

Facultad Infraestructura Hábitat y Creatividad

Desarrollo de un modelo predictivo de Incendios Forestales utilizando algoritmos de Machine Learning en el Distrito Metropolitano de Quito.

María Delia Tiupul Carrillo

Pontificia Universidad Católica del Ecuador

Maestría en Sistemas de Información Mención en Data Science

22/08/2026

1. Agradecimientos

Expreso mi más profundo agradecimiento a Dios por guiar cada paso de este caminar, brindándome la fuerza, fortaleza, la sabiduría, el conocimiento y la perseverancia necesarias para culminar con éxito esta importante etapa de mi formación académica.

A mi familia por su amor incondicional, comprensión, apoyo constante y palabras de aliento durante todo este proceso, su confianza en mí y su motivación permanente fueron el impulso que me permitió superar los desafíos y alcanzar esta meta.

A mis docentes y a todas las personas que, con sus conocimientos, orientación y valiosos aportes, contribuyeron al desarrollo de esta investigación y a mi crecimiento profesional y personal.

Contenido

Capítulo 1: Introducción 3
1.1 Introducción 3
1.2 Objetivos 4
1.2.1 Objetivo general: 4
1.2.2 Objetivos específicos: 4
1.3 Justificación 5
1.4 Preguntas de Investigación 5
Capítulo 2: Desarrollo 6
2.1 Marco Teórico 6
2.2. Metodología 10
2.2.1. Métodos de recolección de datos 10
2.2.2. Técnicas de análisis 12
2.2.3. Justificación de los métodos 14
2.3. Resultados y Análisis 15
2.3.1. Resultados 16
2.3.1.1. Preparación e integración de las fuentes de datos 16
2.3.1.2. Resultados del análisis exploratorio de datos 19
2.3.1.3. Integración de información meteorológica e incendios forestales 29
2.3.1.4. Preparación de los datos para el modelado predictivo 32
2.3.1.5. Resultados de los modelos predictivos 36
2.3.1.6. Optimización y selección del modelo predictivo 40
2.3.1.7. Interpretabilidad del modelo mediante SHAP 46
2.3.1.8 Evaluación del modelo sin componente espacial 53
2.3.2. Análisis 54
2.3.2.1. Análisis comparativo del desempeño de los modelos predictivos 54
2.3.2.2. Análisis de la influencia de las variables meteorológicas y espaciales 57
2.3.2.3. Aporte del modelo a la predicción y prevención de incendios forestales 59
2.3.2.4. Comparación de los resultados con estudios previos 60
Capítulo 3: Conclusiones y Recomendaciones 63
3.1. Conclusiones 63
3.2. Limitaciones 65
3.3. Recomendaciones 66
4. Referencias 68
5. Anexos 70
5.1 Código de programación 70

Capítulo 1: Introducción

1.1 Introducción

Los incendios forestales son una amenaza para el medio ambiente, la biodiversidad y la seguridad de la población, debido a los impactos ambientales, sociales y económicos que se genera. La ocurrencia de incendios forestales depende de diferentes factores, entre ellos las condiciones climáticas, ambientales y las actividades humanas. En el Distrito Metropolitano de Quito (DMQ), los incendios forestales representan un desafío permanente para las instituciones responsables de la gestión del riesgo ya que afectan áreas protegidas, zonas de interfaz urbano-forestal, infraestructura y la calidad del aire, además de comprometer la disponibilidad de recursos para la atención de emergencias.

El desarrollo de herramientas de Ciencia de Datos y Machine Learning ofrece nuevas oportunidades para fortalecer la prevención y la gestión de incendios forestales mediante el análisis de grandes volúmenes de información, los algoritmos de aprendizaje automático permiten identificar patrones complejos entre variables meteorológicas e información histórica de incendios forestales, generando modelos predictivos capaces de estimar la probabilidad de ocurrencia de futuros eventos, estas capacidades convierten al Machine Learning en una herramienta de apoyo para la planificación estratégica, la asignación eficiente de recursos y la implementación de acciones preventivas orientadas a reducir el impacto de los incendios forestales.

Por tal motivo esta investigación desarrolla un modelo predictivo de incendios forestales para el Distrito Metropolitano de Quito utilizando información histórica de incendios forestales registrada por el Cuerpo de Bomberos del Distrito Metropolitano de Quito (CBDMQ), la cual se complementa con registros meteorológicos provenientes de las estaciones meteorológicas institucionales del CBDMQ y de la Empresa Pública Metropolitana de Agua Potable y Saneamiento (EPMAPS), la integración de estas fuentes permitió construir una base de datos unificada con variables como temperatura, humedad relativa, precipitación, radiación solar, velocidad del viento y coordenadas geográficas asociadas con las estaciones meteorológicas, proporcionando información representativa de las condiciones ambientales asociadas a la ocurrencia de incendios.

Para desarrollar el modelo se utilizó la metodología CRISP-DM, que permitió organizar las diferentes etapas de la investigación. Primero se integraron y depuraron los datos para posteriormente realizar el análisis exploratorio. Con la información preparada se

construyeron tres modelos de clasificación: Regresión Logística, Random Forest y XGBoost. Después se optimizaron los hiperparámetros y se comparó el desempeño de los modelos mediante diferentes métricas. Finalmente, se utilizó SHAP (SHapley Additive exPlanations) para analizar el aporte de las variables en las predicciones y facilitar la interpretación de los resultados.

Los resultados obtenidos permitieron identificar las variables meteorológicas y espaciales con mayor influencia en las predicciones y comparar la capacidad predictiva de los algoritmos evaluados. A partir de estos resultados, la investigación proporciona información que puede servir como apoyo para la toma de decisiones relacionadas con la prevención, planificación y gestión del riesgo de incendios forestales en el Distrito Metropolitano de Quito.

1.2 Objetivos

1.2.1 Objetivo general:

Desarrollar un modelo predictivo basado en algoritmos de Machine Learning para la predicción de incendios forestales en el Distrito Metropolitano de Quito mediante el análisis de variables ambientales e información histórica de incendios forestales, con el fin de fortalecer la prevención de incendios forestales y la toma de decisiones en la gestión de riesgos.

1.2.2 Objetivos específicos:

- Identificar las variables climáticas, ambientales y la información histórica de eventos forestales que tienen mayor influencia en la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito.
- Comparar el desempeño de diferentes algoritmos de Machine Learning Random Forest, Regresión Logística y XGBoost, para determinar cuál presenta mejores resultados en la predicción de incendios forestales en el Distrito Metropolitano de Quito.
- Evaluar cómo un modelo predictivo basado en Machine Learning puede contribuir a la toma de decisiones para la prevención y control de incendios forestales en el Distrito Metropolitano de Quito.

1.3 Justificación

La presente investigación se desarrolla ante la relevancia de los incendios forestales en el Distrito Metropolitano de Quito, debido a sus impactos sobre la población, la calidad del aire, la biodiversidad y las áreas protegidas. Estos eventos constituyen un problema ambiental y de gestión del riesgo que evidencia la necesidad de desarrollar herramientas tecnológicas capaces de contribuir a la identificación anticipada de condiciones asociadas con su ocurrencia y, de esta forma, fortalecer las acciones de prevención y respuesta.

En ese sentido la utilización de modelos de Machine Learning permite analizar grandes volúmenes de datos e integrar variables climáticas geográficas e información histórica de eventos forestales, lo cual permite identificar patrones asociados a la ocurrencia de incendios forestales. El desarrollo de un modelo predictivo permitirá estimar condiciones asociadas con una mayor probabilidad de ocurrencia de incendios forestales, contribuyendo a la identificación de escenarios de riesgo y al apoyo de la planificación y priorización de recursos.

