Aplicación de Técnicas de Machine Learning para la Predicción del Consumo Hídrico en la Ciudad de Bogotá

Javier Callejas Cardozo

Xiomara Grisales Henao

Luis Miguel Rios Cuentas

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos / Maestría en Gerencia de Proyectos

Bogotá, Colombia

20/10/2026

Aplicación de Técnicas de Machine Learning para la Predicción del Consumo Hídrico en la Ciudad de Bogotá

Javier Callejas Cardozo

Xiomara Grisales Henao

Luis Miguel Rios Cuentas

Trabajo de grado presentado como requisito para optar al título de:

Magister en Ciencia de Datos / Magister en Gerencia de Proyectos

Directora:

Lizeth Andrea Moreno Amaya

Modalidad:

Monografía

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos / Maestría en Gerencia de Proyectos

Bogotá, Colombia

20/10/2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, D.C., 2026

A Dios, por ser la luz que guía mis pasos aun cuando el camino es incierto y por enseñarme que, incluso en la adversidad, florecen las lecciones que transforman la vida.

A mi esposa, por caminar a mi lado en esta travesía, por su amor incondicional y por demostrarme que los sueños encuentran su verdadero sentido cuando se construyen juntos.

Luis M. Rios.

A Dios, por ser mi fortaleza y guía en cada paso de este camino. Gracias por permitirme alcanzar un sueño que durante muchos años vivió en mi corazón y por enseñarme que la fe y la perseverancia hacen posible lo que un día parecía inalcanzable.

Javier Callejas.

"Lo que sabemos es una gota de agua; lo que ignoramos es el océano."

— Isaac Newton

Xiomara Grisales.

Agradecimientos

Expresamos nuestro sincero agradecimiento a la Universidad EAN y a la Facultad de Ingeniería, por brindarnos una formación académica de excelencia y un entorno que promovió el pensamiento crítico, la investigación y el desarrollo profesional.

Expresamos un agradecimiento especial a la directora de este trabajo de grado, Lizeth Andrea Moreno Amaya, por su orientación, acompañamiento y valiosos aportes durante el desarrollo de la investigación. Asimismo, agradecemos a los docentes que, con su conocimiento y experiencia, contribuyeron significativamente a nuestro crecimiento académico y profesional.

Finalmente, agradecemos a todas las personas e instituciones que, de una u otra manera, hicieron posible la realización de este proyecto y aportaron a la culminación de esta etapa de nuestra formación.

Resumen

Este trabajo de grado analiza el consumo hídrico en la ciudad de Bogotá en un contexto de creciente presión sobre las fuentes de abastecimiento, como los sistemas de Chingaza y Sumapaz, así como de variabilidad climática asociada a fenómenos como El Niño. Como antecedentes, se consideran la crisis hídrica de 2024-2025 y las diferencias en patrones de consumo entre localidades, evidenciando la necesidad de estudios basados en datos a gran escala.

El objetivo principal consistió en desarrollar y validar modelos predictivos basados en técnicas de *Machine Learning* (ML) para estimar el consumo promedio de agua por localidad y periodo bimestral en la ciudad de Bogotá, mediante la integración de variables climáticas, temporales y espaciales. La metodología fue cuantitativa e incluyó análisis descriptivo, depuración de datos, identificación de valores atípicos, modelado predictivo y evaluación del desempeño frente a modelos de referencia. Se implementaron redes neuronales profundas y *XGBoost*, cuyo desempeño se evaluó mediante tres estrategias complementarias de validación: partición aleatoria, validación temporal con entrenamiento en 2020-2023 y prueba en 2024, y validación espacial *Leave-One-Localidad-Out* (LOLO).

Los resultados evidencian que el consumo hídrico presenta alta variabilidad y una marcada dependencia territorial. Bajo partición aleatoria, *XGBoost* con variables espaciales alcanzó un R^2 de 0,8917; sin embargo, las validaciones temporal y espacial demostraron que ese desempeño corresponde a interpolación dentro de las localidades ya observadas y no a capacidad predictiva generalizable. Un modelo entrenado únicamente con latitud y longitud, sin ninguna variable climática, alcanzó un R^2 de 0,7950, lo que evidencia que las coordenadas operan como identificadores territoriales. En la prueba temporal de 2024 el modelo fue superado por un modelo de referencia basado en la media histórica por localidad y bimestre, y en la validación LOLO solo

4 de las 20 localidades alcanzaron un R^2 positivo. En consecuencia, los resultados no aportaron evidencia suficiente para respaldar la hipótesis planteada bajo las estrategias de validación temporal y espacial. El principal aporte del estudio es metodológico: demuestra que la validación espacial y temporal resulta indispensable para distinguir la interpolación territorial de la capacidad real de generalización, y que los patrones históricos por localidad y bimestre constituyen una línea base difícil de superar. Se recomienda avanzar hacia estudios con mayor resolución espacial y temporal que fortalezcan la gestión del recurso, en concordancia con el Objetivo de Desarrollo Sostenible ODS 6.

Palabras clave: Aprendizaje Automático, Ciencia de Datos, Consumo hídrico, Gestión hídrica, Modelos predictivos, Variables climáticas.

Abstract

This thesis examines water consumption in Bogotá in a context of increasing pressure on water supply systems and climate variability associated with phenomena such as El Niño. The study considers the 2024–2025 water crisis and the differences in consumption patterns across the city's localities, highlighting the need for data-driven analyses.

The main objective of this research was to develop and validate predictive models based on Machine Learning (ML) techniques to estimate the average water consumption by locality and two-month period in the city of Bogotá, through the integration of climatic, temporal, and spatial variables. A quantitative methodology was applied, including descriptive analysis, data cleaning, outlier detection, predictive modeling, and benchmarking against reference models. Deep neural networks and XGBoost were implemented and assessed under three complementary validation strategies: random partition, temporal validation (training on 2020-2023 and testing on 2024), and spatial Leave-One-Locality-Out (LOLO) validation.

The results revealed high variability and strong territorial dependence in water consumption. Under random partitioning, XGBoost with spatial variables reached an R^2 of 0.8917; however, the temporal and spatial validations showed that this performance reflects interpolation within already observed localities rather than generalizable predictive capacity. A model trained solely on latitude and longitude, without any climatic variable, reached an R^2 of 0.7950, indicating that coordinates act as territorial identifiers. In the 2024 temporal test the model was outperformed by a reference model based on the historical mean by locality and two-month period, and under LOLO validation only 4 of the 20 localities achieved a positive R^2. The hypothesis was therefore not supported under the more demanding validation strategies. The study's main contribution is methodological: it demonstrates that spatial and temporal validation is essential to distinguish territorial interpolation from genuine generalization, and that historical patterns by locality and period

constitute a strong baseline. Further research at higher spatial and temporal resolution is recommended to strengthen water resource management, in alignment with Sustainable Development Goal (SDG) 6.

Keywords: Data science, Water consumption, Water management, Machine Learning, Predictive models, Climate variables.

Tabla de Contenido

Lista de Figuras **10**

Lista de Tablas **11**

Introducción **13**

1. Objetivos **16**

1.1 Objetivo General 16

1.2 Objetivos Específicos 16

2. Justificación **17**

3. Marco Teórico **20**

3.1 Variabilidad Climática, Sequía y Estrés Hídrico en Bogotá 21

3.2 Consumo Hídrico Residencial: Factores Climáticos, Temporales, Espaciales y Socioeconómicos 22

3.3 Machine *Learning* Aplicado a la Predicción Hídrica 25

3.4 Modelos Climáticos e Hidrológicos Basados en *Machine Learning* 27

3.5 Preprocesamiento, Selección de Variables y Calidad de Datos 29

3.6 IoT, Smart Water Systems y Detección de Anomalías 30

3.7 Evaluación de Modelos: Métricas, Validación y Explicabilidad 31

3.8 Gerencia de Proyectos en el Desarrollo de Modelos Predictivos 33

3.9 Vacíos Teóricos y Aporte del Estudio 36

4. Hipótesis **37**

5. Variables **38**

5.1 Clasificación de las Variables 38

5.1.1 Variable Objetivo *38*

5.1.2 Variables Predictoras *38*

5.2 Definición Conceptual y Operacional de las Variables 39

6. Metodología **43**

6.1 Enfoque de la Investigación 43

6.2 Alcance de la Investigación 44

6.3 Diseño de la Investigación 45

6.4 Fases de la Investigación 46

6.4.1 Fase 1. Identificación y Selección de Variables Asociadas al Consumo Hídrico 46

6.4.2 Fase 2. Construcción de Modelos Predictivos de Machine Learning 47
6.4.3 Fase 3. Validación y Evaluación del Desempeño Predictivo 49
6.5 Integración de la Gerencia de Proyectos en el Desarrollo de la Investigación. 49
6.5.1 Gestión de los Interesados 50
6.5.2 Gestión de Riesgos, Calidad y Seguimiento 51
6.6 Población y Muestra 53
6.7 Instrumentos de Recolección de Información 53
6.8 Validación de los Instrumentos y Control de Calidad de la Información 54
6.9 Técnicas para el Análisis de la Información 55
6.9.1 Análisis Descriptivo 55
6.9.2 Análisis Exploratorio de Datos 55
6.9.3 Modelado Predictivo 56
6.9.4 Evaluación de Desempeño 56
6.10 Herramientas Computacionales Utilizadas 56
6.11 Consideraciones Éticas 56
7. Trabajo de Campo 58
7.1 Caracterización e Integración de la Información 58
7.2 Preparación y Calidad de los Datos 61
7.2.1 Verificación de la Calidad de los Datos 62
7.2.2 Tratamiento de Valores Faltantes y Valores Atípicos 63
7.2.3 Tratamiento de los registros anómalos de Sumapaz 66
7.2.4 Transformación de Variables 67
7.3 Análisis Exploratorio de los Datos (EDA) 67
7.3.1 Caracterización Estadística de la Muestra (Análisis Descriptivo) 69
7.3.2 Dimensión Comercial y de Consumo 69
7.3.3 Análisis de la dispersión y "desfases" comerciales 70
7.3.3.1 Heterogeneidad entre localidades 70
7.3.3.2 Distribución del consumo promedio 70
7.3.4 Dimensión Climática, Atmosférica y de Suelo (NASA POWER) 71
7.3.5 Análisis de las variables climáticas y su estacionalidad 72
7.3.5.1 Estabilidad Térmica 72
7.3.5.2 El desfase en las Lluvias (PRECTOTCORR_SUM) 72

7.3.5.3 Humedad del Suelo y el factor macro (ONI) ... 72
7.3.6 Análisis detallado de la variable objetivo y métricas de demanda ... 72
7.3.7 Análisis Temporal y Distributivo del Consumo Hídrico en Bogotá (2020-2024) ... 73
7.3.7.1 Periodo de Estabilidad y Fluctuación Estacional (2020 - 2023) ... 74
7.3.7.2 El Impacto de la pandemia (2020) ... 74
7.3.7.3 La regularidad entre 2021 y 2023 ... 74
7.3.7.4 El Comportamiento Atípico de Sumapaz en 2022 ... 75
7.3.7.5 La Anomalía Crítica de 2024 ... 75
7.3.8 Centralización Geográfica de la Demanda en la Crisis ... 76
7.3.9 Análisis de Distribución, Formas y Valores Atípicos (EDA Visual) ... 76
7.3.9.1 Forma y Dispersión de la Variable Objetivo (consumo promedio) ... 76
7.3.9.2 Comportamiento Estructural de la Demanda Global (metros cúbicos) ... 78
7.3.9.3 Comportamiento del consumo frente a la dinámica de usuarios y facturación ... 81
7.3.9.4 Crecimiento Vegetativo de la Base de Usuarios y Capacidad de Medición ... 81
7.3.9.5 Contraste Temporal y Evidencia del Choque Estructural en el Consumo ... 83
7.3.9.6 Periodo de Estabilidad Estacionaria (2020 - Bimestre 3, 2024) ... 84
7.3.9.7 Punto de Inflexión y Ruptura Estructural (Bimestre 4, 2024 en adelante) ... 84
7.3.10 Análisis de Variables Climatológicas y su relación con la Demanda Hídrica ... 84
7.3.10.1 Régimen de Precipitación y Disponibilidad Hídrica ... 85
7.3.10.2 Perfil de Temperatura y Confort Térmico ... 86
7.3.10.3 Dinámica de la Humedad Relativa y Evapotranspiración ... 88
7.3.10.4 Balance Energético mediante Radiación Solar ... 89
7.3.10.5 Dinámica de la Velocidad del Viento ... 90
7.3.10.6 Estado Edáfico: Humedad del Suelo Superficial y Profundo ... 91
7.3.11 Análisis de Distribución Espacial y Caracterización por Localidad ... 92
7.3.11.1 Asimetría entre la Base Operativa y el Consumo Agregado ... 93
7.3.11.2 La Paradoja de Sumapaz: Análisis de la Variable Objetivo ... 95
7.3.12 Análisis de la distribución espacial del consumo hídrico por localidad ... 97
7.3.13 Comportamiento del Casco Urbano Consolidado ... 97
7.3.14 Análisis de Correlaciones y Dependencias Lineales Multi-Variable ... 98
7.3.14.1 El Núcleo de la Demanda Hídrica y la Redundancia Multicolineal Comercial ... 98
7.3.14.2 Ortogonalidad y Aislamiento Estadístico de la Variable Objetivo ... 99

7.3.14.3 Dinámica y Sincronización del Sistema Meteorológico ... 101
7.3.14.4 El Impacto Climático Marginal en el Modelo Lineal ... 101
7.3.15 Implicaciones metodológicas para el Modelo Predictivo del EDA ... 102
7.3.16 Prevención de Fuga de Información (Data Leakage) ... 104
7.4 Identificación de Variables Relevantes ... 105
7.5 Diseño de Modelos Predictivos de *Machine Learning* ... 107
7.5.1 Formulación del Problema Predictivo ... 107
7.5.2 Modelo de Redes Neuronales Profundas ... 108
7.5.2.1 Arquitectura del Modelo ... 109
7.5.2.2 Entrenamiento del Modelo ... 110
7.5.2.3 Resultados Obtenidos ... 110
7.5.3 Modelo Basado en XGBoost ... 112
7.5.3.1 Configuración Inicial del Modelo ... 112
7.5.3.2 Optimización del Modelo ... 113
7.5.3.3 Resultados Obtenidos ... 113
7.5.4 Comparación y Selección del Modelo Predictivo Final ... 116
7.6 Validación y Evaluación del Desempeño Predictivo ... 118
7.6.1 Métricas de Evaluación ... 119
7.6.2 Evaluación del Desempeño de los Modelos ... 120
7.6.3 Validación de la Capacidad de Generalización ... 121
7.7 Síntesis del Trabajo de Campo ... 124
8. Discusión ... 125
8.1 Relación entre Variables y Consumo Hídrico ... 125
8.2 Importancia de los Componentes Espaciales ... 127
8.3 Implicaciones para la Gestión del Recurso Hídrico en Bogotá ... 127
8.4 Alcances y Limitaciones Metodológicas ... 129
8.5 Alcances y Limitaciones Metodológicas ... 131
9. Conclusiones y Trabajo Futuro ... 133
Referencias ... 139
Anexos ... 146
Anexo A: Repositorio Digital de Código, Datos y Resultados Complementarios ... 146

Lista de Figuras

Figura 1. Modelo conceptual para la predicción del consumo hídrico residencial en Bogotá 33

Figura 2. Proceso Cuantitativo. 43

Figura 3. Fuentes de Información. 59

Figura 4. Proceso de Preparación y Calidad de los Datos. 62

Figura 5. Distribución del Volumen de Agua Facturada (Metros Cúbicos). 64

Figura 6. Diagrama de Caja del Volumen de Agua Facturada (Metros Cúbicos). 64

Figura 7. Distribución del Consumo Hídrico Promedio. 65

Figura 8. Diagrama de Caja del Consumo Hídrico Promedio. 65

Figura 9. Distribución Porcentual del Consumo Hídrico Indexado por Localidad en Bogotá (2020–2024). 73

Figura 10. Distribución de frecuencias de la variable objetivo (consumo promedio) e índice de asimetría. 77

Figura 11. Diagrama de caja de la variable objetivo con identificación y contraste de valores atípicos severos. 77

Figura 12. Distribución de frecuencias y curva de densidad del volumen global de consumo hídrico en $m3$. 79

Figura 13. Diagrama de caja del volumen de consumo global en m3 y dispersión intercuartílica. 79

Figura 14. Evolución histórica de la base de usuarios facturados, emisión global de facturas y cobertura de dispositivos de micromedición (2020 - 2024). 81

Figura 15. Trayectoria temporal de la variable objetivo e identificación visual del choque estructural. 83

Figura 16. Evolución bimestral del promedio de precipitación corregida frente al volumen neto de consumo hídrico en metros cúbicos (2020 - 2024). 85

Figura 17. Variación del promedio de temperatura a 2 metros de superficie. 87

Figura 18. Comportamiento del promedio de humedad relativa e impacto en los niveles de consumo agregados. 88

Figura 19. Evolución temporal de los niveles de radiación promedio y su contraste con las variables de control hídrico. 89

Figura 20. Comportamiento de la suma de velocidad del viento y su relación con las fluctuaciones de la demanda. 90

Figura 21. Dinámica comparativa de la humedad del suelo y su correlación estructural con el consumo hídrico. 92

Figura 22. Análisis comparativo por localidad: volumen neto distribuido m^3 93

Figura 23. Análisis comparativo por localidad: densidad de la infraestructura comercial. 94

Figura 24. Distribución geográfica de la variable objetivo y contraste del consumo promedio indexado por localidad 95

Figura 25. Evolución de la distribución espacial del consumo hídrico por localidad en Bogotá (Periodo 2020-2024). 96

Figura 26. Matriz de correlación cruzada 98

Figura 27. Matriz de correlación interna para variables climatológicas. 100

Lista de Tablas

Tabla 1. Ejes conceptuales del marco teórico y referencias principales 24

Tabla 2. Modelos de ML relevantes para la predicción del consumo hídrico 28

Tabla 3. Métricas de evaluación de modelos predictivos 32

Tabla 4. Articulación conceptual de la Gerencia de Proyectos y el desarrollo de modelos predictivos. 35

Tabla 5. Definición conceptual y operacional de las variables. 39

Tabla 6. Articulación de la Gerencia de Proyectos con las fases de investigación. 50

Tabla 7. Registro de Interesados del proyecto. 51

Tabla 8. Riesgos metodológicos y medidas de tratamiento. 52

Tabla 9. Dimensiones y variables integradas en la matriz de datos. 54

Tabla 10. Fuentes de Información utilizadas en la Investigación. 59

Tabla 11. Dimensiones de las Variables Incorporadas. 60

Tabla 12. Actividades Desarrolladas Durante el Análisis Exploratorio de los Datos. 68

Tabla 13. Métricas de tendencia central y dispersión para el consumo y facturación de agua. 69

Tabla 14. Caracterización estadística de los parámetros climáticos y de variabilidad atmosférica. 71

Tabla 15. Principales Actividades del Proceso de Preparación de los Datos. 104

Tabla 16. Categorías de Variables Utilizadas para el Entrenamiento de los Modelos Predictivos 106

Tabla 17. Arquitectura del Modelo de Red Neuronal Profunda 109

Tabla 18. Configuración del Entrenamiento del Modelo de Red Neuronal Profunda 110

Tabla 19. Desempeño del Modelo de Red Neuronal Profunda 111

Tabla 20. Evaluación y comparación de métricas de desempeño para las iteraciones del modelo predictivo 111

Tabla 21. Configuración Inicial del Modelo XGBoost 113

Tabla 22. Evolución del Desempeño del Modelo XGBoost 114

Tabla 23. Variables con Mayor Importancia en el Modelo XGBoost 115

Tabla 24. Comparación de la importancia de las variables según tres métodos 116

Tabla 25. Comparación del Desempeño de los Modelos Predictivos Desarrollados 117

Tabla 26. Métricas Utilizadas para la Evaluación del Desempeño Predictivo 119

Tabla 27. Resultados Comparativos de los Modelos Predictivos 120

Tabla 28. Comparación entre las Estrategias de Validación 122

Tabla 29. Comparación de XGBoost y los modelos de referencia en la evaluación temporal 2024.. 123

Tabla 30. Influencia de los grupos de variables sobre la predicción del consumo hídrico..... 126

Tabla 31. Implicaciones de los Modelos Predictivos para la Gestión del Recurso Hídrico..... 128

Tabla 32. Principales Limitaciones Metodológicas de la Investigación y sus Implicaciones..... 130

Tabla 33. Hoja de ruta propuesta para Implementación de modelos predictivos del consumo hídrico.

..... 132

Introducción

El acceso y la gestión sostenible del agua potable se han convertido en uno de los principales desafíos del siglo XXI, especialmente en contextos urbanos donde la presión sobre los recursos hídricos aumenta de manera constante. Factores como el cambio climático, el crecimiento poblacional, la expansión urbana y las variaciones en los patrones de consumo han incrementado la demanda de agua y han puesto en evidencia la necesidad de fortalecer los sistemas de planificación y gestión del recurso hídrico (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2021; *World Wildlife Fund* [WWF], 2024). En este contexto, la ciudad de Bogotá enfrenta retos asociados a la sostenibilidad del abastecimiento de agua, la eficiencia en su distribución y la capacidad de anticipar cambios en los patrones de consumo. Ante esta situación, surge la necesidad de incorporar herramientas tecnológicas que permitan analizar grandes volúmenes de información y generar predicciones confiables para apoyar la toma de decisiones. Por ello, la presente investigación se enmarca en el campo de la gestión de los recursos hídricos y la ciencia de datos aplicada, proponiendo el uso de técnicas de *Machine Learning* (ML) para la predicción del consumo hídrico como una estrategia orientada a fortalecer la planificación, optimizar la gestión del recurso y contribuir al desarrollo sostenible.

El análisis del consumo hídrico en Bogotá adquiere una importancia estratégica. A pesar de que la ciudad cuenta con fuentes de abastecimiento provenientes de ecosistemas de alta montaña como los páramos de Chingaza y Sumapaz, el sistema enfrenta presiones constantes derivadas de la variabilidad climática, la urbanización acelerada y el aumento sostenido de la demanda (Instituto de Hidrología, Meteorología y Estudios Ambientales [IDEAM], 2022; Martínez et al., 2024). Estas condiciones se han hecho más evidentes en eventos recientes, como la crisis hídrica de 2024-2025, en la cual los niveles de los embalses alcanzaron valores críticos, generando medidas de racionamiento y alertas sobre la seguridad hídrica de la ciudad (Alcaldía

Mayor de Bogotá, 2025; Pérez et al., 2024). Sin embargo, el problema no se limita únicamente a la disponibilidad del recurso, sino que también está relacionado con la forma en que este se consume y se gestiona a nivel urbano.

Diversos estudios e informes institucionales han señalado que en Bogotá existen importantes ineficiencias en el sistema de acueducto, reflejadas en altos niveles de agua no contabilizada, así como en patrones de consumo diferenciados entre localidades y grupos socioeconómicos (Empresa de Acueducto y Alcantarillado de Bogotá [EAAB], 2022; Superintendencia de Servicios Públicos Domiciliarios [Superservicios], 2024). Estas dinámicas evidencian que el consumo hídrico no es homogéneo en la ciudad, sino que responde a múltiples factores, como el tamaño del hogar, el estrato socioeconómico, las condiciones de la infraestructura y las variables climáticas (House-Peters y Chang, 2011; Hussien et al., 2016). A pesar de la disponibilidad de información y de los esfuerzos institucionales por mejorar la eficiencia del servicio, persisten limitaciones importantes en la capacidad de anticipar comportamientos de consumo y orientar acciones preventivas de manera oportuna.

En este contexto, el uso de herramientas de analítica de datos y técnicas de *Machine Learning* representa una oportunidad para comprender mejor el comportamiento del consumo hídrico y apoyar la gestión del recurso desde un enfoque más integral. A nivel internacional, distintos estudios han demostrado que estos modelos permiten identificar patrones complejos, integrar múltiples variables y mejorar la capacidad de predicción en sistemas urbanos (Donkor et al., 2012; Candelieri, 2017). Sin embargo, en el caso de Bogotá y, en general, en el contexto colombiano, aún son limitadas las investigaciones que integren análisis descriptivos, variables climáticas y enfoques espaciales en el estudio del consumo de agua a nivel domiciliario. Esta brecha plantea la necesidad de desarrollar investigaciones que no solo describan el fenómeno, sino que también incorporen modelos predictivos y herramientas analíticas que permitan entenderlo de manera más profunda.

A partir de lo anterior, surge la siguiente pregunta de investigación: ¿cómo pueden aplicarse técnicas de *Machine Learning* para predecir el consumo hídrico en la ciudad de Bogotá, a partir de variables climáticas, temporales y espaciales, y cuál es su capacidad de generalización bajo diferentes estrategias de validación? Esta pregunta orienta el desarrollo del presente estudio, cuyo propósito es aportar evidencia empírica y metodológica que contribuya a una mejor comprensión del consumo de agua en la ciudad y al diseño de estrategias más eficientes y sostenibles.

El documento se organiza de la siguiente manera: en primer lugar, se presentan los objetivos, la justificación y el marco teórico que fundamenta la investigación, incluyendo el estado del arte sobre aplicaciones de *Machine Learning* a la predicción de la demanda hídrica. En seguida se exponen la hipótesis, las variables y el diseño metodológico. A continuación, el capítulo de trabajo de campo describe la integración y preparación de los datos, el análisis exploratorio, la identificación de variables relevantes, el diseño de los modelos predictivos y su evaluación mediante tres estrategias de validación. Finalmente, se presentan la discusión de los resultados, las conclusiones organizadas por objetivo, las líneas de trabajo futuro, las referencias bibliográficas y los anexos.

1. Objetivos

1.1 Objetivo General

Desarrollar modelos predictivos basados en técnicas de *Machine Learning* para estimar el consumo promedio de agua por localidad y periodo bimestral en la ciudad Bogotá, mediante la integración de variables climáticas, temporales y espaciales.

1.2 Objetivos Específicos

- Identificar las variables climáticas, temporales y espaciales asociadas al consumo de agua en la ciudad de Bogotá, mediante el análisis de las diferentes fuentes de información sobre el consumo hídrico, con el fin de caracterizar patrones de uso y comportamiento del recurso en la ciudad.
- Comparar el desempeño de modelos predictivos basados en técnicas de *Machine Learning*, considerando modelos de referencia y el control explícito de la fuga de información, a partir del procesamiento de datos relacionados con el consumo hídrico en la ciudad de Bogotá, con el propósito de identificar la alternativa más adecuada para comprender las dinámicas de consumo de agua en la ciudad.
- Validar el desempeño de los modelos predictivos mediante métricas de regresión como MAE, RMSE y R^2, complementadas con WAPE y mediana del error absoluto, aplicando esquemas de validación aleatoria, temporal y espacial, con el fin de determinar la capacidad de los modelos para representar el comportamiento del consumo hídrico y su capacidad de generalización fuera de muestra.

2. Justificación

El consumo hídrico en la ciudad de Bogotá se ha convertido en un tema de alta relevancia debido a la creciente presión sobre las fuentes de abastecimiento, particularmente los sistemas de Chingaza y Sumapaz. Eventos recientes como la crisis hídrica de 2024–2025 evidencian la necesidad de comprender mejor los patrones de consumo y los factores que los determinan, en un contexto marcado por la variabilidad climática y el crecimiento urbano. Desde el punto de vista económico y de gestión del recurso, la optimización del uso del agua representa un desafío estratégico tanto para entidades públicas como para la sostenibilidad de la ciudad.

En este contexto, el presente trabajo resulta conveniente y pertinente, ya que responde a la necesidad de incorporar herramientas analíticas avanzadas en la toma de decisiones. El uso de técnicas de ML permite analizar grandes volúmenes de datos y detectar patrones complejos que no son fácilmente identificables mediante métodos tradicionales. Esto representa un valor agregado frente a publicaciones previas, las cuales han sido limitados en el uso de datos a gran escala y en la integración de variables climáticas, demográficas y espaciales en el análisis del consumo hídrico en la ciudad de Bogotá.

Desde el punto de vista teórico, la investigación contribuye al desarrollo de modelos predictivos aplicados al contexto urbano, integrando enfoques de análisis de datos, aprendizaje automático y análisis espacial. A nivel práctico, los resultados pueden servir como insumo para entidades encargadas de la gestión del recurso hídrico, permitiendo diseñar estrategias más eficientes orientadas al uso sostenible del agua, en concordancia con el Objetivo de Desarrollo Sostenible 6, agua limpia y saneamiento.

Desde la perspectiva de la Gerencia de Proyectos, la investigación adquiere relevancia al integrar criterios de planificación, gestión de riesgos y aseguramiento de la calidad en el desarrollo de modelos predictivos. Esta articulación permite abordar la incertidumbre asociada con los datos y el desempeño de los modelos, así como establecer criterios de evaluación que

contribuyan a determinar su alcance y pertinencia antes de considerar una eventual implementación institucional.

En términos de impacto social, el estudio aporta a la comprensión de las dinámicas de consumo en diferentes localidades de la ciudad, evidenciando desigualdades y comportamientos diferenciados que pueden ser considerados en políticas públicas. Asimismo, el desarrollo y evaluación de modelos predictivos permite analizar hasta qué punto los patrones históricos pueden utilizarse para estimar el consumo futuro e identificar las limitaciones de generalización frente a cambios temporales o territoriales, lo que resulta clave en contextos de escasez y planificación urbana.

En cuanto a la viabilidad, el proyecto cuenta con acceso a bases de datos de consumo hídrico, variables climáticas y demográficas, así como con herramientas computacionales adecuadas para el procesamiento y análisis de la información. El tiempo y los recursos disponibles permiten desarrollar el estudio dentro del alcance planteado, incluyendo análisis descriptivo, modelado predictivo y validación de resultados.

Respecto a la dimensión de las variables, es preciso señalar que durante las etapas iniciales de recolección y exploración de datos se compilaron variables sociodemográficas (como población proyectada, número de hogares y densidad poblacional por localidad). No obstante, estas variables no fueron incorporadas como predictoras directas en las matrices de entrenamiento de los modelos de Machine Learning, utilizándose únicamente para la caracterización del contexto. La razón técnica de esta decisión responde a la incompatibilidad de resoluciones: las variables sociodemográficas se encuentran a escala anual, lo que no ofrecía variabilidad bimensual suficiente para acoplarse dinámicamente con la unidad de análisis del estudio (localidad–año–bimestre). Por consiguiente, las dimensiones empleadas en el modelado predictivo corresponden a variables climáticas, temporales y espaciales, en estricta concordancia con los objetivos planteados.

Finalmente, este trabajo se enmarca en el campo de la analítica de datos y la ingeniería aplicada, dentro de un grupo de investigación orientado al uso de tecnologías para la toma de decisiones, y específicamente en la línea de ML aplicado a problemas urbanos y sostenibilidad. De esta manera, la investigación articula el conocimiento técnico con problemáticas reales de la ciudad, generando un aporte relevante tanto académico como práctico.

3. Marco Teórico

La gestión del recurso hídrico constituye uno de los principales desafíos de las ciudades contemporáneas debido al crecimiento poblacional, la expansión urbana, la presión sobre las fuentes de abastecimiento y los efectos de la variabilidad climática. La UNESCO (2021) plantea que el agua debe entenderse no solo como recurso natural, sino como un bien estratégico para la sostenibilidad social, económica y ambiental. En esta misma línea, Forero et al. (2019) señalan que Colombia enfrenta importantes retos para avanzar en el cumplimiento del Objetivo de Desarrollo Sostenible 6, especialmente en relación con el acceso, la eficiencia en el uso del agua, la calidad del recurso y la gestión integrada.

A nivel nacional, el IDEAM (2022) advierte que la presión sobre el agua en Colombia se ha intensificado por el crecimiento urbano, la variabilidad climática y la concentración de la demanda en grandes centros poblados. Esta situación evidencia la necesidad de incorporar herramientas analíticas que permitan anticipar escenarios de consumo, mejorar la planificación hídrica y fortalecer la capacidad institucional de respuesta. En este sentido, el Ministerio de Ambiente y Desarrollo Sostenible (Minambiente, 2019) reconoce que la estimación de módulos de consumo de agua debe considerar factores climáticos, territoriales, sociales y económicos, lo cual confirma que la demanda hídrica es un fenómeno multicausal.

