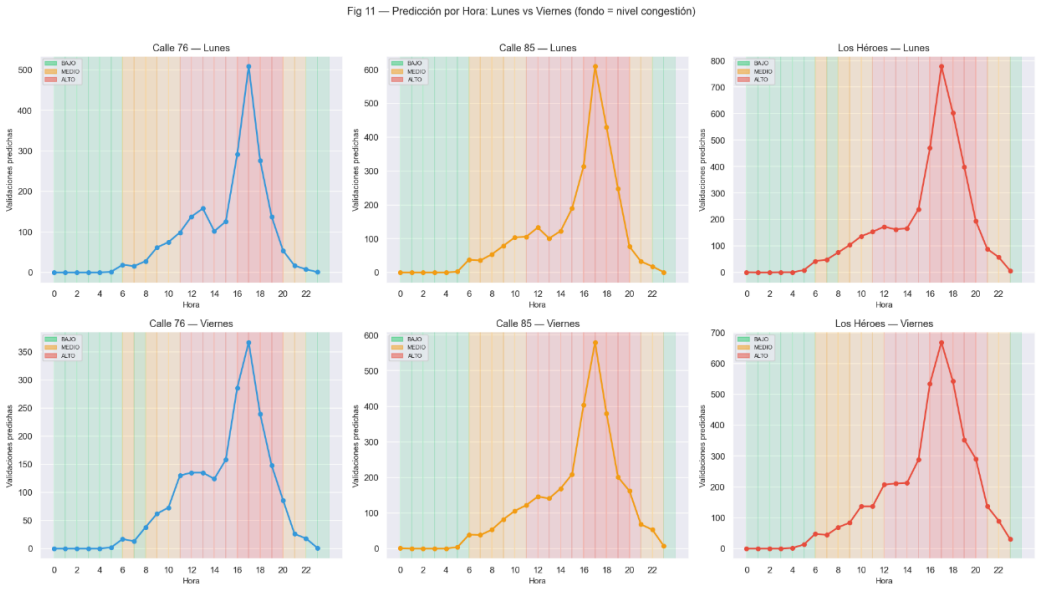
Nota. Elaboración propia. Figura generada mediante el modelo Random Forest Classifier implementado en Python con la librería scikit-learn en Visual Studio Code.

La Figura 4 muestra qué variables fueron más importantes para que el clasificador tomara sus decisiones. En las tres estaciones dominan las variables: RollingMax_4, RollingMean_4, MedianaHistorica, P75Historico y MediaHistorica. Esto muestra que el modelo aprende

principalmente del comportamiento reciente y del historial estadístico acumulado de cada franja horaria.

Figura 5

Predicción por hora: Lunes vs. Viernes



Nota. Elaboración propia. Figura generada a partir de las predicciones del modelo HistGradientBoosting Regressor y Random Forest Classifier implementados en Python con la librería scikit-learn en Visual Studio Code.

La Figura 5 es la más importante para el usuario final debido a que muestra cómo el modelo predice la demanda a lo largo del día, con el fondo coloreado indicando el nivel de congestión esperado (verde = BAJO, naranja = MEDIO, rojo = ALTO).

Análisis de resultados

A partir de los resultados obtenidos, se identifica un buen rendimiento de los algoritmos de aprendizaje automático para el modelado de comportamiento de la demanda en estaciones de TransMilenio. Los resultados de la métrica R”2 superiores a 0.97 indica que el modelo explica más del 97% de la viabilidad en los datos históricos, con una alta relación entre las variables de predicción y la demanda de pasajeros.

Mediante el análisis a las matrices de confusión, las categorías de congestión ALTO y BAJO se identifican con gran precisión, alcanzando valores F1-score superiores a 0.94 en la mayoría de las estaciones. Por otro lado, la categoría MEDIA presenta un desempeño relativamente inferior con respecto a las otras dos, lo que se puede atribuir a la superposición entre los estados de congestión moderada y los limites definidos por las métricas utilizadas para la clasificación.