Esta investigación promueve la aplicación de ciencia de datos e inteligencia artificial en la solución de problemas ambientales que afectan en gran medida a la sociedad en general y a la biodiversidad.

Desde el punto de vista social y ambiental esta investigación busca aportar herramientas tecnológicas que faciliten la toma de decisiones preventivas y oportunas por parte de entidades como el Cuerpo de Bomberos del Distrito Metropolitano de Quito y el Ministerio del Ambiente, Agua y Transición Ecológica, con el objetivo de reducir el impacto ambiental y los riesgos asociados a los incendios forestales. De esta manera, se pretende contribuir a la protección del medio ambiente, la conservación de los ecosistemas y el fortalecimiento de la gestión del riesgo en el Distrito Metropolitano de Quito.

1.4 Preguntas de Investigación

- **Pregunta principal**

  ¿Cómo puede un modelo predictivo basado en algoritmos de Machine Learning contribuir a la predicción de incendios forestales en el Distrito Metropolitano de Quito mediante el análisis de variables climáticas, ambientales e informaciones históricas de incendios forestales?

- **Preguntas Secundarias**
    1. ¿Qué algoritmo de Machine Learning presenta mejores resultados en la predicción de incendios forestales dentro del Distrito Metropolitano de Quito?
    2. ¿Cómo podría contribuir el modelo predictivo a la toma de decisiones para la prevención y control de incendios forestales en el DMQ?
    3. ¿Qué variables climáticas, ambientales e información histórica de eventos forestales tienen mayor influencia en la ocurrencia de incendios forestales en el DMQ?

Capítulo 2: Desarrollo

2.1 Marco Teórico

Esta investigación se basa en diversas teorías que sustentan el desarrollo de modelos predictivos para incendios forestales mediante técnicas de Machine Learning, la teoría del aprendizaje automático establece que los algoritmos pueden identificar patrones a partir de datos históricos para realizar predicciones sobre nuevos eventos, siendo la base para el desarrollo de sistemas inteligentes de apoyo a la toma de decisiones (James et al., 2023; Jiawei Han & Hanghang Tong, 2023), de la misma manera la teoría de la minería de datos menciona que el análisis sistemático de grandes volúmenes de información permite descubrir relaciones y patrones ocultos útiles para la generación de conocimiento, este proceso puede estructurarse mediante la metodología CRISP-DM, ampliamente utilizada en proyectos de Ciencia de Datos (Jiawei Han & Hanghang Tong, 2023). Desde la perspectiva ambiental, la teoría de la gestión del riesgo y los estudios sobre incendios forestales reconocen que la ocurrencia y propagación del fuego dependen de la interacción entre factores climáticos, topográficos, vegetales y antrópicos, siendo variables como la temperatura, la humedad relativa, la velocidad del viento y la precipitación determinantes para identificar el riesgo de incendios (Calvin et al., 2023; Food and Agriculture Organization of the United Nations, 2024). Investigaciones recientes han demostrado que algoritmos como Random Forest y Regresión Logística presentan un alto desempeño para modelar y predecir incendios forestales a partir de datos históricos y meteorológicos, sin embargo la mayoría de estos estudios se han desarrollado en ambientes geográficos diferentes al Distrito Metropolitano de Quito, evidenciando la necesidad de generar modelos predictivos adaptados a las condiciones ambientales y climáticas del Ecuador (Liao et al., 2025; Rodrigues et al., 2014). Estas teorías y antecedentes nos indican el sustento conceptual y metodológico para la presente

investigación, justificando el uso de algoritmos de Machine Learning como herramientas para fortalecer la prevención y la gestión del riesgo de incendios forestales.

El Machine Learning se ha utilizado en diferentes investigaciones para desarrollar modelos de predicción de incendios forestales. Entre los algoritmos empleados se encuentran Random Forest, Regresión Logística, Support Vector Machine (SVM), XGBoost y Redes Neuronales Artificiales. Estos modelos permiten analizar conjuntamente variables meteorológicas, topográficas, ambientales y registros históricos de incendios con el propósito de identificar patrones relacionados con su ocurrencia.

Uno de los estudios más influyentes en la modelación espacial de incendios forestales es el desarrollado por (Rodrigues et al., 2014), quienes analizaron los factores asociados a la ocurrencia de incendios forestales de origen humano en la España continental mediante un modelo de Regresión Logística Geográficamente Ponderada (Geographically Weighted Logistic Regression, GWLR), a diferencia de los modelos de regresión logística tradicionales, que asumen que la relación entre las variables explicativas y el fenómeno estudiado es constante en todo el territorio, la GWLR permite estimar coeficientes locales para cada ubicación geográfica, considerando que la influencia de los factores de riesgo puede variar espacialmente esta metodología resulta especialmente útil para fenómenos complejos como los incendios forestales, cuya ocurrencia depende de características ambientales, topográficas y socioeconómicas propias de cada región.

En su investigación, los autores utilizaron un Sistema de Información Geográfica (SIG) para integrar información histórica de incendios con diversas variables explicativas relacionadas con la actividad humana y las características del territorio, entre las variables analizadas se incluyeron la proximidad a la interfaz urbano-forestal (Wildland–Urban Interface, WUI), la interfaz agrícola-forestal (Wildland–Agricultural Interface, WAI), la distancia a carreteras, líneas eléctricas, vías férreas y caminos forestales, además de otros factores espaciales que pueden favorecer la ignición de incendios provocados por actividades humanas, los resultados evidenciaron que la probabilidad de ocurrencia de incendios no es uniforme en todo el territorio, sino que presenta una marcada diferencia en los patrones según la ubicación geográfica, de modo que un mismo factor puede tener distinta importancia dependiendo de la localización geográfica. Por ejemplo, las zonas de interfaz urbano-forestal y agrícola-forestal mostraron una fuerte asociación con una mayor frecuencia de incendios forestales, mientras que la influencia de infraestructuras

como carreteras o líneas eléctricas, también varió significativamente entre regiones (Rodrigues et al., 2014).

La investigación de (Liao et al., 2025) desarrolló un modelo de predicción del riesgo de incendios forestales en Corea del Sur utilizando datos procesados mediante Google Earth Engine y comparó el desempeño de tres algoritmos de aprendizaje automático los cuales fueron Random Forest, XGBoost y Redes Neuronales Artificiales, para el entrenamiento de los modelos se integraron doce variables ambientales derivadas de sensores remotos, entre ellas factores climáticos, índices de vegetación, características topográficas, densidad poblacional y otras variables relacionadas con el territorio. Antes del modelado realizaron un análisis de correlación para retirar las variables que presentaban alta colinealidad.

Los resultados reportados por (Liao et al., 2025) mostraron que los modelos evaluados permitieron identificar patrones espaciales relacionados con los incendios forestales y elaborar mapas de probabilidad de riesgo. XGBoost presentó el mejor rendimiento global, mientras que Random Forest también mostró un desempeño competitivo. Este estudio es importante para la comparación de algoritmos realizada en esta investigación, aunque los resultados deben interpretarse considerando las diferencias entre las variables, el territorio y la metodología utilizada en cada trabajo.

