

Maestría en

Inteligencia Artificial Aplicada

Trabajo previo a la obtención de título de Magister en Inteligencia Artificial Aplicada

AUTORES:

ORTEGA CAMPUES JHON HENRY

RAMOS MAYORGA OSCAR ANDRÉS

TORRES CHANGO CHRISTIAN DAMIAN

TUTORES

MARCELA AZUCENA VENEGAS ESPINOSA

FERNANDA PAULINA VIZCAINO IMACAÑA

TEMA:

Comparación de modelos de aprendizaje automático para predecir el tiempo de llegada entre paradas en el sistema BRT Ecovía de Quito utilizando registros históricos de GPS.

Certificación de autoría

Nosotros, **Ortega Campues Jhon Henry, Ramos Mayorga Oscar Andrés, Torres Chango Christian Damian**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma
Ortega Campues Jhon Henry



Firma
Ramos Mayorga Oscar Andrés



Firma
Torres Chango Christian Damian

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Ortega Campues Jhon Henry, Ramos Mayorga Oscar Andrés, Torres Chango Christian Damian**, en calidad de autores del trabajo de investigación titulado “**Comparación de modelos de aprendizaje automático para predecir el tiempo de llegada entre paradas en el sistema BRT Ecovía de Quito utilizando registros históricos de GPS**”, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, agosto 2026

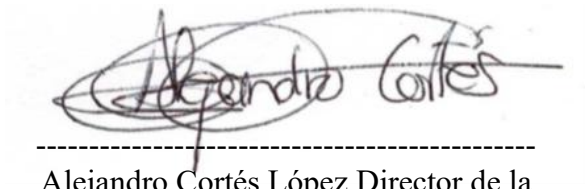
Firma
Ortega Campues Jhon Henry

Firma
Ramos Mayorga Oscar Andrés

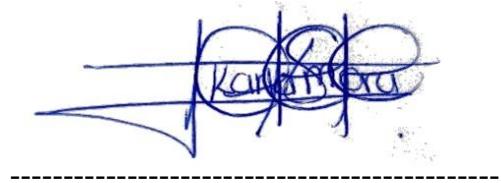
Firma
Torres Chango Christian Damian

Aprobación de dirección y coordinación del programa

Nosotros, **Juan Sebastián Arbeláez** y **Karla Estefanía Mora Cajas** **Coordinador/a UIDE**, declaramos que: **Ortega Campus Jhon Henry, Ramos Mayorga Oscar Andrés, Torres Chango Christian Damian**, son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la
Maestría en Inteligencia Artificial
Aplicada



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

DEDICATORIA

Este trabajo está dedicado a Dios, nuestro creador que nos da sabiduría y salud; y a nuestros padres, quienes a lo largo de toda nuestra vida nos dieron las herramientas para llegar hasta aquí. Su apoyo, su confianza y su presencia constante fueron el respaldo más importante durante este proceso.

Cada esfuerzo que hicieron por nosotros tiene sentido en este logro. Este título es tan nuestro como de ellos.

AGRADECIMIENTOS

Agradecemos a la Universidad Internacional del Ecuador (UIDE) y a los docentes de la Maestría en Inteligencia Artificial Aplicada por la formación recibida durante el programa. Su conocimiento, exigencia y disposición para orientarnos fueron fundamentales para desarrollar este trabajo con solidez académica.

A la empresa Mivilsoft S.A., por facilitar el acceso a los registros históricos del sistema Ecovía para fines académicos. Su colaboración fue clave para que este proyecto trabajara con datos reales del transporte público de Quito.

A nuestras familias y personas cercanas, por el apoyo durante este proceso. Y a nosotros tres como equipo, porque supimos complementarnos y sacar esto adelante juntos.

RESUMEN

El sistema Ecovía de Quito, como parte del transporte público urbano, enfrenta un problema cotidiano: los buses no siempre llegan a las paradas en los tiempos que indica el horario programado, lo que genera incertidumbre en los usuarios y complica la gestión operativa del corredor. Esta variabilidad responde a múltiples factores como el tráfico, las condiciones de la vía y la demanda de pasajeros en cada franja horaria.

Para abordar este problema, el presente trabajo compara seis modelos de aprendizaje automático: promedio histórico, regresión lineal, HistGradientBoosting, XGBoost, LSTM y GRU, entrenados con registros históricos de GPS generados por las propias unidades del sistema. La comparación incluye tanto modelos de redes neuronales recurrentes, que aprenden patrones en secuencias temporales, como modelos de gradient boosting, que utilizan representaciones tabulares de una ventana temporal reciente. El objetivo es determinar cuál familia de modelos se adapta mejor a las condiciones operativas del corredor Ecovía, sin asumir de antemano cuál enfoque es superior.

El proceso parte de la limpieza y preparación del dataset histórico del sistema SAE de Mivilsoft S.A., seguida del análisis exploratorio, la ingeniería de variables y el entrenamiento de los seis modelos con partición temporal hold-out. El desempeño se evalúa con tres métricas estándar: error absoluto medio (MAE, por sus siglas en inglés), raíz del error cuadrático medio (RMSE) y error porcentual absoluto medio (MAPE), y los resultados se presentan en un tablero de visualización web orientado al equipo de operaciones de la EPMTQP.

Palabras clave: aprendizaje automático, comparación de modelos, predicción de tiempos de llegada, transporte BRT, GPS, series temporales, gradient boosting, redes neuronales recurrentes.

ABSTRACT

The Ecovía bus rapid transit system in Quito faces a persistent operational challenge: buses frequently deviate from their scheduled arrival times, generating uncertainty for passengers and complicating corridor management. This variability stems from multiple interacting factors, including traffic congestion, road conditions, and fluctuating passenger demand throughout the day.

To address this challenge, this study compares six machine learning models—historical mean, linear regression, HistGradientBoosting, XGBoost, LSTM, and GRU—trained on historical GPS data collected by the system’s own fleet. The comparison encompasses both recurrent neural network architectures, which capture temporal sequential patterns, and gradient boosting models, which use tabular representations of the recent temporal window rather than recurrent sequence architectures. The central research question is which modeling family best fits the operational characteristics of the Ecovía corridor.

The methodology encompasses data cleaning and preparation of the SAE system records provided by Mivilsoft S.A., exploratory data analysis, feature engineering, and model training using a hold-out temporal split to prevent data leakage. Performance is evaluated using three standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), and results are presented in a web visualization dashboard designed for the EPMTPQ operations team.

Keywords: machine learning, model comparison, arrival time prediction, BRT transport, GPS, time series, gradient boosting, recurrent neural networks.

TABLA DE CONTENIDOS

DEDICATORIA	4
AGRADECIMIENTOS	5
RESUMEN	6
ABSTRACT	7
TABLA DE CONTENIDOS	8
LISTA DE TABLAS	10
LISTA DE FIGURAS	11
CAPÍTULO 1: INTRODUCCIÓN	12
1.1. DEFINICIÓN DEL PROYECTO	12
1.2. JUSTIFICACIÓN E IMPORTANCIA DEL TRABAJO DE INVESTIGACIÓN	13
1.3. ALCANCE	16
1.4. OBJETIVOS.....	17
1.4.1. <i>Objetivo general</i>	17
1.4.2. <i>Objetivos específicos</i>	17
CAPÍTULO 2: REVISIÓN DE LITERATURA	19
2.1. ESTADO DEL ARTE	19
2.1.1. <i>Predicción de tiempos de llegada en transporte público urbano</i>	19
2.1.2. <i>Limitaciones identificadas en la literatura</i>	20
2.1.3. <i>Experiencias en sistemas BRT de América Latina</i>	21
2.2. MARCO TEÓRICO	23
2.2.1. <i>El sistema BRT y las características del corredor Ecovía</i>	23
2.2.2. <i>Redes neuronales recurrentes: LSTM y GRU</i>	28
2.2.3. <i>Modelos de gradient boosting: XGBoost e HistGradientBoosting</i>	31
2.2.4. <i>Modelos de referencia: promedio histórico y regresión lineal</i>	34
2.2.5. <i>Ingeniería de variables para predicción de ETA</i>	35
2.2.6. <i>Métricas de evaluación del modelo</i>	36
2.2.7. <i>Estrategia de validación temporal en series temporales</i>	36
2.3. SÍNTESIS DEL MARCO TEÓRICO Y CRITERIOS PARA LA SELECCIÓN DE MODELOS	39
CAPÍTULO 3: DESARROLLO	44
3.1 DATASET	44
3.1.1 <i>Fuente de los datos</i>	44
3.1.2 <i>Descripción de Dataset</i>	45
3.1.3 <i>Variable objetivo</i>	50
3.1.4 <i>Análisis exploratorio de datos</i>	50
3.2 DESARROLLO DEL TRABAJO	56
3.2.1 <i>Preprocesamiento</i>	56
3.2.2 <i>Diseño de los modelos</i>	58
3.2.3 <i>Entrenamiento y Validación</i>	60
3.2.4 <i>Prototipo</i>	62
CAPÍTULO 4: ANÁLISIS DE RESULTADOS	66
4.1 PRUEBAS DE CONCEPTO	66
4.2 ANÁLISIS DE RESULTADOS	67
4.2.1 <i>Comparación general de modelos</i>	67
4.2.2 <i>Análisis de significancia estadística</i>	71

4.2.3 Análisis por ruta	73
4.2.4 Efecto del tipo de normalización.....	74
4.2.5 Selección del modelo y discusión.....	76
CAPÍTULO 5: CONCLUSIONES Y RECOMENDACIONES	79
5.1 CONCLUSIONES.....	79
5.2 RECOMENDACIONES.....	81
REFERENCIAS BIBLIOGRÁFICAS.....	83
APÉNDICE A.....	88
<i>RUTAS Y TIEMPOS PROGRAMADOS DEL SISTEMA BRT ECOVÍA DE QUITO</i>	<i>88</i>
APÉNDICE B.....	111
<i>REPOSITORIO DE DATOS Y RECURSOS</i>	<i>111</i>
APÉNDICE C	112
<i>VARIABLES EXCLUIDAS DEL MODELO.....</i>	<i>112</i>
APÉNDICE D	115
<i>REPOSITORIO DE CÓDIGO Y NOTEBOOKS DE DESARROLLO</i>	<i>115</i>
APÉNDICE E.....	117
<i>PROTOTIPO WEB DE PREDICCIÓN DEL TIEMPO DE LLEGADA</i>	<i>117</i>

LISTA DE TABLAS

Tabla 1. Descripción de variables del dataset.....	41
Tabla 2. Variables consideradas para el modelo tras el análisis exploratorio	49
Tabla 3. Umbral de eliminación de outliers de eta_next_stop por ruta.....	50
Tabla 4. Pruebas exploratorias de hiperparámetros para modelos recurrentes (LSTM, ruta E1) ...	52
Tabla 5. Ventana temporal óptima seleccionada por ruta para LSTM y GRU	54
Tabla 6. MAE por ruta y modelo - StandardScaler (segundos)	60
Tabla 7. RMSE por ruta y modelo - StandardScaler (segundos)	60
Tabla 8. MAPE por ruta y modelo - StandardScaler (%).....	61
Tabla 9. Ranking final por MAE promedio - StandardScaler (7 rutas).....	62
Tabla 10. Comparación de XGBoost frente a cada modelo — prueba de Wilcoxon e IC 95 %	63
Tabla 11. MAE promedio por modelo - MinMaxScaler vs. StandardScaler (segundos, 7 rutas)...	65
Tabla 12. Tiempos programados de paradas – Ruta E1 (Guamaní – Las Universidades – Guamaní)	79
Tabla 13. Tiempos programados de paradas – Ruta E1M (Guamaní – Marín Central – Guamaní)	83
Tabla 14. Tiempos programados de paradas – Ruta E2_N-S (Río Coca – Quitumbe).....	87
Tabla 15. Tiempos programados de paradas – Ruta E2_S-N (Quitumbe – Río Coca).....	90
Tabla 16. Tiempos programados de paradas – Ruta E3_N-S (Río Coca – Playón de la Marín)	93
Tabla 17. Tiempos programados de paradas – Ruta E3_S-N (Playón de la Marín – Río Coca)	95
Tabla 18. Tiempos programados de paradas – Ruta E4 (Quitumbe – Playón de la Marín – Quitumbe).....	97
Tabla 19. Variables excluidas del modelo	101
Tabla 20. Notebooks que integran el repositorio de desarrollo	104
Tabla 21. Especificaciones técnicas del prototipo	107

LISTA DE FIGURAS

Figura 1. Distribución de eta_next_stop – Tiempo entre paradas.....	46
Figura 2. ETA promedio por hora del día, por ruta, antes de la limpieza de datos.....	46
Figura 3. Correlación de las variables operacionales con eta_next_stop, por ruta (dataset ya preprocesado)	48
Figura 4. Panel del prototipo – comparación de los seis modelos para la ruta E1, viaje 06:34, parada 5.....	56
Figura 5. Panel del prototipo – comparación de los seis modelos para la ruta E4, viaje 06:35, parada 9.....	57
Figura 6. MAE por ruta y modelo, esquema StandardScaler	61
Figura 7. Ranking de modelos por MAE promedio, esquema StandardScaler	62
Figura 8. Comparación de MAE promedio por modelo entre MinMaxScaler y StandardScaler ...	66

CAPÍTULO 1: INTRODUCCIÓN

1.1. Definición del proyecto

Este proyecto propone la comparación de modelos de aprendizaje automático para calcular el tiempo de llegada en cada parada del corredor BRT Ecovía en Quito. Se revisarán seis modelos que pertenecen a tres familias: dos modelos de referencia (el promedio histórico y la regresión lineal), dos modelos de gradient boosting (HistGradientBoosting y XGBoost) y dos modelos que utilizan redes neuronales recurrentes: LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit). De acuerdo con la literatura más reciente, en especial cuando se tiene acceso a grandes volúmenes de datos históricos del GPS (Hosseini et al., 2024; Rashvand et al., 2024), los modelos que se fundamentan en el aprendizaje profundo pueden ser más precisos que las técnicas tradicionales para estimar tiempos de llegada en sistemas de transporte público. Frente a esa tensión en la literatura, este trabajo no asume de antemano cuál enfoque es superior, sino que somete las tres familias de modelos a las mismas condiciones experimentales para determinar cuál se adapta mejor al corredor Ecovía.

Para tomar esa decisión con evidencia, los modelos considerarán variables como la ubicación GPS, la velocidad instantánea, los segmentos previos del recorrido, la hora del día y la categoría del día (laborable, fin de semana o feriado), entre otras que se identifiquen en la fase exploratoria. La comparación entre modelos se realiza sobre el mismo dataset histórico del sistema SAE de Mivilsoft S.A., con las mismas métricas de evaluación: MAE, RMSE y MAPE, y respetando el orden cronológico de los registros mediante una partición temporal (hold-out cronológico).

El aporte principal de este trabajo radica en un hallazgo contrario a la expectativa habitual de la literatura, en un corredor BRT urbano caracterizado por segmentos entre paradas de muy corta duración, los modelos gradient boosting superan a los modelos de redes neuronales recurrentes (LSTM y GRU), pese a que su diseño está específicamente desarrollado para datos

secuenciales y estos suelen ser el enfoque adecuado para la predicción de series temporales. El estudio no solo documenta este resultado, sino que identifica factores que podrían explicar este comportamiento. En particular, la información predictiva parece concentrarse en el estado inmediato del segmento, principalmente en la distancia a la siguiente parada, mientras que la dependencia temporal de largo plazo aporta un valor adicional más limitado bajo las condiciones evaluadas. A esto se suma una contribución de carácter contextual y metodológico: los resultados se obtienen empleando exclusivamente registros GPS institucionales, sin fuentes externas de información, en una ciudad andina de altitud elevada, contexto poco representado en la literatura revisada. En conjunto, este trabajo ofrece evidencia replicable acerca de qué familia de modelos mostró mejor desempeño en el corredor Ecovía bajo las condiciones experimentales estudiadas y qué implicaciones puede tener este resultado para corredores BRT con características similares.

1.2. Justificación e importancia del trabajo de investigación

Relevancia del problema. La previsibilidad del tiempo de llegada (ETA) es fundamental en la decisión de usar el transporte público, superando incluso a la velocidad como prioridad para los usuarios (Hosseini et al., 2024). El Ecovía mueve más de 170.000 personas al día en sus 39 paradas (EPMTPQ, 2024), por lo que cualquier mejora en la precisión del ETA tiene un efecto inmediato sobre una porción significativa de la movilidad quiteña. La importancia de una predicción precisa también se relaciona con la gestión de la operación. Las variaciones en los tiempos de recorrido pueden producir diferencias entre los horarios previstos y el comportamiento real de las unidades, especialmente durante los períodos en los que cambian las condiciones de circulación. Contar con una estimación basada en los registros recientes del vehículo permite representar mejor el estado actual del recorrido que utilizar únicamente valores históricos o tiempos programados.

Desde esta perspectiva, la predicción del ETA puede aportar información tanto para el usuario como para el sistema de transporte. Para el pasajero, una estimación actualizada reduce la incertidumbre asociada al tiempo de espera y facilita la planificación del viaje. Para la operación, los resultados pueden servir como una referencia adicional para analizar el comportamiento de las unidades y reconocer los períodos o segmentos en los que existe una mayor variabilidad. Por tanto, la utilidad del modelo no se limita a obtener un valor numérico de llegada, sino que se relaciona con la posibilidad de utilizar información histórica del sistema para representar su comportamiento operativo. Este dato cobra mayor relevancia si se considera que el 74 % de las mujeres en la ciudad usa este sistema para organizar sus desplazamientos diarios (Quito Cómo Vamos, 2025), convirtiendo la seguridad y el tiempo en factores directamente vinculados a la puntualidad del servicio. A esto se añade la presión competitiva: el Metro de Quito, que sí informa sus horarios de llegada, registra un 89 % de satisfacción, frente al 55 % de la Ecovía (Quito Cómo Vamos, 2025), una diferencia que evidencia con claridad el déficit de satisfacción del corredor.

Justificación académica. Desde el punto de vista investigativo, el trabajo se justifica por las siguientes razones:

- La investigación sobre predicción de ETA con aprendizaje automático está fuertemente concentrada en ciudades de Asia, Europa y América del Norte (Lee y Yoon, 2022; Rashvand et al., 2024). Sus condiciones de tráfico, altitud e infraestructura son poco comparables con las de Quito, lo que limita la aplicabilidad directa de esos modelos al contexto local.

Las condiciones de Quito son únicas: situada a 2.800 msnm y con pendientes marcadas, la ciudad afecta directamente el comportamiento de los buses eléctricos en términos de velocidad y aceleración. Esta variabilidad operativa prácticamente no ha sido estudiada en la literatura de predicción de ETA para sistemas BRT, lo que representa un vacío que este trabajo busca contribuir a llenar.

- La mayoría de estudios que comparan modelos de aprendizaje profundo con modelos de árboles en transporte público se realizan en contextos de alta disponibilidad de datos externos (clima, tráfico en tiempo real, eventos). Este trabajo demuestra que es posible obtener resultados competitivos usando exclusivamente los registros GPS de la propia flota institucional, lo que lo convierte en un caso de estudio replicable para otros corredores de la región.

Justificación tecnológica y social. Un modelo que funcione bien para predecir el ETA tiene varios usos prácticos concretos en la operación del corredor:

- Podría alimentar paneles informativos en las paradas y aplicaciones móviles para que los usuarios planifiquen mejor sus viajes.
- Apoyo a la regulación de frecuencias en tiempo real por parte del área de operaciones de la EPMTPQ.
- La metodología desarrollada aquí también podría utilizarse para otros corredores del sistema Metrobús-Q, incluidos el trolebús y la línea troncal occidental, ya que operan en condiciones similares.
- Por último, este trabajo se basa en los esfuerzos continuos que EPMTPQ ha realizado desde 2025 para renovar su flota con autobuses eléctricos, modernizar las paradas y actualizar el sistema de cobro de tarifas (EPMTPQ, 2025). Añadir la predicción de la hora estimada de llegada sería el siguiente paso lógico.