La importancia de las variables definidas para el entrenamiento de los modelos evidencio que los atributos relacionados con el comportamiento y el historial estadístico fueron los más influyentes en la toma de decisiones del modelo. Variables como RollingMean-4, RollingMax_4 y MediaHistorica influyeron en gran medida la capacidad explicativa, confirmando que la demanda futura depende en su mayoría de los patrones recientes de movilidad.

Asimismo, las gráficas de predicción por hora demuestran que el modelo logra capturar correctamente los picos de demanda significativos de las horas pico laborales, lo que permite la anticipación de escenarios de congestión y por ende generar recomendaciones para los usuarios.

Propuesta de solución

Se propone el diseño de un sistema inteligente de predicción de demanda y participacion ciudadana para tres estaciones estratégicas de TransMilenio (Héroes, Calle 76 y Calle 85), mediante el procesamiento de datos históricos y Machine Learning. Esta solución representa una arquitectura compuesta principalmente por módulos de captura de datos, procesamiento, entrenamiento de modelos y finalmente visualización de predicciones en tiempo real.

Además, integra la participación del usuario mediante reportes reales en formato de comentarios que permitirán complementar la predicción del sistema integrando a su vez dos fuentes de conocimiento, obteniendo como resultado un sistema más robusto y alineado con los principios de ciudades inteligentes.

Esto permite que el sistema pueda anticipar escenarios de congestión en las estaciones, realizar recomendaciones mediante estimación de flujo de pasajeros por estación y franja horaria. Es por esto por lo que este sistema predictivo busca apoyar la toma de decisiones estratégicas frente al sistema de transporte público masivo TransMilenio. En la Tabla 2 se muestra el resumen de los beneficios esperados con la implementación del sistema como propuesta de solución.

Tabla 5

Beneficios e impactos esperados de la propuesta de solución.

Beneficio	Impacto
Optimización de rutas	Reducción de tiempos de espera
Predicción de congestión	Mejora de experiencia del usuario
Análisis de demanda	Mejor toma de decisiones
Recomendación predictiva	Reducción de umbral de demanda

Nota. Elaboración propia. Beneficios esperados de la propuesta de solución.

Arquitectura del sistema web

La arquitectura del sistema web TransmiSmart esta estructura por cinco capas principales, las cuales trabajan de manera integrada para procesar datos, generar predicciones, clasificar niveles de congestión, almacenar información en tiempo real y visualizar los resultados mediante la interfaz web interactiva.

Capa 1 – Carga de datos

En la capa 1 se realiza la obtención de la información utilizada para el modelo. Esta capa conforma la parte de los registros históricos de validaciones de TransMilenio en las estaciones Calle 74, Los Héroes y Calle 85, las cuales componen la base para el entrenamiento del modelo predictivo.

Figura 6

Datos tomados de datos abiertos

Reporte Validaciones por Intervalo Estaciones Troncales y Rutas Duales - Tarjeta Tullave

Periodo Validaciones: al 30 de Noviembre 2025
Fuente: FCS Center

Fase	Línea	Estación	Acceso de Estación	Intervalo	01/11/2025	02/11/2025
Dual	(10006) M84-C84	(12296) MC84_V4	(71476) 195B03|Av Suba - K110A.	7:00	0	0
Dual	(10006) M84-C84	(12296) MC84_V4	(73636) 081A00|AC85 - KR16A	7:45	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(0) (Unknown)	7:30	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(0) (Unknown)	7:45	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(0) (Unknown)	9:15	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(0) (Unknown)	9:30	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(0) (Unknown)	15:30	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(0) (Unknown)	20:00	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	8:45	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	9:00	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	14:00	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	14:45	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	15:00	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	15:15	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	15:30	0	0
Dual	(10006) M84-C84	(12387) MC84_V5	(61410) C001|21 Ángeles A - 3	15:45	0	0

Nota. Elaboración propia. Datos históricos utilizados para el entrenamiento de los modelos predictivos.