(Zhou et al., 2025) proponen un modelo de predicción del riesgo de incendios forestales basado en el algoritmo Extreme Gradient Boosting (XGBoost), denominado Explainable Wildfire Prediction System (EWXS), este sistema integra variables meteorológicas, características topográficas, índices de vegetación y factores relacionados con la actividad humana para construir un modelo predictivo robusto capaz de estimar el riesgo de incendios forestales, una de las principales contribuciones de esta investigación es la incorporación de técnicas de Inteligencia Artificial Explicable (Explainable Artificial Intelligence, XAI), específicamente mediante SHapley Additive exPlanations (SHAP), que permiten interpretar la influencia de cada variable sobre la predicción realizada por el modelo, de esta manera, además de obtener predicciones de alta precisión, el sistema proporciona información transparente sobre los factores que incrementan el riesgo de incendio, fortaleciendo la confianza de los responsables de la gestión forestal en el proceso de toma de decisiones.

Los resultados presentados por (Zhou et al., 2025) mostraron un buen desempeño de XGBoost en la predicción de incendios forestales. Mediante el análisis de importancia de variables, los autores identificaron factores relacionados con la proximidad a asentamientos humanos, la temperatura, la humedad del aire y las características de la vegetación. El estudio también incorporó técnicas de Inteligencia Artificial Explicable para comprender la influencia de estas variables en las predicciones. Este trabajo es un antecedente relevante para la presente investigación, debido a que combina un algoritmo de Machine Learning con técnicas orientadas a facilitar la interpretación de sus resultados.

La presente investigación se fundamenta en la Ciencia de Datos y el Machine Learning para analizar información histórica e identificar patrones relacionados con la ocurrencia de incendios forestales. (Jiawei Han & Hanghang Tong, 2023) mencionan que la minería de datos permite descubrir patrones y relaciones en grandes volúmenes de información, lo que facilita su aplicación en problemas de clasificación y predicción. Para organizar las diferentes etapas de este trabajo se utilizó la metodología CRISP-DM, que comprende la comprensión del problema, comprensión y preparación de los datos, modelado, evaluación e implementación.

En esta investigación, el modelo predictivo se utiliza para analizar los datos históricos disponibles y estimar la probabilidad de ocurrencia de incendios forestales a partir de los patrones identificados durante el entrenamiento.

Desde el punto de vista ambiental, la ocurrencia de incendios forestales está relacionada con diferentes factores climáticos, topográficos, relacionados con la vegetación y antrópicos. Entre las variables meteorológicas utilizadas para evaluar el riesgo de incendios se encuentran la temperatura, la humedad relativa, la velocidad del viento y la precipitación IPCC (2023) y la FAO (2023). Estas variables forman parte de la información considerada para la construcción del modelo predictivo de esta investigación

A pesar de los avances tecnológicos la mayoría de los estudios han sido desarrollados utilizando información de otros países y bajo condiciones climáticas distintas a las del Distrito Metropolitano de Quito esta situación evidencia una brecha de investigación respecto a la construcción de modelos predictivos adaptados al contexto ecuatoriano, tomando en cuenta datos históricos de incendios forestales y registros meteorológicos del CBDMQ. De esta manera la presente investigación busca aportar evidencia científica

mediante la comparación de algoritmos de Machine Learning y la identificación de las variables con mayor influencia en la ocurrencia de incendios forestales, de esta manera fortalecer la toma de decisiones orientadas a la prevención y gestión del riesgo.

2.2. Metodología

La presente investigación tiene un enfoque cuantitativo, debido a que se busca analizar datos históricos y variables meteorológicas mediante técnicas estadísticas y algoritmos de Machine Learning, con el propósito de desarrollar un modelo predictivo que permita estimar la probabilidad de ocurrencia de incendios forestales en el Distrito Metropolitano de Quito (DMQ). Esta investigación es de tipo aplicado, con alcance predictivo y un diseño no experimental, longitudinal y retrospectivo, ya que utiliza información histórica sin manipular las variables de estudio.

2.2.1. Métodos de recolección de datos

La presente investigación utilizó información proveniente de fuentes institucionales oficiales. Los datos fueron obtenidos mediante solicitud formal de información dirigida al Cuerpo de Bomberos del Distrito Metropolitano de Quito (CBDMQ), con el propósito de acceder a los registros históricos de las ocho estaciones meteorológicas institucionales y a la base de datos de eventos de incendios forestales ocurridos en el Distrito Metropolitano de Quito durante el período comprendido entre los años 2021 y 2026.

La información meteorológica de la Empresa Pública Metropolitana de Agua Potable y Saneamiento (EPMAPS) fue obtenida del portal oficial https://paramh2o.aguaquito.gob.ec/, disponible en la plataforma PARAMH2O. Esta plataforma proporciona información de libre acceso a la ciudadanía proveniente de 46 estaciones meteorológicas. Para la presente investigación se recopilaron los registros correspondientes al período comprendido entre los años 2021 y 2026, los cuales fueron utilizados en el proceso de integración, análisis y construcción del modelo predictivo de incendios forestales en el Distrito Metropolitano de Quito.

Las principales fuentes de información fueron:

- Registros históricos de eventos forestales atendidos por el CBDMQ.
- Variables meteorológicas registradas por las ocho estaciones meteorológicas institucionales, como son: temperatura, humedad relativa, velocidad y dirección del viento, precipitación, presión atmosférica y otras variables disponibles.

- Registros Meteorológicos de Empresa Pública Metropolitana de Agua Potable y Saneamiento
- Información geográfica relacionada con la ubicación de los eventos forestales.
- Variables temporales derivadas de la fecha y hora de ocurrencia de cada incendio.

Durante la revisión de los registros institucionales se identificó un cambio en el criterio técnico utilizado para la clasificación de los incendios forestales a partir de 2025. Esta modificación fue considerada durante la preparación, segmentación temporal e interpretación de los datos, debido a sus implicaciones sobre la comparabilidad de los registros históricos.

Posteriormente, los datos fueron integrados en una única base de datos para su procesamiento, durante esta etapa se realizó un proceso de depuración que incluyó:

- Eliminación de registros duplicados.
- Corrección de inconsistencias en los datos.
- Tratamiento de valores faltantes.
- Homologación de formatos de fecha, hora y unidades de medida.
- Unificación de las bases de datos meteorológicas provenientes de las ocho estaciones del CBDMQ y de las estaciones de la Empresa Pública Metropolitana de Agua Potable y Saneamiento (EPMAPS).
- Integración de las variables meteorológicas con los registros históricos de incendios forestales mediante criterios espaciales y temporales.
- Selección y transformación de variables para el modelo predictivo.
- Generación del conjunto de datos final para el entrenamiento.
- Entrenamiento de los modelos predictivos.
- Optimización de los modelos predictivos.
- Evaluación y validación del desempeño de los modelos mediante métricas como Accuracy, Precision, Recall, F1-Score y ROC-AUC.
- Interpretación y explicación del modelo predictivo mediante SHAP.

La Base de Datos Meteorológica Unificada es la integración de los registros meteorológicos provenientes de las estaciones del Cuerpo de Bomberos del Distrito Metropolitano de Quito y de la Empresa Pública Metropolitana de Agua Potable y Saneamiento. Esta integración permitió disponer de información de temperatura, humedad relativa, precipitación, velocidad del viento y radiación solar para el periodo

analizado. La base consolidada fue utilizada posteriormente para integrar la información meteorológica con los registros históricos de incendios forestales y preparar los datos empleados en el modelado predictivo.

Una vez consolidada la base de datos, se efectuó un análisis exploratorio para identificar patrones, distribuciones y posibles relaciones entre las variables.