Viabilidad. La ejecución de este proyecto es viable técnica y operativamente gracias a tres factores fundamentales. Primero, la disponibilidad de datos, habitualmente la principal limitante en este tipo de estudios, está garantizada mediante el acceso a los registros históricos de GPS e indicadores operacionales del Ecovía, facilitados por la Empresa Pública Metropolitana de Transporte de Pasajeros de Quito. Segundo, el desarrollo metodológico se apoya en herramientas de acceso libre como Python, TensorFlow/PyTorch y scikit-learn, que no requieren una

infraestructura de alto costo. Finalmente, la estricta delimitación del alcance a un único corredor asegura que los objetivos propuestos sean alcanzables dentro de los plazos establecidos para la tesis.

1.3. Alcance

La secuencia del proyecto comienza con el procesamiento de los datos y termina con la entrega del prototipo. El dataset histórico de GPS y los registros operacionales de las unidades del corredor Ecovía, que incluyen las siete rutas operativas del sistema (E1, E1M, E2_N-S, E2_S-N, E3_N-S, E3_S-N y E4), son el punto de partida. Se entrena cada modelo por separado para cada una de las siete rutas del corredor Ecovía porque cada ruta tiene un tiempo de recorrido diferente, con un número distinto de paradas y patrones operativos. El identificador de ruta no se utiliza como variable predictora. Su función es dividir los datos según la ruta y permitir el entrenamiento de un modelo independiente para cada una. Antes de construir los modelos, se realiza un análisis exploratorio de datos (AED) para identificar patrones temporales, espaciales y operativos relacionados con el tiempo entre llegadas a las paradas. A partir de este análisis se realiza la ingeniería de variables, incorporando información procedente de la ubicación GPS, la velocidad, los intervalos de tiempo, el día de la semana, los días festivos y la ventana temporal seleccionada para cada ruta. Con estas variables se construyen y evalúan seis modelos.

Como referencias se utilizan el promedio histórico y la regresión lineal. Los modelos HistGradientBoosting y XGBoost corresponden a métodos de potenciación de gradiente, mientras que LSTM y GRU pertenecen al grupo de redes neuronales recurrentes. La comparación entre estos seis enfoques permite determinar cuál presenta un mejor comportamiento para las condiciones operativas del corredor Ecovía. Para evaluar su desempeño se utilizan MAE, RMSE y MAPE. Como entregable final del trabajo se construirá un tablero de visualización web destinado al equipo de operaciones de la EPMTPQ, en el cual los operadores podrán consultar las

predicciones del modelo seleccionado para cada ruta. Este tablero opera como un visualizador exploratorio interno, no como una aplicación de información al pasajero ni como un sistema en producción, y su propósito es demostrar que las predicciones del modelo son interpretables y aprovechables dentro del flujo operativo del corredor.

Fuera del alcance del proyecto se encuentran:

- El desarrollo de una aplicación móvil de información al pasajero en tiempo real.
- La integración en producción con los sistemas de gestión de flota de la EPMT PQ.
- El análisis de otros corredores del Metrobús-Q distintos al Ecovía, como el Trole o la Troncal Occidental.
- La predicción de demanda de pasajeros y la optimización de frecuencias de despacho.
- El manejo de eventos extraordinarios no registrados en los datos, como cierres viales, manifestaciones o emergencias.

1.4. Objetivos

1.4.1. *Objetivo general*

Comparar seis modelos de aprendizaje automático promedio histórico, regresión lineal, XGBoost, HistGradientBoosting, LSTM y GRU para predecir el tiempo de llegada entre paradas del corredor BRT Ecovía de Quito usando registros históricos de GPS del sistema SAE de Mivilsoft S.A., con el fin de identificar el enfoque que mejor se adapta a las condiciones operativas del corredor.

1.4.2. *Objetivos específicos*

1. Con el fin de que los datos representen con precisión las condiciones reales de funcionamiento del corredor, se busca obtener un conjunto de datos históricos del GPS del corredor BRT Ecovía de Quito, cuya calidad haya sido confirmada para la

capacitación de modelos. Esto se logrará mediante la integración, validación y limpieza de los registros del sistema SAE.

2. Para evaluar la precisión predictiva en cada parada del corredor de la Ecovía, se entrenarán seis modelos de aprendizaje automático: regresión lineal, promedio histórico, HistGradientBoosting, XGBoost, GRU y LSTM. La ingeniería de variables y la validación con partición temporal preservarán el orden cronológico de los registros y garantizarán la integridad metodológica.
3. Determinar qué familia de modelos de aprendizaje automático predice la ETA del sistema Quito Ecovía con mayor precisión comparando sistemáticamente seis modelos con MAE, RMSE y MAPE, tanto en todas las rutas como mediante análisis separados para cada ruta.

CAPÍTULO 2: REVISIÓN DE LITERATURA

2.1. Estado del arte

2.1.1. *Predicción de tiempos de llegada en transporte público urbano*

Durante años, predecir cuándo llegaría un bus dependía de promedios históricos y modelos estadísticos tradicionales que asumen patrones estables en el tráfico. El problema es que el tráfico urbano raramente es estable, y esos métodos se quedan cortos cuando la variabilidad aumenta (Rashvand et al., 2024; Kumar et al., 2025; Singh y Kumar, 2022). En respuesta, los modelos de aprendizaje profundo y de potenciación de gradiente han avanzado mediante arquitecturas que aprenden directamente de los datos:

- **Redes neuronales recurrentes (RNN):** LSTM y GRU son referentes establecidos para problemas de predicción temporal. Su capacidad para capturar patrones en diferentes escalas de tiempo les otorga una ventaja que los modelos lineales no pueden igualar (Xie et al., 2021).
- **Modelos de gradient boosting:** Modelos de potenciación de gradiente: En la predicción de ETA, XGBoost y variantes relacionadas han producido resultados competitivos, especialmente cuando las variables más relevantes son instantáneas. También destacan porque requieren menos tiempo de entrenamiento y ofrecen mayor interpretabilidad (Mendoza-Mendoza et al., 2022).

Sin embargo, la arquitectura por sí sola no lo explica todo. El contexto operativo también es importante. En las redes de autobuses de Seúl, Lee y Yoon (2022) utilizaron métodos de centralidad de flujo y lograron un MAPE de tan solo el 1,19 %. En Boston, la incorporación de sensores IoT al modelo redujo el error a menos de 4 minutos (Hosseini et al., 2024).

Finalmente, al aplicar estos sistemas en corredores BRT latinoamericanos similares al de Quito, la literatura evidencia que es posible alcanzar predicciones precisas con un margen de error

de ± 90 segundos en el 85 % de los casos (Mendoza-Mendoza et al., 2022), y que el horario de operación y la distancia al siguiente punto de parada se consolidan como los predictores de mayor impacto.

Los sistemas de BRT tienen características operativas que los diferencian del transporte tradicional, lo cual afecta directamente la precisión de las predicciones. El corredor Ecovía en Quito opera a través de un recorrido exclusivo con paradas fijas y horarios determinados, lo que produce datos GPS más confiables. Sin embargo, la presencia de cruces comunes con tráfico no uniforme en ciertas áreas genera una variabilidad que los modelos deben aprender a manejar. Ranjbari et al. (2022), al investigar este tipo de corredores en ciudades de Estados Unidos, descubrieron que la variabilidad en los intervalos entre paradas aumenta hasta un 40 % en las secciones de carril compartido con respecto a las secciones exclusivas. Esto pone de manifiesto la importancia de contar con modelos capaces de diferenciar entre segmentos del corredor.

En cuanto al volumen de datos necesario para entrenar modelos de aprendizaje automático en transporte público, la literatura sugiere que La disponibilidad de grandes volúmenes de registros históricos puede favorecer la representación de diferentes condiciones operativas y patrones temporales. El dataset del sistema Ecovía proporcionado por Mivilsoft S.A. cuenta con 28 millones de registros históricos originales, de los cuales 24,3 millones corresponden a registros limpios tras el preprocesamiento (Rashvand et al., 2024).

2.1.2. Limitaciones identificadas en la literatura

La revisión de la literatura deja ver varios vacíos que dan sentido a este trabajo:

- El grueso de la investigación en predicción de ETA proviene de ciudades asiáticas, europeas o norteamericanas. Ciudades latinoamericanas como Quito, con altitud elevada y topografía particular, casi no aparecen en esa literatura.

- La mayoría de comparaciones entre modelos de aprendizaje profundo y modelos de gradient boosting se realizan en contextos con acceso a datos externos como condiciones climáticas, eventos y sensores de tráfico. La comparación usando exclusivamente datos GPS institucionales es un enfoque menos explorado.
- El efecto que tiene la renovación tecnológica de la flota sobre modelos entrenados con datos históricos previos al cambio es algo que muy pocos trabajos han analizado de frente, y el Ecovía constituye un caso de estudio adecuado para abordar esa pregunta.
- Otro punto débil frecuente es la validación: muchos modelos se evalúan con particiones aleatorias de los datos, lo que permite que información del futuro filtre hacia el entrenamiento y produzca resultados artificialmente buenos. La partición temporal hold-out, que respeta el orden cronológico, se aplica poco.

La eliminación de registros con velocidad cero introduce un sesgo sistemático significativo. El filtro excluye los autobuses estacionados, pero también elimina las observaciones de congestión severa, cuando un autobús puede estar moviéndose prácticamente a velocidad cero. Como resultado, los modelos entrenados con este conjunto de datos aprenden cómo se comporta el corredor en condiciones de flujo normal y pueden subestimar la hora estimada de llegada durante episodios de congestión extrema. Este sesgo, inherente al diseño del preprocesamiento, no invalida los resultados, pero debe considerarse al interpretar las predicciones en condiciones de tráfico detenido.

2.1.3. Experiencias en sistemas BRT de América Latina

El transporte BRT en ciudades latinoamericanas comparte rasgos estructurales alta densidad de usuarios, variabilidad por tipo de día y condiciones viales heterogéneas que lo

distinguen de los sistemas estudiados en la literatura predominante. Pese a ello, los trabajos que aplican modelos predictivos a estos contextos son escasos y, en su mayoría, recientes.

Mendoza-Mendoza et al. (2022) trabajaron con datos del Metrobús Línea 6 de Ciudad de México, un corredor BRT con características operativas cercanas al Ecovía de Quito en términos de longitud, número de paradas y demanda. Sus resultados mostraron que los modelos de series temporales lograban errores por debajo de 90 segundos en el 85 % de los casos, y que la hora del día explicaba una proporción mayor de la variabilidad que cualquier otra variable operacional. Este hallazgo resulta importante para el avance de esta investigación puesto que el Ecovía opera con horarios que presentan variaciones de demanda considerables entre horas pico y valles.

Un aspecto que ha sido poco analizado a nivel regional es el efecto de la modernización tecnológica de las unidades en los modelos que fueron entrenados con datos históricos. Desde 2025, el Ecovía ha incorporado mejoras en sus estaciones y vehículos eléctricos (EPMTPQ, 2025), lo que convierte a esta ruta en un caso de estudio idóneo para examinar cómo dicha transformación afecta la transferibilidad de los modelos desarrollados con información anterior a la renovación de la flota.

En cuanto a la cobertura regional, los estudios en América Latina sobre la predicción del tiempo estimado de llegada (ETA) en sistemas BRT son notablemente escasos. Aparte del trabajo de Mendoza-Mendoza y colaboradores (2022) con datos del Metrobús de la Ciudad de México, en la revisión realizada para este estudio no se identificaron investigaciones directamente comparables en otros corredores BRT de la región incluyendo países con sistemas BRT consolidados como Colombia, Brasil o Perú que apliquen metodologías de aprendizaje automático para predecir los tiempos de arribo de una parada a otra. Esta ausencia no implica que el problema no exista, sino que la investigación aplicada en estos contextos aún no ha alcanzado el volumen

crítico necesario para generar comparaciones sistemáticas. El presente trabajo, al operar con datos reales de un corredor ecuatoriano, contribuye a llenar ese vacío geográfico y operativo en la literatura.

2.2. Marco Teórico

2.2.1. El sistema BRT y las características del corredor Ecovía

Un sistema BRT combina características específicas de la carretera, estaciones de prepago y embarque a nivel con servicio frecuente para ofrecer una experiencia de tránsito similar al metro, pero con una inversión menor. Los corredores BRT pueden compartir secciones de pista con tráfico mixto, introduciendo variabilidad en los tiempos de tránsito que los modelos predictivos deben capturar. El corredor Ecovía de Quito, operado por la EPMTQP desde 1995, tiene características decisivas para la predicción: su red se extiende desde la Terminal Quitumbe en el sur hasta la Estación Río Coca, con carril segregado por barrera física en la mayor parte del recorrido.

- En 2023, el sistema transportó un promedio de 170.376 pasajeros al día (EPMTQP, 2023), aunque la demanda varió notablemente entre las horas punta y las horas valle, diferencias que el modelo debe aprender a reconocer.
- Desde 2025, Ecovía ha operado aproximadamente 149 vehículos, incluyendo unidades tanto antiguas como nuevas (EPMTQP, 2025). Debido a que ambos tipos de vehículos tienen un rendimiento diferente, los tiempos de viaje reales entre paradas han cambiado, lo que hace necesario recalibrar los modelos entrenados con datos de la flota anterior.

El sistema Ecovía opera mediante 7 rutas con tiempos programados de parada a parada, los cuales se detallan en el Apéndice A.

El sistema BRT se diferencia del transporte público convencional en que opera sobre infraestructura segregada que minimiza las interacciones con el tráfico mixto, lo que reduce la variabilidad del tiempo de recorrido y mejora la predictibilidad del servicio.

Sin embargo, la efectividad de esta división no es la misma en todo el corredor: los tramos con carril exclusivo protegido por una barrera física presentan tiempos de viaje considerablemente más estables que aquellos con carril preferente que admite giros a la derecha, donde la interrupción del tráfico general introduce una variabilidad operativa adicional que los modelos de pronóstico deben aprender a manejar (Ranjbari et al., 2022). En el recorrido de Ecovía en Quito, esta diferencia se refleja en variaciones de hasta el 40 % en la fluctuación del ETA entre paradas ubicadas en secciones exclusivas y las situadas en intersecciones críticas de tráfico mixto, haciendo que el modelo de predicción sea una herramienta de mayor valor, especialmente en los tramos con alta variabilidad donde la planificación tradicional es menos confiable.

La incorporación de autobuses eléctricos en el corredor Ecovía a partir de 2025 introduce un factor importante para los modelos que predicen el Tiempo Estimado de Llegada. Los vehículos eléctricos poseen distintas capacidades de aceleración y frenado en comparación con los autobuses diésel convencionales: su torque instantáneo les posibilita un arranque más dinámico al dejar las paradas, lo que podría reducir los tiempos de espera en las estaciones y modificar los esquemas de velocidad en las salidas. Adicionalmente, el peso particular de las baterías afecta el comportamiento del vehículo en pendientes pronunciadas, las cuales son frecuentes en el trayecto del Ecovía, dado que esta cruza zonas urbanas con desniveles considerables entre el sector norte y el sur del corredor. Estas diferencias operativas implican que los modelos desarrollados únicamente con información de la flota diésel podrían disminuir su exactitud progresivamente a medida que incrementa el número de autobuses eléctricos operativos, haciendo necesario plantear estrategias para reentrenar los modelos de forma periódica o actualizarlos continuamente (EPMTPQ, 2025).

Las siete rutas operativas del sistema Ecovía presentan características estructurales distintas que son relevantes para el diseño del modelo predictivo. La ruta E1 es la de mayor extensión y demanda, con 39 paradas entre el terminal Quitumbe al sur y el terminal Río Coca al norte, operando en ambas direcciones bajo los identificadores E1 y E1M para el ramal al Marín. Las rutas E2_N-S y E2_S-N cubren el tramo principal en ambas direcciones sin desviación al Marín, mientras que E3_N-S y E3_S-N operan el tramo norte del corredor entre el sector de la Universidad Central y el terminal Río Coca. La E4, con su diseño circular y al considerar La Marín como un punto de intercambio, muestra la mayor variabilidad operativa del sistema, algo que se debe a la complejidad del punto de conexión. Esta diversidad de trayectorias justifica la necesidad de entrenar modelos independientes para cada ruta, en lugar de uno solo a nivel global. La razón es que las condiciones del tráfico, las pendientes del camino y los patrones de demanda varían de forma considerable entre las rutas, y un modelo que lo intente abarcar todo terminaría por suavizar estas diferencias en vez de capturarlas con precisión.

2.2.1.1. Caracterización empírica del corredor a partir del análisis exploratorio

El análisis exploratorio de datos (EDA) realizado sobre el dataset del sistema SAE permite caracterizar el comportamiento operativo real del corredor Ecovía con una precisión que no es posible obtener a partir de los registros operacionales publicados ni de los horarios programados. Con base en 24.308.817 registros GPS limpios correspondientes al período septiembre 2025 marzo 2026, el ETA promedio global del corredor es de 99,8 segundos, con una mediana de 78 segundos y una desviación estándar de 68,4 segundos. Esta asimetría positiva en la que la media queda por encima de la mediana, indica que la distribución del tiempo estimado de llegada tiene una cola larga hacia la parte derecha: un gran número de segmentos finalizan por debajo de 80 segundos, pero existen casos de retrasos extremos que elevan la media a casi 100 segundos. Este

tipo de distribución asimétrica afecta directamente a la elección del modelo predictivo, ya que los modelos que buscan reducir el error cuadrático suelen ser más propensos a estos sucesos extremos comparados con los que buscan minimizar el error absoluto.

El análisis por ruta revela diferencias estructurales importantes que justifican el entrenamiento de modelos independientes en lugar de un modelo global unificado. La ruta E3_N-S presenta el ETA promedio más bajo del corredor con 78 segundos, seguida de E3_S-N con 82 segundos, lo que refleja que estos ramales operan sobre tramos más cortos y con menor variabilidad en su trazado. En el extremo opuesto, la ruta E4 de carácter circular con 22,84 kilómetros de recorrido registra un ETA promedio de 114 segundos, explicado por la mayor longitud de sus segmentos inter-parada y por las complejidades operativas del nodo de La Marín. Las rutas E1 y E1M tienen tiempos de llegada promedio intermedios de 106 y 105 segundos, respectivamente, lo que se ajusta a su función como eje principal del corredor. La diferencia entre la ruta más corta y la más larga alcanza el 46%, lo que sugiere que un modelo entrenado con todas las rutas mixtas tendería a realizar predicciones de valor promedio y perdería precisión en las rutas con comportamientos más extremos.

El análisis de correlaciones entre las variables predictoras y la variable objetivo revela que `distance_to_next_stop` es, con amplia diferencia, el predictor más potente del ETA en el corredor Ecovía. La correlación de Pearson entre esta variable y el ETA alcanza valores de hasta +0,80 en la ruta E4 y oscila entre +0,12 y +0,80 según la ruta analizada. Esta relación refleja una realidad operativa obvia pero frecuentemente ignorada en los modelos basados exclusivamente en series temporales: el tiempo que tarda un bus en llegar a la siguiente parada depende fundamentalmente de cuántos metros le faltan para llegar. La velocidad instantánea presenta una correlación negativa baja (-0,16 en promedio), lo que en apariencia contradice la intuición, pero se explica porque reportes de velocidad alta corresponden a segmentos donde el bus ya ha avanzado

significativamente y el ETA restante es pequeño. La variable de tiempo desde la última parada (`time_since_last_stop`) presenta correlaciones positivas moderadas (entre +0,20 y +0,45 según la ruta), lo que sugiere que los buses que llevan más tiempo en movimiento tienden a estar más cerca de la siguiente parada, pero con mayor variabilidad acumulada.

Un hallazgo especialmente significativo del análisis exploratorio es la correlación prácticamente nula entre el tiempo programado de paso (`scheduled_passing_time`) y el ETA real. Los coeficientes de correlación de Pearson entre estas dos variables oscilan entre -0,03 y +0,06 según la ruta, lo que estadísticamente indica ausencia de relación lineal. Este resultado tiene implicaciones operativas directas: el sistema de planificación oficial del corredor Ecovía no refleja el comportamiento real de los buses en operación, y cualquier modelo predictivo que incorpore el horario programado como variable de entrada estaría introduciendo ruido sistemático en lugar de información útil. Adicionalmente, el análisis de las variables de altitud reveló signos contradictorios entre rutas positivo en E3_N-S y negativo en E4 lo que indica que la altitud no tiene un efecto uniforme sobre el ETA en el corredor y que su inclusión como variable global podría degradar la generalización del modelo. Por estas razones, tanto el tiempo programado como la altitud fueron excluidos del conjunto final de variables predictoras.