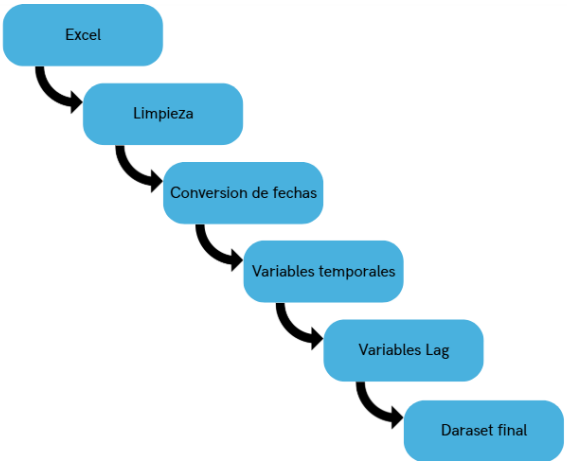
Capa 2 – Transformación de datos y motor de Inteligencia artificial

En esta capa se realiza toda la parte de limpieza y normalización de los datos históricos. De igual forma, se generan variables relacionadas con horarios, días, festivos, horas pico y patrones de movilidad.

Por consiguiente, se implementa los algoritmos (HistGradientBoosting y Random Forest) previamente entrenados para estimar la predicción de la demanda y clasificar los niveles de congestión.

Figura 7

Flujo de transformación y preparación de datos para el modelo predictivo



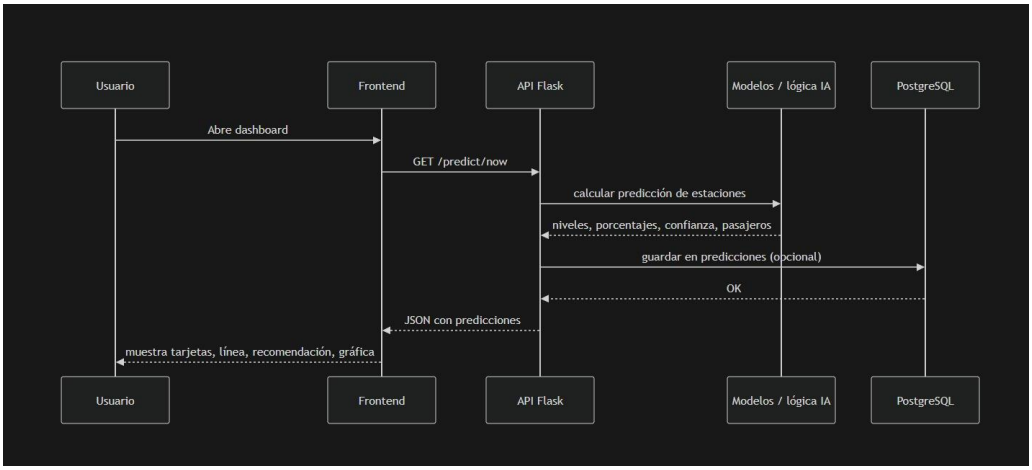
Nota. Elaboración propia.

Capa 3 - Backend y API de servicios

Para el desarrollo el backend se utilizó el lenguaje de programación Python y Flask para crear la API encargada de centralizar la lógica del proyecto. Este archivo se encarga de conectar la base de datos, los modelos ORM, funciones de predicción, cálculos de congestión y consulta de servicios.