En Bogotá, la gestión del agua adquiere especial relevancia por la dependencia de sistemas de abastecimiento estratégicos como Chingaza, Sumapaz y Tibitoc. La EAAB (2022) ha señalado la importancia de fortalecer la infraestructura, optimizar la operación del sistema y reducir pérdidas en la red. Sin embargo, Ortiz (2025) indica que cerca del 37 % del agua producida para Bogotá se pierde por fugas, fallas técnicas, errores comerciales o conexiones ilegales. A su vez, Diago (2025) reporta que en 2024 se registraron más de 25.000 daños en redes de agua, con pérdidas superiores a 35 millones de m^3. Estas cifras muestran que el

problema hídrico urbano no se limita al consumo del usuario final, sino que involucra infraestructura, pérdidas no contabilizadas, gestión institucional y desigualdades territoriales.

Izquierdo et al. (2023) analizan la integración de infraestructura natural dentro del sistema de abastecimiento de agua de Bogotá, destacando el papel de ecosistemas estratégicos como páramos, bosques y áreas de recarga hídrica en la regulación del recurso. Los autores señalan que la conservación de estos ecosistemas contribuye a fortalecer la seguridad hídrica de la ciudad y aumentar la resiliencia frente a eventos de variabilidad climática. Este enfoque complementa la visión tradicional basada exclusivamente en infraestructura gris y resalta la importancia de una gestión integral del recurso hídrico.

En consecuencia, el estudio del consumo hídrico en Bogotá debe abordarse desde una perspectiva integral, donde el consumo se entienda como resultado de la interacción entre condiciones climáticas, comportamiento social, estructura urbana, infraestructura hidráulica y capacidad de gestión pública.

3.1 Variabilidad Climática, Sequía y Estrés Hídrico en Bogotá

La disponibilidad de agua depende de la interacción entre precipitación, temperatura, evapotranspiración, almacenamiento en embalses y regulación ecosistémica. En contextos urbanos, la variabilidad climática puede alterar tanto la oferta como la demanda de agua. Mathivha et al. (2024) evidencian que los eventos de sequía asociados a El Niño reducen significativamente los niveles de embalses y aumentan la vulnerabilidad de los sistemas de abastecimiento. En Colombia, la Comisión Colombiana del Océano (CCO, 2025) destaca que las fases de el Niño-Oscilación del Sur (ENSO) influyen en los patrones de precipitación y temperatura, afectando la disponibilidad hídrica regional.

Bogotá experimentó una crisis hídrica entre 2024 y 2025 asociada a bajos niveles de embalses y reducción de lluvias. Ortega et al. (2026) analizan esta crisis desde una perspectiva

multidimensional, integrando factores climáticos, institucionales y sociales. Los autores resaltan que la ciudad presenta una alta dependencia del sistema Chingaza y que los eventos del Fenómeno del Niño han incrementado el estrés hídrico urbano. Esta referencia es central para el presente estudio porque sitúa el problema de investigación en un contexto reciente de vulnerabilidad climática, racionamiento y necesidad de gestión predictiva.

Los ecosistemas de páramo cumplen una función crítica en la regulación hídrica. Martínez et al. (2024) definen los páramos como una forma de “infraestructura natural”, debido a su capacidad de almacenar y liberar agua de manera gradual. Zapata et al. (2021) complementan esta visión al mostrar la importancia de las lagunas de páramo en Colombia para la regulación hidrológica. A su vez, Gil y Tobón (2016), mediante modelación hidrológica del páramo de Chingaza, evidencian que estos ecosistemas son altamente sensibles a cambios en precipitación, humedad y cobertura vegetal.

Desde una perspectiva crítica, estos estudios coinciden en que la crisis hídrica no puede interpretarse únicamente como resultado de una disminución temporal de lluvias. También responde a una combinación de dependencia estructural de pocas fuentes, pérdidas en redes, crecimiento urbano, deficiencias de monitoreo y baja capacidad de anticipación. Por tanto, los modelos predictivos basados en ML pueden aportar valor al permitir una lectura anticipada de patrones de consumo bajo diferentes condiciones climáticas.

3.2 Consumo Hídrico Residencial: Factores Climáticos, Temporales, Espaciales y Socioeconómicos

El consumo hídrico residencial es un fenómeno dinámico y multifactorial. Jiménez et al. (2017), en un estudio sobre Manizales, Colombia, encontraron que la variabilidad climática afecta la demanda residencial de agua, especialmente durante eventos El Niño y períodos de baja precipitación. Este estudio resulta fundamental porque demuestra, en un contexto urbano

colombiano, que las variables climáticas no solo afectan la oferta hídrica, sino también el comportamiento de consumo en la ciudad.

Brentan et al. (2017) señalan que las variables predictoras del consumo de agua incluyen temperatura, precipitación, día de la semana, estacionalidad, características urbanas y patrones históricos de demanda. De manera similar, Lee y Derrible (2020) muestran que los modelos de aprendizaje estadístico pueden mejorar la predicción de la demanda residencial al integrar variables climáticas, temporales y socioeconómicas. Xenochristou et al. (2020) agregan que la precisión de los modelos depende de la escala espacial de análisis, dado que los patrones de consumo pueden variar significativamente entre zonas, barrios o sectores urbanos.

La dimensión espacial influye tanto en el comportamiento del consumo hídrico como en la capacidad de generalización de los modelos predictivos. Por ello, es importante evaluar su desempeño en contextos geográficos diferentes a los observados durante el entrenamiento (Xenochristou et al., 2020), especialmente en ciudades como Bogotá, donde existen diferencias significativas entre localidades.

En América Latina, Nunes et al. (2021) aplicaron técnicas de ML para modelar la demanda urbana de agua en Fortaleza, Brasil, concluyendo que las variables meteorológicas y temporales mejoran el desempeño predictivo. Sanjur et al. (2025) también evaluaron el uso de aprendizaje automático para predicción del consumo de agua potable en Panamá Metropolitana, lo cual muestra una tendencia regional hacia el uso de modelos inteligentes para la gestión hídrica urbana.

Para Bogotá, la dimensión espacial es especialmente importante. Diago (2025) muestra diferencias notables en pérdidas de agua por localidad, mientras que Ortega et al. (2026) evidencian desigualdades territoriales en vulnerabilidad y adaptación frente al racionamiento. Esto justifica la inclusión de variables espaciales como localidad, estrato, densidad poblacional o ubicación geográfica en modelos predictivos de consumo hídrico residencial.

Tabla 1.

Ejes conceptuales del marco teórico y referencias principales

Eje Conceptual	Referencias Principales	Aporte al Estudio
Gestión hídrica y sostenibilidad	UNESCO (2021), IDEAM (2022), Forero et al. (2019), MinAmbiente (2019)	Fundamentan la importancia social, ambiental e institucional del uso eficiente del agua
Crisis hídrica en Bogotá	EAAB (2022), Ortiz (2025), Diago (2025), Ortega et al. (2026)	Contextualizan el problema local de abastecimiento, pérdidas y racionamiento
Variabilidad climática	CCO (2025), Mathivha et al. (2024), Pérez et al. (2024), Jiménez et al. (2017)	Sustentan la relación entre clima, sequía, ENSO y demanda hídrica
Ecosistemas reguladores	Martínez et al. (2024), Zapata et al. (2021), Gil y Tobón (2016), Izquierdo et al. (2023)	Explican el papel de páramos, embalses e infraestructura natural
ML hídrico	Aminu et al. (2025), Canquil et al. (2026), Zhu et al. (2022), Niazkar et al. (2024)	Sustentan el uso de modelos inteligentes en recursos hídricos
Predicción de consumo	Lee y Derrible (2020), Xenochristou et al. (2020), Nunes et al. (2021), Görenekli y Gülbağ (2024)	Aportan antecedentes directos de *forecasting* de demanda hídrica
Validación y métricas	Palma (2022), Costa et al. (2024), Triana et al. (2023), Rueda et al. (2025)	Fundamentan métricas, validación y evaluación de modelos

Fuente. Elaboración propia a partir de la Revisión Bibliográfica.

La Tabla 1 resume los ejes conceptuales y las referencias principales que fundamentan el presente marco teórico. La literatura revisada evidencia que el consumo hídrico residencial está influenciado por múltiples factores de naturaleza climática, temporal, espacial y sociodemográfica. Variables que han mostrado capacidad para explicar variaciones en la demanda de agua en distintos contextos urbanos (Brentan et al., 2017; Nunes et al., 2021). En consecuencia, la selección de variables incorporadas en la presente investigación se fundamenta tanto en la evidencia reportada por la literatura como en la disponibilidad de información para la ciudad de Bogotá, permitiendo integrar múltiples dimensiones asociadas al comportamiento del consumo hídrico.

3.3 Machine *Learning* Aplicado a la Predicción Hídrica

El ML se ha consolidado como una herramienta relevante para modelar fenómenos hídricos complejos. Zhu et al. (2022) destacan que los algoritmos de aprendizaje automático permiten capturar relaciones no lineales entre variables ambientales, climáticas e hidrológicas, superando en muchos casos las limitaciones de modelos puramente estadísticos o empíricos.

Aminu et al. (2025) revisan una década de implementación de modelos inteligentes para demanda hídrica y concluyen que los métodos de *Ensemble Learning*, redes neuronales, *Random Forest*, *XGBoost* y *Support Vector Machines* son útiles para problemas de predicción de consumo y gestión del agua. Canquil et al. (2026), en una revisión sistemática sobre predicción del consumo doméstico, coinciden en que los modelos basados en árboles y *boosting* presentan resultados robustos, especialmente cuando se incorporan variables climáticas y temporales.

Adamowski y Karapataki (2010) compararon modelos de regresión multivariada y redes neuronales artificiales para la predicción de la demanda máxima urbana de agua, encontrando que las redes neuronales presentaron una mayor capacidad para capturar relaciones no lineales

entre variables meteorológicas y consumo hídrico. Este estudio constituye uno de los antecedentes más relevantes en la aplicación de técnicas de inteligencia artificial para la predicción de la demanda de agua.

Random Forest, propuesto por Breiman (2001), es un método de ensamble basado en múltiples árboles de decisión. Su principal ventaja radica en la reducción de varianza, robustez frente al ruido y capacidad para estimar importancia de variables. Esto lo convierte en un modelo adecuado para problemas donde se busca no solo predecir, sino también identificar factores relevantes del consumo.

Por su parte, *XGBoost* ha mostrado alto desempeño en problemas de ingeniería de recursos hídricos. Niazkar et al. (2024) evidencian que *XGBoost* ha sido aplicado en predicción de caudales, calidad del agua, sequías, niveles de agua, demanda hídrica y gestión de recursos. Wang et al. (2025) aplicaron *XGBoost* a la gestión de aguas subterráneas agrícolas e incorporaron *SHapley Additive exPlanations* (SHAP) para interpretar la importancia de variables climáticas y socioeconómicas. Este enfoque es relevante para el presente estudio porque permite combinar precisión predictiva con explicabilidad del modelo.

Görenekli y Gülbağ (2024), en un estudio sobre predicción del consumo de agua en Kocaeli, Turquía, compararon técnicas como *Random Forest* (RF), *Support Vector Machines* (SVM), Gradient *Boosting Machines* (GBM) *y Artificial Neural Networks* (ANN). Estas últimas, inspiradas en la estructura y funcionamiento del sistema nervioso biológico, corresponden a modelos computacionales conformados por capas de nodos interconectados (neuronas) que procesan información para aprender patrones complejos y relaciones no lineales en los datos.

En su investigación, los autores encontraron que los modelos de Boosting ofrecieron los mayores niveles de precisión en la estimación del consumo. Este antecedente es especialmente pertinente porque se enfoca directamente en la demanda hídrica y demuestra la utilidad de los esquemas comparativos para seleccionar el algoritmo más adecuado., encontrando que los modelos de *Boosting* ofrecieron altos niveles de precisión. Este antecedente es especialmente

pertinente porque se enfoca directamente en consumo de agua y demuestra la utilidad de modelos comparativos para seleccionar el algoritmo más adecuado.

3.4 Modelos Climáticos e Hidrológicos Basados en *Machine Learning*

Aunque el presente estudio se centra en el consumo hídrico de la ciudad de Bogotá, las investigaciones sobre evapotranspiración, precipitación y calidad del agua aportan fundamentos metodológicos valiosos. Grecco et al. (2025) utilizaron *XGBoost* para predicción diaria de precipitación, incorporando variables como temperatura, humedad, radiación solar, presión atmosférica y viento. Este estudio es relevante porque demuestra la utilidad de *XGBoost* en predicción climática con series temporales.

En esta misma línea, la evapotranspiración, entendida como la transferencia conjunta de agua a la atmósfera por evaporación del suelo y transpiración vegetal, resulta clave para entender la demanda hídrica regional. Al respecto, Granata (2019) evaluó el uso de algoritmos como *Support Vector Regression* (SVR) y *Random Forest* para estimar este proceso hidrológico, evidenciando cómo el modelado de la demanda atmosférica mediante técnicas de *Machine Learning* permite capturar patrones climáticos complejos a partir de datos meteorológicos estándar.

Ayus et al. (2023) evaluaron técnicas de ML y *Deep Learning* para la predicción de niveles de agua, comparando modelos como *RF*, *XGBoost*, *Recurrent Neural Network* (RNN) y *Bidirectional Long Short-Term Memory* (Bi-LSTM). Los resultados mostraron que *XGBoost* alcanzó el mejor desempeño predictivo, evidenciando la capacidad de los algoritmos basados en árboles para modelar fenómenos hidrológicos complejos. Aunque el estudio se enfoca en niveles de agua y no en consumo hídrico, sus resultados respaldan el uso de modelos avanzados de aprendizaje automático en la gestión de recursos hídricos.

Rueda et al. (2025) evaluaron modelos de ML para estimar evapotranspiración en el Valle del Cauca, Colombia, utilizando regresión lineal múltiple, regresión polinómica y redes neuronales. Los autores encontraron que combinaciones con radiación solar y velocidad del viento alcanzaron valores de *R-squared* (R^2) cercanos a 0.99, evidenciando la importancia de la selección adecuada de variables climáticas.

Triana et al. (2023) desarrollaron modelos de ML para estimar evapotranspiración en el suroccidente colombiano bajo condiciones de información escasa. Compararon la técnica estadística *Multiple Linear Regression* (MLR), *Artificial Neural Networks* (ANN) y el modelo estadístico *AutoRegressive Integrated Moving Average* (ARIMA), encontrando que ARIMA-M3 obtuvo el mejor desempeño, aunque las ANN fueron útiles en contextos donde no se dispone de la variable objetivo.

Estos estudios no predicen consumo residencial, pero son metodológicamente pertinentes porque trabajan con variables climáticas, series temporales, información escasa, métricas de desempeño y modelos supervisados. Además, se desarrollan en Colombia, lo que fortalece su valor contextual para la monografía. A partir de la literatura revisada, la Tabla 2 resume las principales características, ventajas y limitaciones de los modelos de ML considerados en esta investigación.

Tabla 2.

Modelos de ML *relevantes para la predicción del consumo hídrico*

Modelo	Ventajas	Limitaciones	Referencias
Regresión lineal múltiple	Alta interpretabilidad; útil como línea base	Limitada para relaciones no lineales	Triana et al. (2023), Rueda et al. (2025)
Random Forest	Robusto, maneja no linealidad, estima importancia de variables	Menor interpretabilidad que modelos lineales	Breiman (2001), Görenekli y Gülbağ (2024)

Gradient Boosting / XGBoost	Alta precisión, manejo de relaciones complejas, regularización	Requiere ajuste de hiperparámetros	Niazkar et al. (2024), Wang et al. (2025), Grecco et al. (2025)
ANN	Captura patrones complejos no lineales	Mayor riesgo de sobreajuste y menor explicabilidad	Adamowski y Karapataki (2010), Triana et al. (2023), Rueda et al. (2025)
ARIMA	Adecuado para series temporales	Limitado ante variables externas complejas	Triana et al. (2023), Ayus et al. (2023)
SVM	Buen desempeño en muestras pequeñas o medianas	Sensible a kernel y parámetros	Zhu et al. (2022), Görenekli y Gülbağ (2024)

3.5 Preprocesamiento, Selección de Variables y Calidad de Datos

La calidad de los datos es un elemento decisivo en cualquier modelo predictivo. Costa et al. (2024), en un estudio sobre calidad del agua mediante ML, destacan que el preprocesamiento, la validación cruzada y la selección de variables pueden ser tan importantes como el algoritmo seleccionado. Esta idea coincide con Palma (2022), quien advierte que una alta bondad de ajuste (*Goodness-of-Fit*) no necesariamente implica buena capacidad predictiva si el modelo no ha sido validado adecuadamente.

Además de la calidad de los datos, la literatura sobre aprendizaje automático advierte la necesidad de controlar posibles fenómenos de fuga de información (*Data Leakage*). Este problema ocurre cuando determinadas variables contienen información estrechamente relacionada con la variable objetivo o derivada de ella, generando resultados artificialmente optimistas durante el entrenamiento y la validación del modelo. La identificación y exclusión de

este tipo de variables constituye una práctica fundamental para garantizar la validez y capacidad de generalización de los modelos predictivos.

Triana et al. (2023) emplearon técnicas como el coeficiente de correlación de Spearman, *Variance Inflation Factor* (VIF) y *Principal Component Analysis* (PCA) para seleccionar variables y reducir multicolinealidad. Rueda et al. (2025), por su parte, evaluaron 13 combinaciones de variables climáticas para identificar aquellas con mayor capacidad predictiva, destacando la importancia de radiación solar y velocidad del viento.

En el contexto de Bogotá, el preprocesamiento adquiere especial relevancia porque los datos pueden presentar valores faltantes, cambios por racionamiento, anomalías por fugas, variaciones estacionales y diferencias territoriales. Por ello, la metodología debe contemplar limpieza de datos, detección de valores atípicos, imputación, normalización, codificación de variables categóricas y validación temporal.

3.6 IoT, Smart Water Systems y Detección de Anomalías

La gestión hídrica moderna se orienta hacia sistemas inteligentes capaces de monitorear, predecir y optimizar el consumo en tiempo real. Herrera y Romero (2025) desarrollaron un sistema basado en Internet de las Cosas (IoT) para el monitoreo del consumo de agua en Bogotá, evidenciando que el acceso a información en tiempo real puede modificar hábitos de consumo y reducir desperdicios. Manfrini (2025) también propone una gestión inteligente del agua en Bogotá mediante Inteligencia Artificial (IA) e IoT, articulando sensores, monitoreo y analítica de datos.

Kanyama et al. (2025) aplicaron modelos de *Ensemble Learning* para detección de anomalías en sistemas de medición inteligente, alcanzando altos niveles de precisión. Aunque su enfoque es la detección de anomalías y no la predicción de demanda, el estudio resulta pertinente porque el consumo hídrico puede verse afectado por fugas, daños en redes o comportamientos atípicos.

Acevedo (2023) revisa métodos de detección y predicción de fugas en sistemas de distribución mediante ML, mostrando que ANN*, Random Forest,* SVM y *XGBoost* son ampliamente utilizados en redes urbanas de agua. Este antecedente conecta directamente con el problema de pérdidas reportado en Bogotá por Ortiz (2025) y Diago (2025).

Aunque estos estudios se enfocan principalmente en monitoreo y detección de anomalías, muestran que las tecnologías inteligentes pueden proporcionar información valiosa para el desarrollo de modelos predictivos orientados a la gestión del recurso hídrico.

3.7 Evaluación de Modelos: Métricas, Validación y Explicabilidad

La evaluación rigurosa de los modelos predictivos requiere utilizar múltiples métricas. Palma (2022) advierte que el coeficiente de determinación (R^2) no debe emplearse como único indicador de desempeño, ya que puede aumentar artificialmente al incorporar más variables o no reflejar adecuadamente el error predictivo. Por ello, se recomienda complementar R^2 con Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE), Error Porcentual Absoluto Medio (MAPE), Eficiencia de *Nash-Sutcliffe* (NSE) y validación cruzada (*Cross-Validation*).

Costa et al. (2024), Grecco et al. (2025), Triana et al. (2023) y Rueda et al. (2025) coinciden en utilizar métricas múltiples para comparar modelos. Rueda et al. (2025) utilizaron El coeficiente de correlación de Pearson (R), R^2 y RMSE para evaluar modelos de evapotranspiración. Triana et al. (2023) usaron RMSE, MAE, *Mean Bias Error* (MBE), NSE y R, lo cual ofrece una evaluación más integral del error, sesgo y eficiencia del modelo. La evaluación de modelos es un aspecto fundamental en el análisis del consumo de agua. En la Tabla 3 se presentan las principales métricas utilizadas en la literatura.

Tabla 3.

Métricas de evaluación de modelos predictivos

Métrica	Descripción
MAE - Mean Absolute Error	Error Absoluto Medio
RMSE - Root Mean Square Error	Raíz del Error Cuadrático Medio
R² - R-squared	Representa la Proporción de la Varianza de la Variable Dependiente
MAPE - Mean Absolute Percentage	Error Porcentual Absoluto Medio.

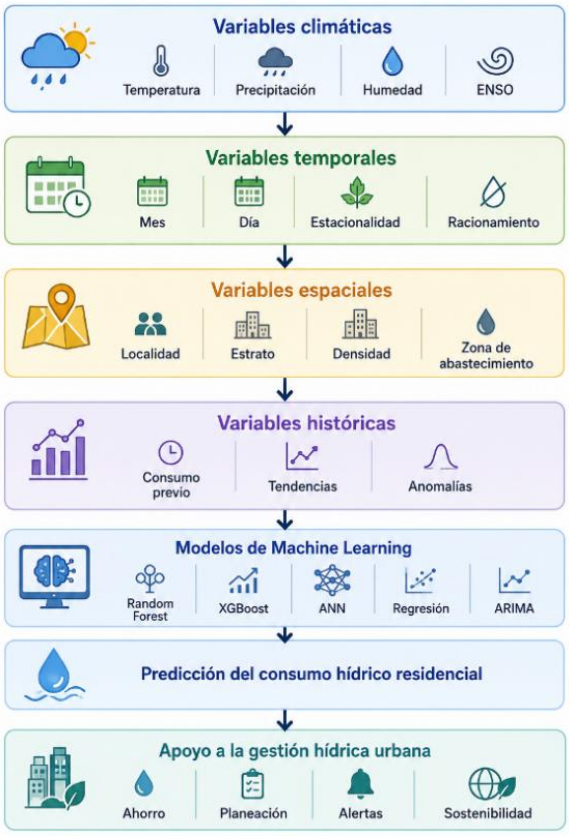
Fuente. Adaptado de Palma (2022); Kanyama et al. (2025).

El coeficiente de determinación (R^2) es ampliamente utilizado; sin embargo, presenta limitaciones importantes. Palma (2022) advierte que este indicador no debe utilizarse de manera aislada, ya que no refleja necesariamente la capacidad predictiva del modelo.

Además de la precisión, la explicabilidad es cada vez más relevante. Wang et al. (2025) emplearon SHAP para interpretar modelos *XGBoost* aplicados a gestión hídrica. Esta perspectiva es importante para el presente estudio porque no basta con predecir el consumo: también es necesario identificar qué variables explican su comportamiento, por ejemplo, precipitación, temperatura, localidad, mes, estrato o patrones históricos. La Figura 1 presenta el modelo conceptual que orienta la investigación y la construcción de los modelos predictivos para estimar el consumo hídrico residencial en Bogotá.

Figura 1.

Modelo conceptual para la predicción del consumo hídrico residencial en Bogotá



3.8 Gerencia de Proyectos en el Desarrollo de Modelos Predictivos

La Gerencia de Proyectos proporciona fundamentos para la planificación, organización, seguimiento y evaluación de iniciativas orientadas al cumplimiento de objetivos específicos, considerando el alcance, los recursos disponibles, la calidad y las condiciones de incertidumbre. De acuerdo con el Project Management Institute Cultura [PMI], 2021, la dirección de proyectos debe adaptarse a las características del entorno y a la naturaleza de los resultados esperados. De manera complementaria, la norma ISO 21502 (ISO, 2020) destaca la importancia de considerar el contexto, la complejidad y los riesgos asociados con cada iniciativa. Estos principios resultan pertinentes para el desarrollo de modelos predictivos del consumo hídrico, debido a que

su construcción requiere integrar información de distintas fuentes, establecer criterios de calidad y organizar actividades técnicas orientadas a obtener estimaciones verificables.

En los proyectos de Ciencia de Datos, la planificación debe considerar las particularidades de la información y la incertidumbre asociada con el desempeño de los algoritmos. Chapman et al. (2000), mediante el modelo Cross-Industry Standard Process for Data Mining (CRISP-DM), plantean un proceso estructurado en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Martínez et al. (2021) señalan que los proyectos contemporáneos de Ciencia de Datos presentan trayectorias exploratorias que requieren flexibilidad y revisión de decisiones durante su desarrollo. Desde esta perspectiva, la Gerencia de Proyectos complementa los procesos técnicos mediante la delimitación del alcance, la planificación de actividades y la gestión de restricciones que pueden afectar el cumplimiento de los objetivos.

En el contexto de la predicción del consumo hídrico en Bogotá, esta articulación adquiere relevancia por la necesidad de integrar información climática, temporal y espacial con registros de consumo de agua, considerando diferencias en disponibilidad, cobertura y resolución de los datos. Asimismo, la naturaleza territorial y temporal del fenómeno plantea desafíos relacionados con la representatividad de la información, la consistencia de los registros y la evaluación de la capacidad predictiva de los modelos. Por tanto, la gestión de riesgos, el seguimiento de la calidad y la definición de criterios de evaluación constituyen elementos que pueden complementar el desarrollo técnico y orientar la toma de decisiones durante este tipo de investigaciones.

La Tabla 4 presenta la relación conceptual entre los principales componentes de la Gerencia de Proyectos y las actividades requeridas para desarrollar modelos predictivos del consumo hídrico en Bogotá.

Tabla 4.

Articulación conceptual de la Gerencia de Proyectos y el desarrollo de modelos predictivos.

Componente Gerencial	Aplicación a la predicción del consumo hídrico	Aporte desde la Gerencia de Proyectos
Gestión del Alcance	Delimitación del problema predictivo, la unidad territorial de análisis y la escala temporal de estimación.	Orientar la definición de objetivos, entregables y límites del proyecto.
Planificación	Organización de las actividades necesarias para integrar información de consumo hídrico, clima y localización geográfica.	Establecer una secuencia de trabajo considerando recursos, dependencias y restricciones.
Gestión de la Calidad	Establecimiento de criterios para verificar la consistencia, integridad y pertinencia de la información utilizada.	Definir condiciones de calidad para los insumos y resultados del proyecto.
Gestión de Riesgos e Incertidumbre	Consideración de problemas de disponibilidad de datos, registros atípicos, fuga de información y sobreajuste.	Identificar condiciones que puedan comprometer el cumplimiento de los objetivos y orientar su tratamiento.
Seguimiento y Control	Definición de indicadores de desempeño y criterios de comparación de alternativas predictivas.	Facilitar la evaluación del avance y la verificación del cumplimiento de criterios establecidos.
Gestión de Cambios y Adaptación	Revisión de decisiones metodológicas ante restricciones de información o hallazgos durante el desarrollo.	Orientar ajustes de manera coherente con el alcance y los objetivos del proyecto.
Evaluación y Cierre	Valoración del cumplimiento de objetivos, documentación de limitaciones y formulación de recomendaciones.	Favorecer la trazabilidad de los resultados y la identificación de aprendizajes para futuras iniciativas.

Así, la Gerencia de Proyectos y la Ciencia de Datos constituyen enfoques complementarios para abordar iniciativas de predicción del consumo hídrico. Mientras la Ciencia de Datos proporciona las técnicas para el procesamiento de información, la construcción de modelos y la evaluación de su capacidad predictiva, la Gerencia de Proyectos aporta fundamentos para estructurar las actividades, gestionar la incertidumbre, establecer criterios de seguimiento y orientar el cumplimiento de los objetivos. Esta integración ofrece una base conceptual para abordar el desarrollo de modelos predictivos del consumo hídrico en Bogotá desde una perspectiva técnica y de gestión.

3.9 Vacíos Teóricos y Aporte del Estudio

La revisión evidencia que existe una amplia literatura internacional sobre ML aplicado a recursos hídricos, calidad del agua, evapotranspiración, precipitación y demanda urbana. Sin embargo, persisten vacíos importantes.

Primero, en Colombia existen estudios relevantes sobre variabilidad climática y consumo residencial, como Jiménez et al. (2017), pero son escasos los trabajos que incorporan ML para predecir consumo hídrico urbano. Segundo, estudios nacionales como Triana et al. (2023) y Rueda et al. (2025) demuestran la utilidad de ML en variables hidroclimáticas, pero no abordan directamente la demanda urbana. Tercero, los estudios sobre Bogotá se han concentrado más en crisis hídrica, estrés hídrico, infraestructura o monitoreo, como Ortega et al. (2026), Ortiz (2025), Diago (2025), Herrera y Romero (2025) y Manfrini (2025), pero aún falta integrar estos enfoques con modelos predictivos robustos.

Cuarto, aunque modelos como *XGBoost*, *Random Forest* y ANN han mostrado alta precisión, aún existe el reto de mejorar su interpretabilidad. En ese sentido, herramientas como SHAP podrían fortalecer la explicación de resultados y aportar información útil para la toma de decisiones.

Adicionalmente, la literatura revisada evidencia una limitada evaluación de la capacidad de generalización espacial de los modelos predictivos de consumo hídrico, ya que la mayoría de los estudios se basan en particiones aleatorias de los datos y pocos analizan su desempeño en territorios no observados durante el entrenamiento. Esta limitación resulta especialmente relevante en ciudades con alta heterogeneidad territorial, como Bogotá.

4. Hipótesis

A partir de los antecedentes revisados y de la evidencia reportada en la literatura sobre la influencia de variables climáticas, temporales, espaciales y demográficas en el comportamiento del consumo hídrico, se formula la siguiente hipótesis de investigación, la cual orienta el desarrollo y la evaluación de los modelos predictivos propuestos:

La integración de variables climáticas, temporales y espaciales proporciona información útil para el desarrollo de modelos de *Machine Learning* orientados a la predicción del consumo hídrico en la ciudad de Bogotá.

5. Variables

De acuerdo con Hernández et al. (2014), una variable es una propiedad que puede fluctuar y cuya variación es susceptible de medirse u observarse. Las variables adquieren relevancia dentro de la investigación científica cuando forman parte de hipótesis o modelos teóricos y permiten describir, explicar o predecir un fenómeno de interés.

En la presente investigación, las variables se derivan de la hipótesis planteada y corresponden a los factores asociados al comportamiento del consumo hídrico en la ciudad de Bogotá. Estas variables incluyen componentes climáticos, temporales y espaciales, los cuales fueron incorporados al proceso de modelado predictivo mediante técnicas de ML con el propósito de analizar su relación con el consumo de agua y estimar su comportamiento.

5.1 Clasificación de las Variables

Aunque el estudio posee un alcance descriptivo-correlacional y predictivo, para efectos del modelado de ML se identificó una variable objetivo o variable de respuesta y un conjunto de variables predictoras.

5.1.1 Variable Objetivo

Corresponde al consumo hídrico en la ciudad de Bogotá medido mediante el consumo promedio de agua registrado para cada localidad y período bimestral.

5.1.2 Variables Predictoras

Las variables predictoras corresponden a los factores potencialmente asociados al comportamiento del consumo hídrico e incluyen variables climáticas, temporales, espaciales y demográficas.

5.2 Definición Conceptual y Operacional de las Variables

La Tabla 5 presenta la definición conceptual y operacional de las principales variables consideradas en la investigación.

Tabla 5.

Definición conceptual y operacional de las variables.