El análisis temporal del ETA confirma los patrones de hora pico esperados en un sistema de transporte urbano, pero con matices específicos del corredor Ecovía. Los horarios con mayor variabilidad son los comprendidos entre las 06:30 y las 08:30, y de 17:00 a las 19:00 en los días laborables, donde el tiempo medio de llegada estimado se incrementa entre un 15% y un 20% respecto a las horas de menor actividad. Por otra parte, la influencia de los sábados, domingos y festivos sobre el tiempo estimado de arribos es menor que la que se pensaba en principio. La diferencia entre el tiempo medio de llegada estimado un día laborable (100 segundos), un sábado o

domingo (100 segundos) y un festivo (98 segundos) es mínima, lo que indica que la ruta funciona de manera bastante regular independientemente del día. Esta homogeneidad entre tipos de día es consistente con el hecho de que el Ecovía opera con frecuencias elevadas durante toda la semana y no reduce drásticamente su servicio en días no laborables, por lo que las condiciones de tráfico y demanda son más estables de lo que se observa en sistemas de menor frecuencia.

2.2.2. *Redes neuronales recurrentes: LSTM y GRU*

Las redes neuronales recurrentes (RNN) son modelos de aprendizaje profundo que tienen como propósito el procesamiento de secuencias temporales, donde la salida de cada paso se utiliza como entrada complementaria para el siguiente. Esta estructura es adecuada para problemas en los que la historia temporal importa, incluida la predicción de la hora estimada de llegada para el transporte público. El proyecto se centra en dos arquitecturas principales:

- ***LSTM (Long Short-Term Memory)***: Con tres compuertas (de entrada, de olvido y de salida), esta arquitectura controla cómo la información se mueve a través de la celda de memoria. Como resultado, puede preservar el contexto relevante a lo largo de secuencias largas y abordar el problema del gradiente evanescente (Waqas y Wannasingha Humphries, 2024). El historial operativo reciente es especialmente importante al predecir el tiempo estimado de llegada (ETA), y LSTM está diseñado para capturar este tipo de dependencia (Xie et al., 2021; Chughtai et al., 2022).
- ***GRU (Gated Recurrent Unit)***: A diferencia de una LSTM, una GRU combina las compuertas de entrada y olvido en una sola. Hwang diseño reduce el número de parámetros y acelera el entrenamiento, sin disminuir significativamente la precisión. Hwang et al. (2022) identifican la GRU como la opción más viable cuando el tiempo de computación es limitado, aunque la LSTM sigue siendo más precisa en escenarios de tráfico con una variabilidad considerable.

La arquitectura LSTM incorpora una celda de estado que permite conservar información relevante a lo largo de una secuencia temporal. El flujo de esta información se regula mediante tres compuertas que se actualizan en cada paso de tiempo. La compuerta de olvido determina qué información almacenada previamente se descarta, mientras que la compuerta de entrada controla la información nueva que se incorpora a la celda. Por su parte, la compuerta de salida determina qué parte del estado interno se transmite a la siguiente capa.

Las LSTM fueron desarrolladas para hacer frente al problema del gradiente evanescente presente en las redes neuronales recurrentes básicas. En este tipo de redes, el gradiente asociado a la función de pérdida puede disminuir considerablemente al propagarse a través de secuencias largas, dificultando el aprendizaje de relaciones temporales que se mantienen durante períodos prolongados. En el caso de la predicción de la hora estimada de llegada (ETA), esto resulta relevante porque la información registrada en las paradas anteriores de una misma ruta puede aportar información para estimar el tiempo hasta las siguientes paradas.

El diseño de la GRU es más sencillo que el de la LSTM porque combina las puertas de olvido y de entrada en una sola puerta de reinicio y elimina las representaciones separadas del estado de la celda y de la salida oculta. Esta reducción en el número de parámetros entrenables en configuraciones equivalentes, una cantidad de parámetros que puede ser menor que la del LSTM hace que GRU converja más rápidamente durante el entrenamiento y sea menos susceptible al sobreajuste cuando el conjunto de datos de entrenamiento es de tamaño moderado (Cho et al., 2014).

En aplicaciones de transporte urbano donde el tiempo de entrenamiento es un factor operativo relevante, por ejemplo, cuando se requiere reentrenar el modelo frecuentemente con datos recientes GRU ofrece una alternativa computacionalmente más eficiente lo que puede ofrecer un equilibrio favorable entre costo computacional y precisión en determinados problemas de series temporales. Chughtai et al. (2022) descubrieron que los modelos GRU con mecanismos

de atención tienen un rendimiento comparable al de los modelos LSTM a la hora de predecir tiempos de viaje. Además, ofrecen una mayor interpretabilidad, ya que pueden identificar los momentos clave utilizados para realizar cada predicción.

Una decisión metodológica clave al aplicar modelos LSTM y GRU a la predicción del tiempo estimado de llegada (ETA) en sistemas de transporte es la elección del horizonte temporal de entrada. Esta ventana establece el número de registros GPS consecutivos que recibe el modelo antes de producir una predicción. Su tamaño apropiado depende tanto de la frecuencia de registro como de la duración habitual de un segmento entre paradas. En el sistema Ecovía, los registros GPS se recopilan aproximadamente cada 11 segundos, mientras que los segmentos duran un promedio de 1,3 minutos. Con una ventana de 10 pasos, el modelo recibe, por lo tanto, unos 110 segundos de contexto temporal, casi la duración de un segmento completo. Sin embargo, el tamaño óptimo de la ventana no es necesariamente el mismo para todas las rutas, ya que el número de informes GPS disponibles por segmento difiere; por lo tanto, cada ruta requiere una calibración individual en lugar de un parámetro global. Waqas y Wannasingha Humphries (2024) revisan críticamente las variantes de LSTM y GRU para la predicción de series temporales y recomiendan calibrar la ventana temporal mediante validación cruzada cronológica en lugar de elegirla heurísticamente. Su recomendación refuerza la necesidad de una estrategia de validación rigurosa que preserve el orden temporal de los datos.

Los hallazgos del análisis exploratorio tienen implicaciones directas sobre la hipótesis de rendimiento relativo entre las familias de modelos. En un sistema donde la variable de mayor poder predictivo es la distancia a la siguiente parada una variable espacial instantánea los modelos que procesan secuencias temporales de múltiples pasos no tienen garantizado un rendimiento superior a los modelos que operan sobre el estado actual del vehículo. Los modelos LSTM y GRU están diseñados para capturar dependencias temporales de largo alcance, por lo que resultan

especialmente útiles cuando el historial reciente de un sistema ayuda a predecir su estado futuro. Sin embargo, esta ventaja es más limitada en segmentos con una duración media de 1,3 minutos. En ciertas rutas, la correlación entre distancia y ETA alcanza 0, lo que indica una fuerte relación lineal y deja menos margen para la explicación de la memoria temporal. Dado que trabajan directamente con variables de estado instantáneas, los modelos de potenciación de gradiente pueden capturar el papel dominante de la distancia de forma más eficiente. Con base en los hallazgos del análisis exploratorio de datos (EDA), esta es la hipótesis que el estudio pondrá a prueba con evidencia empírica del corredor de Ecovía.

2.2.3. Modelos de gradient boosting: XGBoost e HistGradientBoosting

Los modelos de potenciación de gradiente construyen una cadena de árboles de decisión de forma secuencial, donde cada árbol corrige las debilidades del modelo anterior. A diferencia de las redes neuronales recurrentes, no aprenden de secuencias dependientes del tiempo. Operan sobre los valores actuales de las variables de entrada, lo que los hace particularmente eficientes cuando los predictores más informativos son instantáneos. En lugar de procesar una secuencia mediante una arquitectura recurrente, esta investigación proporciona a los modelos variables tabulares de la ventana temporal reciente, aplanadas en un solo vector. Por lo tanto, la información histórica se incluye sin una arquitectura secuencial.

El XGBoost (Extreme Gradient Boosting) (Chen y Guestrin, 2016) es una adaptación optimizada del gradient boosting que tiene como objetivo bajar el tiempo de formación mediante la paralelización y evitar el sobreajuste a través de la regularización. Su capacidad para procesar grandes cantidades de datos con múltiples variables, esto lo ha posicionado como uno de los modelos más utilizados en aplicaciones predictivas sobre movilidad urbana.

HistGradientBoosting es una versión que utiliza histogramas y antes de construir los árboles, agrupa los valores continuos de cada variable en intervalos discretos, gracias a esto el costo computacional disminuye considerablemente.

Esto permite entrenar modelos sobre conjuntos de datos de varios millones de registros en tiempos bastante menores que los métodos tradicionales de gradient boosting (Pedregosa et al., 2011). Los modelos de gradient boosting resultan competitivos cuando la variable objetivo depende principalmente del estado instantáneo del vehículo como la distancia a la siguiente parada o la velocidad actual en lugar de patrones temporales extendidos, en el contexto de predicción de ETA para sistemas BRT.

El modelo de promedio histórico, pese a su aparente simplicidad, constituye un punto de referencia metodológicamente riguroso en la literatura de predicción de tiempo de viaje. Su construcción requiere decisiones no triviales sobre la granularidad temporal con que se definen las condiciones similares: un modelo de promedio histórico que promedia por hora del día y tipo de día laborable, fin de semana, feriado captura más variabilidad que uno que solo promedia por tramo, pero requiere un volumen de datos históricos suficiente para que el promedio condicionado sea estadísticamente representativo en cada celda de la tabla de condiciones. Mendoza-Mendoza et al. (2022) estudiaron el Metrobús de la Ciudad de México utilizando el promedio histórico estratificado por hora y día como referencia. Durante los períodos de máxima demanda, cuando la relación entre las variables predictoras y la ETA se vuelve menos lineal, esta referencia superó a la regresión lineal simple. Este hallazgo respalda el uso del promedio histórico como límite inferior, no como modelo de producción: cualquier enfoque más complejo debería superar su precisión lo suficiente como para justificar el costo computacional adicional.

La regresión lineal, por su parte, introduce la capacidad de ponderar múltiples variables predictoras de manera simultánea, lo que le permite capturar relaciones de primer orden entre

variables como la velocidad instantánea, la hora del día y la distancia a la siguiente parada con el tiempo de llegada. Aunque este modelo asume que estas relaciones son aditivas y lineales una simplificación que raramente se cumple en sistemas de transporte urbano con intersecciones semaforizadas y variabilidad climática su interpretabilidad y la facilidad de diagnóstico de los coeficientes lo convierten en un paso metodológico relevante entre el promedio histórico y los modelos no lineales. Además de servir como modelo de referencia, la regresión lineal proporciona una base interpretable para contrastar el aporte de modelos capaces de capturar relaciones no lineales e interacciones.

El entrenamiento de XGBoost se diferencia del gradient boosting tradicional en tres aspectos principales. Estas propiedades lo hacen idóneo para grandes conjuntos de datos como el sistema Ecovía la complejidad de los árboles individuales como los pesos asignados a las hojas terminales. Esto reduce considerablemente el sobreajuste en comparación con implementaciones anteriores. El método también aplica una expansión de Taylor de segundo orden al gradiente de la función de pérdida, lo que ayuda al modelo a converger de forma más rápida y precisa durante el entrenamiento. XGBoost poda adicionalmente los árboles, eliminando las ramas que no mejoran el rendimiento predictivo; el resultado es un modelo más compacto con mayor capacidad de generalización (Chen y Guestrin, 2016). Esta combinación resulta idónea para grandes conjuntos de datos como el sistema Ecovía, que contiene más de 24 millones de registros. Con datos de esta magnitud, la estabilidad del entrenamiento es tan importante como la precisión final.

HistGradientBoosting también incorpora una optimización durante el preprocesamiento de los datos de entrada. Antes de construir los árboles de decisión, clasifica los valores continuos de cada predictor en un número fijo de intervalos discretos, generalmente entre 128 y 256. Esto reduce el coste computacional de identificar el punto de división óptimo en cada nodo del árbol. Esta discretización permite que el algoritmo opere con matrices de datos comprimidas en memoria, Esta discretización reduce considerablemente el costo computacional de la búsqueda de

divisiones y facilita el entrenamiento sobre conjuntos de datos grandes (Pedregosa et al., 2011). En el contexto del corredor Ecovía, donde el entrenamiento de modelos independientes por ruta implica procesar entre 500.000 y 3.000.000 de registros según la ruta, esta eficiencia computacional convierte a HistGradientBoosting en una alternativa prácticamente equivalente a XGBoost en términos de precisión, pero con ventajas operativas significativas en el tiempo de reentrenamiento periódico del modelo.

2.2.4. Modelos de referencia: promedio histórico y regresión lineal

Para que la comparación entre modelos tenga validez científica, es necesario establecer referencias base contra las cuales medir el desempeño de los modelos más complejos. En este trabajo se utilizan dos modelos de referencia ampliamente aceptados en la literatura de predicción de series temporales en transporte público.

En esta investigación, el promedio histórico se implementa como el promedio de `eta_next_stop` calculado sobre el conjunto de entrenamiento de cada ruta y utilizado como referencia para el conjunto de prueba. Esta implementación constituye un baseline simple que establece una cota de comparación frente a modelos de mayor complejidad. Si un modelo complejo no supera este promedio, no tiene justificación práctica.

La regresión lineal estima la hora estimada de llegada (ETA) expresándola como una función matemática directa de sus factores predictivos, asignando a cada factor un peso específico en una combinación lineal. No se basa únicamente en promedios históricos, sino que examina simultáneamente variables como la hora del día, la distancia a la siguiente parada y la velocidad. Esto permite evaluar la relación entre el rendimiento del autobús y la hora estimada de llegada. Si bien el método no permite capturar patrones temporales complejos ni interacciones no lineales, su rápida implementación y sencilla interpretación lo convierten en un punto de partida útil para

determinar si modelos más sofisticados ofrecen mayores beneficios (Mendoza-Mendoza et al., 2022).

2.2.5. Ingeniería de variables para predicción de ETA

La forma en que ajusta y selecciona las variables de entrada es uno de los factores más importantes que afectan el desempeño de los modelos predictores de ETA. Tras revisar la literatura y realizar un análisis exploratorio del conjunto de datos del sistema Ecovía, se seleccionaron cuatro tipos variables para formar el conjunto final de once predictores utilizados en este estudio.

Las variables espaciales son: identificador de la parada actual (`current_stop_id`), orden de la parada en la ruta (`current_stop_order`) y distancia a la siguiente parada (`distance_to_next_stop`), siendo esta última la que mayor poder predictivo del ETA tiene, con correlaciones de entre 0,12 y 0,80 dependiendo de la ruta (ver sección 3.1.4). El identificador de ruta (`route_id`) se utiliza únicamente como criterio de segmentación para entrenar un modelo independiente por ruta y no se incorpora como variable predictora.

El modelo utiliza la velocidad instantánea(`speed`), la aceleración(`acceleration`) y el tiempo transcurrido desde la última parada (`time_since_last_stop`) como variables operativas. En un informe GPS individual, las tres presentan bajas correlaciones lineales con la hora estimada de llegada (ETA). Aun así, se mantuvieron porque los modelos LSTM y GRU procesan la secuencia completa de informes para cada segmento, lo que les permite detectar patrones que una correlación lineal no percibe. Lee y Yoon (2022) informan que la inclusión de variables de flujo operativas mejora significativamente la precisión en comparación con los modelos basados únicamente en coordenadas.

Variables temporales son, por ejemplo, la hora del día (hour), el día de la semana (day_of_week), si es feriado (is_holiday), si es fin de semana (is_weekend) y si es hora punta (is_peak_hour). Según Mendoza-Mendoza et al. (2022) la hora del día es el predictor que tiene mayor impacto en los corredores BRT de Latinoamérica; este resultado también fue confirmado por el análisis exploratorio para el corredor Ecovía.

2.2.6. Métricas de evaluación del modelo

La evaluación del desempeño de los modelos se realizará con las siguientes métricas estándar para problemas de regresión en series temporales:

- **MAE (Mean Absolute Error):** promedio del valor absoluto de los errores de predicción, expresado en segundos. Es interpretable directamente como el error promedio de llegada.
- **RMSE (Root Mean Squared Error):** eleva cada error al cuadrado antes de promediarlos, lo que hace que los errores grandes pesen desproporcionadamente más que los pequeños. Si el modelo falla mucho en alguna parada o franja horaria concreta, el RMSE lo refleja con más fuerza que el MAE.
- **MAPE (Mean Absolute Percentage Error):** expresa el error como porcentaje del valor real. Su ventaja es que permite comparar el desempeño del modelo entre paradas con tiempos de recorrido muy distintos, algo que las métricas absolutas no capturan bien. En este estudio, su interpretación debe realizarse con cautela debido a la sensibilidad de esta métrica ante valores reales pequeños; los segmentos con tiempos inferiores a un minuto se excluyeron del cálculo de MAPE, por lo que esta métrica no representa todos los registros del conjunto de prueba.

2.2.7. Estrategia de validación temporal en series temporales

La validación de modelos predictivos para series temporales requiere un enfoque metodológico distinto al de la clasificación o regresión estándar, principalmente debido a la fuga de información temporal. Esta fuga se produce cuando los datos de entrenamiento contienen observaciones posteriores a las utilizadas para las pruebas. El modelo puede entonces aprender patrones del futuro, patrones que no estarían disponibles en un entorno de producción real. La partición aleatoria utilizada en la validación cruzada convencional no conserva el orden temporal de los registros, ya que puede incorporar observaciones de distintos períodos tanto en el entrenamiento como en la evaluación. Esto puede producir estimaciones de rendimiento artificialmente optimistas y alejarse del comportamiento que tendría el modelo después de su implementación.

Para evitar este problema, se utiliza una partición cronológica o división temporal, en la que el entrenamiento se realiza únicamente con información disponible antes del período de evaluación. En este caso, los datos se organizan en tres bloques consecutivos y sin solapamiento: entrenamiento, validación y prueba. El primer bloque corresponde a las observaciones más antiguas. A continuación, el bloque de validación reúne el período intermedio y se utiliza para seleccionar los hiperparámetros y establecer el criterio de parada temprana de la red neuronal. Finalmente, el bloque de prueba contiene las observaciones más recientes y permite evaluar el comportamiento del modelo sobre un período posterior al utilizado durante el entrenamiento. Esta estrategia permite preservar el orden temporal y reducir el riesgo de fuga de información, proporcionando una evaluación más cercana a un escenario de predicción futura. En este estudio, las métricas de desempeño MAE, RMSE y MAPE se calculan sobre el conjunto de prueba, y es la que se adopta en el presente trabajo con una partición del 70 % para entrenamiento, 15 % para validación y 15 % para prueba, respetando estrictamente el orden cronológico de los registros.

Es importante destacar que en el contexto del sistema Ecovía, la validación temporal también permite evaluar el modelo sobre un período posterior, con condiciones operativas que pueden diferir de las observadas durante el entrenamiento. El conjunto de prueba cubre el período de febrero a marzo de 2026, mientras que el entrenamiento abarca de septiembre de 2025 a enero de 2026. Esta separación temporal reduce la posibilidad de que la evaluación dependa de observaciones temporalmente cercanas y permite valorar la generalización hacia un período posterior al utilizado para el entrenamiento, lo que es el requisito fundamental para un sistema de predicción de ETA que opere de manera continua y confiable.

Un elemento metodológico particularmente relevante en la ingeniería de variables para sistemas BRT es la gestión de la estacionalidad temporal. En el corredor Ecovía de Quito, la demanda y el comportamiento del tráfico presentan patrones claramente diferenciados según la hora del día, el día de la semana y el período del año. Las franjas de hora pico habitualmente de 06:30 a 08:30 y de 17:00 a 19:00 en días laborables concentran una proporción significativamente mayor de la variabilidad del ETA, mientras que las horas valle presentan tiempos de recorrido más estables y predecibles. La variable `is_peak_hour`, construida a partir de los umbrales operacionales del propio sistema, intenta capturar esta distinción de manera binaria, aunque una representación más granular mediante variables cíclicas como el seno y el coseno de la hora normalizada podría capturar mejor la continuidad de los patrones temporales a lo largo del día. Lee y Yoon (2022) demostraron que la incorporación de variables cíclicas para representar la temporalidad mejora la precisión de los modelos de predicción de ETA en un 8 % en promedio, en comparación con la codificación directa de la hora como variable categórica.