Figura 8

Comunicación entre el frontend y el backend mediante el API



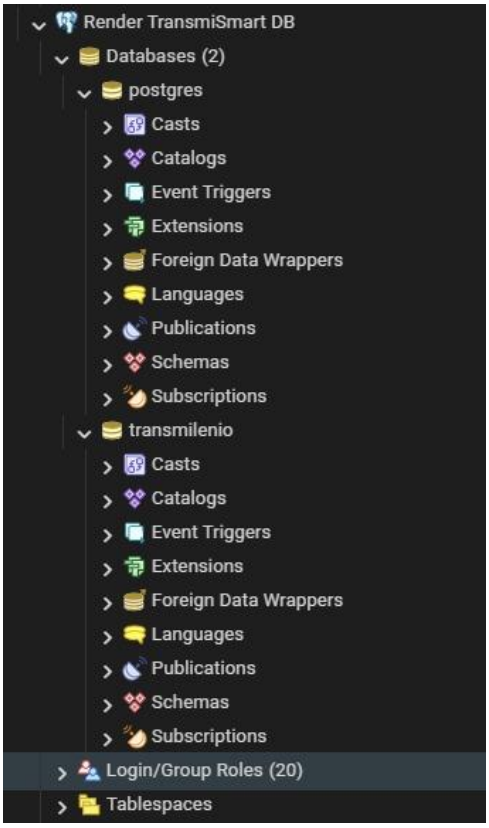
Nota. Elaboración propia.

Capa 4 – Base de datos

La base de datos seleccionada fue PostgreSQL, permite almacenar comentarios de usuarios, historial de predicciones y mecanismos de cache. Al momento de su despliegue se creó una base de datos externa en la nube para consultar los datos reales que utiliza la aplicación web.

Figura 9

Estructura lógica de la base de datos



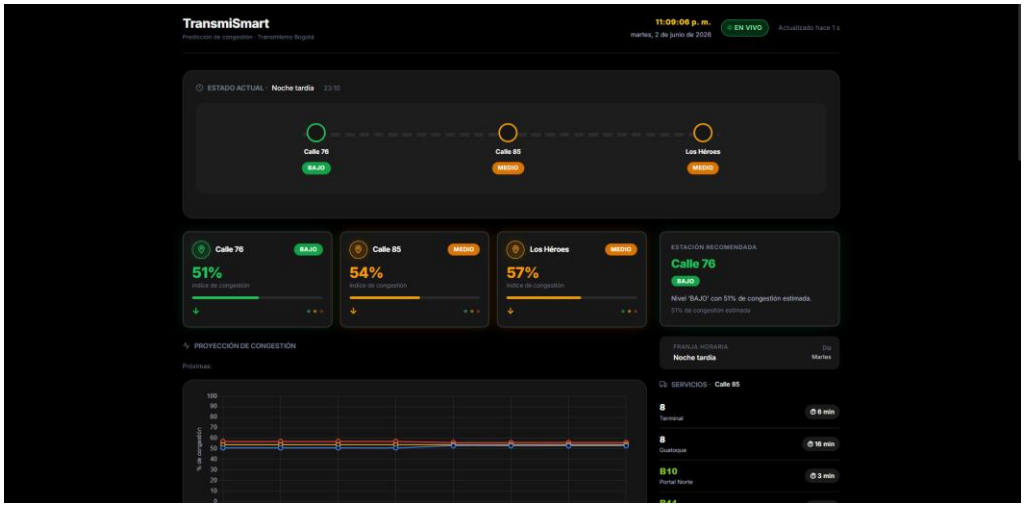
Nota. Elaboración propia.

Capa 5 - Frontend e interacción

En la última capa se permite visualizar el dashboard como aplicación web, consultar los niveles de congestión, predicciones y registrar comentarios de usuarios sobre el esta real de las estaciones de TransMilenio.