2.2.2. Técnicas de análisis

Para el desarrollo del modelo predictivo se seguió la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente utilizada en proyectos de minería de datos y ciencia de datos por proporcionar un proceso estructurado para el desarrollo de modelos predictivos.

Las etapas contempladas son:

Exploración de los datos (Exploratory Data Analysis - EDA)

Se realizó un análisis exploratorio con el propósito de comprender las características del conjunto de datos históricos de incendios forestales y de las variables meteorológicas utilizadas en la investigación, se han empleado técnicas de estadística descriptiva y visualización para identificar la distribución de las variables, detectar valores atípicos, analizar la relación entre las variables independientes y la variable objetivo, así como identificar posibles patrones asociados a la ocurrencia de incendios forestales.

Limpieza y preparación de los datos

Se revisaron los incendios forestales desde el año 2021 a mayo del 2026 teniendo un total de 2615.

La información meteorológica utilizada corresponde a las ocho (8) estaciones del Cuerpo de Bomberos del Distrito Metropolitano de Quito (CBDMQ), las cuales registran variables como temperatura, humedad relativa, velocidad del viento, precipitación y otras condiciones ambientales relacionadas con los incendios forestales.

Los registros históricos de estas estaciones se encuentran disponibles únicamente a partir del año 2025, motivo por el cual fue necesario complementar la información con los registros de las estaciones meteorológicas de la Empresa Pública Metropolitana de Agua Potable y Saneamiento (EPMAPS), con el objetivo de tener datos para el periodo de estudio. Luego, ambas fuentes fueron integradas y cada incendio forestal fue relacionado

con la estación meteorológica más cercana, tomando en cuenta las coordenadas geográficas (latitud y longitud) de los incidentes.

Posteriormente se llevó a cabo el proceso de preparación del conjunto de datos, con el fin de garantizar su calidad antes del entrenamiento de los modelos predictivos. Esta etapa comprendió la eliminación de registros duplicados, el tratamiento de valores faltantes, la corrección de inconsistencias, la transformación de formatos de fecha y hora, la codificación de variables categóricas y la normalización o estandarización de las variables numéricas cuando sea requerida por los algoritmos de aprendizaje automático.

Asimismo, se realizó la selección de las variables más relevantes para el modelo mediante técnicas de análisis estadístico y de importancia de variables, con el objetivo de reducir la dimensionalidad del conjunto de datos y mejorar el desempeño predictivo.

Modelado predictivo

Una vez preparado el conjunto de datos, se desarrollaron diferentes modelos de Machine Learning orientados a estimar la probabilidad de ocurrencia de incendios forestales en el Distrito Metropolitano de Quito. Para el modelado se implementaron tres algoritmos de clasificación: Regresión Logística, Random Forest y XGBoost. Posteriormente, se comparó su desempeño para determinar cuál presentaba mejores resultados con los datos utilizados en esta investigación.

Para evitar fuga de información temporal y evaluar la capacidad de generalización de los modelos en condiciones más cercanas a un escenario predictivo real, se utilizó una partición temporal de los datos, los registros de 2021–2022 se emplearon para el entrenamiento interno, 2023 para la validación y selección de hiperparámetros, y 2024 se mantuvo completamente separado como conjunto de prueba temporal final, los registros correspondientes a 2025–2026 fueron utilizados posteriormente como una evaluación adicional, considerando el cambio identificado en el criterio institucional de registro de incendios forestales.

Evaluación del desempeño del modelo

El rendimiento de los modelos predictivos se evaluó mediante métricas de clasificación ampliamente utilizadas en problemas de predicción binaria, se utilizó la matriz de

confusión, la exactitud (Accuracy), la precisión (Precision), la sensibilidad (Recall), la puntuación F1 (F1-Score) y el área bajo la curva ROC (ROC-AUC).

Finalmente, se comparó el desempeño de los diferentes algoritmos con el fin de seleccionar el modelo que presente la mayor capacidad predictiva para estimar la ocurrencia de incendios forestales. Los resultados obtenidos permitieron identificar el algoritmo que presentó el mejor desempeño entre los modelos evaluados. Esta información puede servir como apoyo para la prevención, planificación y toma de decisiones relacionadas con la gestión del riesgo de incendios forestales.

Asimismo, se utilizaron técnicas de Inteligencia Artificial Explicable (XAI), particularmente SHAP (SHapley Additive Explanations), para identificar la importancia de cada variable en las predicciones del modelo.

Se ha desarrollado un procedimiento automatizado en Python para la integración de la información meteorológica proveniente de las ocho estaciones del CBDMQ y las 40 estaciones de EPMAPS, el proceso incluyó la lectura masiva de archivos, homologación de nombres de variables, normalización de formatos de fecha y hora, incorporación del identificador de estación, tratamiento de registros faltantes y consolidación de una base meteorológica unificada para su posterior integración con el histórico de incendios forestales mediante criterios espaciales y temporales.

Se desarrolló un procedimiento automatizado para la corrección de variables meteorológicas con inconsistencias de escala, detectando valores fuera de los rangos físicos esperados y aplicando una transformación basada en factores de conversión definidos para cada tipo de sensor.

2.2.3. Justificación de los métodos

Se seleccionó un enfoque cuantitativo porque permite analizar la relación entre las variables meteorológicas y los registros históricos de incendios mediante técnicas estadísticas y de Machine Learning. A partir de este enfoque es posible medir el comportamiento de las variables y obtener resultados reproducibles para el desarrollo del modelo predictivo (Creswell & Creswell, 2023).

La utilización de datos históricos permite construir modelos predictivos basados en evidencia objetiva, evitando interpretaciones subjetivas y facilitando la reproducibilidad

de los resultados, el análisis de datos históricos favorece la identificación de patrones y tendencias que apoyan la predicción de eventos futuros mediante técnicas de Machine Learning (James et al., 2023).

Se seleccionó la metodología CRISP-DM porque permite organizar el desarrollo del proyecto en diferentes etapas, desde la comprensión del problema y la preparación de los datos hasta el modelado y la evaluación de los resultados (Chapman, 2000). Esta estructura facilitó el desarrollo ordenado del modelo predictivo de la investigación.

Los algoritmos seleccionados permitieron comparar diferentes enfoques de clasificación, la Regresión Logística se utilizó como modelo base debido a su simplicidad e interpretabilidad para problemas de clasificación binaria (James et al., 2023). Random Forest permitió modelar relaciones no lineales mediante la combinación de múltiples árboles de decisión (Breiman, 2001), por otro lado XGBoost utiliza un enfoque de gradient boosting basado en árboles y permite modelar relaciones complejas entre las variables predictoras (Chen & Guestrin, 2016). La comparación de los tres algoritmos permitió evaluar su desempeño bajo las mismas condiciones de entrenamiento y determinar cuál presentó mejores resultados para los datos utilizados en esta investigación.

Finalmente, la incorporación de técnicas de Inteligencia Artificial Explicable (XAI), como SHAP, permitió interpretar el comportamiento del modelo e identificar cuáles variables meteorológicas tienen mayor influencia en la ocurrencia de incendios forestales, fortaleciendo la transparencia del modelo y facilitando su utilización como herramienta de apoyo para la toma de decisiones del Cuerpo de Bomberos del Distrito Metropolitano Quito.