La variable de corrección de feriados merece una mención especial en el contexto de este estudio. El dataset original del sistema SAE presentaba un fallo en el campo `is_holiday`, que permanecía en False para todos los registros, incluidos los correspondientes a días festivos

nacionales como Navidad, Año Nuevo y Carnaval, así como a las Fiestas de Quito. Este error fue identificado y corregido mediante el cruce del dataset con el calendario oficial de feriados del Ecuador, lo que resultó en la reclasificación correcta de más de 640.000 registros. El impacto de esta corrección en el modelo es relevante porque los feriados presentan patrones operativos completamente distintos a los días laborables menor demanda, diferentes horarios de servicio, ausencia del patrón de hora pico y su inclusión incorrecta como días normales habría introducido ruido sistemático en el entrenamiento del modelo. Esta corrección constituye un aporte metodológico concreto del presente trabajo al proceso de preparación de datos para predicción de ETA en el corredor Ecovía.

2.3. Síntesis del marco teórico y criterios para la selección de modelos

La revisión de la literatura y los fundamentos teóricos presentados en este capítulo permiten establecer los principales elementos que intervienen en la predicción del tiempo estimado de llegada en sistemas de transporte público. El problema combina características propias de las series temporales con variables espaciales y operativas asociadas al desplazamiento de cada unidad. Por esta razón, la selección de los modelos no se limita a comparar algoritmos con diferente nivel de complejidad, sino que considera también la forma en que cada enfoque representa y utiliza la información disponible.

Los modelos de referencia, como el promedio histórico y la regresión lineal, proporcionan un punto de partida para evaluar la utilidad de técnicas más avanzadas. El promedio histórico permite representar el comportamiento esperado a partir de observaciones anteriores, mientras que la regresión lineal permite establecer una relación directa entre las variables de entrada y el tiempo estimado de llegada. Aunque ambos enfoques presentan limitaciones para representar relaciones complejas, su inclusión es importante porque permite determinar si los modelos de aprendizaje automático y aprendizaje profundo generan una mejora suficiente frente a alternativas sencillas.

Los modelos de gradient boosting representan un segundo enfoque dentro de esta comparación. Tanto XGBoost como HistGradientBoosting utilizan conjuntos de árboles de decisión para construir predicciones a partir de relaciones no lineales entre las variables. Una característica relevante para este estudio es que estos modelos pueden trabajar con una representación tabular de la información temporal reciente. De esta manera, las observaciones correspondientes a una ventana temporal pueden transformarse en variables de entrada que permitan al modelo representar el estado reciente del vehículo sin utilizar directamente una arquitectura recurrente.

Por otra parte, las arquitecturas LSTM y GRU fueron consideradas debido a su capacidad para procesar información secuencial. En estos modelos, las observaciones correspondientes a distintos instantes se mantienen como una secuencia y son procesadas de manera ordenada. Esta característica permite estudiar si la información temporal contenida en los registros anteriores aporta una ventaja adicional respecto a una representación tabular de la misma ventana.

La comparación entre estos enfoques es especialmente relevante debido a las características del problema estudiado. El objetivo no consiste únicamente en determinar cuál modelo presenta el menor error, sino en analizar qué tipo de representación permite aprovechar mejor los datos disponibles en el corredor Ecovía. Una arquitectura más compleja puede ser conveniente cuando existen dependencias temporales prolongadas, pero su ventaja puede reducirse cuando la información más reciente del vehículo concentra una parte importante de la señal necesaria para realizar la predicción.

En este sentido, la distancia hasta la siguiente parada constituye una variable especialmente relevante. El desplazamiento del vehículo puede representarse mediante la combinación de esta distancia con la velocidad, la aceleración y el tiempo transcurrido desde la parada anterior. Estas variables permiten describir el estado operativo inmediato de la unidad y, junto con las variables temporales, proporcionan información sobre las condiciones en las que se produce cada

predicción. El análisis exploratorio realizado en el estudio muestra que la importancia de estas variables puede variar entre rutas, por lo que el comportamiento global del corredor no necesariamente representa de manera adecuada a cada segmento.

El identificador de ruta tampoco se utiliza como predictor directo. Su función consiste en separar los datos y permitir el entrenamiento de un modelo independiente para cada ruta. Esta decisión responde a la existencia de diferencias en las características de los recorridos y permite evitar que el modelo tenga que aprender una relación general que podría ocultar particularidades de cada trayecto. El análisis posterior por ruta permite, además, determinar si el desempeño obtenido para el corredor completo se mantiene cuando se consideran de manera independiente las diferentes condiciones operativas.

Otro elemento fundamental es la representación de la temporalidad. Variables como la hora del día, el día de la semana, la condición de feriado y la identificación de períodos de hora punta permiten incorporar cambios sistemáticos en el comportamiento del transporte. Estas variables son especialmente importantes en un sistema urbano, donde las condiciones de circulación no permanecen constantes durante todo el día. La combinación de información temporal con variables relacionadas directamente con el estado del vehículo permite representar tanto patrones recurrentes como condiciones inmediatas de operación.

La ingeniería de variables adquiere, por tanto, un papel central dentro del proceso de modelado. Los modelos no reciben únicamente la posición GPS original, sino un conjunto de variables construidas a partir de la información disponible. La selección de estas variables busca conservar información relevante para la predicción y, al mismo tiempo, evitar incorporar atributos que no aporten información útil al modelo. Esta consideración resulta particularmente importante cuando se comparan modelos de diferentes familias, ya que una representación adecuada de los datos puede tener un efecto tan importante como la elección del algoritmo.

La revisión también permite establecer la importancia de utilizar una estrategia de evaluación coherente con la naturaleza temporal del problema. En una predicción de ETA, los datos futuros no deberían estar disponibles durante el entrenamiento del modelo. Una partición aleatoria puede mezclar observaciones pertenecientes a diferentes períodos y generar una estimación del desempeño que no representa adecuadamente el comportamiento esperado durante la implementación. La división cronológica permite mantener esta relación temporal y evaluar el modelo sobre observaciones posteriores al período utilizado para su entrenamiento.

La utilización conjunta de un conjunto de entrenamiento, uno de validación y uno de prueba permite separar las diferentes etapas del proceso. El entrenamiento se utiliza para ajustar los parámetros del modelo, la validación permite seleccionar configuraciones e hiperparámetros y el conjunto de prueba se reserva para estimar el desempeño final. Esta separación resulta especialmente importante para las redes neuronales, donde las decisiones relacionadas con la arquitectura y el entrenamiento pueden influir significativamente en el resultado obtenido.

Las métricas seleccionadas también responden a diferentes necesidades de interpretación. El MAE permite expresar directamente el error promedio de predicción en segundos, lo que facilita su interpretación desde el punto de vista operativo. El RMSE asigna mayor peso a los errores grandes y permite identificar situaciones en las que el modelo presenta desviaciones importantes. Por su parte, el MAPE proporciona una medida relativa del error, aunque su interpretación requiere precaución cuando los tiempos reales son pequeños. El uso conjunto de estas métricas permite obtener una evaluación más completa que la utilización de una única medida.

En conjunto, los fundamentos revisados permiten establecer un marco para comparar seis modelos bajo condiciones experimentales comunes: promedio histórico, regresión lineal, HistGradientBoosting, XGBoost, LSTM y GRU. Los dos primeros representan referencias de menor complejidad; los siguientes dos corresponden a modelos de gradient boosting; y los últimos

dos corresponden a arquitecturas recurrentes. Esta clasificación permite estudiar no solo cuál obtiene el menor error, sino también si existe una relación entre la complejidad del modelo, la forma de representar la información temporal y el comportamiento particular del corredor.

Finalmente, la revisión realizada evidencia que la elección del modelo debe estar relacionada con las características específicas de los datos y del sistema de transporte. Los resultados obtenidos en otros corredores proporcionan antecedentes útiles, pero no permiten asumir que una arquitectura determinada tendrá el mismo comportamiento en Quito. Las condiciones del corredor Ecovía, la disponibilidad de registros GPS institucionales, la estructura de sus rutas y la duración de los intervalos entre paradas hacen necesario realizar una evaluación específica. Por esta razón, el siguiente capítulo desarrolla el proceso de preparación de los datos, construcción del conjunto experimental, entrenamiento de los modelos y evaluación bajo una partición temporal.

CAPÍTULO 3: DESARROLLO

3.1 Dataset

3.1.1 Fuente de los datos

Los datos utilizados en este trabajo provienen del sistema SAE desarrollado por la empresa Mivilsoft S.A., entidad especializada en soluciones tecnológicas para el transporte público urbano de la ciudad de Quito. En cuanto al uso de la información, Mivilsoft S.A. autorizó expresamente el uso de los registros históricos de operación del sistema BRT Ecovía con fines académicos.

El acceso y tratamiento de información se llevó a cabo siguiendo criterios de confidencialidad y un uso responsable de datos, siendo que su aplicación se limitó específicamente para el desarrollo de este Trabajo de Fin De Máster. El dataset contiene 28011776 filas y 36 columnas de registros históricos pertenecientes al sistema Ecovía, registros de naturaleza operacional y temporal, asociado a la circulación de la flota Ecovía, tales como rutas, velocidades, horarios de paso y variables derivadas. No se incluyen datos personales de usuarios o pasajeros, conductores, asimismo, se define que esta información no se emplea para fines comerciales.

La recolección de información se realiza mediante un equipo GPS de procedencia rusa, instalado en todas las unidades del corredor Ecovía. Este dispositivo transmite posiciones con una frecuencia de 20 segundos; sin embargo, el equipo ajusta de forma dinámica este intervalo ante eventos de giros pronunciados, generando reportes con menor frecuencia. Esta característica explica la variabilidad en el número de reporte disponibles por segmento entre paradas.

La transmisión de datos se realiza a través de una red móvil 3G o 4G dependiendo de la cobertura disponible en el recorrido y la información se la plataforma de Mivilsoft S.A., donde posteriormente se extraen los registros para la construcción del dataset.

Para construir el dataset se realizó un proceso ETL sobre el sistema SAE de Mivilsoft S.A., extrayendo los registros generados entre el 15 de septiembre de 2025 y el 15 de marzo de 2026, siendo que la selección de este periodo se realizó de forma intencional, para disponer datos que

identifiquen condiciones operativas de acuerdo al fenómeno estudiado. Entre estas condiciones se encuentran patrones normales de operación, feriados de noviembre, festividades de Quito en diciembre, navidad, fin de año e inicio de año nuevo y el período de carnaval, con el fin de favorecer una variabilidad de entrenamiento a los modelos seleccionados. El dataset completo en formato CSV está disponible en el Apéndice B.

Los registros históricos corresponden a las 7 rutas operativas del sistema Ecovía, cuyos tiempos programados de referencia se presentan en el Apéndice A.

3.1.2 Descripción de Dataset

El dataset está compuesto por 36 variables organizadas en cuatro categorías: operacional, espacial, temporal y variable objetivo. Los registros corresponden a la recopilación de todos los viajes realizados por las unidades de transporte del sistema Ecovía entre paradas consecutivas, abarcando las diferentes rutas que opera el sistema. La tabla 1 describe cada variable con su tipo de dato y descripción correspondiente.

Cada fila del dataset representa un reporte GPS individual, enviado por cada dispositivo instalado y no a un resumen agregado del segmento, por ende, entre dos paradas consecutivas pueden existir varios reportes, dependiendo la distancia y tiempo existente.

Las variables temporales (hour, day_of_week, is_weekend, is_peak_hour e is_holiday) no forman parte de los registros originales obtenidos de Mivilsoft S.A., sino que fueron calculados en el proceso ETL a partir de la variable timestamp, con la finalidad de incorporar los patrones de estacionalidad horaria en los modelos.

Tabla 1*Descripción de variables del dataset*

Variable	Tipo	Descripción	Categoría
bus_id	Categorico	Identificador único del bus	Operacional
timestamp	Fecha/Hora	Fecha y hora exacta del registro GPS	Temporal
latitude	Flotante	Latitud geográfica de la unidad (WGS84)	Espacial
longitude	Flotante	Longitud geográfica de la unidad (WGS84)	Espacial
altitude	Flotante	Altitud de la unidad en metros sobre el nivel del mar	Espacial
speed	Numérico	Velocidad instantánea de la unidad en km/h	Operacional
course	Numérico	Dirección de desplazamiento en grados	Operacional

accuracy	Numérico	Precisión del registro GPS en metros	Operacional
valid	Binario	Indica si el registro GPS es válido (1) o no (0)	Operacional
trip_id	Categórico	Identificador único del viaje	Operacional
start_time	Fecha/Hora	Hora de inicio del viaje	Temporal
end_time	Fecha/Hora	Hora de fin del viaje	Temporal
date	Fecha	Fecha del registro del servicio	Temporal
route_id	Categórico	Identificador de la ruta operada	Operacional
current_stop_id	Entero	Identificador de la parada actual	Espacial
current_stop_order	Entero	Orden de la parada actual en la ruta	Espacial
bus_stop_lat	Numérico	Latitud de la parada actual	Espacial
bus_stop_lng	Numérico	Longitud de la parada actual	Espacial

step_time	Numérico	Tiempo del segmento entre paradas	Operacional
scheduled_passing_time	Numérico	Tiempo programado de paso por la parada actual	Operacional
real_passing_time	Numérico	Tiempo real de paso por la parada actual	Operacional
last_stop_time	Numérico	Tiempo de paso por la última parada visitada	Operacional
next_stop_order	Entero	Orden de la siguiente parada en la ruta	Espacial
next_stop_id	Entero	Identificador de la siguiente parada	Espacial
next_stop_time	Numérico	Tiempo estimado de llegada a la siguiente parada	Operacional
next_stop_lat	Flotante	Latitud de la siguiente parada	Espacial
next_stop_lng	Flotante	Longitud de la siguiente parada	Espacial

time_since_last_stop	Numérico	Tiempo transcurrido desde la última parada en segundos	Operacional
distance_to_next_stop	Numérico	Distancia a la siguiente parada en metros	Espacial
acceleration	Numérico	Aceleración de la unidad en m/s ² (136 nulos)	Operacional
hour	Numérico	Hora del día del registro	Temporal
day_of_week	Entero	Día de la semana del registro	Temporal
is_weekend	Binario	Indica si es fin de semana (1) o no (0)	Temporal
is_peak_hour	Binario	Indica si es hora pico (1) o no (0)	Temporal
is_holiday	Binario	Indica si es feriado (1) o no (0)	Temporal
eta_next_stop	Numérico	Tiempo estimado de llegada a la siguiente parada en minutos	Variable objetivo

Nota. Dataset proporcionado por (Mivilsoft S.A., 2024).

3.1.3 Variable objetivo

La variable objetivo para cada modelo es **eta_next_stop**, la cual representa el tiempo en minutos que tarda una unidad de transporte en desplazarse desde una parada hasta la siguiente. Esta variable ya viene calculada dentro del dataset, por lo que no requiere ser derivada nuevamente. Los valores están en minutos y muestran cuánto tardó realmente el bus en ir de una parada a la siguiente.

Mediante la perspectiva operativa de Ecovía, el valor de **eta_next_stop** refleja de forma directa la fluidez existente entre cada tramo de las rutas del corredor. Valores bajos y consistentes muestran una operación sin interrupciones significativas, mientras que valores altos suelen estar asociados con fenómenos como congestión, o tiempos prolongados de embarque y desembarque de pasajeros.

Igualmente, la decisión de predecir **eta_next_stop** de manera segmentada, en lugar de estimar el tiempo total del viaje, responde a que este enfoque permite actualizar la predicción de manera continua conforme avanza la unidad su recorrido.

Al revisar el dataset por primera vez se identificó valores inconsistentes en la operación real del sistema, por lo que es necesario realizar un proceso de limpieza antes de entrenar cada modelo, el cual se describe en la siguiente sección.

3.1.4 Análisis exploratorio de datos

Antes de aplicar cualquier transformación, se realizó un análisis exploratorio (EDA) sobre el dataset completo de 28.011.776 registros con el fin de entender el comportamiento de la variable objetivo y decidir qué variables incluir en los modelos. El análisis se organizó en tres bloques: la variable objetivo, las variables temporales y operacionales, y las variables espaciales.

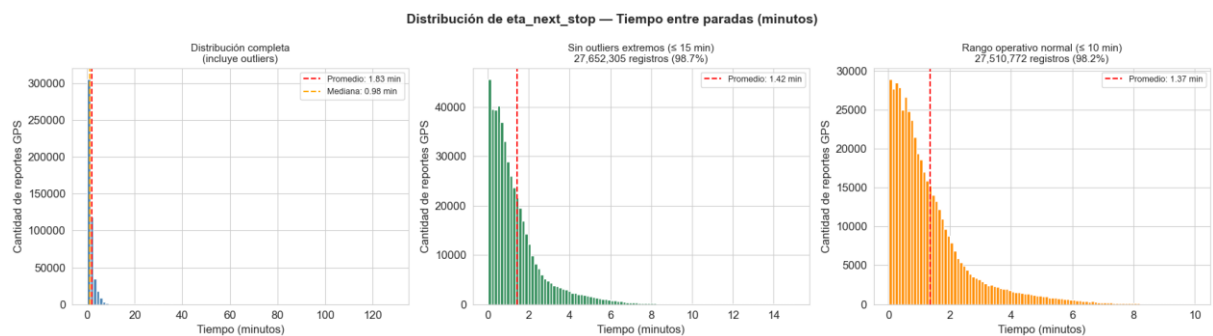
La variable **eta_next_stop** no presentó valores nulos y el tiempo promedio entre paradas fue de 1,83 minutos, con una mediana de 0,98 minutos y un máximo de 138,03 minutos. El 98,21 % de los registros se ubicó en un rango operativo normal (≤ 10 minutos); los valores por encima

de ese umbral (501.004 registros, 1,79 %) y en particular los superiores a 60 minutos (42.341 registros) se identificaron como valores atípicos no representativos de la operación real del corredor, atribuibles a desvíos de ruta, unidades fuera de servicio o fallas de posicionamiento GPS.

Estos valores atípicos corresponden a eventos puntuales que se apartan de la circulación regular del corredor. De no realizarse una limpieza antes del entrenamiento, podrían introducir un sesgo hacia eventos poco representativos de una operación cotidiana, incrementando la varianza del error de los modelos y reduciendo el desempeño habitual del sistema.

Figura 1

Distribución de eta_next_stop – Tiempo entre paradas



Nota. Distribución antes de la limpieza de datos

Al desagregar la eta promedio por hora del día para cada ruta (Figura 2) se observó, antes de la limpieza, un pico anómalo en las primeras horas de operación de la madrugada en varias rutas, coherente con los valores atípicos detectados en la variable objetivo. Descontado este efecto, el eta se mantiene relativamente estable durante el resto del día, con variaciones moderadas entre hora pico y hora valle en 5 de las 7 rutas, y diferencias menores pero consistentes entre días laborables y fines de semana en 6 de las 7 rutas. Por esta razón, hour, is_peak_hour, is_weekend y day_of_week se conservaron como variables predictoras, además de is_holiday, que se corrige en el preprocesamiento (sección 3.2.1).

Figura 2

ETA promedio por hora del día, por ruta, antes de la limpieza de datos



El análisis de correlación confirmó que `distance_to_next_stop` es la variable más predictiva del eta, con correlaciones de entre 0,12 y 0,80 según la ruta. Esta relación resulta consistente: a mayor distancia entre la posición reportada y la siguiente parada, mayor será el tiempo requerido para alcanzarla. La velocidad instantánea presentó una correlación negativa en todas las rutas (a mayor velocidad, menor tiempo entre paradas), aunque de magnitud baja al analizarse sólo en el primer reporte de cada segmento. Las variables de aceleración y tiempo transcurrido desde la última parada presentaron correlaciones bajas sobre el mismo criterio, pero se decidió conservar en los modelos porque LSTM y GRU procesan la secuencia de reportes GPS y de esta forma se

puede aprovechar patrones que la correlación lineal calculada sobre un único reporte, no logra capturar.