Figura 10

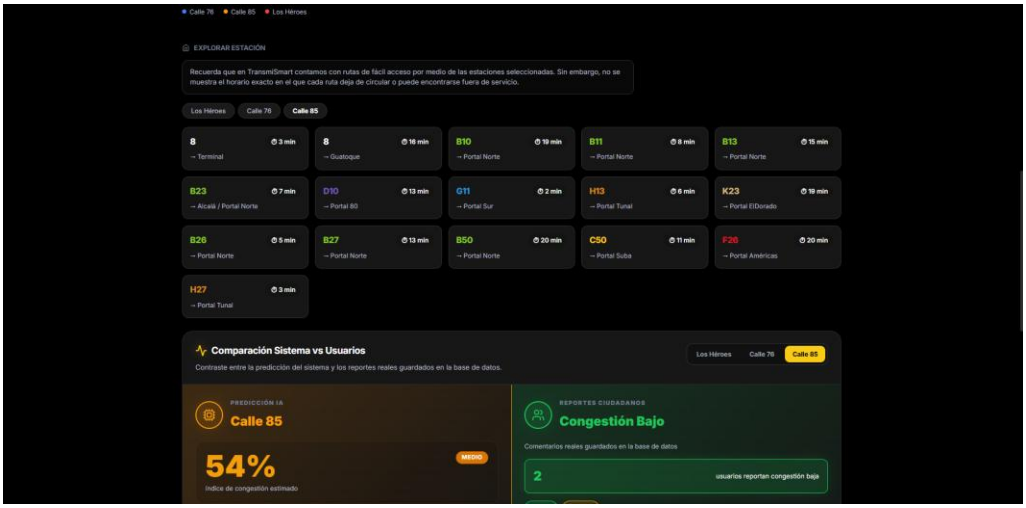
Apartado principal de recomendación de TransmiSmart



Nota. Elaboración propia.

Figura 11

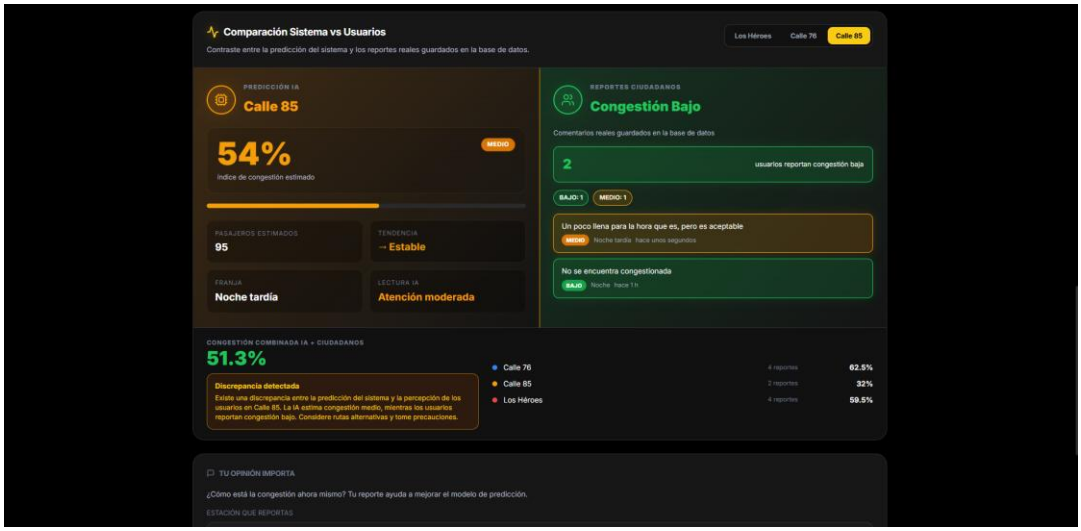
Sección de servicios por estación



Nota. Elaboración propia.