2.3. Resultados y Análisis

En esta sección se presentan los resultados obtenidos durante el desarrollo del modelo predictivo de incendios forestales para el Distrito Metropolitano de Quito, los resultados comprenden el procesamiento e integración de las fuentes de información, el análisis exploratorio de los datos, la preparación del conjunto destinado al modelado, la construcción y evaluación de los algoritmos de Machine Learning, la optimización del modelo seleccionado y su posterior interpretación mediante SHAP.

Los resultados se presentan mediante estadísticas descriptivas, tablas, gráficos y métricas de evaluación. Posteriormente, en la sección de análisis, los hallazgos son interpretados en función de los objetivos y preguntas de investigación, así como de los antecedentes científicos considerados en el estudio.

2.3.1. Resultados

Para facilitar la presentación de los hallazgos obtenidos durante las diferentes etapas del estudio, los resultados se organizaron de acuerdo con el proceso seguido para la construcción y evaluación del modelo predictivo.

2.3.1.1. Preparación e integración de las fuentes de datos

Para la construcción del conjunto de datos utilizado en el desarrollo del modelo predictivo se trabajó con información proveniente de tres componentes principales: registros meteorológicos del Cuerpo de Bomberos del Distrito Metropolitano de Quito (CBDMQ), registros meteorológicos de la Empresa Pública Metropolitana de Agua Potable y Saneamiento (EPMAPS) y registros históricos de incendios forestales atendidos por el CBDMQ.

Al realizar el análisis de las diferentes fuentes meteorológicas presentaban estructuras diferentes, tanto en la denominación de las variables como en su frecuencia temporal y disponibilidad histórica, motivo por el cual antes de realizar la integración con los eventos de incendios forestales fue necesario ejecutar procesos independientes de limpieza, homologación y consolidación.

Preparación de la información meteorológica del CBDMQ

Los registros meteorológicos proporcionados por el CBDMQ se encontraban distribuidos en diferentes archivos y hojas de cálculo correspondientes a las estaciones meteorológicas institucionales, dado que los archivos no mantenían una estructura completamente uniforme, se realizó un proceso de identificación de encabezados, estandarización de nombres de columnas, construcción de la variable fecha_hora, conversión de variables meteorológicas a formatos numéricos y normalización de los nombres de las estaciones.
Entre las principales variables identificadas se encontraron temperatura, humedad relativa, precipitación, radiación solar, velocidad del viento y dirección del viento. También se encontraron variables adicionales, como punto de rocío y ráfaga de viento, además de campos específicos generados por los equipos de las diferentes estaciones.

Como resultado del proceso de consolidación, la base del CBDMQ alcanzó inicialmente 13.709.654 registros, la validación posterior permitió conservar las variables necesarias para el proceso de integración y verificar la disponibilidad de información meteorológica en el período comprendido entre el 1 de enero de 2025 y el 30 de abril de 2026.
Para complementar la dimensión espacial de los registros, se incorporaron las coordenadas geográficas de las estaciones mediante las variables latitud_estacion y longitud_estacion, de esta forma cada observación meteorológica quedó asociada no solamente con una fecha y hora, sino también con la ubicación de la estación de donde se obtuvo el registro de información.

Preparación de la información meteorológica de EPMAPS

La información meteorológica proporcionada por EPMAPS fue procesada de manera independiente debido a las diferencias existentes respecto de la estructura de los registros del CBDMQ, después de la limpieza y homologación se obtuvo una base compuesta por 1.646.215 registros correspondientes a 40 estaciones meteorológicas. No se identificaron registros duplicados al considerar la combinación de estación y fecha-hora.

El período disponible en esta fuente se extendió desde el 1 de enero de 2021 hasta el 5 de julio de 2026**,** permitiendo ampliar considerablemente la cobertura histórica de la información meteorológica utilizada en la investigación.

Las variables consolidadas incluyeron código y nombre de estación, fecha y hora, temperatura, humedad relativa, precipitación, radiación solar, velocidad del viento, dirección del viento y fuente. Posteriormente, se normalizaron los nombres de las estaciones y se incorporaron sus respectivas coordenadas geográficas.

La revisión de completitud evidenció diferencias importantes entre variables. La precipitación presentó 13,63 % de valores nulos**,** mientras que otras variables mostraron porcentajes mayores de ausencia de información: temperatura 72,59 %**,** humedad relativa 74,99 %**,** velocidad del viento 76,04 %**,** radiación solar 76,11 % y dirección del viento 89,13 %**,** esta situación se relacionó con la heterogeneidad de las estaciones disponibles, debido a que no todas registraban el mismo conjunto de variables meteorológicas.

Consolidación de las fuentes meteorológicas

Una vez procesadas ambas fuentes, se realizó la homologación de las variables comunes y posteriormente la concatenación de los registros meteorológicos del CBDMQ y EPMAPS.

La base meteorológica consolidada quedó conformada por 15.355.869 registros, distribuidos de la siguiente manera:

Fuente	Registros	Participación
CBDMQ	13.709.654	89,28%
EPMAPS	1.646.215	10,72%
Total	15.355.869	100,00%

Tabla 1:Distribución de registros de la base meteorológica consolidada.
Nota. Elaboración propia a partir de los registros meteorológicos del CBDMQ y EPMAPS.

La integración permitió obtener una estructura común que incluyó fecha_hora, estación, estación normalizada, latitud y longitud de la estación, temperatura, humedad relativa, precipitación, radiación solar, velocidad del viento, dirección del viento y fuente, además de algunas variables complementarias disponibles en determinados registros.

Este procedimiento permitió ampliar la cobertura temporal de la información meteorológica, debido a que los registros del CBDMQ se encontraban disponibles principalmente desde 2025, mientras que EPMAPS proporcionó información desde 2021.

Preparación de los registros históricos de incendios forestales

La base original de incendios forestales estuvo constituida por 2.623 registros y 11 variables, los campos disponibles se encontraron código de parte, fecha y hora del evento, tipo de evento, causa presuntiva, código de estación, zona, parroquia, barrio, latitud y longitud.

Durante la preparación se estandarizaron los nombres de las variables y se construyó el campo fecha_hora, la conversión temporal fue satisfactoria para los 2.623 registros, sin identificarse fechas inválidas o nulas.

Luego se verificó la calidad de las coordenadas geográficas, inicialmente se identificó un registro con latitud y longitud iguales a cero**,** el cual fue eliminado, después de esta depuración permanecieron 2.622 observaciones, una segunda validación espacial identificó siete registros adicionales con coordenadas fuera de los rangos considerados válidos, por lo que también fueron excluidos.

Como resultado, la base final quedó conformada por 2.615 registros de incendios forestales, lo que representa la conservación del 99,70 % de los registros originales y la

eliminación de ocho observaciones debido a problemas relacionados con sus coordenadas geográficas.

Etapa	Registros	Eliminados
Base original de incendios	2.623	---
Eliminación de coordenadas (0,0)	2.622	1
Validación de rangos geográficos	2.615	7
Base final	2.615	8

Tabla 2: Resultado del proceso de depuración de los registros de incendios forestales
Fuente: Creación de autor, D. Tiupul (2026)

La base final de incendios quedó compuesta por 22 variables, incluyendo información temporal derivada año, mes, día, hora, minuto y día de la semana, información territorial zona, parroquia, barrio, latitud y longitud y características relacionadas con el evento. Adicionalmente, se generó la variable incendio, con valor igual a 1 para identificar los registros correspondientes a eventos reales de incendios forestales. El período de los eventos válidos comprendió desde el 1 de enero de 2021 hasta el 26 de mayo de 2026, sin valores nulos en la variable temporal final.