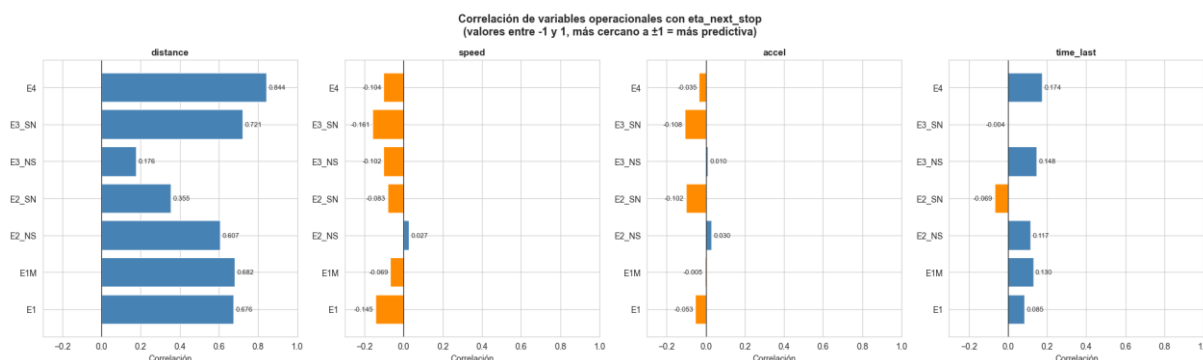
Cabe precisar que, la correlación lineal se utilizó como criterio orientador y no como el único factor de selección, esto dado a que los modelos evaluados, tanto las redes recurrentes como los árboles de decisión capturan relaciones no lineales e interacciones entre variables. Una correlación lineal baja no implica la ausencia de valor predictivo. Por ello, se combinaron tres criterios para las variables seleccionadas (evidencia de correlación, relevancia operativa de cada viaje según la naturaleza del problema, capacidad de generalización entre viajes y rutas).

Un enfoque de selección, basado únicamente en la correlación lineal, hubiese descartado características que, pese a mostrar una relación débil de forma aislada, aportan información importante al procesar de manera conjunta dentro de una ventana temporal. Por esta razón, se priorizó un criterio combinado por sobre un filtro netamente estadístico.

Bajo los criterios mencionados, las variables con correlación baja como speed, acceleration o time_since_last_stop se conservaron porque tienen aporte en la ventana temporal que procesan los modelos, los árboles al combinarlas en la ventana aplanada y las redes recurrentes al recorrer la secuencia, mientras que variables como trip_id o timestamp se descartan por no generalizar y step_time por la fuga de información hacia la variable objetivo.

Figura 3

Correlación de las variables operacionales con eta_next_stop, por ruta (dataset ya preprocesado)



La característica altitud presentó correlaciones de signo contradictorio entre rutas (desde, +0,29 en E3_NS hasta a -0,29 en E4), por lo que se decidió descartar de los modelos al no mostrar consistencia. El tiempo programado (scheduled_passing_time) presentó correlaciones casi nulas con la eta real (entre -0,03 y 0,06), lo que confirma empíricamente que el horario programado del sistema Ecovía no predice el comportamiento real de la operación y justifica, precisamente, la necesidad de una comparación de modelos predictivos basado en datos históricos reales.

A partir de estos hallazgos, del total de 36 variables del dataset original se seleccionaron 11 variables predictoras (Tabla 2) y se excluyeron 23, ya sea por generar fuga de información hacia la variable objetivo (como step_time o next_stop_time), por no generalizar entre viajes (como trip_id o timestamp), o por no mostrar un patrón predictivo consistente (como altitude o scheduled_passing_time). El identificador de ruta (route_id) no se toma en cuenta como variable predictora, ya que esta se emplea únicamente para segmentar los datos y entrenar un modelo por ruta.

Tabla 2*Variables predictoras seleccionadas para los modelos*

Variable	Categoría	Correlación / evidencia
current_stop_id	Espacial	Identifica la parada actual
current_stop_order	Espacial	Correlación baja (< 0,13)
distance_to_next_stop	Espacial	Correlación alta (0,12 a 0,80)
speed	Operacional	Correlación negativa, baja (< 0,16)
acceleration	Operacional	Correlación baja (< 0,13)
time_since_last_stop	Operacional	Correlación baja (< 0,17)
hour	Temporal	Impacto alto
day_of_week	Temporal	Impacto mínimo pero consistente
is_weekend	Temporal	Impacto pequeño (6 de 7 rutas)
is_peak_hour	Temporal	Impacto medio (5 de 7 rutas)
is_holiday	Temporal	Requiere corrección (ver 3.2.1)

Nota. El identificador de ruta (route_id) no se incluye como variable predictora ya que se utiliza para la segmentación de los datos y entrenar un modelo independiente por ruta. El detalle completo de las 23 variables excluidas y su justificación se presentan en el Apéndice C.

3.2 Desarrollo del Trabajo

3.2.1 Preprocesamiento

Con base en los hallazgos del análisis exploratorio (sección 3.1.4), se aplicó un pipeline de preprocesamiento secuencial sobre el dataset completo (28.011.776 registros), compuesto por los siguientes pasos:

1. Eliminación de registros GPS inválidos (`valid = False`).
2. Eliminación de reportes con velocidad igual a cero (unidad detenida), dado que no aportan información sobre el desplazamiento entre paradas.
3. Conversión de la velocidad de nudos a kilómetros por hora ($\times 1,852$).
4. Eliminación de valores atípicos de velocidad superiores a 60 km/h, umbral definido a partir del percentil 99 de todas las rutas (entre 51 y 54 km/h).
5. Eliminación de outliers de `eta_next_stop` aplicando un umbral específico por ruta, basado en el percentil 99 de cada una (Tabla 3), en lugar de un umbral único para todo el sistema.
6. Corrección de la variable `is_holiday`, que en el dataset original se encontraba en `False` para el 100 % de los registros. Se reconstruyó a partir del calendario oficial de feriados nacionales y locales de Ecuador para el período septiembre 2025 - marzo 2026, quedando 925.403 registros (3,8 %) correctamente marcados como feriado.
7. Imputación de 136 valores nulos (0,0005 % del dataset) en la variable `acceleration` con el valor 0.
8. Conversión de las variables booleanas a formato entero (0/1).
9. Selección final de las 11 variables predictoras identificadas en el EDA.

Tabla 3*Umbral de eliminación de outliers de eta_next_stop por ruta*

Ruta	Umbral eta (min)	Percentil 99 observado (min)
E1	7,0	6,5
E1M	6,5	6,0
E2_NS	7,5	7,1
E2_SN	5,0	4,4
E3_NS	6,0	5,7
E3_SN	6,5	6,0
E4	7,0	6,8

Nota. Los valores mostrados corresponden al percentil 99 de eta_next_stop obtenido para cada ruta. Se estableció un límite para eliminar registros extremos sin afectar los tiempos que pueden presentarse de forma natural en la operación.

Se obtuvo una reducción del dataset de 28.011.776 a 24.308.817 registros (86,78% de los datos originales). La desviación estándar de eta_next_stop disminuyó de 4,48 a 1,23 minutos, comprobando que los valores atípicos eliminados correspondían a comportamientos anormales y no a variabilidad real de la operación. La velocidad máxima quedó en 59,80 km/h, un valor consistente con la operación urbana del corredor Ecovía. El dataset limpio se guardó en un archivo independiente (dataset_limpio.csv) para preservar intacto el dataset original.

3.2.2 Diseño de los modelos

Con el dataset limpio se entrenaron y compararon seis modelos de aprendizaje automático, agrupados en tres familias: métodos de referencia (baseline), modelos de gradient boosting y redes neuronales recurrentes. Los seis modelos se entrenaron de forma independiente para cada una de las 7 rutas del corredor Ecovía, ya que cada ruta presenta una duración de recorrido, un número de paradas y patrones operativos distintos.

Modelos baseline: El promedio histórico utiliza el promedio de `eta_next_stop` del conjunto de entrenamiento como predicción única para todos los registros del conjunto de prueba, y sirve como punto de referencia mínimo que cualquier modelo más complejo debe superar para justificar su uso. La regresión lineal es un modelo simplificado que permite constatar si la relación entre las variables predictoras y la variable objetivo es aproximadamente lineal antes de acudir a modelos más complejos.

Modelos de gradient boosting: Se utilizaron dos modelos, el primero `HistGradientBoostingRegressor`, disponible en `scikit-learn` y se configuró con 200 iteraciones, junto con una tasa de aprendizaje de 0,1 y una profundidad máxima de 6. El segundo, `XGBoost` (`XGBRegressor`), se configuró con 200 estimadores, una tasa de aprendizaje de 0,1, junto con una profundidad máxima de 6 y el método 'hist' para acelerar el entrenamiento sobre grandes volúmenes de datos. En ambos casos los modelos reciben como entrada la ventana temporal aplanada, es decir, los reportes de los últimos n pasos de tiempo concatenados en un solo vector, lo que les permite aprovechar el historial reciente sin necesitar una arquitectura secuencial.

Los valores de los hiperparámetros corresponden a las configuraciones de referencia habituales para algoritmos de gradient boosting (Chen & Guestrin, 2016).

Modelos de redes neuronales recurrentes: Se emplearon LSTM y GRU, cada una compuesta con una arquitectura de dos capas recurrentes con 64 y 32 neuronas respectivamente. Para disminuir un riesgo de sobreajuste se incorporó una capa Dropout con tasa 0,3 además de una

regularización L2 de 0,001. Para el entrenamiento se utilizó el optimizador Adam con una tasa de aprendizaje 0,001 y lotes de 1024 muestras, definidos con un máximo de 50 épocas. También se aplicó EarlyStopping con una paciencia de 5 épocas tomando como referencia la pérdida de validación junto con ReduceLROnPlateau configurado con un factor de 0,5 y una paciencia de 3 épocas para reducir la tasa de aprendizaje cuando el modelo deja de mejorar. A diferencia de los modelos de gradient boosting, estos modelos procesan la secuencia completa de reportes GPS del segmento respetando su orden temporal, en lugar de recibir la ventana aplanada.

Para la selección de las configuraciones se apoyó en pruebas exploratorias realizadas sobre la ruta E1, se toma como referencia esta ruta por ser la que mayor volumen de datos tiene en el conjunto de datos, y en las que se compararon distintas arquitecturas y niveles de regularización según su efecto sobre el error y el tiempo de entrenamiento.

Tabla 4

Pruebas exploratorias de los hiperparámetros para modelos recurrentes (LSTM, ruta E1)

Configuración	MAE(s)	Tiempo de entrenamiento(min)
2 capas (64/32), dropout 0,3	21,1	6,2
3 capas (128/64/32), dropout 0,3	22,0	19,8
2 capas (64/32), dropout 0,4	22.8	5,5

Nota. La configuración seleccionada es dos capas (64/32) con un dropout de 0,3, estos por obtener el menor MAE con el menor tiempo de entrenamiento.

La configuración de dos capas con un dropout de 0,3 obtuvo el mejor desempeño en cuanto al entrenamiento, el aumento de la regularización a 0,4 deterioró el error y la arquitectura

de tres capas no mejoró los resultados más bien triplicó el tiempo de entrenamiento. Con base en esta exploración, la configuración fue tomada la de mejor desempeño y se adoptó para LSTM y GRU en todas las rutas.

Para el optimizador Adam y la tasa de aprendizaje de 0,001 se adoptaron por ser los valores de referencia, recomendado en la literatura para redes neuronales (Kingma & Ba, 2015), al igual que es uso del dropout como técnica de regularización para reducir el sobreajuste (Srivastava et al., 2014).

Ventana temporal: El tamaño de la ventana (número de reportes históricos considerados por segmento) se definió de forma independiente para cada ruta según la disponibilidad real de reportes GPS por segmento, con ventanas de 5, 7 o 10 pasos según la ruta (el detalle se presenta en la Tabla 4, sección 3.2.3).

Normalización: Se evaluaron dos esquemas de normalización de las variables de entrada: MinMaxScaler (rango [0, 1]) y StandardScaler (media 0, desviación estándar 1), ajustados únicamente sobre el conjunto de entrenamiento para evitar fuga de información hacia validación y prueba. La comparación entre ambos esquemas permitió evaluar si el tipo de normalización afecta de forma diferenciada a los modelos basados en árboles frente a los modelos recurrentes, cuya función de activación tanh está centrada en cero. Los resultados de esta comparación se presentan en la sección 4.2.3.

3.2.3 Entrenamiento y Validación

El dataset limpio se dividió cronológicamente en tres conjuntos para cada ruta: 70 % para entrenamiento, 15 % para validación y 15 % para prueba, respetando el orden temporal de los registros (del 15 de septiembre de 2025 al 15 de marzo de 2026) en lugar de una partición aleatoria. Esta decisión metodológica es fundamental en problemas de series temporales: una partición aleatoria permitiría que el modelo “vea” implícitamente información del futuro durante

el entrenamiento, produciendo métricas optimistas que no se sostendrían en un escenario real de operación.

El dataset de entrenamiento se utilizó para ajustar cada uno de los modelos seleccionados, mientras que el de validación se separó del proceso de ajuste, con la finalidad de observar el comportamiento de los modelos sobre las observaciones temporales posteriores a las empleadas en el entrenamiento. De esta forma se facilita la detección de posibles problemas de sobreajuste.

Para garantizar la homogeneidad de evaluación sobre los 6 modelos, se utilizó las mismas particiones temporales para cada ruta, con las mismas variables de entrada y variable objetivo. En consecuencia, las diferencias obtenidas en las métricas de desempeño son atribuidas directamente a la habilidad de cada modelo y no a la diferencia en los datos utilizados para su evaluación.

El uso de los datasets de validación y prueba independientes, permiten obtener una estimación más sólida de cada modelo ante nuevos registros de operación.

Tabla 5

Ventana temporal óptima seleccionada por ruta para LSTM y GRU

Ruta	Ventana temporal	Cobertura de segmentos
E1	7	51,3 % con ≥ 7 reportes
E1M	5	64,2 % con ≥ 5 reportes
E2_NS	5	≥ 60 % con ≥ 5 reportes
E2_SN	10	cobertura insuficiente con 5 o 7
E3_NS	10	cobertura insuficiente con 5 o 7
E3_SN	7	≥ 40 % con ≥ 7 reportes
E4	7	≥ 40 % con ≥ 7 reportes

Nota. Para la selección del tamaño de la ventana se analizó la cantidad de reportes GPS en los segmentos de cada ruta. Se utilizaron 7 pasos cuando al menos el 40 % de los segmentos tenía 7 o más reportes, y 5 pasos cuando este porcentaje alcanzaba el 60 %. En los demás casos se utilizaron 10 pasos.

Antes de entrenar cada modelo se verificó la ausencia de valores nulos, eta negativos y velocidades fuera de rango (sección 4.1); las siete rutas completaron esta verificación. El conjunto de datos de prueba reservado, se utilizó específicamente para métricas finales de desempeño. El desempeño de los seis modelos se evaluó sobre el conjunto de prueba, no utilizado, durante el entrenamiento, utilizando tres métricas: MAE (error absoluto medio, en segundos), RMSE (raíz del error cuadrático medio, en segundos) y MAPE (error porcentual absoluto medio). La estandarización del procedimiento de partición y evaluación de todos los modelos fueron establecidas en condiciones justas para determinar el modelo óptimo en cada ruta. Los resultados completos de esta evaluación, desagregados por ruta y modelo, se presentan en el Capítulo 4.

3.2.4 Prototipo

Como resultado final del trabajo se desarrolló e implementó un prototipo web orientado al área de operaciones del sistema Ecovía, publicado como una aplicación pública en Hugging Face Spaces (Torres Chango et al., 2026), disponible en <https://huggingface.co/spaces/ChrissHF/Dev-ML>. El prototipo muestra las predicciones de tiempo de llegada entre paradas alcanzadas por los seis modelos entrenados (promedio histórico, regresión lineal, HistGradientBoosting, XGBoost, LSTM y GRU) de las siete rutas del corredor. Esta herramienta está diseñada como una prueba de concepto, destinada a detallar la predicción de los modelos y no corresponde a una aplicación móvil destinada a un sistema de información al usuario.

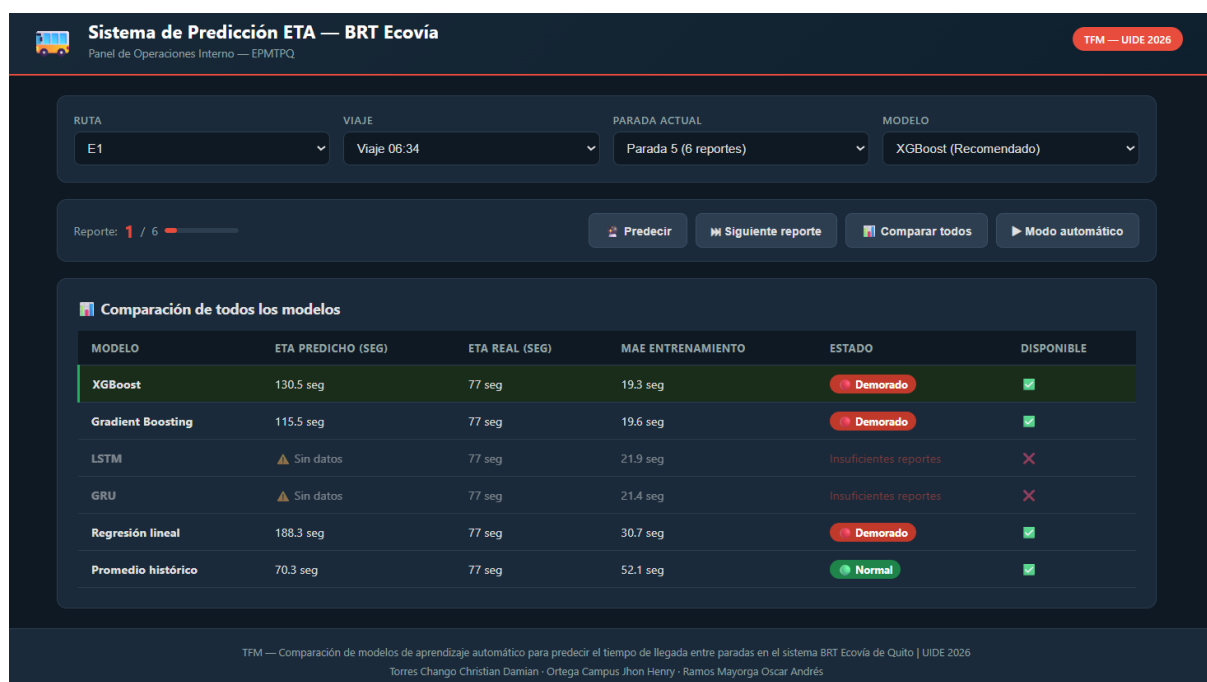
El prototipo se diseñó con FastAPI y se ejecuta dentro de un contenedor Docker, con los seis modelos entrenados por ruta, cargados como artefactos junto con una muestra del dataset histórico de operación. En la interfaz, el usuario puede seleccionar una ruta, un viaje y la parada

actual, posteriormente, elige uno de los seis modelos disponibles y como resultado, se obtiene el tiempo estimado hasta la siguiente parada, junto con el tiempo real medido, la velocidad y la distancia correspondientes a ese segmento.

El objetivo del prototipo consiste en demostrar una prueba de concepto al integrar seis modelos mediante un entorno web, evaluando su desempeño en cada ruta y parada seleccionada. No pretende sustituir a la plataforma SAE de Mivilsoft S.A., sino operar como una herramienta de información. Asimismo, no se incluyó un criterio de aceptación por los operadores de la Ecovía, sino que los resultados se limitan a demostrar la funcionalidad e integración de los modelos a través del prototipo.