Figura 12

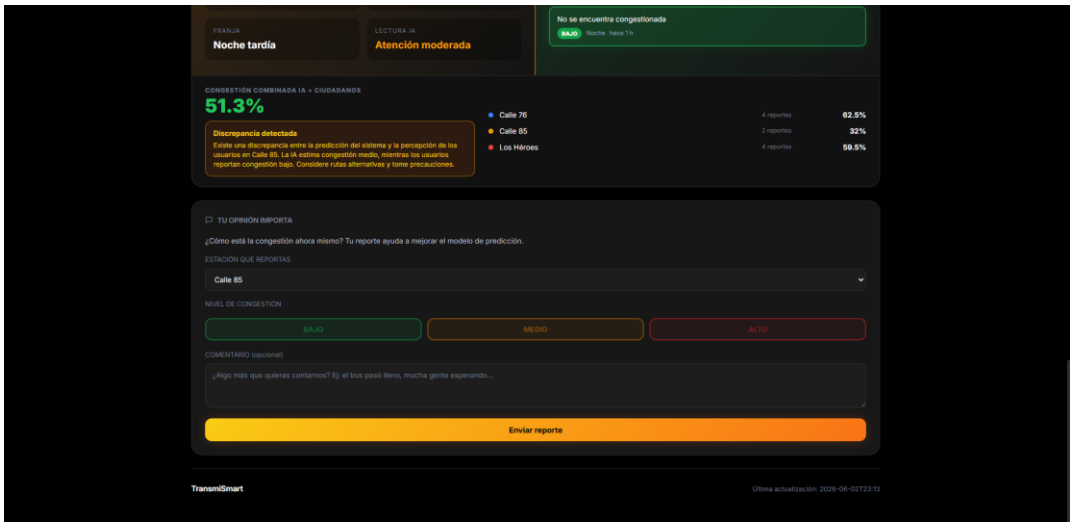
Sección de comparación entre IA vs Usuario



Nota. Elaboración propia.

Figura 13

Formulario de reporte de usuarios.



Nota. Elaboración propia.

Discusión

A base de los resultados obtenidos y la investigación realizada, se logra identificar un gran potencial en la implementación de técnicas de Machine Leraning para lograr fortalecer la gestión inteligente en los sistemas de transporte masivo. Debido a la capacidad predictiva alcanzada por los modelos Random Forest y Gradient Boosting, se puede demostrar que los modelos de ensemble learning son significativamente relevantes para contribuir a la optimización operacional y planificación estratégica de los sistemas de transporte.

Un aspecto a destacar sobre esta investigación es la capacidad predictiva que se logró obtener utilizando únicamente información operacional e histórica de tres estaciones estratégicas de TransMilenio, sin la necesidad de incorporar modelos más complejos de Deep Learning o de mayor demanda en rendimiento computacional. Lo cual, representa una ventaja desde el lado de implementación y escalabilidad.

A pesar de esto, la incorporación de variables como el clima, cambio de nombre a las estaciones y alteraciones operacionales podrían aumentar aún más la capacidad predictiva del modelo. En la investigación se demostró que este tipo de variables afecta directamente en la precisión de las predicciones en escenarios específicos.

Plan de divulgación

Con la finalidad de seguir construyendo y promoviendo este proyecto, se propone un plan de divulgación orientado al ámbito académico, tecnológico e institucional. El objetivo es dar a conocer el potencial de las técnicas de Machine Learning en el área de la predicción de demanda en sistemas de transporte masivo y como fomentar el uso de estas herramientas para la toma de decisiones en entornos urbanos.

Para el ámbito académico, se propone la socialización del resultado del proyecto mediantes presentaciones en espacios institucionales, semilleros de investigación y en lo posible espacios relacionados con ciencia de datos y movilidad inteligente. De igual forma, se pretende la publicación del trabajo en repositorios académicos, revistas y plataformas de acceso abierto para su consulta y reutilización del conocimiento.

Por el lado del ámbito tecnológico, se tiene un enfoque hacia el desarrollo de la documentación detallada del aplicativo web desplegado, incluyendo lujo de transformación de datos, metodología y modelos implementados. Adicionalmente se busca publicar el código fuente y manual de desarrollo en entornos colaborativos como GitHub, con el fin de promover la visibilidad e implementación del proyecto y permitir futuras mejoras por parte de otros usuarios interesados.

En el ámbito institucional, se busca presentar los hallazgos y resultados a entidades relacionadas con movilidad urbana y gestión del transporte público, presentando alternativas como modelos predictivos para la optimización de la planificación operativa, su gestiona miento de rutas y poder anticipar esos escenarios de congestión, mejorando la experiencia de los usuarios.

Finalmente, esta propuesta de divulgación tiene como fin promover el proyecto con la transferencia tecnológica, incentivar futuras investigaciones sobre movilidad urbana y la implementación de inteligencia artificial como solución a construir ciudades más sostenibles, eficiente e inteligentes orientadas al análisis de datos.