Tipo de Variable	Variable	Definición Conceptual	Definición Operacional	Unidad de Medida
Objetivo	Consumo promedio	Cantidad promedio de agua consumida por la unidad de análisis durante un período determinado.	Consumo registrado en las bases oficiales de la EAAB, depuración de registros anómalos, agrupación por localidad, año y período bimestral, y cálculo del consumo promedio ponderado	m^3

Predictora Climática	Temperatura (T2M)	Condición térmica promedio del ambiente.	Temperatura media del aire a 2 metros de altura. Extracción diaria desde NASA POWER por localidad. Agregación temporal a escala bimestral.	°C
Predictora Climática	Precipitación (PRECTOTCORR)	Cantidad de agua precipitada durante un período determinado.	Precipitación acumulada registrada para la localidad y período analizado. Extracción diaria desde NASA POWER por localidad. Agregación temporal a escala bimestral.	mm
Predictora Climática	Humedad relativa (RH2M)	Cantidad de vapor de agua presente en el aire respecto al máximo posible.	Humedad relativa promedio a 2 metros de altura. Extracción diaria desde	%

			NASA POWER por localidad. Agregación temporal a escala bimestral.	
Predictora Climática	Índice ONI	Indicador oceánico asociado al fenómeno ENSO.	Valor mensual del Oceanic Niño Index.	°C (Anomalía)
Predictora Sociodemográfica	Población	Número de habitantes presentes en una localidad.	Población proyectada por localidad.	Habitantes
Predictora Sociodemográfica	Hogares proyectados	Número estimado de hogares por localidad.	Proyección oficial del DANE.	Hogares
Predictora Espacial	Localidad	División administrativa del territorio urbano.	Categoría territorial de cada observación.	Categórica
Predictora Espacial	Latitud	Coordenada geográfica norte-sur.	Coordenada geográfica de la localidad.	Grados decimales
Predictora Espacial	Longitud	Coordenada geográfica este-oeste.	Coordenada geográfica de la localidad.	Grados decimales

Predictora Temporal	Año	Periodo anual de observación.	Año correspondiente al registro analizado.	Entero
Predictora Temporal	Bimestre	Periodo bimestral de observación.	Bimestre correspondiente al registro analizado.	Entero

Adicionalmente, la matriz de datos incorporó otras variables climáticas relacionadas con humedad del suelo, radiación solar, presión atmosférica, velocidad del viento y cobertura nubosa, las cuales fueron utilizadas durante el análisis exploratorio y el proceso de selección de variables. Cabe precisar que las variables sociodemográficas (población, hogares proyectados y personas por hogar) fueron recopiladas e incorporadas a la matriz de datos con fines de caracterización del contexto, pero no formaron parte del conjunto de predictoras utilizado en el entrenamiento de los modelos finales, debido a que su resolución anual no resulta compatible con la unidad de análisis bimestral del estudio. Esta decisión se retoma en el apartado de limitaciones metodológicas.

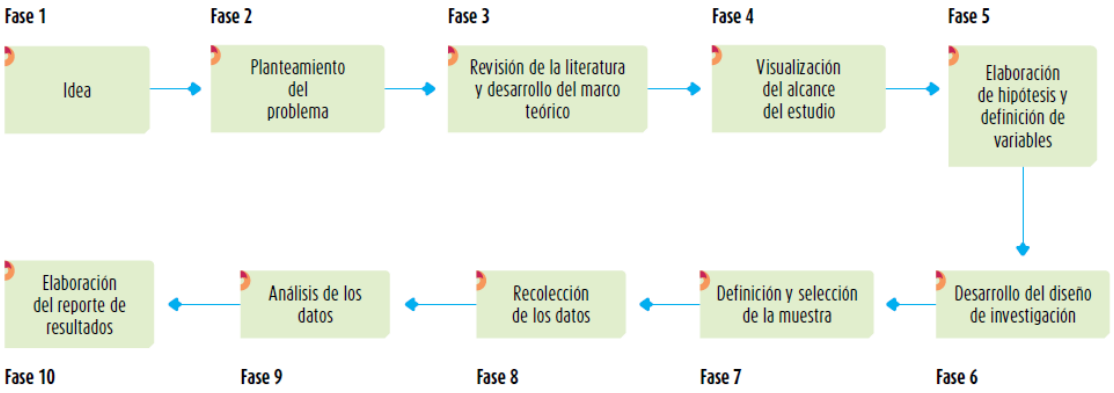
6. Metodología

6.1 Enfoque de la Investigación

La investigación se desarrolla bajo un enfoque Cuantitativo, debido a que se fundamenta en la recopilación, procesamiento y análisis de datos numéricos con el propósito de identificar relaciones entre variables y construir modelos predictivos que permitan estimar el consumo hídrico en la ciudad de Bogotá. De acuerdo con Hernández et al., (2014), el enfoque cuantitativo se caracteriza por utilizar la medición de variables, el análisis estadístico y la contrastación empírica para responder preguntas de investigación y generar conocimiento objetivo y verificable. La Figura 2 presenta las principales etapas del proceso cuantitativo.

Figura 2.

Proceso Cuantitativo.



Fuente. Tomado de Hernández Sampieri et al. (2014, p. 37).

Se emplearon registros históricos de consumo de agua, variables climáticas, variables sociodemográficas y variables espaciales, las cuales son analizadas mediante técnicas estadísticas y algoritmos de ML. La información utilizada corresponde a datos estructurados y cuantificables, por lo que el proceso investigativo se orienta a la identificación de patrones, asociaciones y capacidades predictivas a partir de evidencia empírica.

Asimismo, Ander (2011) señala que la investigación científica constituye un proceso sistemático orientado a la producción de conocimiento verificable mediante procedimientos rigurosos de observación, análisis e interpretación. Así, la investigación integra procedimientos de análisis exploratorio, modelado predictivo y validación estadística que permiten estudiar el comportamiento del consumo hídrico desde una perspectiva objetiva y reproducible.

Por su parte, Perez et al., (2020) destacan que el enfoque cuantitativo resulta apropiado cuando el propósito de la investigación consiste en medir fenómenos, establecer relaciones entre variables y desarrollar modelos explicativos o predictivos basados en datos observables.

6.2 Alcance de la Investigación

De acuerdo con la clasificación propuesta por Hernández et al. (2014), la presente investigación posee un alcance descriptivo-correlacional con orientación predictiva. Descriptivo, ya que busca caracterizar el comportamiento del consumo hídrico en la ciudad de Bogotá a partir del análisis de variables operativas, climáticas, sociodemográficas, temporales y espaciales. Esta caracterización permite identificar patrones de consumo, tendencias temporales, diferencias territoriales entre localidades y comportamientos asociados al uso del recurso hídrico.

Correlacional porque analiza las asociaciones existentes entre las variables independientes y el consumo hídrico, considerado como variable objetivo. Según Hernández et al. (2014), los estudios correlacionales tienen como finalidad determinar el grado de relación existente entre dos o más variables dentro de un contexto específico, sin pretender establecer relaciones causales directas. En este sentido, el estudio busca identificar qué variables climáticas, sociodemográficas, temporales y espaciales presentan una mayor asociación estadística con el comportamiento del consumo de agua en Bogotá.

Además, la investigación incorpora una orientación predictiva mediante la aplicación de técnicas de ML para el desarrollo de modelos capaces de estimar el consumo hídrico a partir de

información histórica y variables explicativas, permitiendo evaluar la capacidad de diferentes algoritmos para modelar su comportamiento.

6.3 Diseño de la Investigación

De acuerdo con Hernández et al. (2014), la presente investigación se desarrolla bajo un diseño no experimental de tipo longitudinal retrospectivo, dado que analiza fenómenos sin manipular las variables de estudio. En este sentido, se empleó información histórica proveniente de fuentes oficiales, incluyendo registros de consumo hídrico, variables climáticas, datos sociodemográficos y componentes espaciales de la ciudad de Bogotá.

Asimismo, el estudio corresponde a un diseño longitudinal ya que analiza observaciones registradas durante múltiples momentos en el tiempo, permitiendo examinar la evolución del consumo hídrico y de las variables asociadas entre los años 2020 y 2024. Según Hernández et al. (2014), los diseños longitudinales permiten identificar cambios, tendencias y patrones temporales mediante la observación repetida de un fenómeno a lo largo de diferentes periodos.

Además, el diseño puede considerarse retrospectivo debido a que la información analizada corresponde a registros históricos previamente generados y almacenados por entidades oficiales, los cuales son utilizados posteriormente para realizar análisis descriptivos, correlacionales y predictivos. Esta característica resulta coherente con lo planteado por Perez et al. (2020), quienes señalan que el diseño metodológico debe responder a la naturaleza de los datos disponibles y a los objetivos de la investigación.

Por último, Ander (2011) plantea que la investigación científica puede desarrollarse mediante el análisis sistemático de hechos y registros existentes, permitiendo generar conocimiento válido sobre fenómenos complejos sin necesidad de intervenir directamente sobre ellos.

6.4 Fases de la Investigación

La investigación se estructuró en tres fases principales que permite desarrollar de manera secuencial el proceso de identificación de variables, construcción de modelos predictivos y validación de resultados, en concordancia con los principios de sistematicidad y rigor científico propuestos por Hernández et al. (2014), Ander (2011) y Perez et al. (2020).

6.4.1 Fase 1. Identificación y Selección de Variables Asociadas al Consumo Hídrico

Fase orientada a identificar las variables relevantes asociadas al consumo hídrico en Bogotá. Se recopilaron e integraron datos de fuentes oficiales relacionadas con el consumo de agua, variables climáticas, información sociodemográfica y componentes espaciales. Las principales fuentes de información incluyeron registros de consumo hídrico y variables operativas del acueducto, información climática y meteorológica, datos poblacionales y variables espaciales asociadas a las localidades de Bogotá. Luego, se desarrolló un proceso de integración de las diferentes bases de datos con el fin de construir un conjunto de información homogéneo que permitiera analizar conjuntamente las distintas dimensiones del fenómeno estudiado.

La información fue sometida a un proceso de preparación y control de calidad de los datos, incluyendo la revisión de consistencia de registros, identificación de valores faltantes, detección de registros duplicados, homologación de formatos y análisis de valores atípicos.

Consolidada la base de datos, se desarrolló un análisis exploratorio mediante estadística descriptiva, histogramas, diagramas de caja, análisis temporal, análisis territorial y matrices de correlación. Esto permitió caracterizar el comportamiento del consumo hídrico, identificar patrones espaciales y temporales, y seleccionar las variables con mayor relevancia para el modelado predictivo. Asimismo, se excluyeron aquellas variables que presentaban riesgo de fuga de información, garantizando la validez de los modelos desarrollados posteriormente.

6.4.2 Fase 2. Construcción de Modelos Predictivos de Machine Learning

Tuvo como propósito diseñar y desarrollar modelos predictivos capaces de estimar el comportamiento del consumo hídrico en Bogotá a partir de las variables identificadas en la fase anterior. Para ello se definió la variable objetivo correspondiente al consumo promedio de agua, sobre la cual se aplicaron transformaciones orientadas a reducir la asimetría y mejorar la estabilidad del proceso de entrenamiento. Asimismo, se realizaron procesos de normalización y preparación de las variables predictoras con el fin de garantizar la compatibilidad con los algoritmos de aprendizaje automático utilizados.

Se dividió la información en conjuntos de entrenamiento y prueba, permitiendo evaluar la capacidad predictiva de los modelos sobre observaciones no utilizadas durante el proceso de aprendizaje. Sobre esta base, se implementaron diferentes enfoques de Machine Learning orientados a problemas de regresión supervisada, incluyendo redes neuronales profundas (Deep Neural Networks) y XGBoost, una técnica de Gradient Boosting ampliamente utilizada en problemas de regresión con datos tabulares. Para cada modelo se realizaron procesos de ajuste y optimización orientados a evaluar su desempeño predictivo.

Inicialmente, se evaluó un escenario que incorporaba latitud y longitud como variables predictoras, con el propósito de analizar el aporte de la dimensión espacial. Esta evaluación se complementó mediante diferentes estrategias de validación, incluyendo partición aleatoria, validación temporal y validación espacial Leave-One-Localidad-Out (LOLO), con el fin de diferenciar la capacidad de interpolación dentro de las localidades observadas de la generalización hacia periodos y territorios no incluidos durante el entrenamiento.

Según Meyer et al. (2018), la validación Leave-Location-Out (LLO) permite evaluar la capacidad de generalización de un modelo hacia ubicaciones no observadas durante el entrenamiento. Este procedimiento consiste en excluir una localización —o un grupo de localizaciones— del conjunto de entrenamiento y utilizarla posteriormente como conjunto de

validación. Los autores señalan que este enfoque proporciona una estimación más realista del desempeño espacial del modelo que una partición aleatoria cuando existen dependencias espaciales en los datos. En esta investigación, este principio se adapta mediante una estrategia (LOLO), en la que cada localidad de Bogotá es excluida de manera iterativa para evaluar la capacidad del modelo de generalizar hacia localidades no utilizadas durante el entrenamiento.

Esta estrategia también se sustenta en Roberts et al. (2017), quienes advierten que la validación cruzada aleatoria puede subestimar el error predictivo cuando los datos presentan estructuras espaciales, debido a la falta de independencia entre las observaciones utilizadas para entrenamiento y validación. Los autores recomiendan utilizar esquemas de validación por bloques o grupos espaciales cuando el objetivo es estimar la capacidad del modelo para generalizar hacia ubicaciones no observadas. En el contexto de esta investigación, las localidades de Bogotá constituyen los grupos espaciales utilizados para realizar dicha evaluación.

Posteriormente, se implementó un escenario adicional de validación temporal en el que el modelo se entrenó con los registros correspondientes al periodo 2020–2023 y se evaluó exclusivamente con los registros de 2024. La selección de hiperparámetros se efectuó mediante validación temporal interna con ventana expansiva dentro del periodo de entrenamiento, manteniendo la información de 2024 aislada hasta la evaluación final.

En este escenario adicional se excluyeron la latitud, la longitud y la localidad del conjunto de variables predictoras de XGBoost, con el propósito de evaluar la contribución de las variables climáticas y temporales sin incorporar identificadores territoriales directos. Asimismo, se excluyeron completamente los registros correspondientes a Sumapaz, por lo que el alcance de esta evaluación quedó delimitado a las 19 localidades restantes. La variable localidad se conservó únicamente como metadato para analizar los errores por territorio y construir los modelos históricos de referencia.

6.4.3 Fase 3. Validación y Evaluación del Desempeño Predictivo

Esta fase estuvo orientada a evaluar el desempeño de los modelos desarrollados mediante métricas de regresión, específicamente el MAE, la RMSE y el R^2, las cuales permitieron cuantificar el error de predicción y la capacidad explicativa de los modelos.

La evaluación se realizó mediante dos estrategias de validación, una validación aleatoria tradicional para analizar el desempeño predictivo dentro del dominio espacial observado durante el entrenamiento y una validación espacial mediante el esquema *Leave-One-Localidad-Out (LOLO)*, con el fin de evaluar la capacidad de generalización de los modelos frente a localidades no observadas durante el proceso de aprendizaje.

Finalmente, se compararon los resultados obtenidos por los algoritmos implementados, identificando fortalezas, limitaciones y capacidad de generalización. Se analizaron las implicaciones de los resultados para la comprensión del consumo hídrico en Bogotá y para el potencial uso de herramientas de ML como apoyo a la gestión sostenible del recurso hídrico.

6.5 Integración de la Gerencia de Proyectos en el Desarrollo de la Investigación.

La aplicación de la Gerencia de Proyectos en el desarrollo metodológico de esta investigación se relacionó con la delimitación del alcance, la organización de las actividades y el seguimiento de los procesos orientados a la construcción y validación de modelos predictivos del consumo hídrico en Bogotá. Para ello, el trabajo se estructuró en tres fases vinculadas con los objetivos específicos: identificación y selección de variables, construcción de modelos predictivos y validación del desempeño. Esta organización permitió articular la recopilación e integración de información de consumo hídrico, climática, temporal y espacial con el desarrollo de los modelos de Machine Learning, considerando las restricciones de disponibilidad y calidad de los datos, así como la necesidad de evaluar progresivamente las alternativas predictivas.

Desde esta perspectiva, la Gerencia de Proyectos aportó criterios de planificación, gestión de incertidumbre, control de calidad y seguimiento, que complementaron los procedimientos técnicos de Ciencia de Datos. Su aplicación se reflejó en la identificación de restricciones metodológicas, la revisión de la consistencia de la información, la prevención de riesgos asociados con la fuga de información y el sobreajuste, y la utilización de diferentes estrategias de validación para orientar las decisiones durante el desarrollo. La Tabla 6 presenta la articulación entre las fases de la investigación y los componentes de Gerencia de Proyectos relacionados con su ejecución.

Tabla 6.

Articulación de la Gerencia de Proyectos con las fases de investigación.

Fase de investigación	Actividades principales	Articulación con la Gerencia de Proyectos
Fase 1. Identificación y selección de variables	Recopilación de información de consumo hídrico, clima y características territoriales; integración, depuración y análisis exploratorio de datos.	Planificación de actividades, identificación de fuentes y restricciones, gestión de calidad de la información y delimitación de los insumos requeridos.
Fase 2. Construcción de modelos predictivos	Preparación de variables, entrenamiento y ajuste de redes neuronales profundas y XGBoost, y evaluación progresiva de alternativas.	Organización del trabajo, seguimiento de actividades y adaptación de decisiones técnicas de acuerdo con las condiciones de los datos y los resultados intermedios.
Fase 3. Validación y evaluación del desempeño	Aplicación de métricas predictivas, validación aleatoria, temporal y espacial, y comparación con modelos de referencia.	Seguimiento del cumplimiento de objetivos, evaluación de la calidad de los resultados e identificación de limitaciones que condicionan su utilización.

6.5.1 Gestión de los Interesados

Se identificaron los interesados que podían influir en el proyecto o verse afectados por sus resultados, y se clasificaron según su nivel de interés e influencia, con el fin de definir una estrategia de involucramiento para cada uno (PMI, 2021). La Tabla 7 presenta el registro de

interesados. Esta identificación tuvo consecuencias directas sobre el diseño del estudio: por ejemplo, la dependencia de los datos publicados por la EAAB condicionó la unidad de análisis (localidad y bimestre), y el interés de las entidades distritales en la planeación motivó la inclusión de la validación temporal como escenario más cercano a un uso real.

Tabla 7.

Registro de Interesados del proyecto.

Interesado	Rol / Interés en el proyecto	Interés	Influencia	Estrategia de gestión
Empresa de Acueducto y Alcantarillado de Bogotá (EAAB)	Fuente de los datos de consumo; usuario potencial del modelo para la planeación de la demanda	Alto	Alta	Gestionar de cerca: uso fiel de sus datos y presentación de resultados con su alcance real
Secretaría Distrital de Ambiente y Alcaldía Mayor de Bogotá	Formulación de políticas de ahorro y gestión ante racionamientos	Alto	Media	Mantener informada: implicaciones para la gestión territorial
NASA POWER, NOAA y DANE	Proveedores de datos climáticos, del índice ONI y sociodemográficos	Bajo	Alta	Monitorear: verificar resolución, cobertura y actualización de las fuentes
Directora del trabajo de grado	Orientación metodológica y aprobación de entregables	Alto	Alta	Gestionar de cerca: reuniones periódicas de seguimiento
Jurados y Universidad Ean	Evaluación de la calidad académica y del cumplimiento de los requisitos de ambas maestrías	Alto	Alta	Gestionar de cerca: atención documentada de la retroalimentación
Ciudadanía de Bogotá	Beneficiaria indirecta de una mejor planeación del recurso hídrico	Medio	Baja	Mantener informada: resultados divulgados en lenguaje accesible
Comunidad académica	Reutilización del protocolo de validación y del repositorio	Medio	Baja	Mantener informada: repositorio abierto y documentado

6.5.2 Gestión de Riesgos, Calidad y Seguimiento

La gestión de riesgos e incertidumbre constituye un componente relevante para el desarrollo de modelos predictivos, debido a que las características de los datos y las

condiciones de evaluación pueden afectar la confiabilidad de las estimaciones. En esta investigación, el tratamiento de tales condiciones se relacionó con la revisión de la disponibilidad y consistencia de los registros, la identificación de valores atípicos, la compatibilidad temporal entre fuentes y la prevención de fugas de información. Asimismo, la consideración de diferentes escenarios de validación permitió examinar riesgos asociados con el sobreajuste y la dependencia de características territoriales. La Tabla 8 sintetiza los principales riesgos metodológicos considerados y las medidas de tratamiento relacionadas con los procedimientos descritos en la investigación.

Tabla 8.

Riesgos metodológicos y medidas de tratamiento.

Riesgo metodológico	Posible efecto sobre el proyecto	Medida de tratamiento
Inconsistencias y registros atípicos	Afectación de la calidad de la información y de la confiabilidad de las estimaciones.	Revisión de integridad, detección de duplicados, identificación de valores faltantes y análisis de observaciones atípicas.
Diferencias de resolución temporal entre fuentes	Dificultades para integrar variables en una unidad de análisis común.	Homologación temporal y selección de variables compatibles con la escala bimestral del estudio.
Fuga de información (*data leakage*)	Sobreestimación del desempeño predictivo por el uso de información relacionada directamente con la variable objetivo.	Identificación y exclusión de variables que pudieran introducir información indebida en el entrenamiento.
Sobreajuste y dependencia espacial	Desempeño favorable en datos conocidos, sin garantía de generalización hacia otros períodos o localidades.	Aplicación de validaciones aleatoria, temporal y espacial, incluida la estrategia *Leave-One-Localidad-Out* (LOLO).
Incertidumbre en el desempeño predictivo	Dificultad para determinar la conveniencia de los modelos desarrollados.	Evaluación mediante métricas de regresión y comparación con modelos de referencia.

En consecuencia, la Gerencia de Proyectos contribuyó a estructurar el desarrollo de la investigación mediante la delimitación del alcance, la organización de las fases, el

reconocimiento de riesgos metodológicos y el seguimiento de las actividades orientadas al cumplimiento de los objetivos. Su integración con la Ciencia de Datos permitió abordar el desarrollo de modelos predictivos del consumo hídrico en Bogotá desde una perspectiva que articula la planificación y la gestión de incertidumbre con el procesamiento de información, el modelado y la validación predictiva.

6.6 Población y Muestra

De acuerdo con Hernández et al. (2014), la población corresponde al conjunto de casos que comparten determinadas características y sobre los cuales se pretende realizar inferencias. En esta investigación, la población estuvo conformada por los registros históricos de consumo hídrico residencial de la ciudad de Bogotá, complementados con variables climáticas, sociodemográficas y espaciales correspondientes al periodo 2020–2024.

La unidad de análisis estuvo constituida por los registros consolidados por localidad y periodo bimestral, integrando información proveniente de las diferentes fuentes de datos en una única estructura de análisis. Dado que se trabajó con la totalidad de los registros disponibles que cumplieron los criterios de calidad establecidos, no fue necesario realizar un proceso de selección muestral. La base final estuvo conformada por 596 observaciones organizadas por localidad y periodo bimestral, resultado de las 600 combinaciones posibles menos los cuatro registros de Sumapaz excluidos por corresponder a artefactos de agregación.

6.7 Instrumentos de Recolección de Información

El principal instrumento de investigación fue una matriz de datos estructurada construida a partir de la integración de fuentes secundarias oficiales (registros de consumo hídrico de la EAAB, información climática de *NASA POWER* y del índice ONI de NOAA, datos

sociodemográficos del DANE y datos geoespaciales de las localidades de Bogotá obtenidos mediante ArcGIS), utilizada como base para el análisis exploratorio, el procesamiento de datos y el desarrollo de modelos predictivos.

Tabla 9.

Dimensiones y variables integradas en la matriz de datos.

Dimensión	Variables Principales	Propósito Dentro del Estudio
Operativas	Metros cúbicos consumidos, Consumo promedio, Facturación, Suscriptores facturados y Suscriptores con medidor.	Caracterizar el comportamiento del consumo hídrico y la operación del servicio de acueducto.
Climáticas y atmosféricas	Temperatura, Precipitación, Humedad relativa, Radiación solar, Humedad del suelo, Velocidad del viento e Índice Oceánico del Niño (ONI).	Analizar la posible influencia de las condiciones climáticas sobre el consumo de agua.
Sociodemográficas	Población, Hogares proyectados y Personas por hogar.	Representar las características demográficas asociadas a la demanda hídrica.
Espaciales y temporales	Localidad, Latitud, Longitud, Año y Bimestre.	Incorporar la dimensión territorial y temporal del consumo hídrico.

6.8 Validación de los Instrumentos y Control de Calidad de la Información

La validación de la información se realizó mediante procedimientos orientados a garantizar la calidad, consistencia y confiabilidad de los datos utilizados en la investigación. Este proceso incluyó la verificación de la consistencia entre las diferentes fuentes de información, el control de registros duplicados, la revisión de la integridad de los datos, la identificaron de valores

faltantes y la detección de observaciones atípicas. Asimismo, se evaluó la coherencia temporal de los registros y se validaron las variables utilizadas en el modelado.

Asimismo, se desarrolló un análisis específico de riesgo de fuga de información (*Data Leakage*), identificando variables que podían generar sobreestimación artificial del desempeño predictivo. Como resultado de este proceso se excluyeron variables directamente relacionadas con el cálculo del consumo promedio, garantizando que los modelos fueran entrenados únicamente con variables predictoras independientes. Siguiendo los planteamientos de Hernández et al. (2014), estos procedimientos contribuyen a fortalecer la validez interna de la investigación y la confiabilidad de los resultados.

6.9 Técnicas para el Análisis de la Información

El análisis de la información se realizó mediante técnicas estadísticas, herramientas de visualización de datos y algoritmos de *Machine Learning*. Las técnicas aplicadas incluyeron:

6.9.1 Análisis Descriptivo

- Medidas de tendencia central.
- Medidas de dispersión.
- Distribuciones de frecuencia.

6.9.2 Análisis Exploratorio de Datos

- Histogramas.
- Diagrama de caja y bigotes *(Boxplots)*.
- Series temporales.
- Mapas temáticos.

- Matrices de correlación.

6.9.3 Modelado Predictivo

- Redes neuronales profundas (*Deep Neural Networks*).
- *XGBoost*.

6.9.4 Evaluación de Desempeño

El desempeño de los modelos se evaluó mediante el Error Absoluto Medio, la Raíz del Error Cuadrático Medio y el Coeficiente de Determinación. Adicionalmente, se implementaron procedimientos de validación aleatoria y validación espacial para analizar la capacidad de generalización de los modelos desarrollados.

6.10 Herramientas Computacionales Utilizadas

El procesamiento, análisis y modelado de los datos se realizó utilizando herramientas especializadas de ciencia de datos. Entre las principales utilizadas se encuentran *Python, Pandas, NumPy, Scikit-Learn, XGBoost, TensorFlow/Keras, Matplotlib, Seaborn y Power BI.* Estas herramientas permitieron desarrollar procesos de integración de datos, análisis exploratorio, entrenamiento de modelos, validación estadística y visualización de resultados.

6.11 Consideraciones Éticas

La investigación se desarrolló utilizando información secundaria proveniente de fuentes oficiales y bases de datos institucionales. No se recopilaron datos personales ni información sensible. El tratamiento de la información se realizó con fines académicos y científicos,

respetando los principios de confidencialidad, integridad de los datos y uso responsable de la información.

Asimismo, todas las fuentes utilizadas fueron citadas y referenciadas conforme a las normas APA vigentes, garantizando el respeto por la propiedad intelectual y los principios éticos de la investigación científica.

7. Trabajo de Campo

Este capítulo presenta el desarrollo del trabajo de campo realizado para la construcción y evaluación de modelos predictivos del consumo hídrico en Bogotá. Se describen las actividades de integración y preparación de datos, análisis exploratorio, identificación de variables, construcción de modelos de Machine Learning y evaluación de su desempeño mediante diferentes estrategias de validación.

Desde la perspectiva de la Gerencia de Proyectos, el desarrollo de estas actividades incorporó criterios de planificación, seguimiento, control de calidad y gestión de riesgos metodológicos, orientados al cumplimiento de los objetivos de investigación. Su aplicación se reflejó en la revisión progresiva de las alternativas de modelado, la identificación de restricciones asociadas con la información disponible y la adaptación de las decisiones técnicas a partir de los resultados obtenidos. De esta manera, la Gerencia de Proyectos complementó los procedimientos de Ciencia de Datos al contribuir a la organización del trabajo, la gestión de la incertidumbre y la evaluación del alcance y las limitaciones de los modelos desarrollados.

7.1 Caracterización e Integración de la Información

La construcción del conjunto de datos permitió consolidar la información necesaria para el análisis exploratorio y el desarrollo de los modelos predictivos. Para esto, se integraron diferentes fuentes secundarias de información relacionadas con el consumo hídrico, las condiciones climáticas y las características espaciales de la ciudad de Bogotá, obteniendo una base de datos unificada que sirvió como insumo para todas las etapas posteriores del estudio.

El proceso de integración comprendió actividades de recopilación, depuración, homologación y consolidación de la información, garantizando la consistencia de las variables provenientes de diferentes fuentes. La integración se realizó utilizando como criterios comunes

la localidad, el período de observación (año y bimestre) y la correspondencia espacial entre los registros, permitiendo relacionar la información operacional del sistema de acueducto con las variables climáticas y geográficas correspondientes. La Figura 3 resume el proceso general seguido para la integración de la Información utilizada durante la investigación.

Figura 3.

Fuentes de Información.



Tabla 10.

Fuentes de Información utilizadas en la Investigación.

Fuente	Información Suministrada	Aplicación en la Investigación
Empresa de Acueducto y Alcantarillado de Bogotá **(EAAB)**	Consumo de agua, suscriptores, facturación y variables operativas	Variable objetivo e indicadores de consumo
NASA POWER	Variables climáticas y atmosféricas	Caracterización de las condiciones ambientales

Infraestructura de Datos Espaciales de Bogotá (IDECA)	Localidades, coordenadas geográficas y límites administrativos	Variables espaciales e integración territorial
National Oceanic and Atmospheric Administration **(NOAA)**	Índice Oceánico del Niño (ONI), serie mensual	Variable de variabilidad climática asociada al fenómeno ENSO
Departamento Administrativo Nacional de Estadística **(DANE)**	Proyecciones de población y hogares por localidad	Caracterización sociodemográfica del contexto; no incorporadas al modelado final

Las fuentes de Información utilizadas en la Investigación se presentan de manera resumida en la Tabla 10. Las variables Incorporadas se agruparon de acuerdo con la dimensión del fenómeno que representan. Esta clasificación facilitó el análisis de la Información y permitió Integrar variables de distinta naturaleza dentro de un mismo conjunto de datos.

Tabla 11.

Dimensiones de las Variables Incorporadas.

Dimensión	Ejemplos de Variables	Propósito
Operativa	Consumo promedio, volumen facturado, suscriptores	Representar el comportamiento del consumo de agua
Climática	Temperatura, precipitación, humedad, presión, radiación, ONI	Caracterizar las condiciones ambientales
Temporal	Año, bimestre	Representar la evolución temporal del consumo
Espacial	Localidad, latitud, longitud	Incorporar la dimensión territorial del estudio

El cruce inicial de las fuentes produjo una matriz de 600 combinaciones posibles, correspondientes a 20 localidades × 5 años × 6 bimestres. Durante la verificación de calidad se identificaron cuatro registros de la localidad de Sumapaz cuyos valores de consumo promedio correspondían a artefactos del proceso de agregación y no a consumo real (2022-Bimestre 1 y 2024-Bimestres 4, 5 y 6); su tratamiento se detalla en el apartado siguiente. Tras su exclusión, la base consolidada quedó conformada por 596 registros y 49 variables, organizada por localidad, año y período bimestral, con 30 observaciones para cada una de las 19 localidades urbanas y 26 para Sumapaz. Esta estructura permitió integrar información operacional, climática, temporal y espacial en una misma unidad de análisis, constituyendo el insumo principal para el análisis exploratorio, la construcción de los modelos predictivos y la evaluación de su desempeño.