Figura 4

Panel del prototipo - comparación de los seis modelos para la ruta E1, viaje 06:34, parada 5

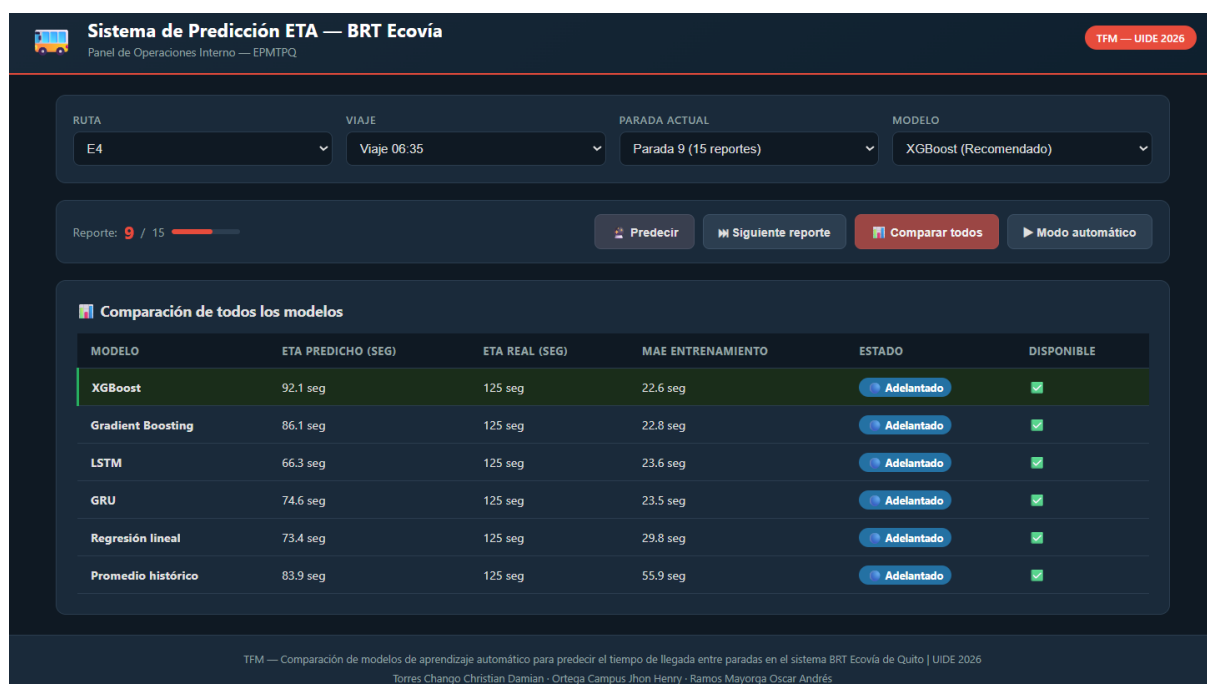


La Figura 4 muestra un ejemplo real de uso del prototipo sobre la ruta E1, viaje de las 06:34, en la parada 5 (segmento con 6 reportes GPS disponibles). Para ese reporte, el panel

compara simultáneamente la predicción de los seis modelos frente al tiempo real medido de 77 segundos: XGBoost, el modelo recomendado, predijo 130,5 segundos; HistGradientBoosting, 115,5 segundos; Regresión lineal, 188,3 segundos; y Promedio histórico, 70,3 segundos. Los modelos LSTM y GRU no generaron predicción para este reporte específico (estado “insuficientes reportes”), ya que la parada 5 de este viaje solo contaba con 6 reportes GPS, por debajo de la ventana temporal de 7 pasos requerida por ambas arquitecturas recurrentes en la ruta E1 (Tabla 5). Este caso ilustra en la práctica una limitación relevante de los modelos secuenciales frente a los modelos de árboles: estos últimos pueden generar una predicción incluso cuando el historial disponible es corto, mientras que LSTM y GRU requieren completar su ventana temporal mínima para operar.

Figura 5

Panel del prototipo - comparación de los seis modelos para la ruta E4, viaje 06:35, parada 9



La Figura 5 presenta un segundo ejemplo, en este caso sobre la ruta E4 (la de mayor longitud y mayor error agregado del corredor, ver sección 4.2.2), viaje de las 06:35, en la parada 9 de un segmento con 15 reportes GPS, cobertura suficiente para que los seis modelos generen

predicción. Para este reporte, el tiempo real medido fue de 125 segundos y los seis modelos lo subestimaron (estado “adelantado” en los seis casos): XGBoost predijo 92,1 segundos (diferencia de 32,9 segundos frente al valor real), HistGradientBoosting 86,1 segundos (38,9 segundos de diferencia), Promedio histórico 83,9 segundos (41,1 segundos), Regresión lineal 73,4 segundos (51,6 segundos), GRU 74,6 segundos (50,4 segundos) y LSTM 66,3 segundos (58,7 segundos, la mayor diferencia de los seis modelos en este reporte puntual).

Este segundo ejemplo permite observar una diferencia entre los resultados generales y los casos individuales que presenta la operación. En el conjunto de prueba, XGBoost y HistGradientBoosting ofrecen menores errores que la regresión lineal y el promedio histórico de acuerdo con los resultados de la Tabla 5. Sin embargo, esto no significa que estos modelos tengan el menor error en cada predicción. En este caso particular, el promedio histórico presentó un valor más cercano al tiempo real que LSTM y GRU. Por esta razón, la evaluación de los modelos se realiza utilizando las métricas calculadas sobre todo el conjunto de pruebas, en lugar de considerar únicamente casos individuales.

CAPÍTULO 4: ANÁLISIS DE RESULTADOS

4.1 Pruebas de Concepto

Antes de ejecutar el entrenamiento completo de los seis modelos sobre las siete rutas, se implementó una etapa de verificación automática por ruta dentro del propio pipeline de entrenamiento. Para cada ruta, antes de iniciar el entrenamiento se comprobó que: (a) no existieran valores nulos en las variables predictoras ni en la variable objetivo, (b) el valor mínimo de `eta_next_stop` fuera mayor o igual a cero, y (c) la velocidad máxima no superará el umbral de 60 km/h definido en el preprocesamiento. Únicamente si la ruta superaba estas tres verificaciones se continuaba con la selección de la ventana temporal y el entrenamiento de los modelos; en caso contrario, la ruta se habría excluido de esa ejecución para evitar propagar errores de datos al entrenamiento.

Adicionalmente, la ventana temporal óptima para los modelos LSTM y GRU no se fijó de forma arbitraria, sino que se determinó para cada ruta a partir de la cobertura real de reportes por segmento: se calculó el porcentaje de segmentos que contaban con al menos 5, 7 o 10 reportes GPS, y se seleccionó la ventana de mayor tamaño que conservará una cobertura aceptable (al menos 40 % de los segmentos con ventana 7, o al menos 60 % con ventana 5; en caso contrario, ventana 10). Este criterio buscó equilibrar la cantidad de historial temporal disponible para el modelo con la cantidad de secuencias de entrenamiento aprovechables.

Las siete rutas superaron todas las verificaciones y se procesaron completamente en ambos esquemas de normalización (MinMaxScaler y StandardScaler), lo que permitió proceder con la fase de entrenamiento completo descrita en la sección 3.2.3 y con el análisis de resultados que se presenta a continuación.

4.2 Análisis de Resultados

4.2.1 Comparación general de modelos

Tras entrenar los seis modelos sobre las siete rutas del corredor Ecovía con los dos esquemas de normalización, se obtuvieron los resultados de las Tablas 6, 7 y 8 (MAE, RMSE y MAPE respectivamente), calculados con StandardScaler, el esquema recomendado para este problema según se sustenta en la sección 4.2.4. Los resultados completos con MinMaxScaler se presentan de forma comparativa en esa misma sección.

Tabla 6

MAE por ruta y modelo - StandardScaler (segundos)

Modelo	E1	E1M	E2_NS	E2_SN	E3_NS	E3_SN	E4	Prom.
Promedio histórico	52,1	59,2	59,6	22,4	42,6	46,1	55,9	48,3
Regresión lineal	30,7	35,4	32,1	15,1	27,3	27,0	29,8	28,2
Gradient Boosting	19,6	20,9	22,1	10,4	19,7	18,2	22,8	19,1
XGBoost	19,3	20,7	21,9	10,4	19,8	18,0	22,6	19,0
LSTM	21,2	22,6	23,5	11,9	22,3	20,1	24,2	20,8
GRU	21,9	22,4	23,8	11,0	21,3	19,3	23,9	20,5

Nota. Ventanas temporales por ruta: E1=7, E1M=5, E2_NS=5, E2_SN=10, E3_NS=10, E3_SN=7, E4=7.

Tabla 7*RMSE por ruta y modelo - StandardScaler (segundos)*

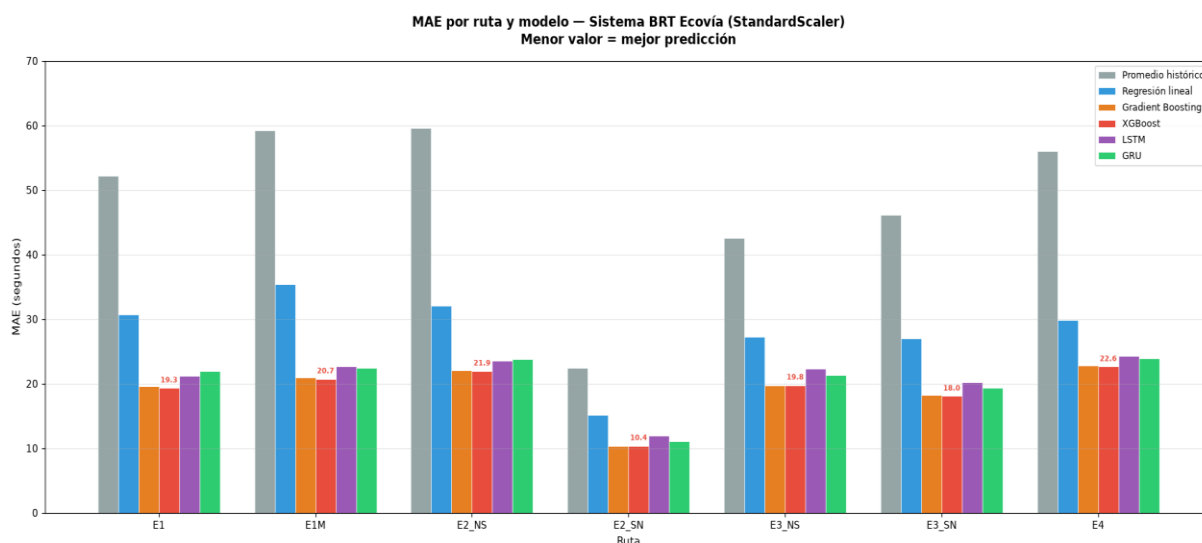
Modelo	E1	E1M	E2_NS	E2_SN	E3_NS	E3_SN	E4	Prom.
Promedio histórico	65,5	72,3	77,6	28,3	52,4	58,1	68,4	60,4
Regresión lineal	40,4	50,1	70,7	20,5	36,1	36,0	42,8	42,4
Gradient Boosting	28,6	30,4	35,0	15,9	28,1	25,9	32,4	28,0
XGBoost	28,3	30,2	34,9	15,9	28,3	25,7	32,2	27,9
LSTM	31,2	32,2	36,8	18,4	32,1	28,7	34,2	30,5
GRU	32,5	33,0	36,0	17,4	30,8	27,4	33,9	30,1

Tabla 8*MAPE por ruta y modelo - StandardScaler (%)*

Modelo	E1	E1M	E2_NS	E2_SN	E3_NS	E3_SN	E4	Prom.
Promedio histórico	36,92	39,29	42,03	66,85	49,96	37,37	32,99	43,63
Regresión lineal	31,38	34,78	32,32	36,53	31,89	29,96	30,27	32,45
Gradient Boosting	23,78	25,23	27,44	28,43	28,34	23,79	24,54	25,94
XGBoost	23,49	25,05	27,26	27,57	27,84	23,64	24,45	25,61
LSTM	24,27	25,63	27,86	35,73	31,59	24,97	24,83	27,84
GRU	24,91	26,17	27,46	33,55	31,18	24,17	25,16	27,51

Figura 6

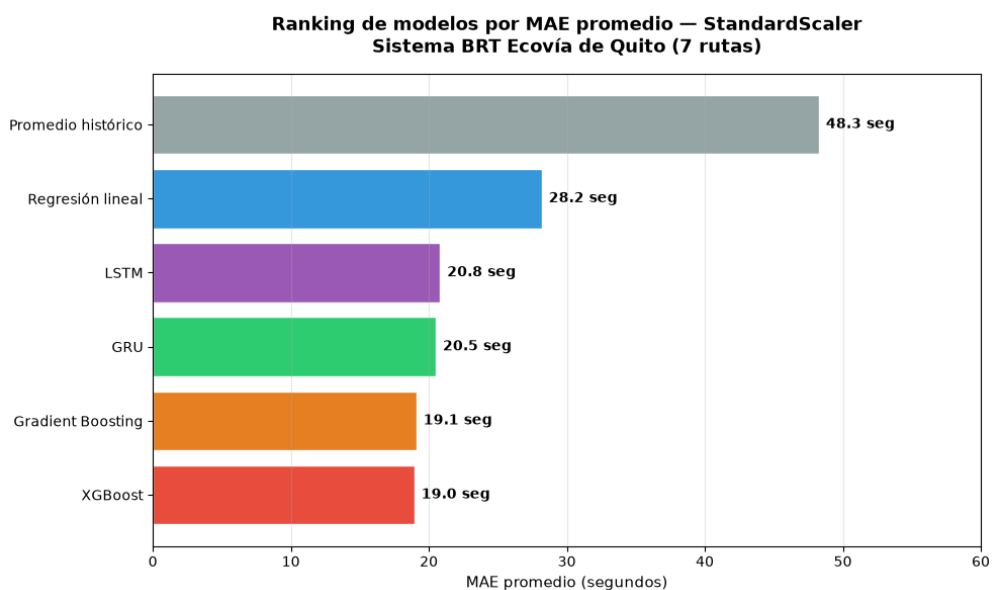
MAE por ruta y modelo, esquema StandardScaler



El modelo XGBoost obtuvo el mejor desempeño promedio, con un MAE de 19,0 segundos, un RMSE de 27,9 segundos y un MAPE de 25,61 %, seguido muy de cerca por HistGradientBoosting (MAE = 19,1 segundos). Los modelos recurrentes GRU y LSTM se ubicaron en un segundo grupo, con un MAE promedio de 20,5 y 20,8 segundos respectivamente. La regresión lineal (MAE = 28,2 segundos) y el promedio histórico (MAE = 48,3 segundos) confirmaron su rol de referencia, siendo ambos superados de forma consistente por los cuatro modelos de aprendizaje automático propuestos, en las siete rutas y en las tres métricas.

Figura 7

Ranking de modelos por MAE promedio, esquema StandardScaler

**Tabla 9**

Ranking final por MAE promedio - StandardScaler (7 rutas)

Posición	Modelo	MAE (s)	RMSE (s)	MAPE (%)
1	XGBoost	19,0	27,9	25,61
2	Gradient Boosting	19,1	28,0	25,94
3	GRU	20,5	30,1	27,51
4	LSTM	20,8	30,5	27,84
5	Regresión lineal	28,2	42,4	32,45
6	Promedio histórico	48,3	60,4	43,63

Es favorable interpretar que los valores en función de la duración promedio entre los segmentos de paradas (aproximadamente 78 segundos, sección 4.2.5). Un MAE de 19 segundos para XGBoost, representa una cuarta parte de la duración típica de un segmento, lo que es coherente para los valores MAPE observados (25,61 % para XGBoost). Esta condición explica cuando un error absoluto parece reducido en términos de segundos, el error relativo resulta comparativamente alto: en tramos tan cortos, cualquier desviación absoluta da sentido a una proporción considerable del tiempo total entre paradas.

4.2.2 Análisis de significancia estadística

Los resultados de la métrica MAE presentados en la Tabla 8, por si solos no permiten afirmar que un modelo supera a otro, ya que existen pequeñas diferencias como los 0,1 segundos entre XGBoost y Gradient Boosting, estos podrían ser causas de las variaciones aleatorias que pueda tener el conjunto de datos de prueba.

Para determinar si las diferencias observadas son estadísticamente significativas entre los modelos, se aplicó la prueba de rangos con signo de Wilcoxon en cuanto a todos los errores absolutos de las 711.177 predicciones realizadas en el conjunto de prueba emparejando cada predicción entre modelos. Se optó por esta prueba no paramétrica porque la distribución de los errores no es normal, lo que desaconseja el uso de una prueba paramétrica como la t de Student.

Como complemento, se estimó un Intervalo de Confianza (IC) del 95% de la diferencia de MAE mediante remuestreo de bootstrap, esto con el fin de cuantificar no sólo la existencia de la diferencia, sino también su magnitud.

Tabla 10

Comparación de XGBoost frente a cada modelo - prueba de Wilcoxon e intervalo de confianza del 95%

Comparación (XGBoost vs.)	Dif. MAE (s)	IC 95 %	p-valor	¿Significativa?	<i>Nota. Dif.</i>
Gradient Boosting	0,25	[0,24 ; 0,25]	< 0,001	Sí	
LSTM	1,82	[1,79 ; 1,84]	< 0,001	Sí	
GRU	2,03	[2,00 ; 2,07]	< 0,001	Sí	
Regresión lineal	11,51	[11,43 ; 11,59]	< 0,001	Sí	
Promedio histórico	34,85	[34,74 ; 34,93]	< 0,001	Sí	

MAE(s) muestra la diferencia entre el MAE de cada modelo y el obtenido por XGBoost. Un valor positivo significa que XGBoost presentó menor error. Las diferencias con intervalos de confianza que no incluyen cero se consideraron significativas. Los p-valores inferiores a 0,001 se muestran como < 0,001.

XGBoost presentó un menor error que la regresión lineal, el promedio histórico, LSTM y GRU. Las diferencias fueron estadísticamente significativas en todas las comparaciones, con intervalos de confianza superiores a cero y p-valores inferiores a 0,001. Frente a LSTM y GRU, la ventaja de XGBoost fue de 1,82 y 2,03 segundos de MAE respectivamente, magnitudes relevantes en el contexto de segmentos cuya duración promedio es de aproximadamente 78 segundos.

En la comparación entre XGBoost y Gradient Boosting conviene realizar una interpretación distinta. La diferencia resultó estadísticamente significativa con ($p < 0,001$), pero su magnitud es de apenas 0,25 segundos, un valor sin relevancia práctica. El análisis por ruta sustenta esta conclusión: XGBoost superó a Gradient Boosting de manera significativa en cinco de las siete rutas, mientras que en las dos restantes (E2_SN y E3_NS) no se encontró diferencia significativa.

Por lo tanto, ambos modelos deben considerarse equivalentes en desempeño, esto hace que constituyan conjuntamente el grupo de mejor rendimiento del estudio.

Este resultado ilustra una distinción relevante, con un tamaño muestral de 711.177 predicciones emparejadas, la prueba de Wilcoxon posee una capacidad muy alta para detectar diferencias mínimas entre modelos, incluso cuando estas carecen de valor práctico para la operación de la Ecovía. Bajo este sentido, el p-valor por si solo no debe utilizarse como criterio único de decisión; el intervalo de confianza y la magnitud de la diferencia en las unidades del problema, componen criterios más adecuados para determinar las diferencias preferidas por cada modelo.

La selección de XGBoost como modelo final no se sustenta en una superioridad determinante sobre el modelo Gradient Boosting, sino en su ligera ventaja promedio y en su robustez consistente a lo largo del corredor.

4.2.3 Análisis por ruta

El análisis desagregado por ruta muestra que el desempeño de los modelos no es homogéneo en todo el corredor. La ruta E2_SN registró el menor error para todos los modelos (MAE = 10,4 segundos con XGBoost y con Gradient Boosting), resultado que podría estar asociado a ser la ruta más corta del sistema (9,29km) y, a presentar segmentos entre paradas más cortos y homogéneos. En contraste, la ruta E4 presentó el mayor error (MAE de 22,6 segundos con XGBoost), lo que posiblemente se relaciona con ser la ruta más larga del corredor (22,84 km) y con recorrido circular (Quitumbe-Playón de la Marín-Quitumbe), condición que podría introducir mayor variabilidad operativa a lo largo del trayecto.

En las siete rutas analizadas, XGBoost igualó o superó el desempeño de Gradient Boosting y obtuvo menores errores que LSTM y GRU. En promedio, la diferencia de MAE fue de aproximadamente 1,82 segundos frente a LSTM y de 2,03 segundos frente a GRU. Estos

resultados son consistentes con el análisis de significancia estadística presentado en la sección 4.2.2. En dos rutas, E2_SN y E3_NS, no se encontró una diferencia estadísticamente significativa entre XGBoost y Gradient Boosting; en las cinco restantes, XGBoost mantuvo una ligera ventaja.