Resultado de la preparación de las fuentes

El proceso permitió obtener dos conjuntos depurados que luego fueron utilizados para la integración espacio-temporal: una base meteorológica consolidada con 15.355.869 observaciones provenientes del CBDMQ y EPMAPS, y una base histórica depurada con 2.615 registros de incendios forestales.

Base procesada	Registros finales	Período disponible	Resultado
Meteorología CBDMQ	13.709.654	01/01/2025 – 30/04/2026	8 estaciones, Consolidada y homologada
Meteorología EPMAPS	1.646.215	01/01/2021 – 05/07/2026	40 estaciones, sin duplicados estación-fecha
Meteorología unificada	15.355.869	2021–2026	CBDMQ + EPMAPS
Incendios forestales	2.615	01/01/2021 – 26/05/2026	Registros temporales y espaciales válidos

Tabla 3:Resumen de las fuentes de información preparadas para el estudio
Fuente: Creación de autor, D. Tiupul (2026)

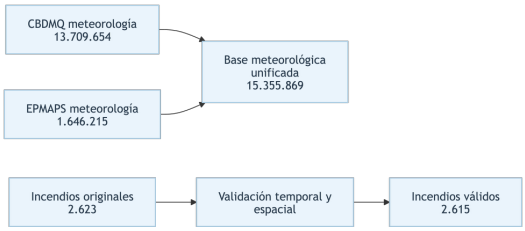


Figura 1:Diagrama de proceso de preparación
Fuente: Creación de autor, D. Tiupul (2026)

2.3.1.2. Resultados del análisis exploratorio de datos

Una vez preparadas y consolidadas las fuentes de información se realizó el análisis exploratorio de datos con el propósito de caracterizar el comportamiento temporal,

espacial y meteorológico de los registros disponibles, así como identificar valores faltantes, posibles valores atípicos y relaciones entre las variables consideradas para el desarrollo posterior del modelo predictivo.

Análisis exploratorio de los incendios forestales

La base depurada estuvo conformada por 2.615 registros de incendios forestales y 22 variables, correspondientes al período comprendido entre el 1 de enero de 2021 y el 26 de mayo de 2026, la revisión de calidad no identificó valores nulos ni registros duplicados en esta base.

La distribución anual de los eventos presentó diferencias importantes durante el período analizado, el mayor número de incendios se registró en 2021, con 992 eventos, seguido de 2023 con 868, 2022 con 375 y 2024 con 334, para 2025 y 2026 se registraron 40 y 6 eventos, respectivamente.

Año	Incendios forestales
2021	992
2022	375
2023	868
2024	334
2025	40
2026*	6
Total	**2.615**

Tabla 4:Distribución anual de los incendios forestales registrados
Nota. El año 2026 contiene información disponible hasta el 26 de mayo. Elaboración D. Tiupul (2026)

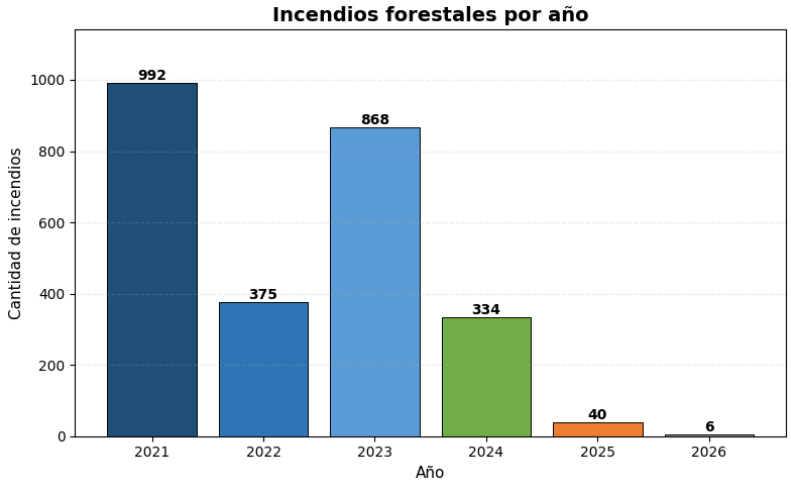


Figura2:Distribución de incendios forestales por año.
Fuente: Creación de autor, D. Tiupul (2026, a partir de los registros históricos de incendios forestales del CBDMQ.

Cambio en el criterio institucional de clasificación de incendios forestales

La reducción observada en el número de incendios forestales registrados a partir de 2025 debe interpretarse considerando un cambio en el criterio técnico institucional utilizado para clasificar estos eventos, de acuerdo con el procedimiento institucional aplicado en el

registro de emergencias, antes de 2025 la clasificación se encontraba asociada principalmente con la superficie afectada: los eventos con un área quemada igual o superior a 100 m² eran registrados como incendios forestales, mientras que aquellos con una superficie inferior eran clasificados como quemas de desechos.

A partir de 2025 la clasificación de incendios forestales dejó de depender exclusivamente del área quemada y pasó a considerar principalmente las características del material involucrado en la combustión, en especal el tipo de vegetación afectada, por lo tanto un evento ya no se clasifica como incendio forestal únicamente por superar un determinado número de metros cuadrados.

El cambio de este criterio técnico institucional genera una ruptura en la comparabilidad de los registros históricos, siendo el motivo por el cual la disminución observada a partir de 2025 no debe interpretarse como una disminución de incendios forestales dentro del DMQ, debido a que parte de esta diferencia está relacionada con la modificación del criterio utilizado para su clasificación.

Distribución mensual de los incendios forestales

En el análisis mensual evidenció una marcada concentración de los registros durante los meses de julio, agosto y septiembre, donde septiembre presentó el mayor número de eventos, con 781 incendios, seguido de agosto con 738 y julio con 394.

En conjunto, estos tres meses acumularon 1.913 de los 2.615 registros, equivalentes aproximadamente al 73,2 % de los incendios analizados.

Los restantes meses presentaron cantidades considerablemente inferiores: enero 85, febrero 72, marzo 37, abril 50, mayo 52, junio 87, octubre 159, noviembre 79 y diciembre 81.

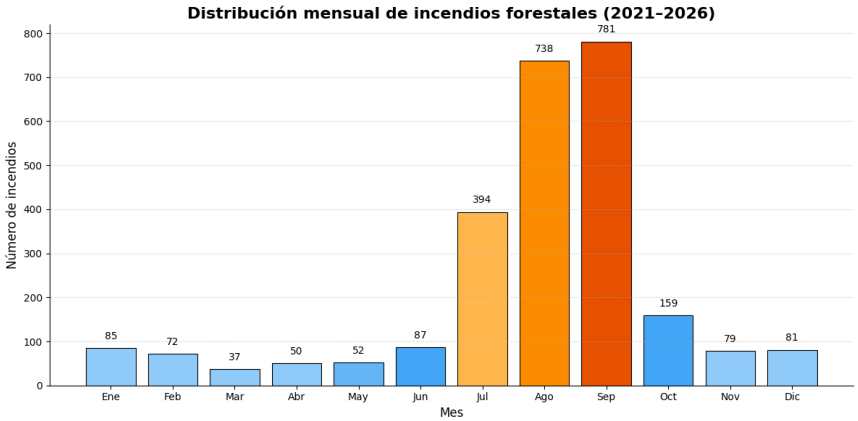


Figura 3:Distribución mensual de incendios forestales.
Fuente: Creación de autor, D. Tiupul (2026)

Distribución horaria

La distribución por hora también evidenció una concentración temporal, el mayor número de registros ocurrió a las 12:00, con 330 incendios, seguido de las 11:00 con 303, 13:00 con 296, 14:00 con 251 y 15:00 con 241.En contraste durante la madrugada se registraron frecuencias considerablemente menores, entre las 00:00 y las 05:00 se contabilizaron entre 5 y 12 eventos por hora, los resultados muestran una mayor concentración de los registros durante las horas del mediodía.