7.2 Preparación y Calidad de los Datos

Consolidada la base de datos, se realizó un proceso de preparación y verificación de la calidad de la información con el propósito de garantizar que los datos utilizados en el entrenamiento y evaluación de los modelos predictivos fueran consistentes, comparables y adecuados para su procesamiento mediante técnicas de ML. Etapa fundamental ya que la calidad de los datos determina la confiabilidad de los resultados. La preparación de los datos se desarrolló mediante una serie de actividades orientadas a garantizar su calidad para el modelado. La Figura 4 presenta el resumen de este proceso.

Figura 4.

Proceso de Preparación y Calidad de los Datos.



7.2.1 Verificación de la Calidad de los Datos

Se realizó una revisión integral de la base de datos con el fin de verificar la calidad de la información proveniente de las diferentes fuentes utilizadas en la investigación. Este proceso incluyó la comprobación de la consistencia entre las variables operativas, climáticas, temporales, espaciales y sociodemográficas, así como la revisión de formatos, tipos de datos y unidades de medida.

Adicionalmente, se verificó la correspondencia entre las localidades de Bogotá, los períodos bimestrales y los años analizados, garantizando que todas las observaciones conservaran una estructura homogénea después del proceso de integración. Asimismo, se revisó la existencia de registros duplicados y posibles inconsistencias derivadas de la unificación de las diferentes bases de datos.

Como resultado, se obtuvo un conjunto de datos estructurado y consistente, conformado por 596 registros y 49 variables, el cual constituyó la base para el desarrollo del análisis exploratorio y el entrenamiento de los modelos predictivos.

7.2.2 Tratamiento de Valores Faltantes y Valores Atípicos

Posteriormente se realizó la evaluación de valores faltantes y valores atípicos presentes en las principales variables del estudio. Este procedimiento permitió identificar posibles inconsistencias que pudieran afectar el desempeño de los modelos predictivos y determinar el tratamiento más adecuado para cada caso.

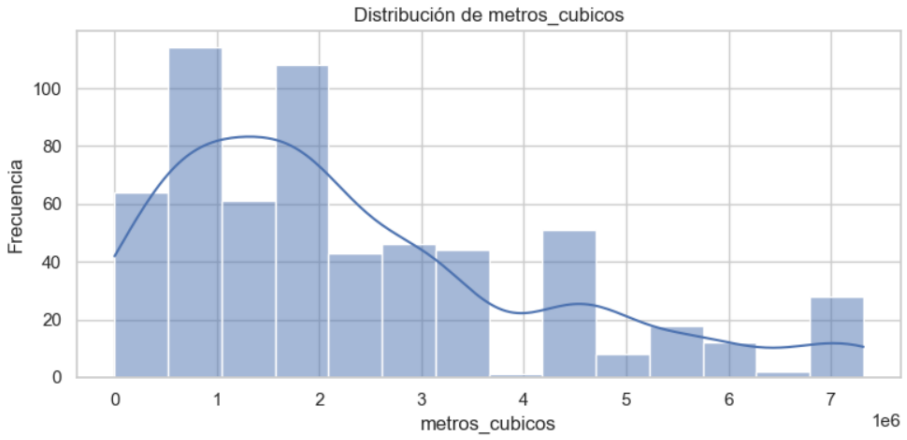
Inicialmente se verificó la existencia de valores faltantes en las variables provenientes de las diferentes fuentes de información. El proceso de integración y depuración previa permitió obtener un conjunto prácticamente completo, sin registros ausentes en las variables operativas de la EAAB ni en las variables espaciales de IDECA. No obstante, para garantizar la compatibilidad con los algoritmos empleados, se incorporó al flujo de modelado un procedimiento de imputación por mediana mediante *SimpleImputer*.

Este imputador se ajustó exclusivamente con el conjunto de entrenamiento y se aplicó posteriormente a los conjuntos de validación y prueba, evitando cualquier transferencia de información entre particiones (*data leakage*). El mismo criterio se aplicó al escalamiento de las variables numéricas. La implementación puede verificarse en los cuadernos del Anexo A.

En cuanto a los valores atípicos, se efectuó un análisis descriptivo utilizando histogramas y diagramas de caja (*boxplots*), herramientas que permiten visualizar la distribución de los datos, su dispersión y la presencia de observaciones extremas.

Figura 5.

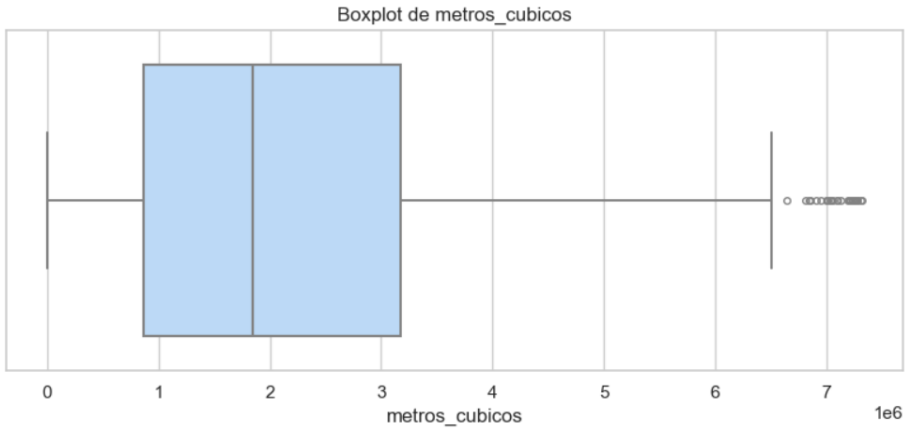
Distribución del Volumen de Agua Facturada (Metros Cúbicos).



La Figura 5 muestra que la variable metros cúbicos presenta una distribución asimétrica positiva, con una mayor concentración de observaciones en rangos bajos y una disminución progresiva de la frecuencia hacia valores superiores. Este comportamiento evidencia la existencia de diferencias importantes entre localidades y períodos de observación.

Figura 6.

Diagrama de Caja del Volumen de Agua Facturada (Metros Cúbicos).

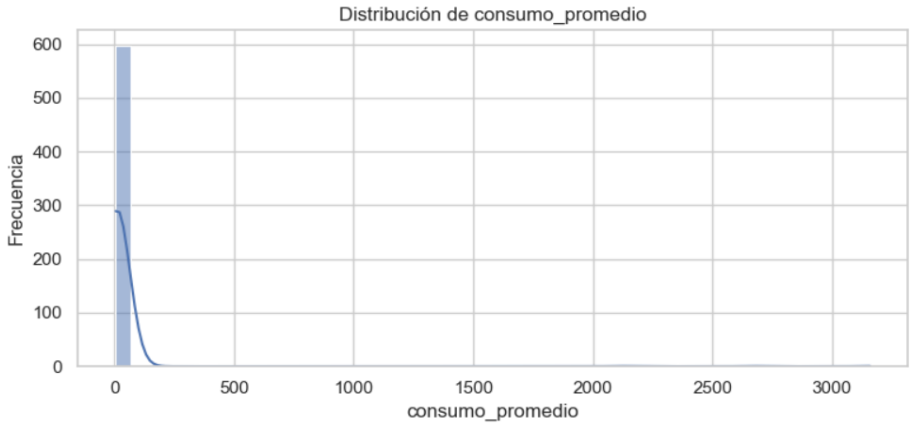


El diagrama de caja de la Figura 6 confirma la presencia de valores atípicos superiores. Sin embargo, la revisión de la información permitió establecer que estas observaciones corresponden a registros reales del sistema de acueducto y reflejan diferencias propias del

comportamiento del consumo entre localidades, por lo que fueron conservadas durante el proceso de modelado. De manera similar, se evaluó la distribución de la variable objetivo consumo promedio, utilizada para el entrenamiento de los modelos predictivos.

Figura 7.

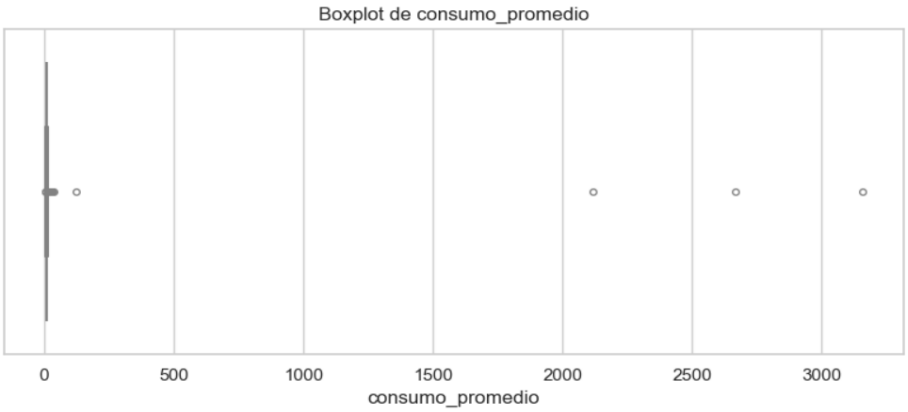
Distribución del Consumo Hídrico Promedio.



La Figura 7 evidencia una distribución asimétrica hacia la derecha, caracterizada por una alta concentración de observaciones en valores bajos y un número reducido de registros con consumos considerablemente superiores. Esto confirma que la variable objetivo no sigue una distribución normal y presenta una variabilidad importante entre las diferentes observaciones.

Figura 8.

Diagrama de Caja del Consumo Hídrico Promedio.



El análisis mediante diagramas de caja permitió identificar observaciones atípicas en la variable objetivo. La revisión de estos registros evidenció que los valores extremos asociados a Sumapaz correspondían a artefactos derivados del proceso de agregación y registro, por lo que fueron excluidos antes del entrenamiento de los modelos. Las demás observaciones se conservaron al no corresponder a las anomalías identificadas, preservando la variabilidad presente en el conjunto de datos depurado.

7.2.3 Tratamiento de los registros anómalos de Sumapaz

Durante la revisión de la variable objetivo se identificaron registros asociados a Sumapaz con valores considerablemente superiores al comportamiento general del conjunto de datos. El contraste con la información de origen permitió establecer que estas observaciones no representaban incrementos reales del consumo hídrico, sino artefactos derivados del proceso de agregación y registro. Debido a que afectaban directamente la variable consumo_promedio, su conservación habría introducido distorsiones en el entrenamiento, las métricas de evaluación y la interpretación de los modelos.

Por esta razón, los cuatro registros anómalos fueron excluidos antes de realizar las divisiones de entrenamiento, validación y prueba. Concretamente, se retiraron las observaciones de Sumapaz correspondientes a 2022-Bimestre 1 y a 2024-Bimestres 4, 5 y 6, que son los mismos periodos identificados como anómalos durante el análisis exploratorio. La localidad no fue eliminada por completo: conservó sus 26 registros válidos dentro de la base depurada de 596 observaciones, cuyo valor máximo de consumo promedio es 38,41 y no 3.158,67. En la evaluación temporal con entrenamiento en 2020-2023 y prueba en 2024, Sumapaz fue excluida completamente con el fin de analizar el desempeño sobre las 19 localidades restantes sin la influencia de sus particularidades territoriales y de registro. En consecuencia, ninguno de los modelos fue entrenado ni evaluado utilizando los valores atípicos previamente identificados.

Todos los modelos fueron desarrollados a partir del conjunto de datos depurado, sin incluir los registros anómalos de Sumapaz identificados durante el análisis exploratorio.

7.2.4 Transformación de Variables

Para facilitar el aprendizaje de los algoritmos de ML, algunas variables fueron transformadas antes del proceso de entrenamiento. Variables categóricas, como la localidad, fueron codificadas mediante técnicas apropiadas para su utilización por los diferentes modelos.

Adicionalmente, la variable bimestre fue transformada utilizando funciones trigonométricas seno y coseno, permitiendo representar adecuadamente su naturaleza cíclica y conservar la continuidad temporal entre el último y el primer período del año. Esta transformación evita que los algoritmos interpreten los bimestres como valores ordinales independientes y favorece la identificación de patrones estacionales en el consumo hídrico.

Asimismo, cuando el algoritmo lo requirió, se aplicaron procesos de normalización o escalamiento sobre las variables numéricas con el fin de mejorar la estabilidad durante el entrenamiento y facilitar la convergencia de los modelos, particularmente en aquellos sensibles a la magnitud de las variables.

7.3 Análisis Exploratorio de los Datos (EDA)

Una vez preparado y validado el conjunto de datos, se desarrolló el análisis exploratorio de los datos, con el propósito de identificar patrones, tendencias y relaciones presentes en la información antes del entrenamiento de los modelos predictivos. Esta etapa permitió comprender el comportamiento del consumo hídrico desde una perspectiva temporal, espacial y climática, así como explorar las asociaciones preliminares entre las variables predictoras y la variable objetivo.

Dicho análisis exploratorio tuvo como finalidad interpretar el comportamiento del fenómeno estudiado y generar evidencia para la selección de variables, el diseño de los modelos predictivos y la interpretación de sus resultados.

En esta sección se realiza una aproximación estadística y visual al conjunto de datos integrado (596 registros y 49 variables). El objetivo es entender el comportamiento de la demanda hídrica en Bogotá, identificar patrones espaciales y temporales, y evaluar la relación entre el consumo de agua y las variables hidroclimáticas, sirviendo como fundamento científico para la fase de modelado predictivo. Las principales actividades desarrolladas durante el EDA se resumen en la Tabla 12.

Tabla 12.

Actividades Desarrolladas Durante el Análisis Exploratorio de los Datos.

Actividad	Objetivo
Caracterización estadística de la muestra (Análisis Descriptivo)	Analizar las distribuciones y el comportamiento general de las variables del estudio.
Análisis detallado de la variable objetivo y métricas de demanda	Evaluar la evolución y las métricas clave de la variable de consumo hídrico.
Comportamiento del consumo frente a la dinámica de usuarios y facturación	Identificar tendencias del gasto de agua en relación con los suscriptores y facturas emitidas.
Análisis de variables climatológicas	Explorar la influencia de las condiciones meteorológicas sobre el consumo de agua.
Análisis de distribución espacial y caracterización por localidad	Examinar las diferencias territoriales e identificar anomalías geográficas en la demanda.
Análisis de Correlaciones	Evaluar el grado de asociación matemática e interdependencia entre todas las variables.

7.3.1 Caracterización Estadística de la Muestra (Análisis Descriptivo)

Resulta necesario comprender la naturaleza, el comportamiento y las anomalías de los datos recopilados entre 2020 y 2024. El conjunto de datos consolidado consta de 596 observaciones, correspondientes al seguimiento bimestral de las 20 localidades de Bogotá durante cinco años.

Los análisis temporales y espaciales presentados en los apartados siguientes, en particular los relativos al comportamiento atípico de Sumapaz en 2022 y 2024, se apoyan en la base previa a la depuración. Se conservan en el documento porque constituyen la evidencia que justificó la exclusión de esos cuatro registros; los estadísticos descriptivos y todos los modelos, en cambio, se calcularon sobre la base depurada de 596 observaciones.

7.3.2 Dimensión Comercial y de Consumo

Se Agrupa las variables que reflejan la demanda de agua y el tamaño del mercado facturado en la capital. Los estadísticos que se presentan corresponden a la base depurada de 596 registros, es decir, posterior a la exclusión de los cuatro registros anómalos de Sumapaz. Los datos revelan una dispersión considerable en los volúmenes agregados y una asimetría moderada hacia la derecha en el consumo promedio por suscriptor.

***Tabla 13*.**

Métricas de tendencia central y dispersión para el consumo y facturación de agua.

Variable	Media	Desviación Estándar (σ)	Mínimo	Mediana (50%)	Máximo
Metros Cúbicos m^3	2.344.162,77	1.848.705,37	38,00	1.846.355,00	7.319.036,00
Facturas	109.646,17	93.252,48	1,00	82.314,50	368.794,00

Consumo Promedio	11,71	3,01	5,08	10,85	38,41
Suscriptores Facturados	109.384,41	93.135,47	3,00	81.990,00	368.407,00

7.3.3 Análisis de la dispersión y "desfases" comerciales

7.3.3.1 Heterogeneidad entre localidades: la variable metros cúbicos presenta una media de 2.344.162,77 m³, una desviación estándar de 1.848.705,37 m³ y una mediana de 1.846.355,00 m³. Esta dispersión refleja diferencias en el volumen total facturado entre localidades, relacionadas con la cantidad de suscriptores y la escala territorial de cada unidad de análisis. Los valores se distribuyen entre un mínimo de 38 m³ y un máximo de 7.319.036 m³.

7.3.3.2 Distribución del consumo promedio: sobre la base depurada, la variable objetivo presenta una media de 11,71, una mediana de 10,85, un tercer cuartil de 11,78, un mínimo de 5,08 y un máximo de 38,41, con una desviación estándar de 3,01. Estos valores contrastan con los obtenidos antes de la depuración, donde la media alcanzaba 25,07 y el máximo 3.158,67 frente a una mediana de 10,87; la diferencia evidencia el peso desproporcionado que ejercían los cuatro registros anómalos sobre los estadísticos globales. La depuración resultó determinante porque dichas anomalías afectaban directamente la variable utilizada para entrenar y evaluar los modelos. Los registros válidos conservan una distribución asimétrica hacia la derecha, aunque mucho menos pronunciada, por lo que se aplicó la transformación log1p durante el entrenamiento y se utilizó expm1 para calcular las métricas en la escala original.

7.3.4 Dimensión Climática, Atmosférica y de Suelo (NASA POWER)

La cobertura geográfica del estudio abarca el Distrito Capital, delimitada estadísticamente en la muestra por los extremos de latitud [4.20, 4.76] y longitud [-74.22, -74.04]. Este rango espacial captura la diversidad territorial, donde el límite inferior de latitud (4.20) está determinado por la extensión rural de la localidad de Sumapaz en el extremo sur, mientras que el límite superior (4.76) corresponde a la frontera norte de las localidades urbanas de Suba y Usaquén. Por su parte, el eje longitudinal encierra desde la franja de los Cerros Orientales (Longitud -74.04) hasta la llanura aluvial del Río Bogotá en el occidente de la ciudad (Longitud -74.22)

A diferencia de las variables comerciales, las variables climáticas capturadas en la zona geográfica de Bogotá (Latitud: 4.20 a 4.76; Longitud: -74.22 a -74.04) muestran un comportamiento mucho más controlado, homogéneo y de baja dispersión, con algunas excepciones ligadas a la lluvia y el viento.

***Tabla 14*.**

Caracterización estadística de los parámetros climáticos y de variabilidad atmosférica.

Variable	Significado Técnico	Media	Mínimo	Mediana	Máximo
T2M (°C)	Temperatura del aire a 2 metros	18,71	14,38	18,91	22,86
T2M_MAX (°C)	Temperatura máxima del aire	25,79	19,80	25,84	32,66
PRECTOTCORR_SUM	Suma de precipitación corregida	151,38	32,88	141,59	430,23
GWETPROF	Humedad del suelo (perfil completo)	0,71	0,51	0,71	0,95
ONI	Índice Oceánico del Niño	-0,17	-1,23	-0,45	1,94

7.3.5 Análisis de las variables climáticas y su estacionalidad

7.3.5.1 Estabilidad Térmica: La temperatura media de la muestra (T2M) se sitúa en los 18.7 °C, con una desviación estándar muy baja (1.32 °C). No obstante, debe señalarse que el registro de temperatura máxima (`T2M_MAX`) llegó a tocar los 32.66 °C. Este pico térmico es atípico para la serie histórica de la Sabana de Bogotá y coincide con las temporadas secas intensas bajo la influencia de variaciones climáticas globales.

7.3.5.2 El desfase en las Lluvias (PRECTOTCORR_SUM): La precipitación acumulada por bimestre es, por naturaleza, la variable climática más inestable de la muestra. Registra un mínimo de 32.88 mm (periodo de sequía extrema) frente a un máximo de 430.23 mm (periodo de intensas lluvias). Esta amplitud de variación puede asociarse a cambios en los hábitos de consumo de agua y la disponibilidad del recurso en el suelo.

7.3.5.3 Humedad del Suelo y el factor macro (ONI): Los tres indicadores de humedad de suelo (GWETPROF, GWETROOT, GWETTOP) muestran un comportamiento consistente, con medias en torno al 0.67 y 0.71. Sin embargo, el índice ONI (Índice Oceánico del Niño) oscila fuertemente entre un mínimo de -1.23 (Fase extrema de La Niña / exceso de lluvias) y un máximo de 1.94 (Fase fuerte de El Niño / sequía severa). La presencia de un ONI de 1.94 en la muestra aporta un elemento de contexto relevante para interpretar la crisis hídrica que obligó al racionamiento en la capital en 2024.

7.3.6 Análisis detallado de la variable objetivo y métricas de demanda

Una vez caracterizadas las propiedades generales de la base de datos, se procedió a analizar el comportamiento de la variable objetivo (consumo_promedio), correspondiente al

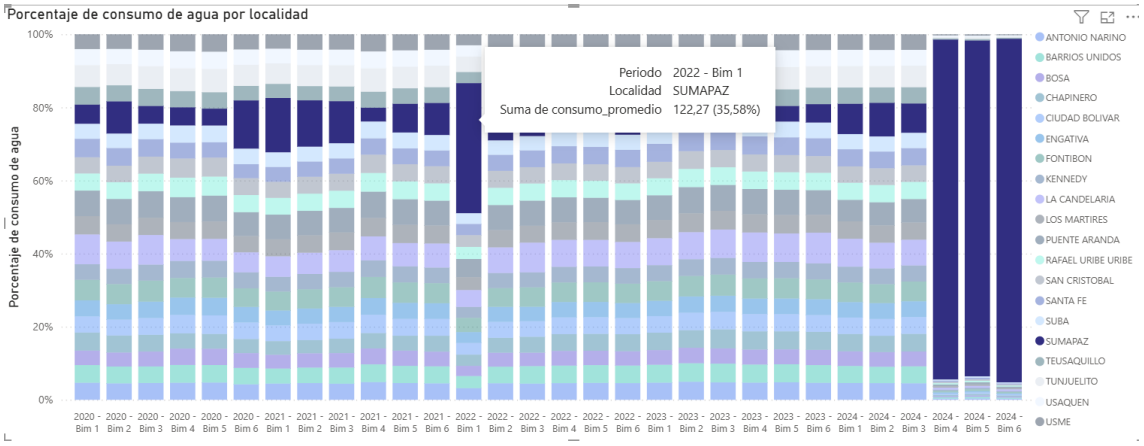
consumo hídrico promedio indexado, así como de su variable correlativa directa, metros_cubicos. Este examen no solo permite identificar la evolución de la demanda durante el período de estudio (2020–2024), sino también reconocer patrones estadísticos de dispersión, asimetría y anomalías que los algoritmos de ML deberán procesar posteriormente.

7.3.7 Análisis Temporal y Distributivo del Consumo Hídrico en Bogotá (2020-2024)

Al evaluar cronológicamente la suma del consumo promedio en Bogotá, se observan dinámicas que transitan de una relativa estabilidad estacional a quiebres estructurales en la serie histórica.

Figura 9.

Distribución Porcentual del Consumo Hídrico Indexado por Localidad en Bogotá (2020–2024).



Para comprender de qué manera se distribuye el gasto del recurso hídrico al interior del Distrito Capital y evaluar si las variaciones agregadas afectan por igual a todo el territorio, se estructuró un análisis de composición mediante un gráfico de barras 100% apiladas. La Figura 9 ilustra la participación porcentual con la que cada una de las 20 localidades aporta al consolidado de consumo bimestral de la ciudad durante el periodo analizado.

Esta visualización es clave para el desarrollo metodológico del estudio, ya que permite aislar el comportamiento histórico de las zonas urbanas frente a las dinámicas de la periferia, poniendo en evidencia los momentos exactos en que se fracturó la equidad distributiva de la demanda debido a choques climáticos o institucionales. A partir de este mapeo, se destacan de forma inmediata tres momentos clave:

7.3.7.1 Periodo de Estabilidad y Fluctuación Estacional (2020 - 2023): Durante los primeros cuatro años analizados, la suma del consumo promedio de agua en la ciudad mantuvo una tendencia mayoritariamente lineal y controlada, oscilando en un rango general entre los 213 y los 262 puntos de consumo.

7.3.7.2 El Impacto de la pandemia (2020): En el año 2020, coincidiendo con las medidas de confinamiento por COVID-19, el consumo inició en 247.4 (Bimestre 1) y experimentó un descenso progresivo hasta su punto más bajo en el Bimestre 5 (213.2). Esto sugiere que, a pesar del traslado de la población a las dinámicas de teletrabajo en los hogares, el cese total de actividades industriales, comerciales y educativas presenciales generó una reducción neta en la demanda agregada de la ciudad.

7.3.7.3 La regularidad entre 2021 y 2023: En este lapso, la ciudad recuperó su ritmo habitual. Se observan picos estacionales sistemáticos en los primeros bimestres de cada año (por ejemplo, el repunte a 258.6 en 2021-Bim 1). Estos incrementos responden comúnmente a temporadas de menores lluvias (periodos secos) y al retorno vacacional. Fuera de estos hitos, la distribución porcentual entre las localidades urbanas se mantuvo prácticamente idéntica Bimestre a Bimestre, demostrando un hábito de consumo rígido y predecible por parte de la ciudadanía.

7.3.7.4 El Comportamiento Atípico de Sumapaz en 2022: Dentro de este periodo de aparente normalidad, el análisis distributivo revela una anomalía estadística significativa en el Bimestre 1 de 2022. En este punto, la localidad de Sumapaz se aparta del patrón general de la gráfica y concentra el 35.58% del consumo promedio total indexado de la ciudad (equivalente a 122.27 unidades dentro del ecosistema de medición).

Al tratarse de una localidad eminentemente rural y de conservación ambiental, este pico no responde a dinámicas demográficas residenciales convencionales (como las de Suba o Kennedy), sino que plantea la hipótesis de variaciones en las actividades agropecuarias de la zona, necesidades de riego por sequía local o un ajuste excepcional en el registro de los macro-medidores de alta montaña durante el inicio de ese año.

7.3.7.5 La Anomalía Crítica de 2024: El comportamiento histórico y la estabilidad de los datos presentan un cambio abrupto en el año 2024. Tras mantener una línea base estable de 237.1 en el Bimestre 3 de 2024, el registro presenta un incremento de gran magnitud:

Bimestre 4 (2024): El consumo promedio asciende hasta los 2,867.5 puntos.

Bimestre 6 (2024): Tras un leve descenso en el quinto bimestre, la serie alcanza su valor máximo al cierre del año, con 3,300.5 puntos.

Este incremento coincide temporalmente con la crisis de desabastecimiento que afrontó Bogotá debido a los niveles críticos en el Sistema de Chingaza y la implementación de esquemas de racionamiento por turnos. El volumen registrado puede explicarse por fenómenos de sobreabastecimiento preventivo (almacenamiento excesivo de agua en los hogares en las horas previas a los cortes de servicio), o bien por un cambio técnico y metodológico en la indexación de los flujos de agua por parte de la empresa de acueducto ante la contingencia.

7.3.8 Centralización Geográfica de la Demanda en la Crisis

Al contrastar el incremento de la curva de 2024 con el gráfico de barras 100% apiladas, se observa que en los bimestres 4, 5 y 6 de 2024 la distribución por localidades se altera por completo: casi la totalidad de la representación gráfica del consumo de la ciudad (más del 90%) pasa a ser concentrada exclusivamente por la localidad de Sumapaz, y la participación del resto de las 19 localidades se reduce a franjas apenas perceptibles en la base de la gráfica.

Este comportamiento sugiere que el impacto de la crisis hídrica de 2024 y las medidas institucionales de restricción no afectaron el territorio por igual. Mientras las localidades urbanas registraron caídas pronunciadas en su consumo facturado y restricciones por el racionamiento, el flujo, la demanda o el registro hídrico se concentró en el registro del sector rural y de cuencas de la capital. Esto sugiere una desconexión entre el comportamiento del consumo residencial urbano y las necesidades de flujo/abastecimiento de la periferia rural en periodos de sequía extrema, marcando un hito metodológico para la formulación de futuras políticas de ordenamiento hídrico en el Distrito Capital.

7.3.9 Análisis de Distribución, Formas y Valores Atípicos (EDA Visual)

Para complementar la perspectiva temporal, se evalúa formalmente la morfología estadística de las variables de consumo mediante histogramas de frecuencia (con estimación de densidad de Kernel - KDE) y diagramas de caja (*boxplots*), revelando las asimetrías y los desfases inherentes al conjunto de datos.

7.3.9.1 Forma y Dispersión de la Variable Objetivo (consumo promedio): Al examinar el histograma de frecuencias y el *boxplot* de la variable consumo promedio, se evidencia un comportamiento de asimetría positiva extrema (sesgo a la derecha).

Figura 10.

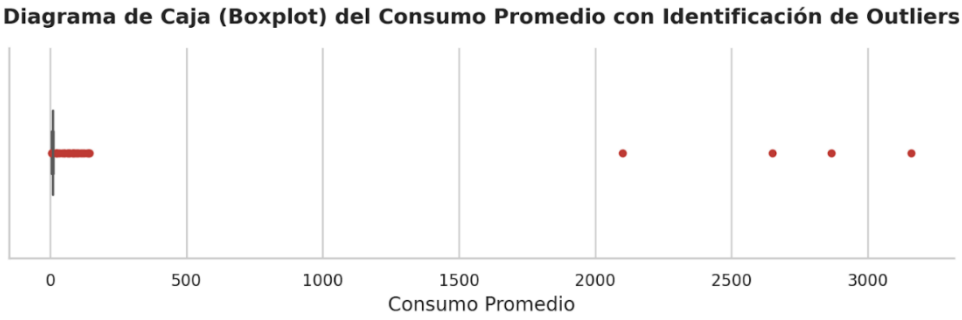
Distribución de frecuencias de la variable objetivo (consumo promedio) e índice de asimetría.



La Figura 10 describe el perfil probabilístico de la variable objetivo en la base previa a la depuración, es decir, antes de excluir los cuatro registros anómalos. El histograma revela una distribución con asimetría positiva extrema, donde la frecuencia absoluta se concentra en el intervalo inferior a las 15 unidades de consumo promedio. La caída de la curva de densidad indica que la dinámica habitual de la ciudad es homogénea y de baja magnitud. El gráfico expone una cola muy extendida a lo largo del eje horizontal, que evidencia la existencia de anomalías en los registros de la serie histórica y que motivó la depuración descrita en el apartado anterior.

Figura 11.

Diagrama de caja de la variable objetivo con identificación y contraste de valores atípicos severos.



El diagrama de caja de la variable consumo promedio permite dimensionar las distorsiones temporales y geográficas presentes en la base sin depurar. El cuerpo central de la distribución se encuentra completamente comprimido cerca del límite inferior, lo que indica que más del 75 % de los registros se sitúa en un rango estrecho de valores. Sin embargo, el análisis detecta una serie de valores atípicos severos (*outliers*), resaltados estratégicamente en color rojo, que fracturan el supuesto de normalidad geométrica. Estos puntos aislados, que se sitúan en los rangos de 2,100, 2,600 y más de 3,100 unidades, corresponden de manera exacta a los choques externos identificados en la periferia rural de Sumapaz durante la crisis hídrica de 2024, imponiendo la necesidad de aplicar transformaciones robustas no lineales en la etapa de ingeniería de características.

La gran mayoría de los registros se concentra en una franja muy estrecha, por debajo del umbral de las 12 unidades, y la curva KDE desciende de forma abrupta, lo que evidencia que el consumo promedio habitual en Bogotá es de magnitud baja y estable.