El patrón observado entre la longitud de la ruta y el error del modelo, indica que la variabilidad geométrica del recorrido, es un factor relevante en el desempeño. E2_SN, con el menor error, es también la ruta más corta y con una distribución más homogénea de distancias entre paradas, en contraste a la ruta E4, que contiene mayor error con una mayor longitud que introduce condiciones de tráfico.

4.2.4 Efecto del tipo de normalización

Para evaluar si el tipo de normalización afecta el desempeño de los modelos, se compararon los resultados obtenidos con MinMaxScaler y con StandardScaler (Tabla 9).

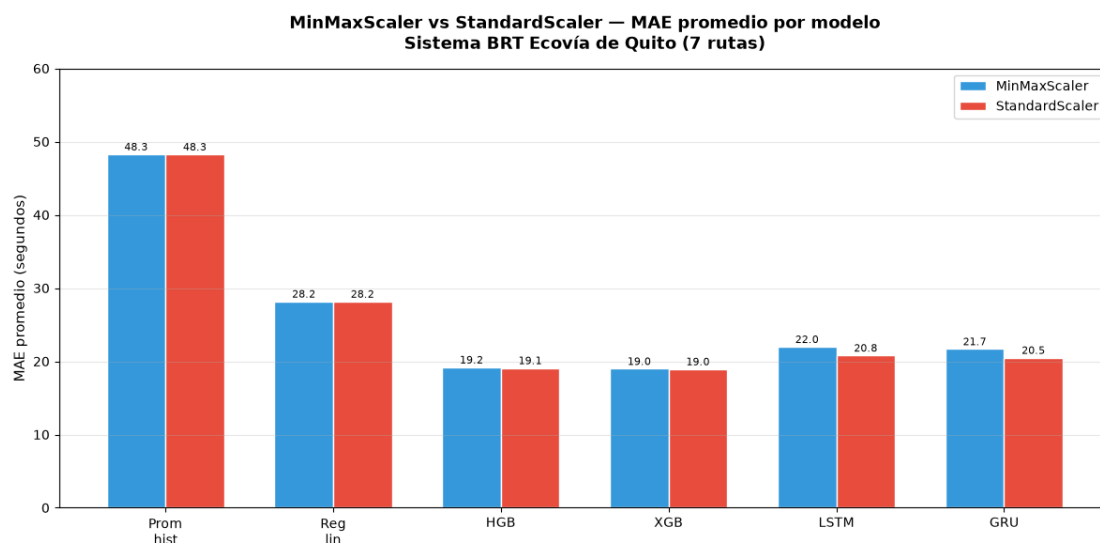
Tabla 11

MAE promedio por modelo - MinMaxScaler vs. StandardScaler (segundos, 7 rutas)

Modelo	MinMaxScaler	StandardScaler	Diferencia
Promedio histórico	48,3	48,3	sin cambio
Regresión lineal	28,2	28,2	sin cambio
Gradient Boosting	19,2	19,1	↓ 0,1 s
XGBoost	19,0	19,0	sin cambio
LSTM	22,0	20,8	↓ 1,2 s
GRU	21,7	20,5	↓ 1,2 s

Figura 8

Comparación de MAE promedio por modelo entre MinMaxScaler y StandardScaler



La normalización utilizada no ocasionó cambios relevantes en el desempeño de XGBoost y Gradient Boosting. Esto se esperaba, ya que los modelos basados en árboles no requieren que las variables de entrada se encuentren en una escala determinada. En modelos recurrentes sí se observó una mejora con StandardScaler. El MAE promedio de LSTM pasó de 22,0 a 20,8 segundos, mientras que el de GRU disminuyó de 21,7 a 20,5 segundos. Este comportamiento puede asociarse con la transformación de las variables a una escala centrada en cero, condición que resulta adecuada para el uso de funciones de activación como tanh en estos modelos. A pesar de esta mejora, XGBoost presentó el menor MAE promedio en ambos esquemas de normalización.

Al utilizar StandardScaler de forma uniforme para cada modelo, se garantiza el mejor desempeño para las arquitecturas recurrentes, sin perjudicar a los modelos de árboles, evitando la necesidad de mantener dos pipelines de procesamiento diferentes según el tipo de modelo utilizados por el usuario en la interfaz web.

4.2.5 Selección del modelo y discusión

Aunque la hipótesis inicial del trabajo consideraba que los modelos recurrentes (LSTM y GRU) podrían aprovechar mejor la dependencia temporal de los datos de GPS, los resultados muestran que XGBoost superó consistentemente a ambas arquitecturas recurrentes en las siete rutas y con los dos esquemas de normalización. Esto se explica principalmente por la corta duración promedio de los segmentos entre paradas del corredor Ecovía (1,3 minutos en promedio): en intervalos tan breves, la variable `distance_to_next_stop` concentra la mayor parte de la información predictiva (correlación de hasta 0,80 en la ruta E4), un patrón que un modelo de árboles de decisión captura de forma directa y eficiente, sin requerir el modelado explícito de dependencias temporales de largo plazo para el cual están diseñados LSTM y GRU.

En consecuencia, se selecciona XGBoost como modelo final para el prototipo de predicción de ETA descrito en la sección 3.2.4, con un MAE promedio de 19,0 segundos sobre las siete rutas del corredor. Este resultado responde directamente a la pregunta planteada en el Capítulo 1: para el corredor Ecovía, bajo las condiciones evaluadas, los modelos de gradient boosting se adaptan mejor que las arquitecturas recurrentes, principalmente por la naturaleza de segmentos cortos que caracteriza la operación del corredor, y no porque las arquitecturas recurrentes sean, en general, inadecuadas para la predicción de series temporales.

No obstante, los resultados deben ser interpretados considerando ciertas limitaciones. La primera limitación se encuentra relacionada con la cantidad de datos que se utilizó en los diferentes modelos. LSTM y GRU para su eficiencia en el funcionamiento necesitan consecutivamente de varios reportes GPS para crear una secuencia temporal antes de realizar la predicción. Este fue el motivo por el cual en las rutas con menor cantidad de reportes se perdieron más registros al construir las secuencias. En cambio, XGBoost utilizó todos los registros disponibles para realizar la predicción. Por tanto, en algunas rutas los modelos LSTM y GRU, se entrenaron con una cantidad de datos menor que XGBoost. En consecuencia, el menor desempeño

de las arquitecturas recurrentes no se atribuye únicamente a la naturaleza del problema como tal, sino también a la restricción estructural derivada de su propio diseño.

Esta limitación estructural sugiere que la comparación entre familias de modelos, no se debe entender como una evaluación en igualdad, sino como una comparación bajo restricciones inherentes al diseño de cada arquitectura frente a las características reales del corredor Ecovía.

En segundo lugar, si bien se realizaron pruebas exploratorias para la selección de los hiper parámetros, descritas en la sección 3.2.2 , no se efectuó una búsqueda detallada para todas las arquitecturas y todas las rutas. No obstante, dado a que la limitación de los modelos recurrentes en este problema es estructural (corta duración de los segmentos) y no de configuración, no se espera que una optimización adicional revista los resultados observados. Una búsqueda óptima de estos hiper parámetros queda planteada como trabajo futuro.

En tercer lugar, en el preprocesamiento se eliminaron registros de la variable (velocidad) igual a cero, lo que implica que en los momentos de detención de las unidades en los semáforos o en congestiones quedaron fuera del entrenamiento. Esta decisión está orientada a descartar reportes sin desplazamiento, introduce un posible sesgo ya que el modelo predomina su aprendizaje sobre las unidades que se encuentran en movimiento y podría subestimar los tiempos en las condiciones de tráfico detenido o movimiento nulo.

En cuarto lugar, se tiene un desbalance del conjunto de datos entre las rutas, ya que la ruta E1 aportó cerca de 290.000 predicciones al conjunto de prueba y la ruta E3_NS aportó menos de 2.000, es por ello que los modelos de las rutas con menor cantidad de datos se entrenaron con menos información y sus resultados deben interpretarse con mayor cuidado.

Por último, los resultados que se obtuvieron corresponden al corredor de la Ecovía comprendido entre el periodo septiembre de 2025 y marzo de 2026. Por esta razón, no se puede garantizar que el desempeño será el mismo para otros periodos o en otros sistemas BRT. Para comprobarlo sería necesario realizar pruebas en nuevos períodos y especialmente realizarlo con

sistemas que tengan mayores distancias entre las paradas. Asimismo, el cálculo de la métrica MAPE se realizó excluyendo los segmentos con tiempos menores a un minuto, para evitar inestabilidad de esta métrica con denominadores muy pequeños, lo que acota su interpretación.

CAPÍTULO 5: CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

El presente trabajo comparó seis modelos de aprendizaje automático -promedio histórico, regresión lineal, HistGradientBoosting, XGBoost, LSTM y GRU- para predecir el tiempo de llegada entre paradas del corredor BRT Ecovía de Quito, utilizando exclusivamente los registros históricos de GPS del sistema SAE de Mivilsoft S.A., sin depender de fuentes de datos externas. Los resultados alcanzados permiten declarar que la información extraída del SAE, son fuente suficiente para la predicción de tiempos de llegada entre paradas de la Ecovía.

La recopilación y el preprocesamiento del dataset histórico permitieron construir un conjunto de datos limpio, a partir de la eliminación de registros GPS inválidos, la corrección de la variable `is_holiday` y la eliminación de valores atípicos de velocidad y de tiempo entre paradas definidos por ruta. Este proceso confirmó que era posible obtener un dataset representativo de la operación real del corredor usando únicamente los registros GPS institucionales, sin recurrir a datos de tráfico, clima o eventos externos.

El análisis exploratorio mostró que la distancia hasta la siguiente parada fue la variable con mayor relación con el tiempo de llegada, evidenciando la importancia del componente espacial para la predicción. También se logró observar relaciones con variables como la hora del día y la condición de hora pico. Por otro lado, la altitud y el tiempo programado de paso no presentaron un comportamiento consistente entre las rutas, por lo que no fueron consideradas en los modelos finales. En el caso del horario programado, los resultados sugieren que este no representa de manera adecuada los tiempos reales de operación del corredor.

A diferencia de la hipótesis inicial, que esperaba que las arquitecturas recurrentes aprovecharán mejor la dependencia temporal de los datos de GPS, los resultados demostraron que el modelo XGBoost obtuvo el mejor desempeño en las siete rutas del corredor, con un MAE promedio de 19,0 segundos. HistGradientBoosting presentó un desempeño muy próximo, mientras

que los modelos recurrentes GRU y LSTM se ubicaron en un segundo grupo, sin superar los modelos de árboles. Esto se justifica porque la duración promedio entre paradas del corredor Ecovía es de apenas 1,3 minutos, un intervalo en el que la distancia a la siguiente parada concentra la mayor parte de la información predictiva y este patrón se capturan de forma más directa y eficiente en los modelos de árboles de decisión que las arquitecturas recurrentes.

El análisis por ruta expuso que el desempeño de los modelos no es homogéneo en todo el corredor: la ruta E2_SN, la más corta del sistema, registró el menor error agregado, mientras que la ruta E4, la más larga y con recorrido circular, presentó el mayor error. Esta información es directamente utilizable para la aplicación de modelos ETA, porque permite definir que el desempeño debe evaluarse de forma desagregada por las características de cada ruta y no a una métrica global del corredor.

Finalmente, se desarrolló y publicó un prototipo web funcional que permite visualizar y comparar las predicciones de los seis modelos entrenados para cada ruta, viaje y parada del corredor (<https://huggingface.co/spaces/ChrissHF/Dev-ML>), siendo que el prototipo constituye una prueba de concepto para demostrar la viabilidad técnica de integrar modelos desarrollados y revisar sus predicciones a través de un entorno web. Su finalidad es transferir los resultados de un análisis experimental a una herramienta funcional, sin sustituir un sistema de producción.

Más allá de la comparación de desempeño, la principal contribución en esta investigación es demostrar con evidencia empírica y respaldo estadístico, que en los sistemas BRT de segmentos cortos los modelos de gradient boosting constituyen una alternativa más precisa y operativamente más factible que las arquitecturas recurrentes, a pesar de que los modelos recurrentes suelen considerarse el estándar para la predicción de series temporales. Este hallazgo discute la elección por defecto de modelos de aprendizaje profundo para este tipo de problemas y aporta un criterio práctico para el diseño de sistemas de predicciones de ETA para corredores con características similares. A ello se suma que el resultado se obtuvo utilizando exclusivamente registros GPS

institucionales, sin fuentes externas, en una ciudad andina de altitud elevada, un contexto escasamente abordado en la literatura previa, esto amplía la aplicabilidad del enfoque a entornos con una limitante en la disponibilidad de los datos.

5.2 Recomendaciones

A partir de los hallazgos de este trabajo, se plantean las siguientes recomendaciones para el sistema Ecovía de Quito y para futuras investigaciones en el área.

En primer lugar, se recomienda extender la metodología desarrollada a los otros corredores del sistema Metrobús-Q, como el Trole y la Troncal Occidental, aprovechando que Mivilsoft S.A. opera con el mismo sistema SAE en dichos corredores y que el pipeline de preprocesamiento, ingeniería de variables y entrenamiento por ruta desarrollado en este trabajo es directamente reutilizable.

En segundo lugar, dado que XGBoost y HistGradientBoosting superaron a las arquitecturas recurrentes precisamente porque los segmentos entre paradas del corredor Ecovía son cortos, se recomienda evaluar en trabajos futuros si esta ventaja se mantiene en corredores con distancias entre paradas mayores, donde la dependencia temporal de largo plazo que LSTM y GRU están diseñados para capturar podría tener un peso mayor.

En tercer lugar, se recomienda incorporar variables externas adicionales en futuras versiones del modelo, como condiciones climáticas y eventos especiales de la ciudad, ya que estos factores no fueron evaluados en este trabajo por no estar disponibles en los registros del sistema SAE, pero podrían mejorar aún más la precisión de las predicciones, especialmente en las rutas y franjas horarias donde el error actual es mayor.

En cuarto lugar, se recomienda monitorear el desempeño del modelo XGBoost seleccionado a medida que avance la renovación de la flota del corredor Ecovía hacia buses eléctricos, y reentrenarlo periódicamente con registros más recientes, dado que el modelo fue

entrenado con datos del período septiembre 2025 - marzo 2026 y el comportamiento de la flota puede cambiar conforme avance esa transición.

Finalmente, se recomienda que el prototipo desarrollado en este trabajo (sección 3.2.4) se utilice como punto de partida para una eventual integración directa con el sistema SAE de Mivilsoft S.A., de manera que los operadores del sistema Ecovía tengan disponible la predicción de tiempos de llegada dentro de la misma plataforma que ya usan día a día para gestionar la flota, en lugar de una aplicación independiente (Mivilsoft S.A., 2024).

REFERENCIAS BIBLIOGRÁFICAS

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. En A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1179>
- Chondrodima, E., Georgiou, H., Pelekis, N., & Theodoridis, Y. (2022). Public transport arrival time prediction based on GTFS data. En *Machine Learning, Optimization, and Data Science. Lecture Notes in Computer Science*, vol. 13164 (pp. 467–481). Springer. https://doi.org/10.1007/978-3-030-95470-3_36
- Chughtai, J., Ul Haq, I., & Muneeb, M. (2022). An attention-based recurrent learning model for short-term travel time prediction. *PLOS ONE*, 17(12), e0278064. <https://doi.org/10.1371/journal.pone.0278064>
- Empresa Pública Metropolitana de Transporte de Pasajeros de Quito. (2023). *Informe de gestión del corredor Ecovía: indicadores operacionales 2022–2023*. EPMTQP. <https://pasajeros.quito.gob.ec/>
- Empresa Pública Metropolitana de Transporte de Pasajeros de Quito. (2024). *Indicadores operacionales – Corredor Ecovía 2023*. EPMTQP. <https://pasajeros.quito.gob.ec/>

Empresa Pública Metropolitana de Transporte de Pasajeros de Quito. (2025). *En 2025 arrancó la modernización integral del transporte público municipal*. EPMTPQ.

<https://pasajeros.quito.gob.ec/?p=13185>

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Hosseini, S. S., Rashvand, N., Azarbayjani, M., & Tabkhi, H. (2024). Real-time bus departure prediction using neural networks for smart IoT public bus transit. *IoT*, 5(4), 650–665.

<https://doi.org/10.3390/iot5040029>

Hwang, G., Hwang, Y., Shin, S., Lee, J., & Kim, H. (2022). Comparative study on the prediction of city bus speed between LSTM and GRU. *International Journal of Automotive Technology*, 23, 983–992. <https://doi.org/10.1007/s12239-022-0085-z>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. En *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 3146–3154). Curran Associates.

Kingma, D. P., & Ba, J. (2015). Adam: *A method for stochastic optimization*. 3rd International Conference on Learning Representations (ICLR).

Kumar, B. A., Singh, R., Shaji, H. E., & Vanajakshi, L. (2025). Bus arrival time prediction: A comprehensive review. *IEEE Transactions on Intelligent Transportation Systems*, 26(6), 7362–7379.

<https://doi.org/10.1109/TITS.2025.3545695>

- Lee, C., & Yoon, Y. (2022). A novel bus arrival time prediction method based on spatio-temporal flow centrality analysis and deep learning. *Electronics*, 11(12), 1875.
<https://doi.org/10.3390/electronics11121875>
- Mendoza-Mendoza, A., García-Macías, J. A., Aguayo-Torres, M. C., & Espinoza-Ruiz, A. (2022). Comprehensive model based on time series for the generation of traffic knowledge for bus transit rapid line 6 of México City. *Electronics*, 11(19), 3036.
<https://doi.org/10.3390/electronics11193036>
- Mivilsoft S.A. (2024). *Sistema de ayuda a la explotación SAE para transporte público urbano*. Mivilsoft S.A. <https://mivilsoft.com/>
- Nuha, H., Azhari, A., & Kuspriyanto, K. (2026). Ensemble machine learning approach for traffic congestion and travel time prediction in urban BRT systems: A case study of Trans Metro Bandung. *IoT*, 7(1), 22. <https://doi.org/10.3390/iot7010022>
- Ortega Campues, J. H., Ramos Mayorga, O. A., & Torres Chango, C. D. (2026). *Dataset histórico de registros GPS del sistema BRT Ecovía de Quito [Conjunto de datos]*. Google Drive.
<https://drive.google.com/file/d/1kNSW47HKJLAWoNRWvGq3fMrHMi8CnoNH/view?usp=sharing>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
<https://jmlr.org/papers/v12/pedregosa11a.html>

Quito Cómo Vamos. (2025). *Movilidad en Quito 2024: Factsheet ICV 2024 y EPC 2024*.

https://quitocomovamos.org/wp-content/uploads/2025/02/07Factsheet_Movilidad2025.pdf

Ranjbari, A., Shaon, M. R. R., & Zhao, L. (2022). Bus travel time prediction using machine learning: A case study in an urban BRT corridor. *Transportation Research Record*, 2676(9), 652–663. <https://doi.org/10.1177/03611981221083218>

Rashvand, N., Hosseini, S., Azarbayjani, M., & Tabkhi, H. (2024). Real-time bus arrival prediction: A deep learning approach for enhanced urban mobility. En *Proceedings of the 13th International Conference on Operations Research and Enterprise Systems (ICORES)* (pp. 123–132). SciTePress. <https://doi.org/10.5220/0012365500003639>

Singh, N., & Kumar, K. (2022). A review of bus arrival time prediction using artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 12(4), e1457. <https://doi.org/10.1002/widm.1457>

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.