Conclusiones

En conclusión, el desarrollo de la aplicación web TransmiSmart permitió demostrar que técnicas como el Machine Learning son efectivas en entornos de predicción de demanda en sistemas de transporte masivos. Los modelos implementados lograron identificar patrones de movilidad y generar predicciones precisas con respecto al comportamiento futuro de los usuarios en las estaciones de TransMilenio.

Debido a los resultados obtenidos, se logró identificar un alto rendimiento predictivo por parte de los modelos, alcanzando valores de precisión mayores al 90% y coeficientes de determinación cercanos a 1. Lo que confirma una capacidad significativa en los algoritmos utilizados Random Forest y HistGradientBoosting para modelar fenómenos asociados a la movilidad urbana en el transporte público.

La implementación de variables temporales, históricas, indicadores de congestión y estadísticas permitieron mejorar la calidad del modelo predictivo, demostrando la importancia de una correcta ingeniera de variables dentro de proyectos de analítica de datos.

Asimismo, el proyecto avanzó hacia una solución tecnológica compuesta por un modelo predictivo, una API que conecta el backend con el frontend, una base de datos en PostgreSQL y una interfaz en una aplicación web tipo dashboard que permite visualizar las estadísticas y los reportes en formato de comentarios ciudadanos.

Por último, los principales aportes de TrrsnsmiSmart es la integración de Inteligencia Artificial y la participación ciudadana. Mediante la posibilidad de realizar registros de reportes de usuarios, el sistema permite complementar las predicciones obteniendo como resultado un

proyecto más robusto. Lo cual, favorece a que el sistema está más cerca de las condiciones reales de cada estación y con una gran oportunidad de mejora continua.

Referencias

TransMilenio. (2025, 17 julio). Estadísticas de oferta y demanda del Sistema Integrado de Transporte Público - SITP - marzo 2025. TransMilenio. https://www.transmilenio.gov.co/publicaciones/154662/estadisticas-de-oferta-y-demanda del-sistema-integrado-de-transporte-publico-sitp-marzo-2025/

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies, 54*, 187–197. https://www.sciencedirect.com/science/article/abs/pii/S0968090X15000935?via%3DihubBanco

Mundial. (2021). Movilidad sostenible en América Latina. Washington, D.C.: Banco Mundial. https://www.bancomundial.org/es/topic/transport/overview

Ortúzar, J. de D., & Willumsen, L. G. (2001). Modelling Transport (3rd ed.). Wiley. https://books.google.com.do/books?id=qWa5MyS4CiwC&printsec=frontcover#v=onepage&q&f

Zhang, J., Zheng, Y., & Qi, D. (2017). Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.10735

Velásquez, J. M. (2009). Análisis de factores que inciden en la demanda del sistema TransMilenio en Bogotá, Colombia. Revista de Ingeniería, (30), 56–64. Universidad de los Andes. https://www.redalyc.org/pdf/1210/121015710012.pdf

Velásquez, J. M. (2009). Análisis de factores que inciden en la demanda del sistema TransMilenio en Bogotá, Colombia. *Revista de Ingeniería, (30)*, 56–64. Universidad de los Andes. https://www.redalyc.org/pdf/1210/121015710012.pdf

Cruz Gil, J. D. (2021). Aprendizaje Automático en la estimación de pasajeros en una aerolínea de bajo costo. https://repositorio.uniandes.edu.co/items/bfc9b6d6-edbf-4310-8bde-b003f35dac7e

Rodríguez Peraza, C. I. Detección y conteo de evasores en el sistema de transporte Transmilenio. https://repositorio.unal.edu.co/items/61f9bd68-1dd1-41c2-b1fd-28eacbfcef85

Bru Diaz, M. M. (2021). Modelos de aprendizaje supervisado para predecir la cantidad de pasajeros que saldrán de la Terminal de Transporte Norte de Medellín a otras regiones de País. https://bibliotecadigital.udea.edu.co/challenge