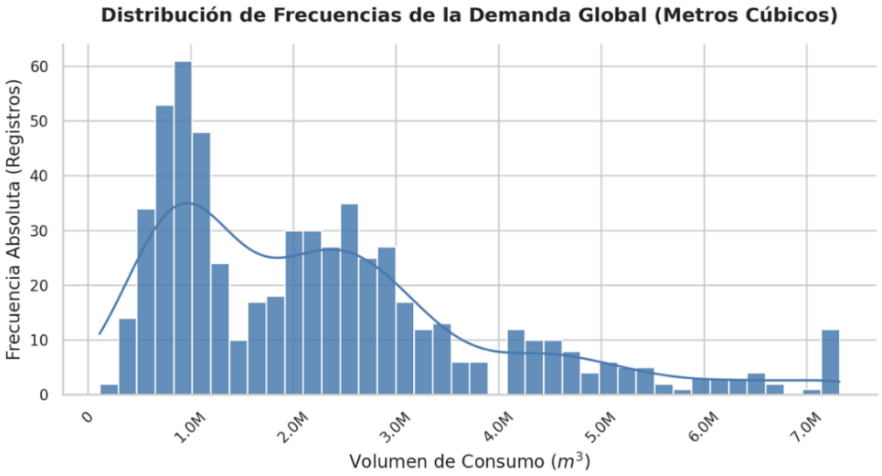
Sin embargo, tanto la cola extendida del histograma como los puntos aislados del boxplot exponen la presencia de registros atípicos situados en los rangos de 2.100, 2.600 y por encima de 3.100 unidades. Que la mediana se ubicara en 10,87 y el máximo alcanzara 3.158,67 confirmaba una distribución fuertemente no gaussiana, atribuible por completo a cuatro observaciones de Sumapaz. Tras su exclusión, el máximo de la variable objetivo se reduce a 38,41 y la distribución conserva una asimetría moderada hacia la derecha, por lo que se mantuvo la transformación logarítmica log1p durante el entrenamiento con el fin de estabilizar la varianza y evitar que los valores más altos dominaran la penalización del error cuadrático.

7.3.9.2 Comportamiento Estructural de la Demanda Global (metros cúbicos):

Un comportamiento distinto se observa al analizar la variable metros cúbicos, la cual cuantifica el volumen neto de agua consumido por localidad.

Figura 12.

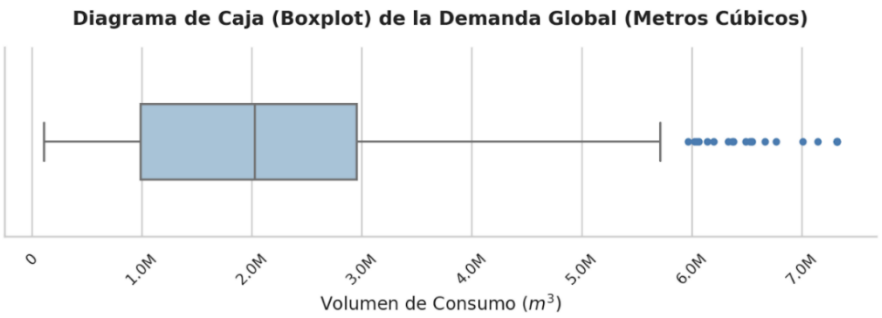
Distribución de frecuencias y curva de densidad del volumen global de consumo hídrico en m^3.



La Figura 12 presenta la distribución de frecuencias absolutas para la variable metros cúbicos a lo largo del periodo de estudio. A través del histograma y la curva de estimación de densidad de Kernel (KDE), se identifica un comportamiento marcadamente bimodal. El primer pico de concentración se ubica entre los 0.5M y 1.2M de m^3, mientras que el segundo segmento de alta frecuencia se consolida en el rango de 1.6M a 2.1M de m^3. Esta morfología bimodal evidencia la segmentación estructural de la demanda en el Distrito Capital, reflejando la coexistencia de localidades con baja densidad de usuarios frente a macro-localidades de alto volumen poblacional. Asimismo, se observa una cola extendida hacia el extremo derecho que denota registros atípicos que superan los 6.0M de m^3.

Figura 13.

Diagrama de caja del volumen de consumo global en m3 y dispersión intercuartílica.



El diagrama de caja permite evaluar la dispersión y los umbrales cuartílicos del volumen neto de agua consumido. El rango intercuartílico (RIC), representado por el cuerpo de la caja azul, exhibe una amplitud considerable con una mediana que se posiciona de forma equilibrada cerca de los 2.0M de m3, lo que confirma la variabilidad esperada debido al tamaño demográfico de las divisiones administrativas de la ciudad. El bigote superior se extiende de manera estable hasta el umbral de los 5.7M de m3. Más allá de este límite, se identifican valores atípicos moderados (puntos azules aislados) que alcanzan un máximo de 7.3M de m3, los cuales corresponden a bimestres de alta demanda en las localidades con mayor cantidad de cuentas de servicio activas.

A diferencia del consumo promedio, el volumen en m^3 muestra una distribución de carácter bimodal y mucho más balanceada. El histograma exhibe claramente dos grandes crestas o concentraciones de frecuencia principales: un primer bloque prominente entre los 0.5 y 1.2 millones de m^3, y un segundo pico muy marcado alrededor de los 1.6 a 2.1 millones de m^3.

Esta morfología bimodal refleja la segmentación demográfica y estructural de Bogotá. La ciudad está dividida entre un grupo de localidades geográficamente pequeñas o con menor densidad de usuarios facturados, y otro grupo de "macro-localidades" (como Suba, Kennedy o Engativá) que, por su enorme volumen poblacional, arrastran de forma natural la demanda global hacia magnitudes superiores.

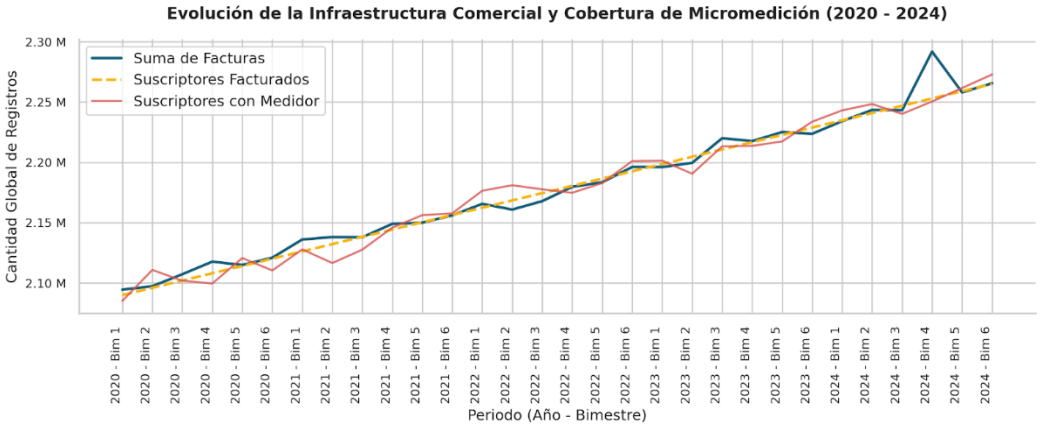
El *boxplot* de metros_cubicos ratifica esta lectura: el cuerpo de la caja es amplio, abarcando desde los 0.86 millones de m^3 (Q1) hasta los 3.17 millones de m^3 (Q3), con una mediana equilibrada de 1.84 millones de m^3. Aunque se observan algunos puntos atípicos sobrepasando los 6.5 millones de m^3 (llegando al máximo de 7.3 millones de m^3), estos *outliers* son moderados y esperables dentro de la dinámica de crecimiento urbano, situándose a una distancia estadística razonable del bigote superior.

7.3.9.3 Comportamiento del consumo frente a la dinámica de usuarios y facturación: Para comprender la naturaleza estructural de los datos y contextualizar el comportamiento de las variables de consumo analizadas previamente, resulta necesario examinar la evolución histórica de la base operativa del servicio. A continuación, se evalúa la trayectoria temporal de la emisión de facturas y la cobertura de suscriptores (tanto facturados como micromedidos) frente al comportamiento del consumo promedio indexado entre los años 2020 y 2024.

7.3.9.4 Crecimiento Vegetativo de la Base de Usuarios y Capacidad de Medición: El comportamiento de las variables operativas fundamentales —Suma de facturas, Suma de suscriptores facturados y Suma de suscriptores con medidor— exhibe una tendencia macroeconómica monótona creciente y altamente cohesionada a lo largo del quinquenio analizado. Al inicio del periodo (primer bimestre de 2020), la base de usuarios partió de un umbral aproximado de 2.09 millones de suscriptores, expandiéndose de forma sostenida hasta alcanzar cerca de 2.27 millones hacia el cierre de 2024.

Figura 14.

Evolución histórica de la base de usuarios facturados, emisión global de facturas y cobertura de dispositivos de micromedición (2020 - 2024).

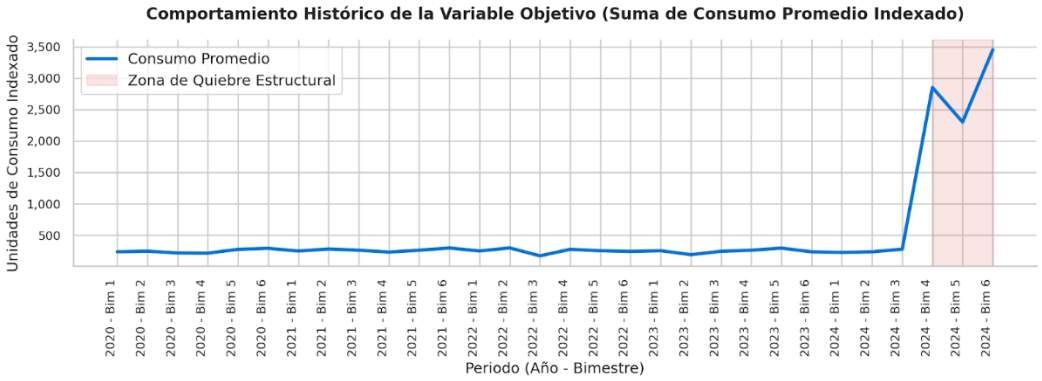


Esta gráfica ilustra el comportamiento longitudinal de la infraestructura operativa y comercial de la entidad prestadora del servicio durante el quinquenio analizado. A nivel macroeconómico, se observa un crecimiento vegetativo y continuo que expande la red desde aproximadamente 2.09 millones de usuarios en el primer bimestre de 2020 hasta consolidarse cerca de los 2.27 millones al cierre de 2024. La alta cohesión y paralelismo entre la curva de suscriptores facturados y la emisión global de facturas valida la consistencia interna del sistema de información comercial. Asimismo, las leves fluctuaciones observadas en la trayectoria de suscriptores con medidor (particularmente la aceleración en el primer semestre de 2022) aportan evidencia sobre campañas institucionales orientadas a la regularización y renovación de la cobertura de micromedición, logrando una convergencia del 1:1 con la base de usuarios activos para los periodos más recientes.

Este incremento continuo responde al crecimiento demográfico y urbanístico inercial de la ciudad, manteniendo una alta correspondencia entre las tres métricas. Un hallazgo relevante en esta serie es la convergencia progresiva entre el número de suscriptores facturados y aquellos que cuentan con un dispositivo de micromedición técnica. Las oscilaciones menores en la curva de suscriptores con medidor (como los ligeros escalones observados entre finales de 2021 y principios de 2022) sugieren campañas institucionales de regularización o renovación de parques de medidores, logrando estabilizar la cobertura de medición real en una relación directa de 1:1 respecto a la base total facturada para los últimos periodos.

Figura 15.

Trayectoria temporal de la variable objetivo e identificación visual del choque estructural.



La serie temporal de la variable de consumo promedio indexado revela un quiebre estructural de gran magnitud hacia el final del periodo de estudio. Entre el primer bimestre de 2020 y el tercer bimestre de 2024, el comportamiento de la demanda agregada se mantiene en un estado de equilibrio estacionario estricto, operando de manera plana y predecible. No obstante, a partir del cuarto bimestre de 2024, se registra un cambio abrupto en el que la variable asciende hasta valores que superan las 3,000 unidades indexadas. Al contrastar de forma directa este gráfico con el comportamiento de la infraestructura comercial de la Figura 14, se comprueba que el incremento no guarda relación con una expansión de la base de usuarios o errores masivos de facturación, ratificando analíticamente que el fenómeno responde a una anomalía exógena y severamente focalizada en puntos específicos de la red de distribución.

7.3.9.5 Contraste Temporal y Evidencia del Choque Estructural en el Consumo: Al correlacionar visualmente la expansión de la infraestructura de usuarios con el comportamiento de la variable objetivo (consumo_promedio), se hace evidente una desconexión analítica crítica que justifica los hallazgos de los histogramas y diagramas de caja previos:

7.3.9.6 Periodo de Estabilidad Estacionaria (2020 - Bimestre 3, 2024): Durante más de cuatro años, a pesar de que la base de suscriptores incorporó a más de 150,000 nuevos usuarios a la red, el volumen del consumo promedio indexado permaneció en un estado de equilibrio estricto, plano y cercano a la línea base de cero anomalías estructurales. Esto indica que la demanda global de la ciudad se mantuvo estable y sujeta a dinámicas de consumo predecibles y ordinarias.

7.3.9.7 Punto de Inflexión y Ruptura Estructural (Bimestre 4, 2024 en adelante): A partir del cuarto bimestre de 2024, la curva de consumo promedio presenta un quiebre pronunciado y asciende hasta superar el umbral de las 3,000 unidades en su pico más alto (Bimestre 6, 2024).

Dado que la primera gráfica evidencia que el número de facturas y suscriptores no sufrió ninguna alteración, salto o decremento anómalo en esos mismos bimestres de 2024, se descarta que el pico del consumo promedio sea el resultado de un error de agregación generalizado, una duplicidad en la base de datos o un crecimiento repentino de usuarios en el sistema.

Por lo tanto, este comportamiento sugiere que el fenómeno observado responde a un factor externo y focalizado: un cambio pronunciado en los patrones de consumo o fallas de medición en puntos específicos de la red (como el caso de la periferia de Sumapaz), donde un grupo reducido de registros alteró de forma desproporcionada los estadísticos descriptivos globales del sistema hídrico sin modificar la infraestructura comercial de la entidad.

7.3.10 Análisis de Variables Climatológicas y su relación con la Demanda Hídrica

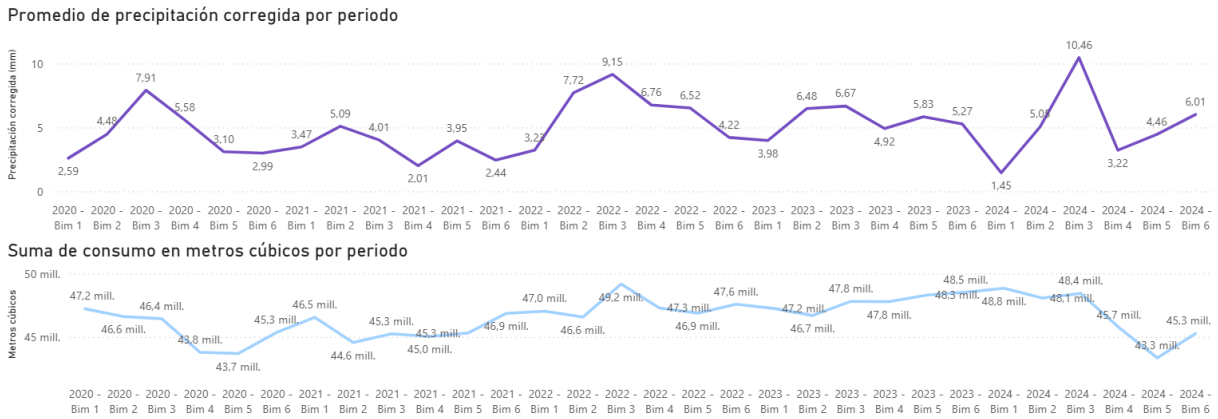
Para robustecer el alcance analítico del modelo predictivo, resulta pertinente evaluar la dinámica exógena del entorno. Las variaciones en el consumo de agua potable no solo responden a dinámicas de crecimiento demográfico, sino también a las fluctuaciones en las

condiciones atmosféricas que modifican el comportamiento de la población y el estado de la infraestructura. A continuación, se examinan de forma aislada e interconectada los seis rubros climáticos recopilados frente a la evolución del consumo hídrico global (metros cúbicos) y unitario (consumo promedio).

7.3.10.1 Régimen de Precipitación y Disponibilidad Hídrica: El comportamiento de las lluvias representa una de las fuerzas más dinámicas al evaluar la demanda hídrica urbana y rural, dado que mitiga o incrementa de forma directa las necesidades de riego y el abastecimiento superficial.

Figura 16.

Evolución bimestral del promedio de precipitación corregida frente al volumen neto de consumo hídrico en metros cúbicos (2020 - 2024).



La Figura 16 presenta el comportamiento del régimen de lluvias indexado en contraste con el volumen de agua distribuido. Se identifica una estructura cíclica estacional bimodal, típica de la región andina colombiana, donde los periodos de menor precipitación coinciden con repuntes significativos en la curva de metros cúbicos globales, la cual alcanza un máximo histórico intermedio cercano a los 49 millones de m^3 en 2022-Bim 3 ante un marcado descenso

de lluvias. Sin embargo, al observar el cierre de 2024, la precipitación muestra una recuperación moderada hacia el bimestre 6; un escenario que contradice el comportamiento de la variable objetivo de consumo promedio, sugiriendo que el pico extremo finalizado ese año no posee una raíz netamente pluviométrica.

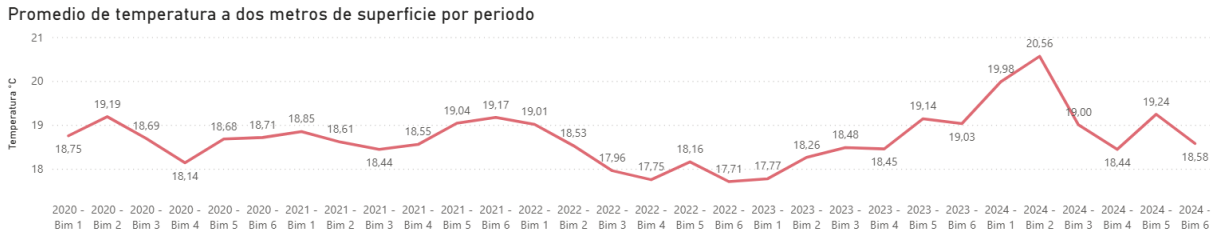
El análisis longitudinal de la precipitación revela que la demanda de la red responde a patrones estacionales bien definidos durante los primeros cuatro años del estudio. Existe una correlación inversa tradicional: cuando el promedio de precipitación desciende, la suma de metros cúbicos consumidos por la red tiende a incrementarse debido a los requerimientos térmicos y de conservación urbana.

El punto de mayor interés metodológico ocurre entre el periodo 2024-Bim 1 y 2024-Bim 6. Durante el primer trimestre de 2024 se observa una reducción pronunciada de la precipitación, alcanzando su punto más bajo en toda la serie temporal (cercano al valor de 1.5 unidades), lo que coincide con los impactos del fenómeno de El Niño a nivel nacional. Este periodo de baja precipitación coincidió con un descenso sostenido en la suma global de metros cúbicos hacia finales de año (llegando a su mínimo histórico por debajo de los 44 millones de m3 en 2024-Bim 5). No obstante, es crítico notar que mientras los metros cúbicos globales disminuían por las restricciones hídricas, la variable consumo promedio presentó un incremento de magnitud excepcional. Esto sugiere que el pico de consumo promedio indexado de fines de 2024 no fue causado por un incremento generalizado del consumo ciudadano ante la sequía, sino por un fallo o distorsión localizada en la medición de sectores específicos.

7.3.10.2 Perfil de Temperatura y Confort Térmico: La temperatura del aire ejerce un control directo sobre los patrones de confort de la población y regula la demanda de agua para actividades cotidianas, refrigeración e higiene personal.

Figura 17.

Variación del promedio de temperatura a 2 metros de superficie.



La Figura 17 describe la fluctuación de la temperatura media ambiental a lo largo de la serie histórica. Se observa un comportamiento relativamente homogéneo con una media oscilante entre los 18°C y 19.5°C durante la mayor parte del periodo. La variación más pronunciada se registra en el primer semestre de 2024 (Bimestres 1 y 2), donde la temperatura rompe su barrera histórica superando los 20.5°C. Este incremento anómalo precede directamente a la caída del volumen global de metros cúbicos y al posterior disparo de la variable objetivo de consumo indexado, evidenciando un desfase temporal entre el estímulo meteorológico y el registro de los sistemas de facturación.

El comportamiento de la temperatura a 2 metros muestra una estabilidad climática relativa entre 2020 y 2023, caracterizada por micro-oscilaciones anuales que no superaban un rango de 1.5°C. Esta estabilidad térmica es consistente con el hecho de que el consumo promedio indexado permaneció plano y sin alteraciones estructurales durante los cuatro años.

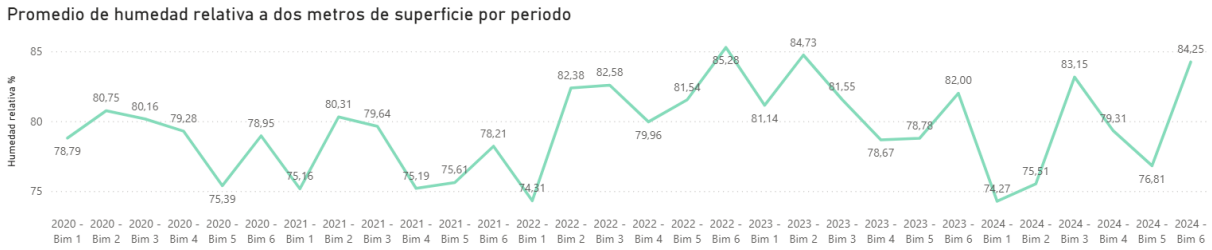
La ruptura del patrón climático ocurre de forma nítida a inicios de 2024. El repunte térmico que alcanza los 20.5°C en 2024-Bim 2 coincide con el calentamiento anómalo asociado a la crisis climática del periodo. En términos de ingeniería de características (*feature engineering*), este incremento de temperatura actúa como un predictor de alerta temprana. Aunque la suma global de metros cúbicos se contrajo en los meses posteriores debido a políticas de racionamiento o pérdida de presión en la red (explicando el valle en la curva inferior en 2024-Bim 5), las presiones por desabastecimiento en zonas rurales periurbanas provocaron consumos descontrolados u

anomalías de medición en parches críticos de la periferia, lo que podría relacionarse con el desfase del consumo promedio observado en los bimestres subsiguientes.

7.3.10.3 Dinámica de la Humedad Relativa y Evapotranspiración: La humedad relativa complementa el análisis térmico al actuar como el factor determinante en la sensación de confort y en la tasa de evaporación del agua superficial.

Figura 18.

Comportamiento del promedio de humedad relativa e impacto en los niveles de consumo agregados.



Esta representación expone la dinámica de la humedad relativa del aire en escala porcentual. El perfil exhibe un comportamiento inversamente simétrico respecto a la temperatura; los periodos de mayor sequedad atmosférica se ubican a mediados de 2020 y a inicios de 2024, registrando mínimos cercanos al 74%. El gráfico muestra que las reducciones pronunciadas de humedad relativa preceden a los cambios en el comportamiento de la red, actuando como un indicador inverso del estrés hídrico ambiental en el área de estudio.

La humedad relativa presenta un comportamiento inverso al de la temperatura dentro del conjunto de datos analizado. Durante el período 2020-2023, esta variable registró valores elevados, oscilando entre el 77 % y un máximo del 85 % en el bimestre 6 de 2022, correspondiente a una temporada de mayor precipitación. Estas condiciones reflejan una alta disponibilidad de vapor de agua en la atmósfera y constituyen un factor climático potencialmente asociado al comportamiento del consumo hídrico.

El desplome de la humedad relativa al 74% en el primer bimestre de 2024 confirma la severidad de la temporada seca que experimentó la región. Un ambiente con baja humedad acelera la deshidratación de coberturas vegetales y eleva la demanda instantánea de agua. Estadísticamente, la humedad relativa aporta un valor de varianza clave para los modelos predictivos: su marcado incremento hacia el final de 2024 (alcanzando cerca del 84% en 2024-Bim 6) indica que la atmósfera ya se había saturado de humedad debido al regreso de las lluvias, lo que no respalda la hipótesis de que el pico de 3.4 mil unidades en el consumo promedio hídrico final se debiera a una sequía persistente en ese bimestre.

7.3.10.4 Balance Energético mediante Radiación Solar: La radiación solar dicta la cantidad de energía directa que recibe el territorio, incidiendo sin intermediarios sobre la evaporación de los cuerpos de agua y el comportamiento térmico general.

Figura 19.

Evolución temporal de los niveles de radiación promedio y su contraste con las variables de control hídrico.



La ilustración mapea los niveles de radiación promedio capturados por las estaciones meteorológicas. A diferencia de las variables térmicas, la radiación presenta una alta frecuencia de oscilación a corto plazo, con un rango que fluctúa de manera constante entre las 14 y las 19 unidades. Destaca el pico de alta radiación registrado a comienzos de 2024, el cual coincide con

los picos de temperatura y los valles de humedad analizados con anterioridad, cerrando el perfil del choque climático de dicho año.

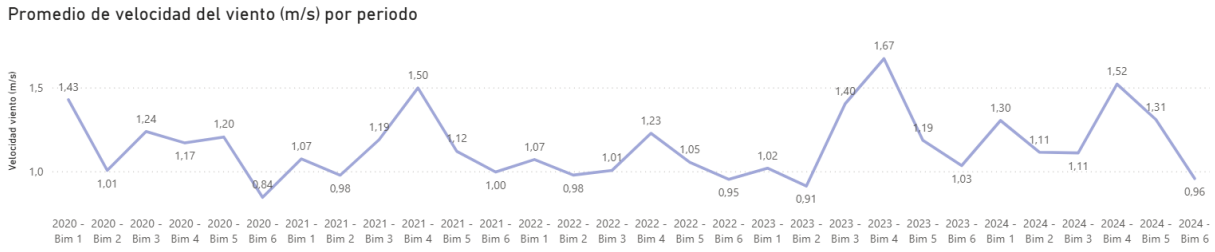
La radiación promedio permite entender el balance energético del modelo. Durante el histórico de la serie, se observan picos recurrentes que superan las 18 unidades (como en 2020-Bim 1 y 2020-Bim 5), los cuales coinciden con fluctuaciones normales y controladas en la curva global de metros cúbicos.

En el año 2024, la radiación describe un comportamiento en "M": presenta un primer pico elevado a inicios de año (coincidente con el aumento de temperatura y la reducción de la humedad), desciende a mitad de año y vuelve a presentar un repunte considerable en 2024-Bim 5 (alrededor de 18 unidades). Este segundo pico de radiación es clave; coincide exactamente con el punto más bajo en la disponibilidad de agua global (metros cúbicos en su nivel mínimo de 43.5 millones). La energía solar extra en un momento donde el volumen global de agua en la red venía disminuyendo coincide con el estrés del sistema de distribución, aislando el impacto del consumo en las terminales de medición vulnerables.

7.3.10.5 Dinámica de la Velocidad del Viento: El viento actúa como un elemento de transporte mecánico que acelera las tasas de evaporación superficial y desplaza las masas de humedad, alterando el microclima local.

Figura 20.

Comportamiento de la suma de velocidad del viento y su relación con las fluctuaciones de la demanda.



Esta gráfica detalla el comportamiento dinámico de la velocidad del viento a lo largo de los periodos evaluados. La variable presenta un patrón estacional nítido, caracterizado por picos de alta velocidad que rozan o superan las 30 unidades de manera recurrente a mitad de cada año (Bimestres 3 y 4). Se evidencia que las rachas de viento sostenidas coinciden con los periodos de transición climática de la ciudad, actuando como un factor de distribución ambiental regular.

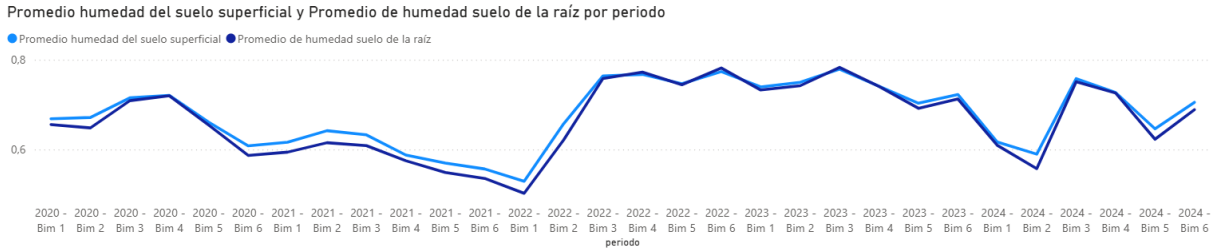
La velocidad del viento muestra el comportamiento más regular e inercial de todo el conjunto de variables climáticas. Los picos observados en 2020-Bim 1, 2021-Bim 4, y de mayor magnitud en 2023-Bim 4 (máximo de la serie, superando las 33 unidades) indican que el régimen de vientos de la región responde a ciclos geográficos fijos de la atmósfera alta y baja.

Cuando se cruza esta métrica con el comportamiento del consumo, se comprueba que el viento opera de manera independiente a las crisis del sistema de agua. Hacia finales de 2024, mientras el consumo promedio indexado presenta su quiebre, la velocidad del viento desciende de forma suave y predecible hacia las 19 unidades en 2024-Bim 6. Esta independencia estadística resulta relevante para los algoritmos de ML, ya que confirma que la velocidad del viento funciona como una variable de control o ruido ambiental regular, carente de correlación con las anomalías críticas del consumo hídrico de la urbe.

7.3.10.6 Estado Edáfico: Humedad del Suelo Superficial y Profundo: La humedad del suelo permite caracterizar las condiciones hídricas del territorio al integrar los efectos de la precipitación, la temperatura y la evaporación. Como se observa en la figura, la humedad superficial y la humedad en la zona de la raíz presentan comportamientos estrechamente relacionados durante el período analizado, con descensos importantes hacia 2022 y comienzos de 2024.

Figura 21.

Dinámica comparativa de la humedad del suelo y su correlación estructural con el consumo hídrico.



Durante 2022 se observa una reducción marcada de la humedad del suelo, coincidente con un incremento del volumen de agua consumido en la ciudad, que alcanzó aproximadamente 49 millones de m³. A comienzos de 2024 se presenta nuevamente una disminución de ambas variables edáficas, temporalmente coincidente con el período de mayor estrés hídrico registrado en Bogotá.

Hacia el sexto bimestre de 2024, ambas variables muestran una recuperación asociada al aumento de las precipitaciones. Sin embargo, esta recuperación contrasta con el incremento del consumo_promedio observado al final de la serie. Esta divergencia sugiere que dicho aumento no puede atribuirse únicamente a las condiciones de humedad del suelo, por lo que deben considerarse otros factores o posibles particularidades de los registros.

7.3.11 Análisis de Distribución Espacial y Caracterización por Localidad

El análisis exploratorio de datos requiere un examen de la dimensión geográfica para identificar patrones de concentración, disparidades en el servicio y posibles anomalías en el territorio. En este apartado se evalúa la distribución espacial de las variables comerciales (suscriptores y facturación) y operativas (metros_cubicos y consumo_promedio), cruzándolas con los perfiles microclimáticos de las distintas localidades de la ciudad.

7.3.11.1 Asimetría entre la Base Operativa y el Consumo Agregado: El primer hallazgo del análisis espacial es la disparidad cuantitativa entre la concentración de la infraestructura y el volumen total de agua facturada en el territorio.

Figura 22.