Torres Chango, C. D., Ortega Campus, J. H., & Ramos Mayorga, O. A. (2026). *Sistema de predicción ETA — BRT Ecovía de Quito [Aplicación web]*. Hugging Face Spaces. <https://huggingface.co/spaces/ChrissHF/Dev-ML>

Waqas, M., & Wannasingha Humphries, U. (2024). A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX*, 13, Article 102946. <https://doi.org/10.1016/j.mex.2024.102946>

- Xie, H., Dong, H., Hu, H., & Gao, Y. (2021). Multistep prediction of bus arrival time with the recurrent neural network. *Mathematical Problems in Engineering*, 2021, Article 6636367. <https://doi.org/10.1155/2021/6636367>
- Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324. <https://doi.org/10.1016/j.trc.2015.02.022>
- Zhu, L., Shu, S., & Zou, L. (2022). XGBoost-based travel time prediction between bus stations and analysis of influencing factors. *Wireless Communications and Mobile Computing*, 2022, Article 3504704. <https://doi.org/10.1155/2022/3504704>

APÉNDICE A

Rutas y tiempos programados del sistema BRT Ecovía de Quito

Las siguientes tablas presentan los tiempos programados de parada a parada para cada una de las 7 rutas que opera el sistema BRT Ecovía de Quito, según los registros del sistema SAE de (Mivilsoft S.A., 2024). Esta información constituye el marco operativo de referencia utilizado para la construcción del dataset histórico empleado en el entrenamiento de los modelos.

Tabla 12
Tiempos programados de paradas - Ruta E1 (Guamaní - Las Universidades - Guamaní)

Código de ruta: E1 | Distancia: 36,00 km | Tiempo por vuelta: 1h 50 min| ID
Traccar: 1

Parada	Tiempo programado	Tipo
Estación Guamaní	00:00	Salida
Santo Tomás	00:04	Llegada
Caupicho	00:06	Llegada
El Beaterio	00:08	Llegada
La Bretaña	00:11	Llegada
Guayanay Ñan	00:14	Llegada

Capulí	00:17	Llegada
Pacarillacta	00:18	Llegada
Puente Guajaló	00:20	Llegada
San Cristóbal	00:22	Llegada
Ayapamba	00:24	Llegada
El Comercio	00:26	Llegada
San Bartolo	00:27	Llegada
Epiclachima	00:29	Llegada
Pujilí	00:30	Llegada
El Recreo	00:33	Llegada
Estadio Chimbacalle	00:38	Llegada
Teatro México	00:39	Llegada

Colegio Montúfar	00:40	Llegada
------------------	-------	---------

El Playón de la Marín	00:45	Llegada
-----------------------	-------	---------

Marín Central	00:48	Llegada
---------------	-------	---------

Av. 12 de Octubre	00:49	Salida
-------------------	-------	--------

Simón Bolívar	00:51	Llegada
---------------	-------	---------

Eugenio Espejo	00:53	Llegada
----------------	-------	---------

De las Universidades	00:54	Llegada
----------------------	-------	---------

Galo Plaza	00:56	Llegada
------------	-------	---------

Casa de la Cultura	00:58	Llegada
--------------------	-------	---------

Eugenio Espejo	01:02	Llegada
----------------	-------	---------

Simón Bolívar	01:04	Llegada
---------------	-------	---------

Marín Central	01:07	Llegada
---------------	-------	---------

El Playón de la Marín	01:09	Llegada
Colegio Montúfar	01:14	Llegada
Teatro México	01:15	Llegada
Estadio Chimbacalle	01:17	Llegada
El Recreo	01:22	Llegada
Pujilí	01:25	Llegada
Epiclachima	01:26	Llegada
San Bartolo	01:27	Llegada
El Comercio	01:29	Llegada
Ayapamba	01:31	Llegada
San Cristóbal	01:32	Llegada
Puente Guajaló	01:34	Llegada

Pacarillacta	01:35	Llegada
Capulí	01:37	Llegada
Guayanay Ñan	01:39	Llegada
La Bretaña	01:41	Llegada
El Beaterio	01:43	Llegada
Caupicho	01:45	Llegada
Santo Tomás	01:47	Llegada
Terminal Guamaní	01:50	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

Tabla 13
Tiempos programados de paradas — Ruta E1M (Guamaní – Marín Central – Guamaní

Código de ruta: E1M | Distancia: 30,30 km | Tiempo por vuelta: 1h 30 min| ID
Traccar: 3

Parada	Tiempo programado	Tipo
Estación Guamaní	00:00	Salida
Av. Maldonado	00:01	Salida
Santo Tomás	00:04	Llegada
Caupicho	00:06	Llegada
El Beaterio	00:08	Llegada
La Bretaña	00:11	Llegada
Guayanay Ñan	00:14	Llegada
Capulí	00:17	Llegada
Pacarillacta	00:18	Llegada
Puente Guajaló	00:20	Llegada
San Cristóbal	00:22	Llegada

Ayapamba	00:24	Llegada
El Comercio	00:26	Llegada
San Bartolo	00:27	Llegada
Epiclachima	00:29	Llegada
Pujilí	00:30	Llegada
El Recreo	00:33	Llegada
Estadio Chimbacalle	00:38	Llegada
Teatro México	00:39	Llegada
Colegio Montúfar	00:40	Llegada
El Playón de la Marín	00:45	Llegada
Marín Central	00:48	Llegada
El Playón de la Marín	00:50	Llegada

Colegio Montúfar	00:55	Llegada
Teatro México	00:56	Llegada
Estadio Chimbacalle	00:58	Llegada
El Recreo	01:03	Llegada
Pujilí	01:06	Llegada
Epiclachima	01:07	Llegada
San Bartolo	01:08	Llegada
El Comercio	01:10	Llegada
Ayapamba	01:12	Llegada
San Cristóbal	01:13	Llegada
Puente Guajaló	01:14	Llegada
Pacarillacta	01:15	Llegada

Capulí	01:17	Llegada
Guayanay Ñan	01:19	Llegada
La Bretaña	01:21	Llegada
El Beaterio	01:23	Llegada
Caupicho	01:25	Llegada
Santo Tomás	01:27	Llegada
Terminal Guamaní	01:30	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

Tabla 14

Tiempos programados de paradas - Ruta E2_N-S (Río Coca – Quitumbe)

Código de ruta: E2_N-S | Distancia: 20,55 km | Tiempo por vuelta: 1h 13 min|
ID Traccar: 12

Parada	Tiempo programado	Tipo
--------	----------------------	------

Estación Río Coca	00:00	Salida
Jipijapa	00:02	Llegada
Sauces	00:04	Llegada
24 de Mayo	00:05	Llegada
Naciones Unidas	00:06	Llegada
Benalcázar	00:08	Llegada
Eloy Alfaro	00:10	Llegada
Bellavista	00:11	Llegada
San Martín	00:13	Llegada
La Paz	00:15	Llegada
Orellana	00:17	Llegada
Baca Ortiz	00:19	Llegada

Manuela Cañizares	00:21	Llegada
Galo Plaza	00:22	Llegada
Casa de la Cultura	00:25	Llegada
Eugenio Espejo	00:28	Llegada
Simón Bolívar	00:29	Llegada
Marín Central	00:32	Llegada
El Playón de la Marín	00:35	Llegada
Manuela Santa Cruz y Espejo	00:36	Salida
Colegio Montúfar	00:40	Llegada
Teatro México	00:42	Llegada
Estadio Chimbacalle	00:44	Llegada
El Recreo (Moraspungo)	00:47	Llegada

Pujilí	00:50	Llegada
Epiclachima	00:51	Llegada
San Bartolo	00:52	Llegada
El Comercio	00:54	Llegada
Ayapamba	00:56	Llegada
San Cristóbal	00:58	Llegada
Puente Guajaló	01:01	Llegada
Pacarillacta	01:03	Llegada
Capulí	01:05	Llegada
Otoya	01:08	Llegada
Quillallacta	01:10	Llegada
Terminal Quitumbe	01:13	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

Tabla 15

Tiempos programados de paradas - Ruta E2_S-N (Quitumbe – Río Coca)

Código de ruta: E2_S-N | Distancia: 20,55 km | Tiempo por vuelta: 1h 13 min|
ID Traccar: 11

Parada	Tiempo programado	Tipo
Terminal Quitumbe	00:00	Salida
Quillallacta	00:02	Llegada
Otoya	00:03	Llegada
Capulí	00:05	Llegada
Pacarillacta	00:07	Llegada
Puente Guajaló	00:08	Llegada
San Cristóbal	00:09	Llegada
Ayapamba	00:11	Llegada

El Comercio	00:12	Llegada
San Bartolo	00:14	Llegada
Epiclachima	00:16	Llegada
Pujilí	00:18	Llegada
El Recreo	00:20	Llegada
Estadio Chimbacalle	00:25	Llegada
Teatro México	00:27	Llegada
Colegio Montúfar	00:29	Llegada
El Playón de la Marín	00:33	Llegada
Av. 12 de Octubre	00:34	Salida
Marín Central	00:35	Llegada
Simón Bolívar	00:37	Llegada

Eugenio Espejo	00:39	Llegada
Casa de la Cultura	00:41	Llegada
Galo Plaza	00:43	Llegada
Manuela Cañizares	00:45	Llegada
Baca Ortiz	00:47	Llegada
Orellana	00:49	Llegada
La Paz	00:51	Llegada
San Martín	00:54	Llegada
Bellavista	00:56	Llegada
Eloy Alfaro	00:59	Llegada
Benalcázar	01:01	Llegada
Naciones Unidas	01:04	Llegada

24 de Mayo	01:06	Llegada
Sauces	01:08	Llegada
Jipijapa	01:10	Llegada
Estación Río Coca	01:13	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

Tabla 16
Tiempos programados de paradas - Ruta E3_N-S (Río Coca – Playón de la Marín)

Código de ruta: E3_N-S | Distancia: 9,29 km | Tiempo por vuelta: 38 min | ID
Traccar: 9

Parada	Tiempo programado	Tipo
Estación Río Coca	00:00	Salida
Av. Río Coca	00:01	Salida
Jipijapa	00:02	Llegada
Sauces	00:04	Llegada

24 de Mayo	00:05	Llegada
Naciones Unidas	00:07	Llegada
Benalcázar	00:09	Llegada
Eloy Alfaro	00:11	Llegada
Bellavista	00:12	Llegada
San Martín	00:14	Llegada
La Paz	00:16	Llegada
Orellana	00:18	Llegada
Baca Ortiz	00:20	Llegada
Manuela Cañizares	00:22	Llegada
Galo Plaza	00:24	Llegada
Casa de la Cultura	00:27	Llegada

Eugenio Espejo	00:30	Llegada
Simón Bolívar	00:32	Llegada
Marín Central	00:35	Llegada
El Playón de la Marín	00:38	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

Tabla 17

Tiempos programados de paradas - Ruta E3_S-N (Playón de la Marín – Río Coca)

Código de ruta: E3_S-N | Distancia: 9,29 km | Tiempo por vuelta: 38 min | ID
Traccar: 10

Parada	Tiempo programado	Tipo
El Playón de la Marín	00:00	Salida
Av. Pichincha (E3)	00:01	Salida
Marín Central	00:02	Llegada

Simón Bolívar	00:05	Llegada
Eugenio Espejo	00:07	Llegada
Casa de la Cultura	00:09	Llegada
Galo Plaza	00:11	Llegada
Manuela Cañizares	00:13	Llegada
Baca Ortiz	00:15	Llegada
Orellana	00:17	Llegada
La Paz	00:19	Llegada
San Martín	00:22	Llegada
Bellavista	00:24	Llegada
Eloy Alfaro	00:26	Llegada
Benalcázar	00:28	Llegada

Naciones Unidas	00:30	Llegada
24 de Mayo	00:32	Llegada
Sauces	00:34	Llegada
Jipijapa	00:35	Llegada
Estación Río Coca	00:38	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

Tabla 18
Tiempos programados de paradas - Ruta E4 (Quitumbe – Playón de la Marín – Quitumbe)

Código de ruta: E4 | Distancia: 22,84 km | Tiempo por vuelta: 1h 04 min| ID
Traccar: 6

Parada	Tiempo programado	Tipo
Terminal Quitumbe	00:00	Salida
Av. Cóndor Ñan	00:02	Salida

Quillallacta	00:02	Llegada
Otoya	00:03	Llegada
Capulí	00:05	Llegada
Pacarillacta	00:07	Llegada
Puente Guajaló	00:08	Llegada
San Cristóbal	00:09	Llegada
Ayapamba	00:11	Llegada
El Comercio	00:12	Llegada
San Bartolo	00:14	Llegada
Epiclachima	00:16	Llegada
Pujilí	00:17	Llegada
El Recreo	00:20	Llegada

Estadio Chimbacalle	00:25	Llegada
Teatro México	00:26	Llegada
Colegio Montúfar	00:28	Llegada
El Playón de la Marín	00:33	Llegada
Colegio Montúfar	00:37	Llegada
Teatro México	00:38	Llegada
Estadio Chimbacalle	00:40	Llegada
El Recreo (Moraspungo)	00:43	Llegada
Pujilí	00:45	Llegada
Epiclachima	00:46	Llegada
San Bartolo	00:47	Llegada
El Comercio	00:49	Llegada

Ayapamba	00:51	Llegada
San Cristóbal	00:53	Llegada
Puente Guajaló	00:55	Llegada
Pacarillacta	00:57	Llegada
Capulí	00:59	Llegada
Otoya	01:01	Llegada
Quillallacta	01:02	Llegada
Terminal Quitumbe	01:04	Llegada

Nota. Tiempos programados proporcionados por Mivilsoft S.A. mediante el sistema SAE (2024).

APÉNDICE B

Repositorio de datos y recursos

Dataset histórico GPS - Sistema BRT Ecovía de Quito

El dataset utilizado en este trabajo contiene: (28011776 filas, 36 columnas) de registros históricos de operación del sistema Ecovía correspondientes al período septiembre 2025 – marzo 2026, obtenidos mediante un proceso ETL sobre el sistema SAE de MivilSoft S.A. El archivo en formato CSV está disponible en:

<https://drive.google.com/file/d/1kNSW47HKJLAWoNRWvGq3fMrHMi8CnoNH/view?usp=sharing>

APÉNDICE C

Variables excluidas del modelo

Ruta	Umbral eta (min)	Percentil 99 observado (min)
bus_id	Identificación	Identifica el bus físico; los modelos no deben aprender comportamientos de buses específicos, que pueden cambiar de ruta o ser reemplazados.
trip_id	Identificación	Identifica el viaje específico; no se repite en el tiempo, por lo que no permite generalizar.
timestamp	Temporal	Fecha y hora exacta del reporte; no se repite. Sus patrones ya se capturan con hour y day_of_week.
start_time	Temporal	Hora de inicio del viaje; no aporta información adicional a la que captura hour.
end_time	Temporal	Hora de fin del viaje; no disponible al momento de predecir.
date	Temporal	Fecha específica; no se repite. Capturada por day_of_week e is_holiday.
latitude	Espacial	Posición GPS exacta; no generaliza. Capturada por current_stop_id.
longitude	Espacial	Misma razón que latitude.
altitude	Espacial	Correlación contradictoria entre rutas (E3_NS: +0,29; E4: -0,29); sin patrón consistente.

course	Operacional	Dirección en grados; capturada por route_id y current_stop_order.
accuracy	Operacional	Precisión del GPS; solo se usa para filtrar registros.
valid	Operacional	Validez del registro GPS; solo se usa para filtrar registros inválidos.
step_time	Operacional	Directamente relacionada con eta_next_stop; generaría fuga de información.
scheduled_passing_time	Operacional	Correlación casi nula con la eta real (-0,03 a 0,06); el horario programado no refleja la operación real.
real_passing_time	Operacional	Hora real de llegada; no disponible en tiempo real en producción.
last_stop_time	Operacional	Momento de llegada a la parada actual; capturado por time_since_last_stop.
bus_stop_lat	Espacial	Coordenada de la parada actual; capturada por current_stop_id.
bus_stop_lng	Espacial	Misma razón que bus_stop_lat.
next_stop_order	Espacial	Equivale a current_stop_order + 1; redundante.
next_stop_id	Espacial	Los modelos predicen la eta sin necesidad de identificar la siguiente parada.
next_stop_time	Operacional	Directamente relacionada con eta_next_stop; generaría fuga de información.

next_stop_lat	Espacial	Capturada por distance_to_next_stop.
next_stop_lng	Espacial	Misma razón que next_stop_lat.

Nota. Son 23 variables excluidas que, sumadas a las 11 variables predictoras (Tabla 2), la variable objetivo (eta_next_stop) y el identificador de segmentación (route_id), completan las 36 variables del dataset original.

APÉNDICE D

Repositorio de código y notebooks de desarrollo

El desarrollo experimental de este trabajo de titulación se implementó en Python, mediante notebooks de Jupyter, estos notebooks se encuentran organizados de manera secuencial para reproducir cada etapa del proceso que son:

- Exploración inicial
- Procesamiento del dataset
- Entrenamiento de modelos
- Evaluación comparativa de resultados

En cada notebook se documentan las decisiones metodológicas adoptadas para este desarrollo, los parámetros empleados y las salidas obtenidas, de modo que un tercero pueda hacer uso de ellas y replicar el estudio a partir del dataset original.

El código fuente completo se encuentra disponible en el repositorio público de GitHub en el siguiente enlace:

<https://github.com/ChrissT94/TFM-Ecovia-Notebooks>

Tabla 20

Notebooks que integran el repositorio de desarrollo

Notebook	Contenido	Salida
01_TFMEcovia_AnalisisExploratorio	Análisis de la variable objetivo <code>eta_next_stop</code> y de las variables temporales, operacionales y espaciales sobre los 28.011.776 registros originales. Identificación de patrones, valores atípicos y decisiones de preprocesamiento.	Gráficas de distribución y matrices de correlación por ruta.

02_TFMEcovia_ ProcesamientoDa tos	Aplicación de las nueve transformaciones definidas en el análisis exploratorio: depuración de registros inválidos, conversión de unidades, eliminación de valores atípicos por ruta, corrección de feriados y selección de variables.	dataset_limpio.csv (24.308.817 registros)
---	---	--

03_TFMEcovia_ Modelos_MinMa xScaler	Entrenamiento y evaluación de los seis modelos sobre las siete rutas, con partición temporal y normalización MinMaxScaler ajustada únicamente al conjunto de entrenamiento.	Métricas MAE, RMSE y MAPE por ruta y modelo.
---	---	--

04_TFMEcovia_ Modelos_Standar dScaler	Réplica del pipeline anterior con StandardScaler, con el fin de evaluar el impacto del tipo de normalización sobre las redes recurrentes.	Métricas MAE, RMSE y MAPE por ruta y modelo
---	---	---

Nota. Todos los notebooks se ejecutaron sobre el mismo dataset y con la misma semilla aleatoria, lo que garantiza la comparabilidad de los resultados presentados en el Capítulo 4.

APÉNDICE E

Prototipo web de predicción del tiempo de llegada

Con la finalidad de validar la aplicabilidad de los modelos entrenados, se desarrolló un prototipo web el cual permite consultar el tiempo de llegada entre paradas a partir de los datos históricos reales del sistema Ecovía contenidos en el dataset.

Cabe mencionar que este prototipo no constituye un producto de despliegue definitivo, sino una prueba de concepto orientada a demostrar que los modelos responden en condiciones cercanas a la operación real.

La aplicación se encuentra desplegada en la plataforma Hugging Face; para una mejor interacción es recomendable esperar a que las librerías se carguen correctamente, & en casos puntuales recargar la página.

Se encuentra disponible públicamente en:

<https://huggingface.co/spaces/ChrissHF/Dev-ML>

El usuario debe seleccionar la ruta, el viaje, la parada origen y el modelo con el que desea realizar la predicción. La interfaz amigable devuelve el tiempo estimado hasta la siguiente parada, & permite comparar de forma simultánea el resultado de los seis modelos mediante la opción “Comparar todos”. Adicionalmente, la opción “Modo automático” genera una predicción de todos los registros de la parada seleccionada en un lapso de 2 segundos, mientras que la opción “Siguiendo reporte” permite avanzar de forma manual.

Para los modelos LSTM y GRU, cuando no se dispone del número de registros necesario para completar la ventana temporal, la interfaz muestra un indicador de advertencia en lugar de la predicción, lo que evidencia de forma directa la limitación estructural de estos modelos descrita en el capítulo 4.

Tabla 21*Especificaciones técnicas del prototipo*

Componente	Tecnología
Servicio de predicción	FastAPI (Python)
Interfaz de usuario	HTML, CSS y JavaScript
Empaquetado	Docker
Plataforma de despliegue	Hugging Face Spaces
Modelos integrados	Promedio histórico, regresión lineal, HistGradientBoosting, XGBoost, LSTM y GRU

Nota. Las Figuras 4 y 5 del Capítulo 3 presentan capturas de la interfaz en funcionamiento para las rutas E1 y E4.