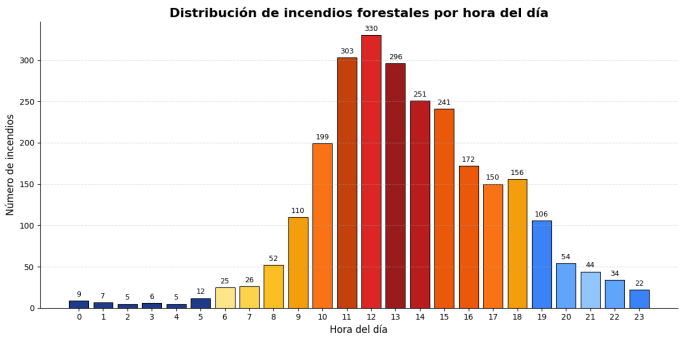


Figura 4:Distribución horaria de los incendios forestales.
Fuente: Creación de autor, D. Tiupul (2026)

Distribución espacial de los incendios

El análisis territorial mostró que los 2.615 registros se distribuyeron entre 68 parroquias. La mayor frecuencia correspondió a El Condado con 182 eventos, seguida de Calderón (148), Conocoto (124), Pifo (119), Carcelén (100), Pintag (96) y Alangasí (90).

Las diez parroquias con mayor cantidad de registros fueron:

Parroquia	Incendios
El Condado	182
Calderón	148
Conocoto	124
Pifo	119
Carcelén	100
Pintag	96
Alangasí	90
Tumbaco	80
Puengasí	77
El Quinche	73

Tabla 5:Parroquias con mayor número de incendios forestales

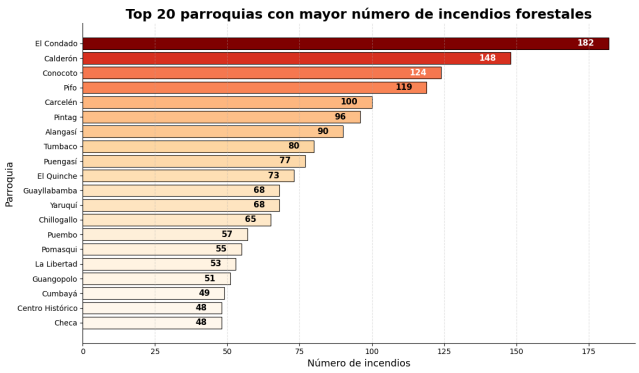


Figura 5:Top 20 parroquias con mayor número de incendios forestales.

Distribución geográfica de los incendios

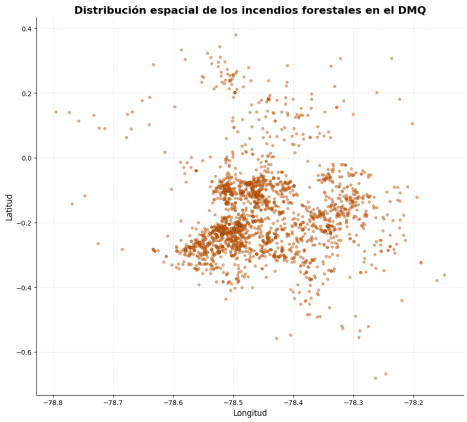


Figura 6:Distribución espacial de los incendios forestales.

El mapa evidencia que los incendios se distribuyen a lo largo del Distrito Metropolitano de Quito.

Mapa de densidad

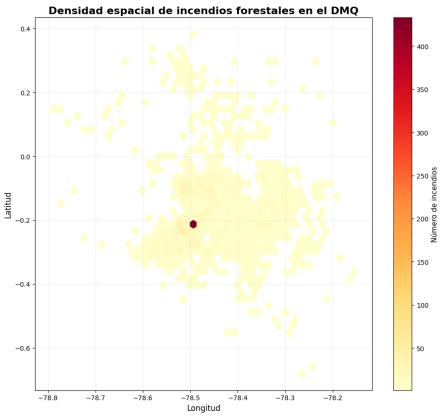


Figura 7:Mapa de densidad de incendios forestales.

El mapa de densidad permite identificar los sectores donde históricamente se registra una mayor concentración de incendios forestales, estas áreas representan zonas de mayor exposición al riesgo y constituyen información relevante para la planificación de acciones preventivas y la asignación de recursos operativos.

Causas presuntivas

Con el propósito de identificar las principales causas asociadas a la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito, se analizó la variable causa presuntiva**,** cuyos resultados se presentan

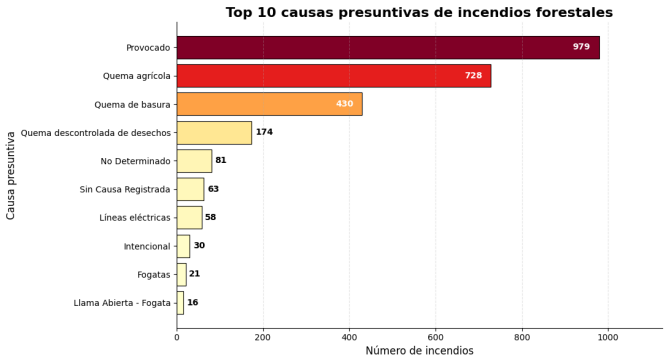


Figura 8: Top 10 de Causas Presuntivas

En la figura de Causas presuntivas se visualiza que la causa presuntiva con mayor frecuencia corresponde a Provocado, con 979 registros**,** constituyéndose en la principal causa identificada durante el período de estudio, en segundo lugar se ubica la Quema agrícola, con 728 eventos, seguida de Quema de basura, con 430 registros, y Quema descontrolada de desechos, con 174 eventos**.**

Estos resultados son consistentes con diversos estudios realizados en Ecuador y América Latina, los cuales señalan que la mayoría de los incendios forestales son originados por actividades humanas, especialmente quemas agrícolas, quema de residuos, fogatas y acciones negligentes o intencionales(Ríos Hervas, 2023). En este contexto, la evidencia obtenida en la presente investigación refuerza la necesidad de implementar programas de prevención y sensibilización dirigidos a reducir las causas antrópicas de los incendios forestales.

Estaciones de atención

Finalmente, se evaluó la distribución de los incendios según la estación operativa responsable de la atención.

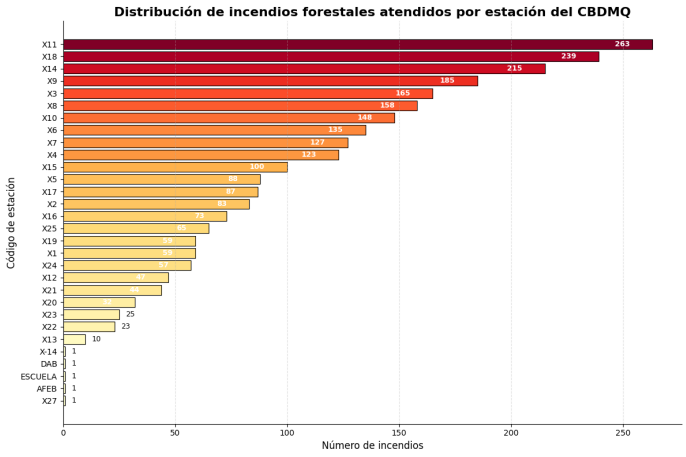


Figura 9:Incendios atendidos por estación.