Análisis comparativo por localidad: volumen neto distribuido m^3

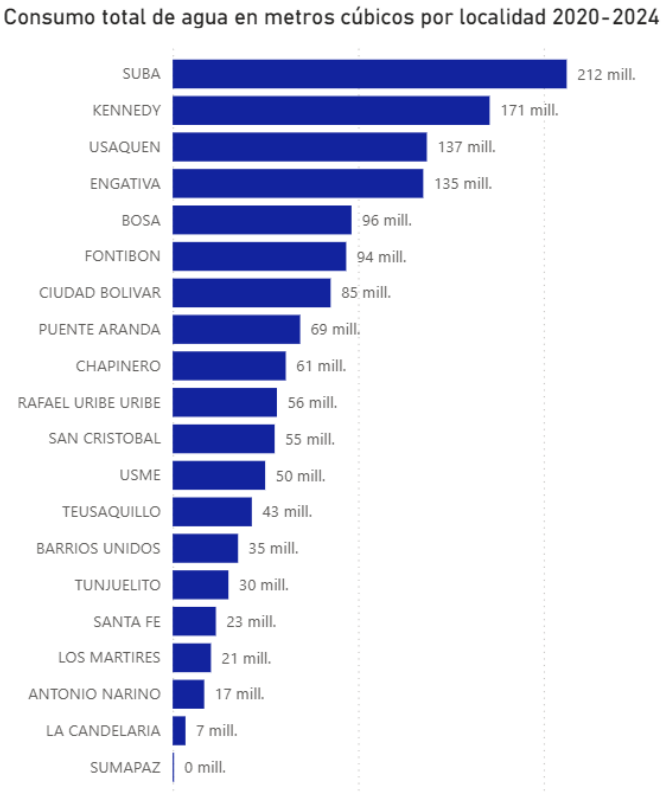
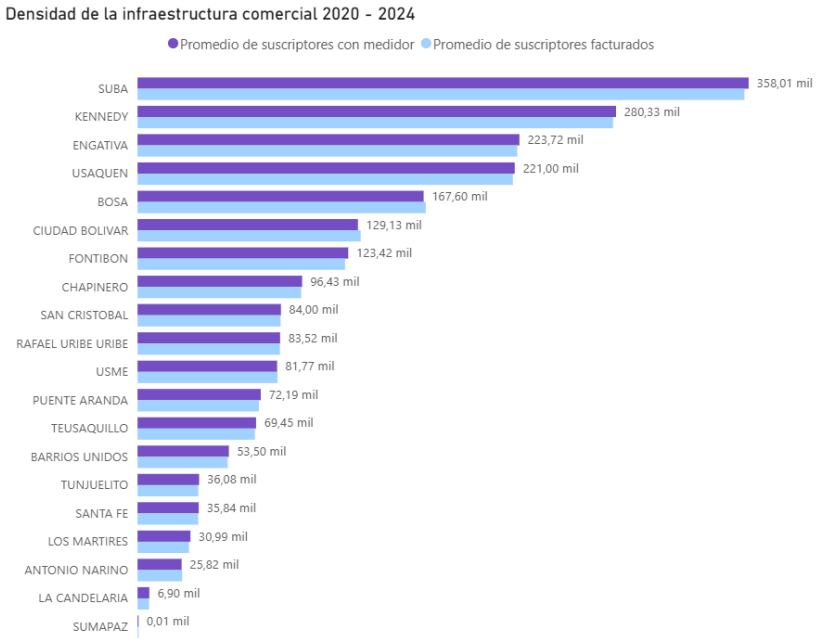


Figura 23.

Análisis comparativo por localidad: densidad de la infraestructura comercial.



Al evaluar las variables de volumen y de infraestructura comercial (Figura 22 y 23), se evidencia que la demanda agregada de la ciudad está fuertemente concentrada en un grupo reducido de nodos urbanos de alta densidad socioeconómica y residencial. Las localidades de Suba, Kennedy, Usaquén y Engativá lideran de manera unánime el sistema hídrico:

Suba se consolida como el principal núcleo operativo, superando los 212 millones de metros cúbicos globales y concentrando la mayor barra de suscriptores facturados, suscriptores con medidor y emisión de facturas de toda la ciudad.

Kennedy y Usaquén se ubican inmediatamente después, manteniendo una proporcionalidad casi exacta (relación 1:1) entre el tamaño de su infraestructura y los metros cúbicos que demandan de la red.

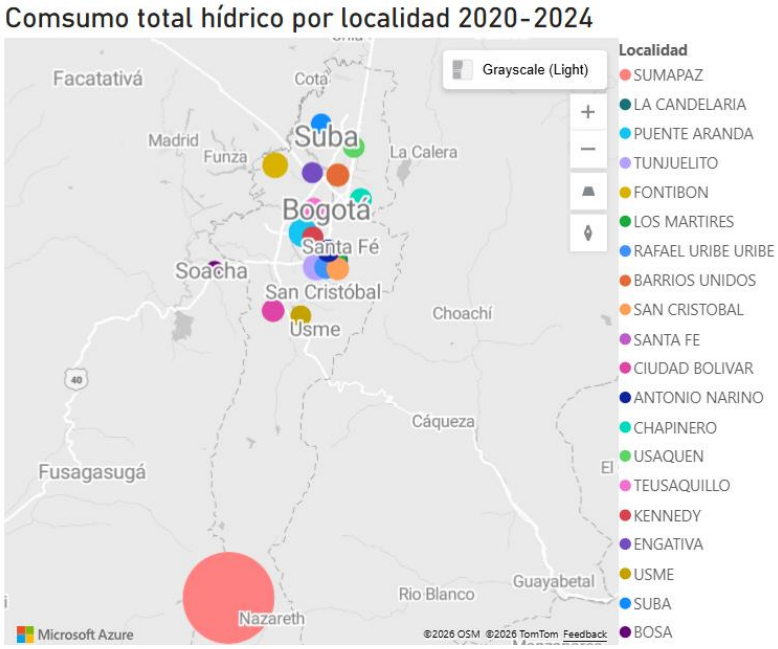
En el extremo opuesto se sitúan zonas como La Candelaria y Sumapaz, las cuales registran los valores más bajos en la suma de metros cúbicos globales y cuentan con una base de usuarios y facturas que geométricamente se aproxima a cero en comparación con los macro-

distritos del norte y el suroccidente. Hasta este punto, el sistema comercial exhibe un comportamiento lógico y predecible: a mayor número de suscriptores y facturas emitidas, mayor es el volumen neto de metros cúbicos inyectados y facturados en la localidad.

7.3.11.2 La Paradoja de Sumapaz: Análisis de la Variable Objetivo: Al introducir en el análisis la variable objetivo calculada (Promedio de consumo_promedio por Localidad), el patrón de proporcionalidad urbana se rompe de manera radical, revelando una anomalía estadística de gran interés metodológico.

Figura 24.

Distribución geográfica de la variable objetivo y contraste del consumo promedio indexado por localidad.



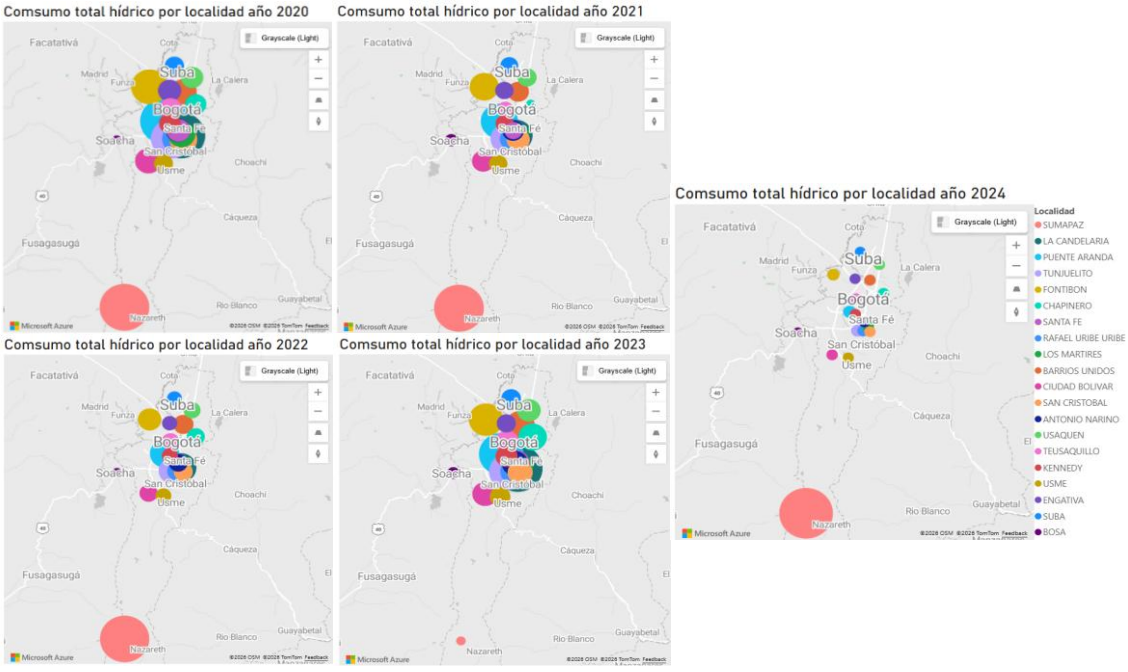
El gráfico central de la Figura 23 y el detalle analítico de la Figura 24 exponen una contradicción evidente: Sumapaz, la localidad con menor cantidad de usuarios, menor número de facturas y el volumen de m^3 agregados más bajo de la serie, registra un promedio de consumo indexado exponencialmente más alto que el resto de la ciudad, fijando un valor atípico exacto de 284.44 unidades.

Para dimensionar la magnitud de esta anomalía, basta con contrastar este número frente a las localidades de mayor consumo agregado: mientras que en Suba y Bosa el promedio de consumo indexado por registro se sitúa en los niveles mínimos de la escala (9.95 y 9.50 unidades respectivamente), el indicador en Sumapaz es casi 30 veces superior.

Dado que ya se demostró en los capítulos previos que la base de facturas de Sumapaz es marginal, este fenómeno no puede atribuirse a un gasto real masivo de agua por parte de la población residente. Estadísticamente, este comportamiento es el resultado de un efecto de distorsión por agregación (efecto escala) o de capturas anómalas en parches de medición específicos. Al tener muy pocos suscriptores registrados, cualquier error de lectura, fuga no detectada en nodos de control de la periferia o alteración en los sistemas de indexación de la base de datos altera de forma desproporcionada el promedio local, disparando el indicador hacia el extremo superior y aislando a esta región como el principal foco de ruido atípico (*outlier* espacial) que el modelo predictivo deberá tratar con especial cuidado.

Figura 25.

Evolución de la distribución espacial del consumo hídrico por localidad en Bogotá (Periodo 2020-2024).



7.3.12 Análisis de la distribución espacial del consumo hídrico por localidad

La dimensión espacial de la demanda de agua en Bogotá revela patrones territoriales asimétricos y expone singularidades en el registro de datos que requieren una interpretación crítica y contextualizada. Al cartografiar el consumo total indexado por localidades a lo largo del quinquenio analizado (Figura 25), sobresalen dos fenómenos de naturaleza opuesta: la relativa estabilidad del denso núcleo urbano del norte, occidente y suroccidente de la ciudad, y la recurrente anomalía de consumo sobredimensionado observada en el extremo sur de la capital.

7.3.13 Comportamiento del Casco Urbano Consolidado

Entre 2020 y 2023, Suba, Kennedy, Engativá y Bosa presentaron patrones de consumo relativamente estables, asociados con la concentración de usuarios residenciales y la continuidad de la demanda en zonas urbanas consolidadas. En 2024 se observó una reducción generalizada en el tamaño de las burbujas, coincidente con las medidas de racionamiento y las campañas institucionales de ahorro implementadas durante la crisis hídrica de Bogotá.

En contraste, Sumapaz, particularmente el nodo de Nazareth, presentó un comportamiento atípico, con valores elevados y variaciones considerables entre años, pese a su carácter rural y baja densidad poblacional. Estas diferencias podrían estar relacionadas con procesos de agregación o atribución de registros, particularidades en los sistemas de medición de acueductos rurales o comunitarios y características del uso agropecuario del agua, susceptible a variaciones climáticas.

Estas explicaciones se consideran hipótesis interpretativas y no relaciones comprobadas. Por ello, los registros de Sumapaz recibieron un tratamiento particular durante el análisis y preparación de los datos, con el fin de reducir posibles distorsiones en la evaluación y generalización de los modelos predictivos.

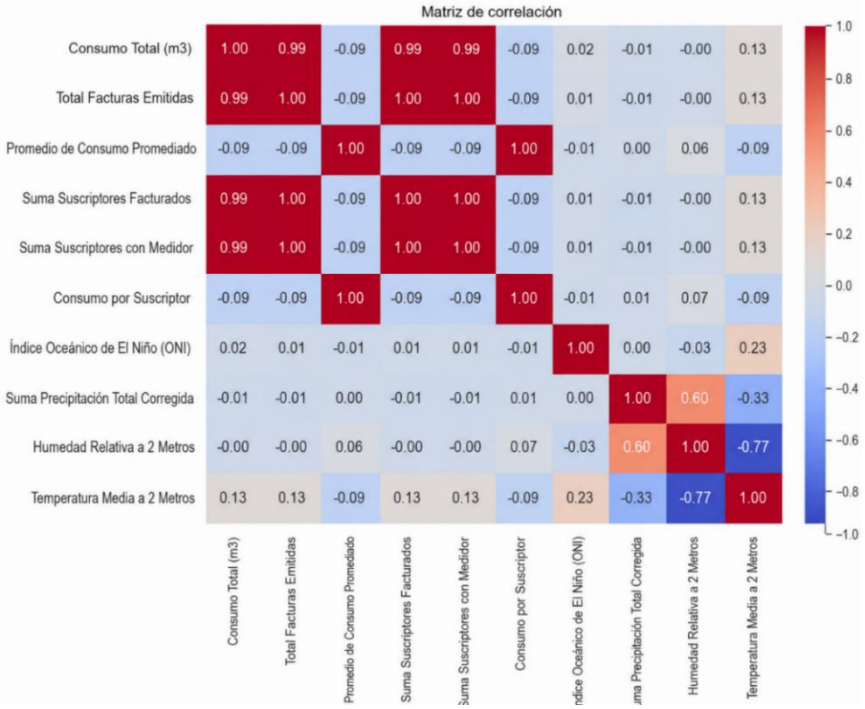
7.3.14 Análisis de Correlaciones y Dependencias Lineales Multi-Variable

Una vez caracterizadas las dimensiones temporales, climáticas y geoespaciales de forma aislada, resulta necesario evaluar las relaciones de interdependencia matemática mediante el cálculo de coeficientes de correlación de Pearson. Este procedimiento permite cuantificar la fuerza y dirección de las asociaciones lineales entre las variables del sistema, proporcionando el sustento analítico para la selección de características (*feature selection*) y previniendo problemas de colinealidad en las etapas subsiguientes de modelado predictivo.

7.3.14.1 El Núcleo de la Demanda Hídrica y la Redundancia Multicolineal Comercial: El examen de la matriz de correlación global revela un bloque de variables operativas fuertemente cohesionado que define la dinámica macro del acueducto, pero que a su vez introduce un desafío metodológico por redundancia de información.

Figura 26.

Matriz de correlación cruzada



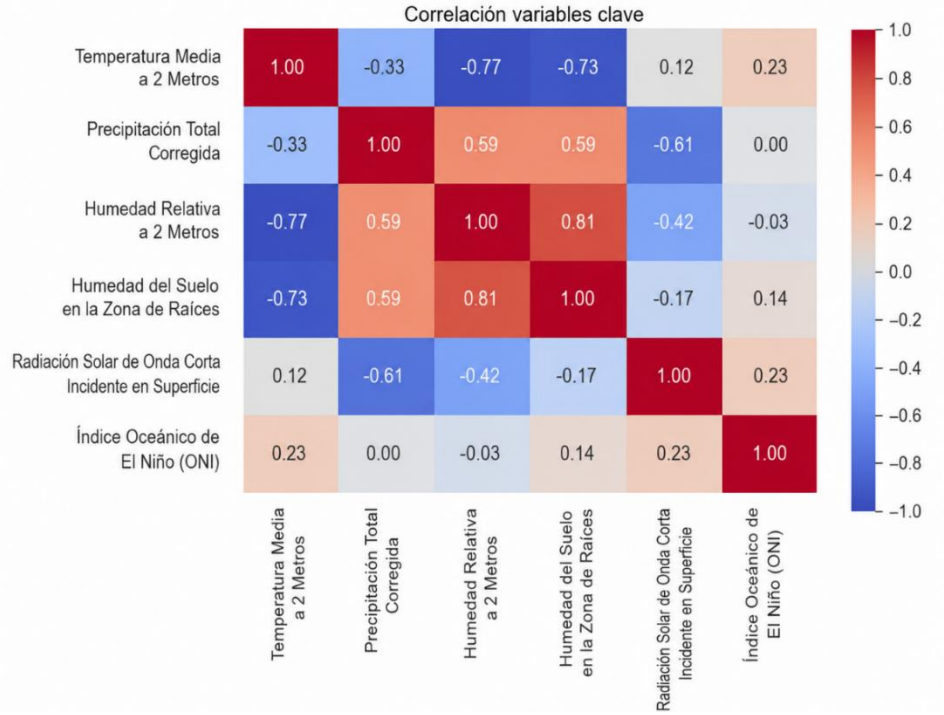
Al analizar el comportamiento de la demanda agregada, se observa un acoplamiento lineal muy elevado (r = 0.99) entre la variable metros cúbicos y el bloque de variables de infraestructura comercial compuesto por facturas, suscriptores facturados y con medidor. Este coeficiente tan cercano a la unidad indica que el volumen neto de agua inyectado y facturado por la red responde con estricta linealidad al tamaño físico y administrativo de la infraestructura de usuarios en la ciudad.

Sin embargo, desde una perspectiva de ciencia de datos, la correlación prácticamente unitaria (r = 1.00) existente de forma cruzada entre facturas, suscriptores facturados y suscriptores con medidor denota una colinealidad estructural intrínseca. Este hallazgo matemático implica que estas tres variables comparten exactamente la misma varianza y aportan información redundante al sistema. Mantener los tres predictores en un algoritmo de aprendizaje automático causaría inestabilidad en la estimación de parámetros (inflación de la varianza). Por lo tanto, el análisis matemático justifica la necesidad de realizar una reducción de dimensiones o seleccionar a un único representante (como suscriptores con medidor) para actuar como variable de control demográfico en los modelos predictivos.

7.3.14.2 Ortogonalidad y Aislamiento Estadístico de la Variable Objetivo: A diferencia de la estrecha correspondencia observada en el volumen global de la red, la variable objetivo-calculada exhibe un comportamiento completamente desacoplado del entorno comercial tradicional.

Figura 27.

Matriz de correlación interna para variables climatológicas.



El comportamiento de la variable objetivo consumo promedio (y su contraparte directa consumo por suscriptor, con la cual posee una correlación de r = 1.00) desarticula cualquier supuesto de dependencia directa con la infraestructura urbana. Su coeficiente de correlación frente a metros cúbicos, facturas y el volumen de suscriptores se fija de manera persistente en un valor negativo y marginal de r = -0.09.

Esto confirma que las fluctuaciones o picos registrados en el promedio unitario de consumo no están ligados a la expansión o contracción de la red comercial ni al volumen total de agua movilizada en los distritos de mayor densidad. La práctica ortogonalidad (cercanía a cero) de la variable objetivo frente a la base instalada sugiere que su dinámica responde a choques o anomalías localizadas (como el caso de Sumapaz o fallas específicas de registro identificadas en el componente espacial), lo que valida su tratamiento como un indicador indexado independiente y descarta el uso de modelos lineales simples para su pronóstico global.

7.3.14.3 Dinámica y Sincronización del Sistema Meteorológico: El análisis de las variables climatológicas evidencia asociaciones consistentes entre temperatura, humedad, precipitación y radiación solar. La temperatura a 2 metros (T2M) presenta correlaciones negativas con la humedad relativa (RH2M, r = -0.77) y la humedad en la zona de la raíz (GWETROOT, r = -0.73), lo que refleja un comportamiento inverso entre la temperatura y las condiciones de humedad atmosférica y del suelo.

Por su parte, GWETROOT presenta correlaciones positivas con la precipitación total corregida (PRECTOTCORR, r = 0.59) y, especialmente, con la humedad superficial del suelo (GWETTOP, r = 1.00), evidenciando una estrecha relación entre las variables asociadas a la disponibilidad de agua en el suelo. Asimismo, la radiación solar de onda corta incidente (ALLSKY_SFC_SW_DWN) muestra correlaciones negativas con la precipitación (r = -0.61) y la humedad relativa (r = -0.42).

En conjunto, estas correlaciones muestran una alta interdependencia entre las variables meteorológicas, coherente con las relaciones físicas esperadas entre temperatura, humedad, precipitación y radiación solar.

7.3.14.4 El Impacto Climático Marginal en el Modelo Lineal: El hallazgo más significativo para el diseño de la arquitectura predictiva surge al cruzar el subsistema climático frente al bloque de consumo y la variable objetivo. A pesar de la sólida consistencia física e interna del clima, sus coeficientes de correlación lineal directa con las variables de demanda hídrica tienden a ser notablemente bajos en toda la matriz:

- La temperatura (T2M) apenas registra un coeficiente de r = 0.13 frente a metros cúbicos y de r = -0.09 frente a consumo promedio.
- La precipitación total acumulada (PRECTOTCORR_SUM) exhibe valores cercanos a cero frente al volumen global (r = -0.01) y al consumo promedio (r = 0.00).

- El índice oceánico de El Niño (ONI), que mide macro-anomalías climáticas a nivel nacional, se mantiene aislado con una correlación débil de r = 0.02 ante los metros cúbicos globales y de r = -0.01 frente al consumo promedio unitario.

Estos hallazgos no indican que el clima carezca de relación con el consumo hídrico, sino que su asociación puede no ser lineal ni inmediata. Los análisis gráficos e históricos mostraron variaciones del consumo coincidentes con condiciones térmicas extremas, como las observadas a inicios de 2024.

Para complementar el análisis de Pearson, se calcularon correlaciones de Spearman, información mutua y correlaciones con rezagos de uno a tres bimestres. Spearman no identificó asociaciones monotónicas fuertes adicionales, mientras que la información mutua sugirió dependencias potenciales, principalmente en variables espaciales y algunas climáticas. Ningún rezago cumplió simultáneamente los criterios de magnitud, significancia ajustada, tamaño de muestra y mejora respecto a la asociación contemporánea, por lo que no se incorporaron variables rezagadas al modelo final.

La selección de *XGBoost* se apoyó además en su comparación con regresión lineal múltiple y Ridge bajo la misma validación temporal. Para 2024, incluyendo latitud y longitud, *XGBoost* alcanzó MAE = 1,1212, RMSE = 2,0006 y R^2 = 0,2935, superando a los modelos lineales. Sin las coordenadas, Ridge obtuvo un desempeño ligeramente superior a *XGBoost*. Los modelos lineales no fueron descartados únicamente por las bajas correlaciones de Pearson y la ventaja del enfoque no lineal se observó principalmente al incorporar información espacial.

7.3.15 Implicaciones metodológicas para el Modelo Predictivo del EDA

El análisis exploratorio evidenció diferencias temporales y espaciales en el comportamiento del consumo hídrico entre períodos bimestrales como entre localidades de Bogotá. Asimismo, las variables climáticas presentaron variaciones consistentes con las

condiciones meteorológicas registradas durante el período de estudio, lo que justificó su incorporación como variables predictoras.

De igual forma, el análisis preliminar de correlación mostró que ninguna variable, considerada de manera individual, explica completamente el comportamiento del consumo hídrico, lo que sugiere la existencia de relaciones complejas entre los diferentes factores analizados. Estos hallazgos respaldan la utilización de algoritmos de ML capaces de modelar interacciones no lineales y constituyeron el fundamento para el desarrollo y evaluación de los modelos predictivos presentados en el apartado siguiente. El análisis descriptivo deja tres conclusiones críticas para el diseño del modelo matemático de predicción:

1. Necesidad de Depuración o Escalado: El desfase masivo en consumo_promedio y metros_cubicos afectará negativamente a modelos basados en distancias o regresiones lineales si no se aplica una transformación previa (como una transformación logarítmica o un escalado robusto que atenúe el peso de los valores atípicos).
2. Multicolinealidad en Datos Climáticos: Variables como la dirección del viento (WD10M y WD2M) o las temperaturas del aire y del suelo (T2M, TS) muestran rangos y comportamientos casi idénticos. Para el modelo, incluir todas estas variables generaría redundancia, por lo que el EDA sugiere una selección previa de predictores.
3. Representatividad Geográfica: el conjunto final contiene 596 registros de las 600 combinaciones posibles de localidad, año y bimestre. Las 19 localidades urbanas cuentan con las 30 observaciones esperadas y Sumapaz con 26, tras la exclusión de sus cuatro registros anómalos. El panel es, por tanto, prácticamente completo y ligeramente desbalanceado en favor de las localidades urbanas, lo que debe tenerse en cuenta al interpretar los resultados de la validación espacial.

7.3.16 Prevención de Fuga de Información (Data Leakage)

Durante el proceso de preparación se evaluó la posible existencia de fuga de información, entendida como la incorporación de variables cuyo contenido pudiera revelar directa o indirectamente la variable objetivo durante el entrenamiento del modelo. Con base en este análisis, se excluyeron las variables operativas que presentaban una relación directa con el consumo promedio y que podían generar una sobreestimación artificial del desempeño predictivo. Esto garantizo que la evaluación de los modelos reflejara su verdadera capacidad para identificar patrones a partir de información disponible antes del período de predicción.

Tabla 15.

Principales Actividades del Proceso de Preparación de los Datos.

Actividad	Propósito
Verificación de Estructura	Garantizar la consistencia del conjunto de datos integrado.
Revisión de Registros Duplicados	Evitar redundancia en la información utilizada para el entrenamiento.
Evaluación de Valores Faltantes	Identificar y tratar información incompleta.
Detección de Valores Atípicos	Analizar observaciones extremas antes del modelado.
Transformación de Variables	Adaptar la información a los requerimientos de los algoritmos.
Prevención de *Data Leakage*	Evitar la incorporación de variables que pudieran sesgar la evaluación de los modelos.

La Tabla 15 resume las principales actividades desarrolladas durante el proceso de preparación del conjunto de datos. Como resultado se obtuvo un conjunto de datos preparado para el desarrollo del análisis exploratorio y el entrenamiento de los modelos predictivos, permitiendo reducir posibles fuentes de sesgo, mejorar la calidad de la información y garantizar que la evaluación de los algoritmos se realizara bajo condiciones metodológicas adecuadas.

7.4 Identificación de Variables Relevantes

Realizado el análisis exploratorio de los datos, se procedió a identificar las variables con mayor potencial para representar el comportamiento del consumo hídrico en la ciudad de Bogotá. Esta etapa tuvo como propósito seleccionar un conjunto de variables predictoras que aportaran información relevante al proceso de aprendizaje de los modelos, evitando la incorporación de atributos redundantes o con escasa capacidad explicativa.

La selección de variables se fundamentó en el conocimiento del dominio del problema, los hallazgos obtenidos durante el análisis exploratorio y la revisión de la literatura especializada sobre predicción de la demanda de agua. Bajo este enfoque, se priorizaron aquellas variables que, de acuerdo con estudios previos y con el comportamiento observado en los datos, podrían contribuir a explicar las variaciones del consumo hídrico.

Las variables finalmente incorporadas al proceso de modelado se agruparon en cuatro categorías: operativas, climáticas, temporales y espaciales, las cuales permitieron representar de manera integral las diferentes dimensiones asociadas al fenómeno estudiado.

Tabla 16.

Categorías de Variables Utilizadas para el Entrenamiento de los Modelos Predictivos

Categoría	Variables Principales	Justificación
Operativas (excluidas del entrenamiento)	Facturas, suscriptores facturados, suscriptores con medidor	Se analizaron durante el análisis exploratorio por representar el comportamiento operativo del sistema de acueducto, pero fueron excluidas del entrenamiento por presentar dependencia directa con la variable objetivo y generar fuga de información (*data leakage*).
Climáticas	Temperatura, precipitación, humedad relativa, presión atmosférica, velocidad del viento, radiación solar, ONI, entre otras	Permiten incorporar la influencia de las condiciones ambientales sobre el consumo de agua.
Temporales	Año, bimestre, variables cíclicas (seno y coseno del bimestre)	Capturan tendencias interanuales y patrones estacionales del consumo.
Espaciales	Localidad, latitud y longitud	Representan la heterogeneidad territorial del consumo hídrico entre las diferentes localidades de Bogotá.

Con el fin de garantizar la validez del proceso predictivo, se descartaron aquellas variables que presentaban una dependencia directa con la variable objetivo y que podían introducir fuga de información durante el entrenamiento de los modelos. Esta decisión permitió que los algoritmos aprendieran únicamente a partir de información disponible antes del momento de la predicción, evitando una sobreestimación artificial del desempeño. Como resultado se obtuvo el conjunto definitivo de variables utilizado para el entrenamiento y evaluación de los modelos de ML desarrollados en la presente investigación.

7.5 Diseño de Modelos Predictivos de *Machine Learning*

El desarrollo de los modelos predictivos constituyó la etapa central de la investigación, cuyo propósito fue estimar el consumo hídrico en la ciudad de Bogotá mediante técnicas de ML. A partir del conjunto de datos previamente integrado, depurado y analizado, se implementaron diferentes algoritmos de aprendizaje supervisado para identificar el modelo con mayor capacidad predictiva bajo las condiciones del problema estudiado.

El proceso de modelado comprendió la formulación del problema predictivo, el entrenamiento de los algoritmos, el ajuste de hiperparámetros, la evaluación mediante métricas de regresión y la comparación de los resultados obtenidos. Esto permitió analizar tanto el desempeño predictivo como la capacidad de generalización de los modelos desarrollados.

7.5.1 Formulación del Problema Predictivo

La predicción del consumo hídrico se formuló como un problema de aprendizaje supervisado de regresión, debido a que la variable objetivo es continua. La unidad de análisis correspondió a cada combinación de localidad, año y bimestre, conformando un conjunto de datos tabular con las variables predictoras y el consumo_promedio observado para cada período.

La variable objetivo fue consumo_promedio, definida como el consumo hídrico promedio por localidad y bimestre. Debido a la asimetría de su distribución, en los modelos que lo requirieron se aplicó la transformación log1p, posteriormente revertida mediante expm1, con el fin de calcular las métricas e interpretar las predicciones en la escala original.

Las variables predictoras incluyeron información temporal, climática y espacial seleccionada durante las etapas previas. El bimestre se representó mediante transformaciones seno y coseno para conservar su carácter cíclico, mientras que latitud y longitud representaron la dimensión territorial. Las variables operativas directamente relacionadas con el consumo

fueron excluidas para prevenir fuga de información (*data leakage*) y garantizar que los modelos utilizaran únicamente información disponible antes de la predicción.

El desempeño se evaluó mediante MAE, RMSE y R^2. Como referencia, se implementaron modelos base basados en la media global, media histórica por localidad, media histórica por localidad y bimestre, y un pronóstico estacional ingenuo del mismo bimestre del año anterior, calculados exclusivamente con información histórica disponible en cada partición.

En la evaluación temporal de 2024, la media histórica por localidad y bimestre presentó el mejor desempeño (MAE = 0,7131; RMSE = 0,9911; R^2 = 0,8266), mientras que *XGBoost* con variables espaciales obtuvo MAE = 1,1212; RMSE = 2,0006; R^2 = 0,2935. En consecuencia, *XGBoost* registró un MAE 57,22 % mayor y un RMSE 101,87 % mayor que el mejor modelo de referencia.

La validación temporal móvil también mostró un mejor desempeño de la media histórica por localidad frente a *XGBoost*. Aunque este algoritmo alcanzó mejores resultados bajo partición aleatoria, dicho esquema puede favorecer la interpolación de patrones territoriales previamente observados. En conjunto, las validaciones temporal y espacial no aportaron evidencia suficiente para respaldar la hipótesis predictiva, mostrando que, para los datos analizados, los patrones históricos por localidad y bimestre proporcionaron predicciones más consistentes que los modelos de Machine Learning evaluados.

7.5.2 Modelo de Redes Neuronales Profundas

Como primera aproximación para la predicción del consumo hídrico se implementó un modelo de Red Neuronal Profunda (*Deep Neural Network, DNN*) utilizando la biblioteca PyTorch. La elección de este algoritmo obedeció a su capacidad para modelar relaciones no lineales entre múltiples variables predictoras, característica que resulta de interés en problemas donde intervienen simultáneamente factores climáticos, temporales y espaciales.

7.5.2.1 Arquitectura del Modelo: El modelo fue implementado como un Perceptrón Multicapa (*Multilayer Perceptron*, MLP) para datos tabulares. La arquitectura definitiva, correspondiente a la versión predic_consumo_agua_v2.ipynb, estuvo conformada por una capa de entrada de 22 variables predictoras, tres capas ocultas con 64, 32 y 16 neuronas y una capa de salida con una única neurona encargada de estimar el valor continuo del consumo_promedio. En las capas ocultas se utilizó la función de activación Unidad Lineal Rectificada (ReLU), y en la de salida, una activación lineal, útil para problemas de regresión.

En el proceso de ajuste, la arquitectura inicial fue simplificada mediante la reducción del número de neuronas y la eliminación de capas *Dropout,* con el propósito de disminuir la complejidad del modelo y adaptarlo al tamaño del conjunto de datos disponible. Esta modificación buscó mejorar la estabilidad del entrenamiento sin incrementar el riesgo de sobreajuste.

Tabla 17.

Arquitectura del Modelo de Red Neuronal Profunda

Capa	Configuración	Función de Activación	Propósito
Entrada	22 variables predictoras	—	Variables climáticas y temporales preprocesadas. La red neuronal no incorporó latitud ni longitud.
Capa oculta 1	Linear (22, 64)	ReLU	Primera extracción de relaciones no lineales.
Capa oculta 2	Linear (64, 32)	ReLU	Reducción de dimensionalidad y aprendizaje de patrones.
Capa oculta 3	Linear (32, 16)	ReLU	Representación interna previa a la predicción.
Salida	Linear (16, 1)	Lineal	Estimación del consumo hídrico promedio.

Fuente: Elaboración propia a partir de la Implementación en *predic_consumo_agua_v2.ipynb*.

7.5.2.2 Entrenamiento del Modelo: El entrenamiento se realizó utilizando el optimizador Adam y la función de pérdida Huber Loss, seleccionada por su mayor robustez frente a valores atípicos. Para controlar el sobreajuste se implementó la estrategia *Early Stopping*, con un máximo de 200 épocas y una paciencia de 20 épocas, conservando automáticamente el modelo con el menor error de validación. El entrenamiento se efectuó mediante mini-lotes (*batch size*) de 32 observaciones, mientras que el conjunto de validación estuvo conformado por 89 registros.

Cabe señalar que aspectos como la transformación log1p de la variable objetivo, el escalamiento de variables y la imputación de datos ya fueron descritos en el apartado Preparación y Calidad de los Datos, por lo que no se reiteran en esta sección.

Tabla 18.

Configuración del Entrenamiento del Modelo de Red Neuronal Profunda

Parámetro	Valor Utilizado
Framework	PyTorch
Función de pérdida	Huber Loss
Optimizador	Adam
Learning Rate	0.0005
Batch Size	32
Máximo de épocas	200
Early Stopping	Paciencia = 20
Registros de validación	89

Fuente: Elaboración propia a partir de la Implementación en *predic_consumo_agua_v2.ipynb*.

7.5.2.3 Resultados Obtenidos: La versión final de la red neuronal (predic_consumo_agua_v2.ipynb) obtuvo un MAE de 1,5487, un RMSE de 2,3295 y un R^2 de −0,0675, evidenciando una capacidad limitada para representar la variabilidad del

consumo hídrico. Un R^2 negativo indica que el modelo se desempeña por debajo de una predicción basada en la media del conjunto de prueba.

El análisis de las predicciones mostró que la red neuronal reprodujo adecuadamente valores cercanos al promedio, pero tendió a subestimar los consumos más elevados, concentrando las estimaciones en un rango reducido. Asimismo, las curvas de entrenamiento y validación evidenciaron que la pérdida de validación dejó de mejorar tras las primeras épocas, indicando dificultades para generalizar a datos no observados durante el entrenamiento.

Tabla 19.

Desempeño del Modelo de Red Neuronal Profunda

Métrica	Valor	Interpretación
MAE	1,5487	Error absoluto medio sobre el conjunto de prueba.
RMSE	2,3295	Evidencia errores de mayor magnitud en algunos registros.
R^2	−0,0675	Capacidad explicativa insuficiente para representar el consumo hídrico.

Con el fin de evaluar la evolución del proceso de desarrollo, la Tabla 20 resume el desempeño obtenido por las dos versiones implementadas del modelo de red neuronal.

Tabla 20.

Evaluación y comparación de métricas de desempeño para las iteraciones del modelo predictivo.

Versión	MAE	RMSE	R^2	Observación
Predic_consumo_agua.ipynb	1,5357	2,3378	−0,0752	Arquitectura inicial con mayor complejidad y uso de *Dropout*.
Predic_consumo_agua_v2.ipynb	1,5487	2,3295	−0,0675	Arquitectura simplificada; ligera mejora, pero con desempeño aún insuficiente.

En conjunto, los resultados muestran que, aunque la red neuronal permitió explorar el potencial del aprendizaje profundo para la predicción del consumo hídrico, su capacidad predictiva fue limitada para las características del conjunto de datos utilizado. Este comportamiento motivó la evaluación de modelos basados en árboles de decisión, particularmente *XGBoost*.

7.5.3 Modelo Basado en XGBoost

Se implementó un modelo basado en *Extreme Gradient Boosting*, algoritmo ampliamente utilizado en problemas de regresión sobre datos tabulares debido a su capacidad para modelar relaciones no lineales, manejar interacciones entre variables y controlar el sobreajuste mediante técnicas de regularización. La selección de este algoritmo respondió a las características del conjunto de datos empleado y a la necesidad de mejorar la precisión en la predicción del consumo hídrico en la ciudad de Bogotá.

El desarrollo del modelo se realizó de forma iterativa mediante cuatro versiones sucesivas (predic_consumo_agua_v3, v4, v5 y v6), en las cuales se incorporaron progresivamente estrategias de optimización y validación para mejorar el desempeño predictivo y evaluar la capacidad de generalización del modelo.

7.5.3.1 Configuración Inicial del Modelo: La primera implementación utilizó *XGBRegressor* con la función objetivo *reg:squarederror*, apropiada para problemas de regresión. La configuración inicial incluyó un conjunto de hiperparámetros orientados a controlar la complejidad del modelo y reducir el riesgo de sobreajuste, manteniendo un equilibrio entre capacidad predictiva y estabilidad durante el entrenamiento.

Tabla 21.

Configuración Inicial del Modelo XGBoost.

Hiperparámetro	Valor Inicial	Función
objective	reg:squarederror	Función objetivo para regresión.
n_estimators	300	Número máximo de árboles.
learning_rate	0.05	Tasa de aprendizaje.
max_depth	4	Profundidad máxima de los árboles.
min_child_weight	3	Control de complejidad.
subsample	0.8	Muestreo de observaciones por árbol.
colsample_bytree	0.8	Muestreo de variables por árbol.
reg_lambda	1.0	Regularización L2.
random_state	42	Reproducibilidad.
n_jobs	-1	Uso de todos los núcleos disponibles.

Fuente: Elaboración propia a partir de la Implementación en *predic_consumo_agua_v3.ipynb*.

7.5.3.2 Optimización del Modelo: Se realizó un proceso de ajuste de hiperparámetros con el propósito de incrementar la capacidad predictiva del modelo y reducir el error de estimación. Esta optimización permitió refinar la configuración inicial y mejorar progresivamente las métricas de desempeño obtenidas durante las diferentes iteraciones del desarrollo experimental.

Una vez optimizado el modelo, se incorporaron las variables espaciales latitud y longitud, con el fin de representar explícitamente la heterogeneidad territorial del consumo hídrico entre las diferentes localidades de Bogotá. Esta modificación produjo la mejora más importante observada durante el proceso de modelado y dio lugar a la versión definitiva utilizada bajo validación aleatoria (predic_consumo_agua_v5.ipynb).

7.5.3.3 Resultados Obtenidos: La Tabla 22 resume la evolución del desempeño alcanzado durante las diferentes versiones desarrolladas del modelo *XGBoost*.

Tabla 22.

Evolución del Desempeño del Modelo XGBoost

Versión	MAE	RMSE	R^2	Principal Modificación
predic_consumo_agua_v3	1,5562	2,1344	0,1038	Configuración Inicial.
predic_consumo_agua_v4	1,5143	2,0808	0,1482	Ajuste de Hiperparámetros.
predic_consumo_agua_v5	0,4348	0,7418	0,8917	Incorporación de variables espaciales.
predic_consumo_agua_v6	2,3063	2,5436	−53,4863	Validación espacial LOLO (promedio de los 20 folds).

Bajo la partición aleatoria, *XGBoost* con variables espaciales obtuvo un MAE de 0,4348, RMSE de 0,7418 y R^2 de 0,8917. Este desempeño refleja principalmente capacidad de interpolación entre localidades ya observadas, dado que latitud y longitud pueden actuar como identificadores territoriales al existir registros de las mismas localidades en entrenamiento y prueba. Esto se evidenció al entrenar un modelo únicamente con coordenadas, que alcanzó MAE de 0,6126, RMSE de 1,0208 y R^2 de 0,7950. Por tanto, las variables espaciales explican gran parte del desempeño del modelo completo, mientras que la incorporación de variables climáticas y temporales incrementó el R^2 en menos de 0,10 puntos.

En la prueba temporal de 2024, el modelo de referencia basado en la media histórica por localidad y bimestre obtuvo un MAE de 0,7131, un RMSE de 0,9911 y un R^2 de 0,8266, superando a *XGBoost* con variables espaciales, que alcanzó un MAE de 1,1212, un RMSE de 2,0006 y un R^2 de 0,2935. Asimismo, en la validación espacial LOLO, el desempeño promedio de los veinte folds fue de un MAE de 2,3063 y un RMSE de 2,5436, con una mediana de R^2 de −2,76 y valores individuales que oscilan entre +0,71 (Antonio Nariño) y −888,63 (Usme); únicamente 4 de las 20 localidades alcanzaron un R^2 positivo. Así, el desempeño observado bajo la partición aleatoria representa una de interpolación dentro del dominio territorial conocido.

Tomando como referencia el experimento v7, se realizó un análisis orientado a interpretar el comportamiento del modelo *XGBoost.* Este análisis no se limitó a la importancia interna proporcionada por el algoritmo, debido a que dicha medida puede favorecer variables continuas o con alta cardinalidad, como la latitud y la longitud. Por esta razón, se incorporaron la importancia por permutación y los valores SHAP como métodos complementarios de interpretabilidad. La importancia por permutación permitió cuantificar la pérdida de desempeño ocasionada al alterar cada predictor, mientras que SHAP permitió estimar la contribución de las variables tanto en las predicciones individuales como en el comportamiento global del modelo.

En el experimento v7, los resultados evidenciaron una contribución predominante de las coordenadas geográficas bajo la partición aleatoria. Sin embargo, la latitud y la longitud pueden funcionar como identificadores indirectos de las localidades y favorecer la interpolación de patrones territoriales previamente observados.

Tabla 23.

Variables con Mayor Importancia en el Modelo XGBoost

Ranking	Variable	Importancia
1	Latitud	0.110163
2	BIMESTRE_sin	0.089269
3	PS	0.072403
4	T2M	0.071296
5	T2M_MIN	0.069054
6	ONI	0.062949
7	Longitud	0.059661
8	GWETROOT	0.048467
9	WS2M	0.043586
10	WS10M	0.040352

Tabla 24.

Comparación de la importancia de las variables según tres métodos

Variable	Rankings (nativa / permutación / SHAP)	Presencia en top 10
Latitud	1 / 1 / 1	3 de 3
Longitud	7 / 2 / 2	3 de 3
BIMESTRE_sin	2 / 4 / 3	3 de 3
PS	3 / 5 / 7	3 de 3
T2M	4 / 8 / 6	3 de 3
ONI	5 / 10 / 5	3 de 3

La Tabla 24 contrasta la posición de cada variable según los tres métodos de importancia; en los rankings, un número menor indica mayor importancia. El análisis evidencia que las variables espaciales y temporales desempeñaron un papel relevante dentro del proceso de predicción, complementadas por variables climáticas relacionadas con temperatura, presión atmosférica, humedad del suelo y velocidad del viento.

7.5.4 Comparación y Selección del Modelo Predictivo Final

Con el fin de identificar el algoritmo con mayor predictiva del consumo hídrico en Bogotá, se realizó una comparación entre los modelos desarrollados durante la investigación. La Tabla 25 resume el desempeño alcanzado por las arquitecturas evaluadas.

Tabla 25.

Comparación del Desempeño de los Modelos Predictivos Desarrollados

Modelo	MAE	RMSE	R^2	Observación
Red Neuronal Profunda (v2)	1,5487	2,3295	−0,0675	Bajo desempeño predictivo.
***XGBoost* (v3)**	1,5562	2,1344	0,1038	Configuración inicial.
***XGBoost* (v4)**	1,5143	2,0808	0,1482	Optimización de hiperparámetros.
***XGBoost* (v5)**	0,4348	0,7418	0,8917	Mejor desempeño bajo partición aleatoria; corresponde a interpolación.
***XGBoost* solo con latitud y longitud (v6)**	0,6126	1,0208	0,7950	Sin variables climáticas ni temporales; evidencia el efecto de identificación territorial.
Referencia: media histórica por localidad y bimestre	0,6877	0,9618	0,7362	Modelo de referencia bajo validación temporal 2024 (19 localidades).
Referencia: media histórica por localidad	0,7355	1,0637	0,6774	Modelo de referencia bajo validación temporal 2024 (19 localidades).

Los resultados evidencian una mejora progresiva en las métricas de los modelos *XGBoost* conforme avanzó el proceso de optimización. Mientras que la red neuronal profunda presentó una capacidad limitada para representar la variabilidad del consumo hídrico, las diferentes versiones de *XGBoost* redujeron progresivamente los errores de predicción bajo partición aleatoria. No obstante, esta comparación debe leerse con cautela: corresponde a un único esquema de validación y no incorpora por sí sola las evaluaciones temporal y espacial, que son las que permiten juzgar la capacidad de generalización.

El modelo *XGBoost* v5 obtuvo el menor MAE y RMSE y el mayor valor de R^2 bajo partición aleatoria. Sin embargo, la incorporación de los modelos de referencia y del modelo entrenado

únicamente con coordenadas modifica sustancialmente esta lectura: el modelo de solo latitud y longitud alcanza un R^2 de 0,7950 sin utilizar información climática, y el modelo de referencia basado en la media histórica por localidad y bimestre supera a *XGBoost* en la validación temporal de 2024. Por tanto, el mejor desempeño de la versión v5 no puede atribuirse a una mayor capacidad de aprendizaje del fenómeno, sino a la identificación de las localidades ya observadas.

La selección del modelo tampoco se fundamentó únicamente en su precisión bajo partición aleatoria. Se consideraron además su estabilidad durante el entrenamiento, la capacidad para modelar relaciones no lineales y la posibilidad de interpretar la contribución relativa de las variables mediante importancia por permutación y valores SHAP, aspectos que resultan pertinentes para el análisis del consumo hídrico aun cuando el desempeño predictivo fuera de muestra sea limitado.

En consecuencia, el modelo *XGBoost* v5 fue seleccionado como el modelo de referencia del estudio para los análisis de interpretabilidad y validación, con la precisión explícita de que su aplicabilidad se restringe a la interpolación dentro de las localidades representadas en el conjunto de entrenamiento. Para la predicción de periodos futuros, el modelo basado en los patrones históricos por localidad y bimestre resultó más consistente.

7.6 Validación y Evaluación del Desempeño Predictivo

La validación del desempeño busco evaluar la capacidad predictiva y de generalización de los modelos desarrollados para estimar el consumo hídrico en Bogotá. Se realizó mediante métricas de regresión y dos estrategias complementarias de validación, partición aleatoria del conjunto de datos (*Hold-Out*) y validación espacial *LOLO*, diseñada para analizar el comportamiento del modelo frente a localidades no utilizadas durante el entrenamiento.

7.6.1 Métricas de Evaluación

El desempeño de los modelos se evaluó mediante tres métricas ampliamente utilizadas en problemas de regresión. Debido a que algunos modelos fueron entrenados utilizando la transformación log1p sobre la variable objetivo, las métricas se calcularon después de aplicar la transformación inversa (expm1), garantizando que los resultados fueran interpretables en la escala original del consumo hídrico.

Tabla 26.

Métricas Utilizadas para la Evaluación del Desempeño Predictivo

Métrica	Propósito	Valor óptimo
MAE	Cuantifica el error absoluto promedio entre los valores observados y predichos.	0
RMSE	Penaliza con mayor intensidad los errores de gran magnitud.	0
R^2	Mide la proporción de la variabilidad explicada por el modelo.	1
WAPE	Expresa el error absoluto agregado como proporción del volumen total observado; es robusto ante la heterogeneidad entre localidades.	0
Mediana del error absoluto	**MedAE -** Resume el error típico sin verse afectada por observaciones extremas.	0

Estas métricas permitieron comparar objetivamente los modelos y evaluar la precisión de las predicciones y su capacidad para representar el comportamiento del consumo hídrico.

7.6.2 Evaluación del Desempeño de los Modelos

La Tabla 27 resume el desempeño obtenido por los modelos evaluados bajo las diferentes estrategias implementadas durante la investigación.

Tabla 27.

Resultados Comparativos de los Modelos Predictivos

Modelo	Estrategia de Validación	MAE	RMSE	R^2
Red Neuronal Profunda (v2)	Hold-Out	1,5487	2,3295	−0,0675
***XGBoost* (v5)**	Hold-Out	0,4348	0,7418	0,8917
***XGBoost* (v6)**	LOLO	2,3063	2,5436	−53,4863 (promedio); mediana −2,76
***XGBoost* (evaluación adicional)**	Temporal 2020–2023/2024, sin coordenadas y sin Sumapaz	1,5829	1,9325	−0,0648
***XGBoost* solo latitud y longitud (v6)**	Hold-Out	0,6126	1,0208	0,7950
Referencia: media histórica por localidad y bimestre	Temporal 2020–2023/2024, sin Sumapaz	0,6877	0,9618	0,7362

Los resultados evidenciaron diferencias importantes entre los modelos y las estrategias de validación evaluadas. La red neuronal profunda presentó un desempeño limitado, reflejado en un R^2 negativo y errores relativamente elevados. Por su parte, *XGBoost* con variables espaciales alcanzó bajo partición aleatoria un MAE de 0,4348, un RMSE de 0,7418 y un R^2 de 0,8917. Este resultado indica una alta capacidad de interpolación entre registros pertenecientes a localidades representadas tanto en el entrenamiento como en la prueba, pero no demuestra por sí mismo

una capacidad de generalización temporal o territorial. La comparación con el modelo entrenado exclusivamente con coordenadas (R^2 de 0,7950) confirma que la mayor parte de ese desempeño proviene de la identificación de la localidad y no de las variables climáticas.

En la validación espacial LOLO, el desempeño disminuyó considerablemente cuando el modelo fue evaluado sobre localidades completamente excluidas del entrenamiento, obteniendo un MAE promedio de 2,3063 y un RMSE promedio de 2,5436. Los valores de R^2 por localidad oscilaron entre +0,71 y −888,63, con una mediana de −2,76, y solo 4 de las 20 localidades alcanzaron un valor positivo. Esto evidencia que los patrones aprendidos en las localidades conocidas no se transfieren adecuadamente a territorios no observados.

En la validación temporal, el modelo se entrenó con los registros de 2020–2023 y se evaluó con los datos de 2024, excluyendo Sumapaz y sin utilizar latitud, longitud ni localidad como variables predictoras. Así, *XGBoost* obtuvo un MAE de 1,5829, un RMSE de 1,9325 y un R^2 de −0,0648. Además, fue superado por el modelo de referencia basado en la media histórica por localidad y bimestre, que alcanzó un MAE de 0,6877, un RMSE de 0,9618, un R^2 de 0,7362, un WAPE de 0,0629 y una mediana del error absoluto de 0,5024. En conjunto, estos resultados muestran que el alto desempeño observado bajo partición aleatoria corresponde principalmente a la interpolación dentro del dominio territorial conocido, mientras que la generalización temporal y espacial continúa siendo limitada. Para la predicción de 2024, los patrones históricos por localidad y bimestre ofrecieron resultados más consistentes que *XGBoost*.

7.6.3 Validación de la Capacidad de Generalización

Con el propósito de evaluar el comportamiento del modelo en escenarios más exigentes, se implementó la estrategia *LOLO*, en la cual cada localidad fue excluida del entrenamiento y utilizada posteriormente para la evaluación. Esta metodología permitió analizar la capacidad del modelo para realizar predicciones sobre territorios no representados previamente.

La comparación entre las estrategias de validación evidenció que la elevada precisión obtenida bajo partición aleatoria disminuye considerablemente cuando el modelo enfrenta localidades completamente nuevas. El detalle por fold resulta más informativo que el promedio: los valores de R^2 oscilaron entre +0,71 (Antonio Nariño) y −888,63 (Usme), con una mediana de −2,76, y únicamente cuatro localidades (Antonio Nariño, Santa Fe, Fontibón y Chapinero) alcanzaron un R^2 positivo. El promedio aritmético del R^2 entre *folds* resulta poco representativo, ya que se encuentra dominado por un único *fold* con error extremo. Adicionalmente, al repetir la validación excluyendo Sumapaz el desempeño se deterioró en varias localidades (Engativá pasó de −48,07 a −81,09), lo que indica que la dificultad de generalización no se explica únicamente por la presencia de la localidad rural. En consecuencia, los resultados indican que las variables espaciales fortalecen la predicción dentro de localidades previamente observadas, pero no garantizan una generalización robusta hacia contextos territoriales diferentes.

Tabla 28.

Comparación entre las Estrategias de Validación

Aspecto	Hold-Out	LOLO	Validación temporal adicional
Tipo de evaluación	Aleatoria	Espacial	Temporal
Periodo de entrenamiento	Registros mezclados temporalmente	Registros de las localidades restantes	2020–2023
Periodo de prueba	Registros aleatorios	Localidad excluida	2024
Variables espaciales	Latitud y longitud	Latitud y longitud	Excluidas
Tratamiento de Sumapaz	Incluida	Evaluada dentro de LOLO	Excluida completamente
Localidades analizadas	20	20	19

Capacidad evaluada	Interpolación	Generalización espacial	Generalización temporal
Resultado principal	Desempeño elevado	Generalización limitada	*XGBoost* no superó el mejor *baseline*

Estos resultados delimitan el alcance del modelo desarrollado y evidencian la importancia de incorporar información territorial adicional para fortalecer su capacidad de generalización en futuras investigaciones.

Tabla 29.

Comparación de XGBoost y los modelos de referencia en la evaluación temporal 2024

Modelo	MAE	RMSE	R^2	Posición
Media histórica por localidad y bimestre	0,6877	0,9618	0,7362	1
Media histórica por localidad	0,7355	1,0637	0,6774	2
Pronóstico estacional basado en 2023	0,8233	1,2154	0,5788	3
XGBoost sin coordenadas y sin Sumapaz	1,5829	1,9325	−0,0648	4
Media histórica global	1,6155	1,9702	−0,1067	5

La Tabla 29 compara el desempeño de *XGBoost* con cuatro modelos históricos de referencia en la predicción de 2024. El *baseline* de media histórica por localidad y bimestre obtuvo los menores errores y el mayor R^2, ubicándose como la alternativa con mejor desempeño. *XGBoost* ocupó la cuarta posición y presentó un R^2 negativo. Estos resultados indican que los patrones históricos por localidad y bimestre fueron más consistentes para estimar el consumo futuro. Por tanto, *XGBoost* no superó al mejor modelo de referencia bajo esta validación temporal.

7.7 Síntesis del Trabajo de Campo

El desarrollo del trabajo de campo permitió integrar información operativa, climática, temporal y espacial para construir un conjunto de datos adecuado para la aplicación de técnicas de ML orientadas a la predicción del consumo hídrico en Bogotá. A partir del proceso de preparación de datos, análisis exploratorio, selección de variables, diseño de modelos y validación experimental, *XGBoost* presentó el mejor desempeño entre los modelos de *Machine Learning* evaluados bajo partición aleatoria; sin embargo, su capacidad de generalización disminuyó sustancialmente bajo validación temporal y espacial, llegando a ser superado por el modelo de referencia basado en patrones históricos por localidad y bimestre.

Asimismo, la incorporación de variables espaciales incrementó de forma notable las métricas obtenidas bajo partición aleatoria. Sin embargo, las evaluaciones complementarias demostraron que dicho incremento no corresponde a una mejora real de la capacidad predictiva: un modelo entrenado solo con coordenadas alcanza un R^2 de 0,7950, la validación LOLO evidenció una capacidad muy limitada para generalizar hacia localidades no representadas durante el entrenamiento, y en la validación temporal de 2024 el modelo fue superado por un modelo de referencia construido a partir de los promedios históricos por localidad y bimestre. Este contraste constituye el principal hallazgo empírico del trabajo de campo.

8. Discusión

Se analizan e interpretan los resultados obtenidos, contrastándolos con la literatura revisada y los objetivos de la investigación. Asimismo, se discuten sus implicaciones para la predicción del consumo hídrico en la ciudad de Bogotá y las principales limitaciones del estudio.

8.1 Relación entre Variables y Consumo Hídrico

El análisis realizado evidenció que el consumo hídrico en la ciudad de Bogotá puede representarse mediante la interacción de variables espaciales, temporales y climáticas. Estas relaciones corresponden a asociaciones predictivas identificadas por los modelos de ML y no implican relaciones de causalidad, dado el diseño no experimental de la investigación. En este contexto, los modelos permiten identificar patrones en la información histórica y estimar el comportamiento del consumo de agua.

El modelo *XGBoost*, que obtuvo el mejor desempeño entre los algoritmos evaluados bajo partición aleatoria, identificó como variables de mayor contribución la latitud, la transformación cíclica del bimestre (BIMESTRE_sin), la presión atmosférica (PS), la temperatura media (T2M), la temperatura mínima (T2M_MIN), el Índice Oceánico del Niño (ONI) y la longitud. La preeminencia de las coordenadas debe interpretarse con cautela, ya que operan como identificadores territoriales más que como factores explicativos del consumo. Los resultados sugieren que la dimensión territorial y la estacionalidad tienen un peso mayor que las condiciones climáticas en la representación del consumo hídrico agregado por localidad y bimestre.

Con el fin de sintetizar la contribución de los diferentes grupos de variables incorporados al modelo predictivo, en la Tabla 30 se presenta una descripción de la influencia observada durante el proceso de modelado.

Tabla 30.

Influencia de los grupos de variables sobre la predicción del consumo hídrico.

Grupo de Variables	Influencia Observada Sobre el Consumo Hídrico
Operativas	Presentaron una alta relación con la variable objetivo; sin embargo, fueron excluidas del entrenamiento debido al riesgo de fuga de información (*data leakage*), con el fin de preservar la validez del modelo predictivo.
Climáticas	Variables como temperatura, precipitación, humedad, presión atmosférica, viento y el índice ONI aportaron una contribución marginal al desempeño predictivo. Su relación con el consumo agregado por localidad y bimestre no resultó lineal ni suficiente para mejorar la predicción fuera de muestra.
Espaciales	Presentaron la mayor contribución aparente dentro del modelo bajo partición aleatoria. Sin embargo, un modelo entrenado únicamente con latitud y longitud alcanzó un R^2 de 0,7950, lo que evidencia que estas variables operan como identificadores territoriales y favorecen la interpolación, sin aportar capacidad de generalización hacia localidades no observadas.
Temporales	La transformación cíclica del bimestre permitió representar adecuadamente la estacionalidad y los patrones intraanuales del consumo hídrico.

Los hallazgos son parcialmente consistentes con investigaciones previas que identifican la influencia de variables climáticas, temporales y espaciales sobre el consumo de agua (Jiménez et al., 2017; Brentan et al., 2017; Lee y Derrible, 2020; Xenochristou et al., 2020). No obstante, a diferencia de esos estudios, en el presente trabajo la integración de dichas variables no se tradujo en una mejora de la capacidad predictiva fuera de muestra: la hipótesis planteada no fue respaldada bajo las estrategias de validación temporal y espacial, ya que el modelo fue superado por un modelo de referencia basado en promedios históricos y mostró una generalización limitada hacia localidades no observadas. Esta divergencia puede atribuirse al nivel de agregación empleado (localidad y bimestre), que atenúa la variabilidad climática de corto plazo, y a la ausencia de variables socioeconómicas y de infraestructura.

8.2 Importancia de los Componentes Espaciales

Los componentes espaciales desempeñaron un papel relevante en la capacidad predictiva del modelo desarrollado para estimar el consumo hídrico en Bogotá. Las variables latitud y longitud se incorporaron como predictores y se ubicaron entre las de mayor importancia en el modelo *XGBoost,* mientras que la variable localidad se utilizó para la validación espacial mediante el esquema LOLO. Estos resultados evidencian que la localización geográfica aporta información relevante para representar la variabilidad territorial del consumo.

Este comportamiento es consistente con House-Peters y Chang (2011), quienes señalan que el consumo urbano de agua está influenciado por factores que operan en diferentes escalas espaciales, y con Xenochristou et al. (2020), quienes demostraron que la precisión de los modelos y la importancia de las variables dependen de la escala espacial de análisis.

Sin embargo, las variables espaciales deben interpretarse como indicadores del contexto territorial y no como factores causales del consumo. Aunque mejoraron el desempeño del modelo bajo validación aleatoria, la estrategia LOLO mostró una disminución de la capacidad predictiva al evaluar localidades no observadas durante el entrenamiento. Esto indica que las variables espaciales favorecen principalmente la interpolación dentro del dominio conocido, pero no garantizan una generalización robusta hacia nuevos territorios.

En consecuencia, futuras investigaciones podrían incorporar variables territoriales, como densidad poblacional, estrato socioeconómico e infraestructura de abastecimiento, con el fin de fortalecer la capacidad de generalización de los modelos predictivos.

8.3 Implicaciones para la Gestión del Recurso Hídrico en Bogotá

Los resultados obtenidos evidencian que las técnicas de ML constituyen una herramienta útil para apoyar la gestión del recurso hídrico en Bogotá. Si bien los modelos desarrollados no

sustituyen los procesos de planificación y toma de decisiones, sí proporcionan información que puede contribuir a anticipar escenarios de consumo y fortalecer la gestión basada en datos.

En este sentido, la incorporación de variables climáticas, temporales y espaciales permite representar el comportamiento del consumo hídrico desde una perspectiva integral, facilitando la identificación de patrones asociados a la variabilidad climática y a las diferencias territoriales entre localidades. Este enfoque resulta coherente con lo planteado por la UNESCO (2021), que destaca la necesidad de fortalecer la gestión del agua mediante herramientas que favorezcan la planificación y la sostenibilidad del recurso y con House-Peters y Chang (2011), quienes señalan que el análisis conjunto de múltiples factores mejora la comprensión de la demanda urbana. Con base en los resultados obtenidos, la Tabla 31 resume las principales implicaciones que los modelos predictivos pueden tener para la gestión del recurso hídrico en la ciudad de Bogotá.

Tabla 31.

Implicaciones de los Modelos Predictivos para la Gestión del Recurso Hídrico.

Aspecto	Aporte Potencial (condicionado a las mejoras señaladas en Trabajo Futuro)
Planeación del Recurso Hídrico	Aportar un marco de referencia para anticipar escenarios de consumo, una vez se incorporen mayor resolución espacial y variables socioeconómicas.
Monitoreo y Seguimiento	Identificar cambios en los patrones de consumo y apoyar el seguimiento de su evolución.
Gestión Territorial	Reconocer diferencias entre localidades para orientar estrategias de gestión acordes con las características del territorio.
Toma de decisiones	Proporcionar información cuantitativa que complemente la formulación de políticas y acciones de gestión hídrica.
Investigación futura	Servir como base para desarrollar modelos que integren nuevas variables y amplíen la capacidad predictiva.

Estos resultados muestran que los modelos predictivos basados en ML pueden constituir un complemento para la gestión sostenible del recurso hídrico, siempre que su alcance se declare de forma explícita. El aporte principal no es la predicción operativa, para la cual los promedios históricos por localidad y bimestre resultaron más confiables, sino la caracterización de la heterogeneidad territorial y la demostración de que la evaluación mediante validación espacial y temporal es indispensable antes de incorporar este tipo de modelos a procesos de decisión. Su aplicación debe entenderse como herramienta de apoyo y no como mecanismo de decisión autónomo, dado que la dinámica del consumo también depende de factores sociales, económicos, institucionales y ambientales no incorporados en el presente estudio.

8.4 Alcances y Limitaciones Metodológicas

La Investigación abordó la predicción del consumo hídrico en Bogotá como un problema de regresión supervisada, integrando variables climáticas, temporales y espaciales en un conjunto de datos estructurado para el desarrollo de modelos de ML. Este enfoque permitió implementar un proceso metodológico que incluyó preprocesamiento de datos, selección de variables, ajuste de hiperparámetros y comparación entre diferentes algoritmos predictivos.

Uno de los principales aportes metodológicos del estudio fue la implementación de la estrategia de validación espacial LOLO, la cual permitió evaluar la capacidad de generalización de los modelos bajo un escenario más exigente que la validación aleatoria convencional. Esta aproximación contribuyó a delimitar con mayor precisión el alcance de los resultados obtenidos.

Sin embargo, la investigación presenta limitaciones asociadas principalmente a la cobertura del conjunto de datos, su nivel de agregación espacial y temporal y la disponibilidad de variables explicativas, aspectos que condicionan el alcance de los resultados y constituyen oportunidades para futuras investigaciones.

Tabla 32.

Principales Limitaciones Metodológicas de la Investigación y sus Implicaciones.

Limitación Metodológica	Implicación para los Resultados
Cobertura del Conjunto de Datos (596 registros)	Limita la representatividad de la información disponible y el entrenamiento de modelos con mayor complejidad.
Cobertura Temporal (2020–2024)	Los resultados representan únicamente las condiciones observadas durante el periodo de estudio y no contemplan cambios en horizontes temporales diferentes.
Nivel de Agregación por Localidad y Bimestre	Reduce la representación de la variabilidad espacial y temporal existente dentro de cada localidad.
Uso de Información Secundaria Proveniente de Fuentes Oficiales	La calidad y disponibilidad de los datos dependen de los procesos de recolección y actualización realizados por las entidades responsables.
Ausencia de Variables Socioeconómicas, Demográficas y de Infraestructura	Limita la incorporación de factores que podrían explicar con mayor detalle las diferencias en el consumo hídrico entre localidades.
Generalización espacial limitada	El modelo mostró dificultades para extrapolar hacia localidades no observadas durante la validación *LOLO*, por lo que su aplicación debe restringirse al dominio representado por los datos.

Estas limitaciones no afectan la validez de la investigación, sino que delimitan el alcance de sus resultados e identifican oportunidades para futuros estudios. En este sentido, la incorporación de nuevas variables, una mayor resolución espacial y temporal, y la ampliación del análisis a otros contextos territoriales podrían fortalecer el desarrollo de modelos predictivos para la gestión del recurso hídrico.

8.5 Alcances y Limitaciones Metodológicas

Los resultados de la investigación permiten reconocer la importancia de integrar criterios de Gerencia de Proyectos en el desarrollo de soluciones predictivas, particularmente en la definición del alcance, la gestión de riesgos, el aseguramiento de la calidad y la evaluación de los entregables. Desde esta perspectiva, el aporte del estudio trasciende la construcción de modelos de *Machine Learning*, al proporcionar elementos para orientar decisiones sobre la viabilidad y pertinencia de futuras aplicaciones en la gestión del recurso hídrico.

Uno de los principales aprendizajes para la Gerencia de Proyectos consiste en la necesidad de establecer criterios de aceptación acordes con el propósito de uso de las soluciones analíticas. En este sentido, la evaluación de un modelo no debería limitarse a su desempeño experimental, sino considerar su capacidad de generalización, su utilidad institucional y las condiciones requeridas para una eventual implementación. Esto permite reducir el riesgo de destinar recursos a soluciones cuyo desempeño no represente una mejora frente a las alternativas disponibles.

A partir de estas consideraciones, se propone una hoja de ruta orientativa para futuros proyectos de implementación de modelos predictivos del consumo hídrico en Bogotá, estructurada en etapas que permitan evaluar progresivamente su viabilidad, calidad y utilidad operativa. La Tabla 33 presenta las etapas propuestas y los criterios generales para orientar su desarrollo.

Tabla 33.

Hoja de ruta propuesta para Implementación de modelos predictivos del consumo hídrico.

Etapa	Propósito	Criterio para avanzar
1. Formulación	Definir la necesidad institucional, el alcance y los interesados.	Viabilidad y patrocinio institucional.
2. Fortalecimiento de datos	Mejorar la calidad, cobertura y resolución de la información.	Disponibilidad de datos adecuados para el modelado.
3. Validación piloto	Evaluar la solución frente a modelos históricos de referencia.	Desempeño consistente y utilidad demostrada.
4. Implementación	Integrar la solución en los procesos institucionales.	Condiciones técnicas y operativas verificadas.
5. Seguimiento y mejora	Monitorear el desempeño y realizar ajustes cuando se requieran.	Cumplimiento de los criterios de desempeño establecidos.

Esta hoja de ruta constituye una propuesta para futuras iniciativas y no una implementación realizada durante el estudio. Su contribución desde la Gerencia de Proyectos radica en orientar la toma de decisiones mediante etapas, entregables y criterios de evaluación que permitan gestionar la incertidumbre antes de comprometer recursos institucionales.

9. Conclusiones y Trabajo Futuro

El desarrollo de la Investigación permitió obtener evidencia suficiente para responder al problema planteado, contrastar la Hipótesis formulada y establecer las principales contribuciones del estudio. En este capítulo se presentan las conclusiones derivadas de los resultados alcanzados y las oportunidades de Investigación futura.

El primer objetivo específico se cumplió mediante la evaluación de la contribución predictiva de las variables temporales, climáticas y espaciales sobre la estimación del consumo promedio de agua. Tras el proceso de integración y depuración de la información, la comparación entre la importancia nativa de XGBoost, la importancia por permutación y los valores SHAP evidenció que la latitud, la longitud, el componente seno del bimestre, la presión atmosférica, la temperatura a 2 metros y el Índice Oceánico de El Niño se posicionaron de manera consistente como los atributos de mayor relevancia bajo una partición aleatoria.

Sin embargo, la elevada importancia de la latitud y la longitud debe interpretarse con cautela, debido a que pueden funcionar como identificadores territoriales y favorecer la interpolación de patrones de las localidades conocidas. Asimismo, la relevancia de las variables climáticas representa una contribución al proceso predictivo y no demuestra una relación causal con el consumo hídrico. En consecuencia, el primer objetivo específico se cumplió al identificar y contrastar la contribución de las variables mediante diferentes técnicas de interpretabilidad, reconociendo las limitaciones de generalización evidenciadas por las validaciones temporal y espacial.

Se dio cumplimiento al segundo objetivo específico mediante el diseño, entrenamiento, comparación y selección de modelos de Machine Learning. La comparación entre los modelos desarrollados permitió establecer que *XGBoost* presentó un desempeño superior al de la red neuronal profunda para el conjunto de datos utilizado. Mientras la arquitectura neuronal mostró limitaciones para representar la variabilidad del consumo, obteniendo un R^2 negativo, el modelo

basado en *Gradient Boosting* redujo los errores de predicción bajo partición aleatoria. Este resultado confirma que, para problemas de regresión sobre conjuntos tabulares de tamaño moderado, los modelos basados en árboles potenciados constituyen una alternativa más adecuada que las redes neuronales profundas. No obstante, la incorporación de modelos de referencia mostró que ninguno de los dos algoritmos superó a la media histórica por localidad y bimestre en la validación temporal de 2024. En consecuencia, se dio cumplimiento al segundo objetivo específico mediante el diseño, entrenamiento, comparación y selección de modelos, reconociendo que el mejor desempeño corresponde a un modelo de referencia y no a los algoritmos de aprendizaje automático evaluados.

El tercer objetivo específico se cumplió al evaluar la capacidad de generalización del modelo bajo esquemas rigurosos de validación. Mediante diferentes estrategias de validación se permitió establecer que el desempeño depende críticamente del escenario de aplicación. Bajo partición aleatoria el modelo alcanzó métricas elevadas, pero la validación espacial LOLO evidenció una caída drástica al evaluarlo sobre localidades no observadas, y la validación temporal mostró que tampoco superaba a un modelo de referencia al predecir un año futuro. Este comportamiento demuestra que el modelo aprende a identificar las localidades presentes en el conjunto de entrenamiento más que la dinámica del consumo. En consecuencia, el modelo desarrollado no debe emplearse para estimar el consumo de localidades sin información histórica previa ni como herramienta de pronóstico para periodos futuros, ámbito en el cual los promedios históricos por localidad y bimestre resultaron más confiables. Con ello se dio cumplimiento al tercer objetivo específico, cuyo valor reside precisamente en haber delimitado con evidencia el alcance real del modelo.

El objetivo general se cumplió mediante el desarrollo, comparación y evaluación de modelos predictivos basados en técnicas de *Machine Learning* para estimar el consumo promedio de agua por localidad y periodo bimestral en Bogotá, integrando variables climáticas, temporales y espaciales. La evaluación permitió establecer que el desempeño de los modelos

depende de la estrategia de validación y del grado de representación territorial presente en los datos de entrenamiento.

XGBoost con variables espaciales presentó un desempeño favorable bajo la partición aleatoria, pero esta capacidad se limitó principalmente a la interpolación entre registros pertenecientes a localidades previamente observadas, como lo confirma el hecho de que un modelo entrenado únicamente con latitud y longitud alcance un R^2 de 0,7950. En la evaluación temporal de 2024, el modelo fue superado por el modelo de referencia de media histórica por localidad y bimestre, mientras que en la validación LOLO la mediana del R^2 por localidad fue de −2,76 y únicamente 4 de las 20 localidades alcanzaron un valor positivo, evidenciando una capacidad muy limitada para estimar el consumo de territorios excluidos del entrenamiento.

En consecuencia, la investigación no demuestra que la incorporación de variables climáticas, temporales y espaciales garantice una mejora general de la capacidad predictiva. En cambio, permite concluir que los patrones históricos específicos de cada localidad y bimestre proporcionaron predicciones más consistentes en los escenarios temporales evaluados, y que la generalización hacia territorios no observados continúa siendo una limitación central del enfoque desarrollado.

El principal aporte de esta investigación es de carácter metodológico. El estudio demuestra que la evaluación de modelos predictivos aplicados a datos de panel espacio-temporal exige esquemas de validación coherentes con su estructura, y que la partición aleatoria puede producir métricas engañosamente favorables cuando las variables espaciales permiten identificar la unidad territorial. La comparación entre el modelo completo, el modelo entrenado solo con coordenadas y los modelos de referencia históricos permitió cuantificar ese efecto y distinguir la interpolación de la generalización. Este procedimiento, junto con la *validación Leave-One-Localidad-Out*, constituye una contribución replicable para futuras investigaciones sobre modelado predictivo aplicado a la gestión del recurso hídrico.

Desde la perspectiva de la Gerencia de Proyectos, la investigación permitió reconocer la importancia de integrar criterios de planificación, seguimiento, gestión de riesgos y aseguramiento de la calidad en el desarrollo de soluciones predictivas. La organización del trabajo por fases y la revisión progresiva de las alternativas de modelado facilitaron la identificación de restricciones metodológicas y la adaptación de las decisiones técnicas a partir de los resultados obtenidos. Asimismo, la evaluación mediante diferentes estrategias de validación y modelos de referencia evidenció la importancia de establecer criterios de aceptación acordes con el propósito de uso de las soluciones analíticas. En este sentido, el aporte de la Gerencia de Proyectos consistió en complementar los procedimientos de Ciencia de Datos con una perspectiva orientada a la gestión de la incertidumbre, la confiabilidad de los entregables y la toma de decisiones fundamentadas, proporcionando consideraciones aplicables a futuras investigaciones sobre predicción del consumo hídrico.

En relación con la hipótesis planteada, los resultados indican que se encuentra *parcialmente respaldada*. Si bien la integración de variables climáticas, temporales y espaciales aportó información estructural para el modelado en particiones aleatorias, dicho aporte no se tradujo en una mayor capacidad predictiva fuera de muestra bajo validaciones temporales y espaciales estrictas, donde los patrones históricos locales representaron la línea base más sólida.

Desde una perspectiva aplicada, los resultados delimitan el alcance actual de este tipo de herramientas. Para la estimación del consumo promedio por localidad y período bimestral en horizontes futuros, los patrones históricos por localidad y bimestre resultaron más confiables que los modelos de aprendizaje automático evaluados, lo que constituyen la referencia mínima que cualquier desarrollo posterior debe superar. El aporte práctico del estudio se orienta, por tanto, a la identificación de comportamientos atípicos, a la caracterización de la heterogeneidad territorial del consumo y a la definición de un protocolo de validación que las entidades responsables pueden aplicar antes de incorporar modelos predictivos a sus procesos de planificación.

Trabajo Futuro

Los resultados obtenidos durante esta investigación evidencian diversas oportunidades para ampliar y fortalecer el modelo predictivo desarrollado, tanto desde el punto de vista metodológico como desde su aplicación práctica.

Como primera línea de investigación, se recomienda ampliar el conjunto de datos incorporando series históricas de mayor duración y aumentando la resolución espacial del análisis. La utilización de información a nivel de sectores censales, barrios o unidades de servicio permitiría representar con mayor precisión la heterogeneidad territorial del consumo hídrico y mejorar la capacidad de aprendizaje de los modelos predictivos.

En segundo lugar, resulta conveniente incorporar nuevas variables explicativas relacionadas con factores socioeconómicos, infraestructura del sistema de acueducto, tarifas, usos del suelo, densidad poblacional, características de las edificaciones y variables asociadas a políticas de ahorro del agua. La inclusión de estos atributos permitiría disminuir la dependencia de las coordenadas geográficas como representación indirecta de diferencias territoriales y mejorar la capacidad de generalización del modelo.

Como tercera línea de investigación, se propone evaluar modelos adicionales de aprendizaje automático y aprendizaje profundo, incluyendo *LightGBM, CatBoost, Random Forest,* modelos híbridos y arquitecturas especializadas para series temporales, tales como *Long Short-Term Memory* (LSTM) y *Transformers*. La comparación de estos enfoques permitiría determinar si es posible incrementar la precisión predictiva bajo escenarios de validación espacial y temporal más exigentes.

Asimismo, se recomienda implementar estrategias de validación más robustas que integren simultáneamente componentes espaciales y temporales. La combinación de esquemas *Leave-One-Localidad-Out* con validaciones basadas en períodos futuros permitiría evaluar la

capacidad del modelo para predecir cambios del consumo bajo condiciones reales de operación, fortaleciendo su confiabilidad para aplicaciones institucionales.

Finalmente, como proyección aplicada, se propone desarrollar un sistema de apoyo a la toma de decisiones que integre el modelo predictivo dentro de plataformas de monitoreo del recurso hídrico. La incorporación de información climática en tiempo real, datos operativos del sistema de acueducto y herramientas de visualización geográfica permitiría generar estimaciones dinámicas del consumo, apoyar la planeación de la demanda y fortalecer la gestión sostenible del recurso hídrico en Bogotá. Para alcanzar este propósito será necesario realizar procesos continuos de actualización del modelo, recalibración periódica y validación con nuevos datos, garantizando así su estabilidad y aplicabilidad en escenarios operativos reales.

Referencias

Acevedo, A. (2023). *A review of leak detection and prediction methods in water distribution systems using machine learning.* Universidad de los Andes. https://hdl.handle.net/1992/65925

Adamowski, J. & Karapataki, C. (2010). Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. *Journal of Hydrologic Engineering.* 15(10):729-743. https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000245

Alcaldía Mayor de Bogotá. (2025). *Racionamiento de agua en Bogotá*. https://bogota.gov.co/racionamiento-agua-bogota/?utm_source

Aminu, I., Bashir, A., Sunusi, A., Malami, S., Usman, A., Abba, S. & Ozsahin, D. (2025). State-of-art implementation of computing intelligent models for water demand modelling: A decade review and future direction. *Journal of Water and Land Development.* No. 66 (VII–IX): 214–224. https://doi.org/10.24425/jwld.2025.155316

Ander, E. (2011). *Aprender a Investigar. Nociones Básicas para la Investigación Social*. Editorial Brujas. (1ª ed.).

Ayus, I., Narayanan, N. & Gupta, D. (2023). Prediction of water level using machine learning and deep learning techniques. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*. vol.47. pp. 2437–2447. https://doi.org/10.1007/s40996-023-01053-6

Breiman, L. (2001). Random Forests. *Machine Learning* 45(1). pp. 5-32. https://doi.org/10.1023/A:1010933404324

Brentan, B., Meirelles, G., Herrera, M., Luvizotto, E. & Izquierdo, J. (2017). Correlation Analysis of Water Demand and Predictive Variables for Short-Term Forecasting Models. *Mathematical Problems in Engineering.* pp. 1-10. https://doi.org/10.1155/2017/6343625

Candelieri, A. (2017). Clustering and support vector regression for water demand forecasting and anomaly detection. *Water*, 9(3), 224. https://doi.org/10.3390/w9030224

Canquil, D., San Martín, D., Inoquio, I. & Leger, P. (2026). A Systematic Literature Review on Machine Learning Techniques for Predicting Household Water Consumption. *Water Resources Management*. vol. 40(83). https://doi.org/10.1007/s11269-026-04496-4

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc. https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf

Comisión Colombiana del Océano (CCO). (2025). Seguimiento de los Fenómenos EL NIÑO – LA NIÑA Comunicado No. 05 de 2025. https://cco.gov.co/seguimiento-de-los-fenomenos-el-nino-la-nina-comunicado-no-05-de-2025/

Costa, D., Bayissa, Y., Vianna, K., Villas, M., Bawa, A., Lugon, J., Silva, A. & Srinivasan, R. (2024). Water quality estimates using machine learning techniques in an experimental watershed. *Journal of Hydroinformatics*. https://doi.org/10.2166/hydro.2024.132

Diago, D. (2025). *Localidades con más reportes por desperdicio de agua en Bogotá.* Concejo de Bogotá. https://concejodebogota.gov.co/localidades-con-mas-reportes-por-desperdicio-de-agua-en-bogota/cbogota/2025-01-27/174833.php

Donkor, E., Mazzuchi, T., Soyer, R., y Roberson, J. (2012). Urban water demand forecasting: Review of methods and models. *Journal of Water Resources Planning and Management*, 140(2), 146–159. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000314

Empresa de Acueducto y Alcantarillado de Bogotá (EAAB). (2022). *Informe de Gestión. EAAB*. https://www.acueducto.com.co/wps/portal/EAB2/Home/transparencia_informacion_publica/planeacion_presupuesto_informes/informes-de-gestion

Forero, J., Castellanos, J. & Imbol, W. (2019). Colombia, ¿alineándose al cumplimiento del objetivo de desarrollo sostenible n°6 sobre agua limpia y saneamiento? *Revista Producción + Limpia*. 14(2). https://doi.org/10.22507/pml.v14n2a3

Gil, E. y Tobón, C. (2016). Hydrological modelling with TOPMODEL of Chingaza páramo, Colombia. *Revista Facultad Nacional de Agronomía Medellín*, 69(2), 7919–7933. https://doi.org/10.15446/rfna.v69n2.59137

Görenekli, K. & Gülbag, A. (2024). Comparative Analysis of Machine Learning Techniques for Water Consumption Prediction: A Case Study from Kocaeli Province. *Sensor*, 24 (17). https://doi.org/10.3390/s24175846

Granata, F. (2019). Evapotranspiration evaluation models based on machine learning algorithms - A comparative study. *Agricultural Water Management*, vol. *217*, 303–315. https://www.sciencedirect.com/science/article/abs/pii/S0378377418312800

Grecco, R., Sanches, R., Dos Santos, B., Martins, R., De Figueiredo, G., Bourscheidt, V. & Toledo, P. (2025). Using Xgboost models for daily rainfall prediction*. Anales de Geografía de la Universidad Complutense*, 45, 75-92 https://doi.org/10.5209/aguc.98944

Hernández, R., Fernández, C. & Baptista, P. (2014). *Metodología de la investigación*. (6ª ed.). McGraw-Hill.

Herrera, J. y Romero, B. (2025). *Sistema IoT para monitoreo en tiempo real del consumo de agua.* Universidad Distrital Francisco José de Caldas. http://hdl.handle.net/11349/94922

House-Peters, L. y Chang, H. (2011). Urban water demand modeling: Review of concepts, methods, and organizing principles. *Water Resources Research*, 47(5). https://doi.org/10.1029/2010WR009624

Hussien, W., Memon, F. y Savic, D. (2016). Assessing and modelling the influence of household characteristics on per capita water consumption. *Water Resources Management*, 30(1), 2931-2955. https://doi.org/10.1007/s11269-016-1314-x

Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). (2022). *Estudio Nacional del Agua 2022 (Resumen ejecutivo)*. Instituto de Hidrología, Meteorología y Estudios Ambientales. https://www.ideam.gov.co/sites/default/files/prensa/boletines/2024-08-23/resumen_ejecutivo_estudio_nacional_del_agua_2022_0.pdf

International Organization for Standardization. (ISO) (2020). *Project, programme and portfolio management — Guidance on project management* (ISO 21502:2020).

Izquierdo, S., Restrepo, D., Ozment, S., Acero, N., Bulbena, L., Camacho, J., Villegas, L., Arjona, F., Bejarano, P., Mora, M. y García, M. (2023). *Integrando Infraestructura Natural al Sistema de Abastecimiento de Agua de Bogotá.* World Resources Institute. https://doi.org/10.46830/wrirpt.22.00046

Jiménez, D., Orrego, S., Cossio, D., Vásquez, F. & Ponce, R. (2017). Efecto de la variabilidad climática sobre la demanda de agua para uso residencial urbano: el caso de la ciudad de Manizales, Colombia. *Desarrollo y Sociedad*, 1(79). https://doi.org/10.13043/dys.79.3

Kanyama, M., Bhunu, F., Gamundani, A. y Hartmann, A. (2025). AI-Driven Anomaly Detection in Smart Water Metering Systems Using Ensemble Learning. *Water*, 17(13), 1933. https://doi.org/10.3390/w17131933

Lee, D. & Derrible, S. (2020). Predicting Residential Water Demand with Machine-Based Statistical Learning. *Journal of Water Resources Planning and Management.* 146(1). https://doi.org/10.1061/(ASCE)WR.1943-5452.0001119

Manfrini, D. (2025). *Gestión Inteligente del Agua en Bogotá: Monitoreo y Optimización del Consumo con IA e IoT*. Universidad EAN. https://hdl.handle.net/10882/14882

Martínez, F., Contreras, L., Ferri, C., Hernández, J., Kull, M., Lachiche, N., Ramírez, M. J. y Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, vol. 33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Martínez, S., Cottyn, H., Garrido, A., Kirshner, J. y O'Bryen, R. (2024). Water Factories of the High Colombian Mountains: Páramo as 'Infrastructural Nature'. *Water Alternatives*. 17(1), 167-186 https://www.water-alternatives.org/index.php/alldoc/articles/vol17/v17issue1/741-a17-1-8/file

Mathivha, F., Mabala, L., Matimolane, S. y Mbatha, N. (2024). El Niño-induced drought impacts on Reservoir water Resources in South Africa. *Atmosphere*, 15(3), 249. https://doi.org/10.3390/atmos15030249

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software, 101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Ministerio de Ambiente y Desarrollo Sostenible (Minambiente). (2019). *Metodología Para el Diseño de Módulos de Consumo del Agua.* Dirección de Gestión Integral del Recurso Hídrico Grupo de Administración del Recurso Hídrico. https://www.minambiente.gov.co/wp-content/uploads/2021/10/5.-Anexo-5-Metodologia-Estimacion-Modulos-de-Consumo-del-Agua.pdf

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P. & Righetti, M. (2024). Applications of XGBoost in water resources engineering, A systematic literature review. *Environmental Modelling and Software*. vol. 174. https://doi-org.bdbiblioteca.universidadean.edu.co/10.1016/j.envsoft.2024.105971

Nunes, T., De Souza, F. & Costa, V. (2021). Urban Water Demand Modeling Using Machine Learning Techniques: Case Study of Fortaleza, Brazil. *Journal of Water Resources Planning and Management.* https://doi.org/10.1061/(ASCE)WR.1943-5452.0001310

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2021). *Informe Mundial de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos 2021: El valor del agua.* UNESDOC. https://doi.org/10.54679/MOMU6572

Ortega, A., Hinojosa, I., Martínez, E., Ramírez, E. & Sánchez, A. (2026). Assessment and Response to Water Stress in Bogotá: A Multidimensional Analysis. *Water*. 18(10). https://doi.org/10.3390/w18101155

Ortiz, J. (2025). *37 % del agua en Bogotá se pierde.* Universidad de los Andes. https://www.uniandes.edu.co/es/noticias/ingenieria/37-del-agua-en-bogota-se-pierde

Palma, R. (2022). Análisis crítico del coeficiente de determinación (r^2), como indicador de la calidad de modelos lineales y no lineales. *MATEMÁTICA ESPOL - FCNM Journal.* vol. 20, No. 2. https://orcid.org/0000-0001-9934-7343

Pérez, C., Alonso, S., Van, N., Garcia, M. y Werner, M. (2024). Disentangling the socio-natural dynamics of drought and water scarcity in Colombia's Tropical Andes. *Journal of Hydrology: Regional Studies.* 56. https://doi.org/10.1016/j.ejrh.2024.102068

Perez, L., Perez, R. y Seca, M. (2020). *Metodología de la Investigación Científica*. Editorial Maipue. (1ª ed.).

Project Management Institute. (PMI) (2021). *Guía de los fundamentos para la dirección de proyectos (Guía del PMBOK)* (7.ª ed.). Project Management Institute.

Roberts, D., Bahn, V., Ciuti, S., Boyce, M., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J., Schröder, B., Thuiller, W., Warton, D., Wintle, B., Hartig, F. & Dormann, C. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Rueda, J., Cañon, D., Mendoza, C., Rodríguez, A. & Cortés, C. (2025). Evaluating machine learning models for estimating evapotranspiration in Colombia's Cauca River Valley. *Scientific Reports*. 16, 278. https://doi.org/10.1038/s41598-025-29514-0

Sanjur, A., Deng, A., Bethancourth, M. & Araúz, R. (2025). Evaluación del uso del aprendizaje automático para la predicción del consumo del agua potable en Panamá metropolitana. *Revista de Iniciación Científica*, 11(2). https://doi.org/10.33412/rev-ric.v11.2.4180

Superintendencia de Servicios Públicos Domiciliarios (Superservicios). (2024). *Informe Sectorial de los Servicios Públicos Domiciliarios de Acueducto y Alcantarillado. Vigencia 2023*. Link.

Triana, J., Ocampo, C., Carvajal, Y., Torres, W., Triana, J. & Canchala, T. (2023). Estimación de evapotranspiración de referencia con información escasa utilizando machine learning en el Suroccidente Colombiano. *Meteorológica*. https://doi.org/10.24215/1850468Xe024

Wang, S., Chen, Y., Hsu, S., Kao, Y., Kimura, M., Chang, L., Pan, T. & Ni, C. (2025). A case study on the application of a data-driven (XGBoost) approach on the environmental and socio-economic perspectives of agricultural groundwater management. *Agricultural Water Management,* vol. 318. https://doi.org/10.1016/j.agwat.2025.109729

World Wildlife Fund (WWF). (2024). *Watering Resilience Case Studies of Nature-Based Solutions for Climate Adaptation in Action.* Link.

Xenochristou, M., Hutton, C., Hofman, J. & Kapelan, Z. (2020). Water Demand Forecasting Accuracy and Influencing Factors at Different Spatial Scales Using a Gradient Boosting Machine. *Water Resources Research.* 56(8). https://doi.org/10.1029/2019WR026304

Zapata, A., Rivera, C., Valoyes, D., Muñoz, C., Mejía, M. y Catalan, J. (2021). Páramo lakes of colombia: An overview of their geographical distribution and physicochemical characteristics. *Water*, 13(16), 2175. https://doi.org/10.3390/w13162175

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. y Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, 1(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001

Anexos

En esta sección se presentan los materiales complementarios que apoyan y respaldan el desarrollo de la investigación. Se incluye información adicional relacionada con las fuentes de datos, el análisis exploratorio, la construcción de los modelos predictivos y los resultados obtenidos, con el fin de facilitar la comprensión y verificación del proceso desarrollado.

Anexo A: Repositorio Digital de Código, Datos y Resultados Complementarios

Enlace al Repositorio Oficial:

https://github.com/ccjaviers777/TRABAJO_GRADO_CIENCIA_DATOS.git

Guía de Estructura del Repositorio Digital: El repositorio está organizado estratégicamente en dos ramas de desarrollo para separar el ciclo de procesamiento de datos de la fase de modelado y evaluación de resultados.

Rama Principal (main): Contiene los cuadernos de experimentación de algoritmos, los pesos de los modelos optimizados y las métricas de evaluación final.

Modelos Entrenados (.pth): Archivos de pesos (best_tabular_regressor.pth y versiones bajo pérdida Huber) correspondientes a las arquitecturas de redes neuronales optimizadas.

Cuadernos de Modelado (predic_consumo_agua.ipynb y v2 a v6.ipynb): iteraciones del proceso de entrenamiento, pruebas de regularización, sintonización de hiperparámetros y validación espacial. La versión v2 corresponde a la red neuronal profunda final; las versiones

v3 a v5 documentan la evolución de XGBoost; la versión v6 contiene la validación LOLO, su variante sin Sumapaz y el modelo entrenado exclusivamente con latitud y longitud.

Cuadernos de ajuste metodológico (ajuste_variable_objetivo_v7.ipynb y siguientes): saneamiento de la variable objetivo, validación temporal con entrenamiento en 2020–2023 y prueba en 2024, cálculo de los modelos de referencia, análisis de asociaciones mediante Spearman e información mutua, evaluación de rezagos e interpretabilidad mediante importancia por permutación y valores SHAP. Resultados de validación (.csv, .txt): reportes detallados de la validación cruzada espacial (LOLO), comparación del impacto de incluir o excluir los registros anómalos de Sumapaz, métricas de ajuste de *XGBoost* y el registro experimentos.txt.

Documentación de Soporte: Diccionario formal de variables (diccionario_variables.xlsx) y matriz consolidada final de consumo unificado con datos climáticos (df_consumo_bime_clima_v3.xlsx).

Rama de Datos (data): Contiene el flujo completo de ingeniería de datos, organizado bajo el paradigma clásico de analítica de datos en tres etapas de madurez:

Carpeta raw_db (Datos Crudos): Almacena la información original sin procesar, segregada de forma individual en carpetas para cada una de las 20 localidades del Distrito Capital y datos globales de Bogotá suministrados por las fuentes oficiales.

Carpeta processed_db (Datos Procesados y Limpios): Bases de datos estructuradas tras la limpieza física y la imputación de registros. Incluye los archivos .csv depurados individualmente por localidad, la serie mensual del Índice Oceánico de El Niño

(oni_mensual.xlsx), la matriz climática limpia y las tablas integradas listas para consumo de Business Intelligence (base_pbi.xlsx) o combinación total (base_unida_total.csv).

Carpeta notebooks (Scripts de Ingeniería de Datos): Código fuente enfocado exclusivamente en la fase de extracción y transformación:

Base_Datos_Bime.ipynb: Automatización del procesamiento y consolidación de las bases de datos bimestrales.

Analisis_descriptivo_bime_cons.ipynb: Rutinas de análisis exploratorio inicial y generación de estadísticas descriptivas.

Scripts por Localidad: Cuadernos específicos individuales encargados de la depuración temporal, tratamiento de datos y estandarización geográfica para cada territorio de la ciudad (desde Usaquén hasta Sumapaz).