La distribución de los eventos por estación refleja diferencias en la carga operativa de las unidades del Cuerpo de Bomberos del Distrito Metropolitano de Quito, información que resulta útil para el análisis de cobertura territorial y la planificación de recursos institucionales.

Análisis exploratorio de la información meteorológica

Con el objetivo de caracterizar las condiciones meteorológicas asociadas a los incendios forestales, se realizó un análisis exploratorio de la base meteorológica consolidada, integrada con los registros provenientes de las estaciones meteorológicas del Cuerpo de Bomberos del Distrito Metropolitano de Quito (CBDMQ) y de la Empresa Pública Metropolitana de Agua Potable y Saneamiento (EPMAPS). Este análisis permitió evaluar la calidad de los datos, describir el comportamiento de las principales variables meteorológicas e identificar posibles relaciones entre ellas antes del proceso de modelado.

Calidad de los datos

Como primera etapa se evaluó la integridad de la información mediante el análisis de valores faltantes y registros duplicados, la identificación de datos ausentes permitió conocer el nivel de disponibilidad de cada variable meteorológica y determinar aquellas que requerían un tratamiento durante la etapa de preparación de datos.

Variable	Valores válidos	Valores faltantes	% faltante
Temperatura	14.008.795	1.347.074	8,77%
Humedad relativa	732.620	14.623.249	95,23%
Precipitación	15.102.826	253.043	1,65%
Radiación solar	755.970	14.599.899	95,08%
Velocidad del viento	14.066.938	1.288.931	8,39%
Dirección del viento	508.798	14.847.071	96,69%

Tabla 6:Valores faltantes de las variables meteorológicas.

Los resultados muestran que la disponibilidad de información varía entre las diferentes variables meteorológicas, la dirección del viento presentó el mayor porcentaje de valores faltantes, seguida de la velocidad del viento, la radiación solar, la humedad relativa y la temperatura, en cambio la precipitación registró el menor porcentaje de datos ausentes. Esta situación se encuentra asociada a diferencias en la disponibilidad de sensores y a la cobertura temporal de las estaciones meteorológicas utilizadas en la investigación.

Como parte del proceso de preparación de los datos, se verificó la calidad de la información meteorológica mediante el análisis de valores faltantes y la validación de la consistencia de las variables que se utilizaran en el modelo predictivo.

Estadísticos descriptivos

Posteriormente se calcularon las principales medidas descriptivas de las variables meteorológicas consideradas en el estudio, incluyendo temperatura, humedad relativa, precipitación, radiación solar, velocidad del viento y dirección del viento.

Variable	Media	Mediana	Desv. estándar	Mínimo	Máximo
Temperatura	14,89	13,72	5,44	-1,535	100
Humedad relativa	82,58	88,89	17,96	0	100,05
Precipitación	0,0158	0	0,2831	0	47,2
Radiación solar	180,22	4,33	279,82	0	1.279,00
Velocidad del viento	4,16	2,4	4,86	0	54,1

Tabla 7:Estadísticos descriptivos de las variables meteorológicas.

Las estadísticas descriptivas permiten caracterizar el comportamiento general de las variables empleadas en el modelo predictivo, se presentan medidas como el número de observaciones, media, desviación estándar, valores mínimo y máximo, así como los cuartiles correspondientes a cada variable, estas medidas constituyen una referencia para comprender la variabilidad de las condiciones meteorológicas registradas durante el período de estudio.

Distribución de las variables meteorológicas

Con el fin de analizar el comportamiento de las variables meteorológicas se elaboraron histogramas para cada una de ellas.

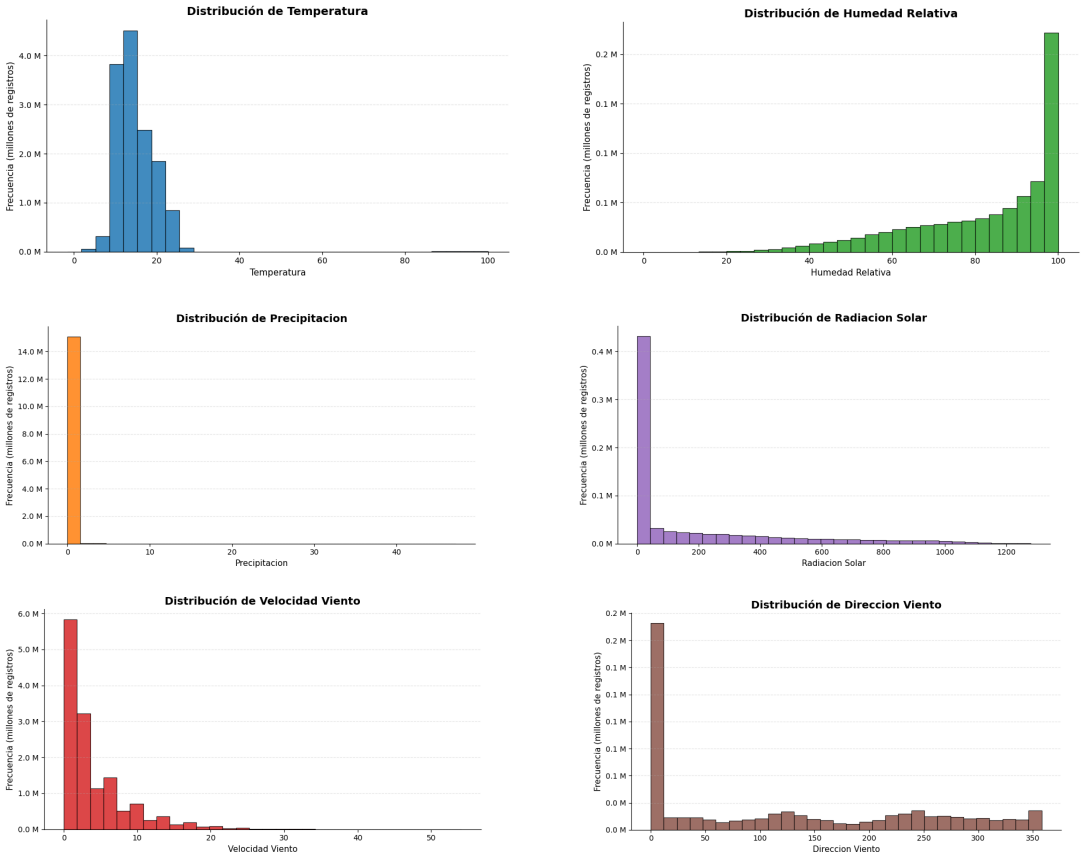


Figura 10:Distribución de las variables meteorológicas.

Las distribuciones muestran que las variables meteorológicas presentan comportamientos coherentes con las condiciones climáticas del Distrito Metropolitano de Quito. La precipitación se caracteriza por la predominancia de valores nulos, la radiación solar refleja el ciclo diario entre horas nocturnas y diurnas, la velocidad del viento se concentra en velocidades bajas y la dirección del viento exhibe una amplia dispersión angular con una elevada frecuencia de registros en 0°.

Identificación de valores atípicos

Con el propósito de identificar observaciones extremas se construyeron diagramas de caja para las principales variables meteorológicas.

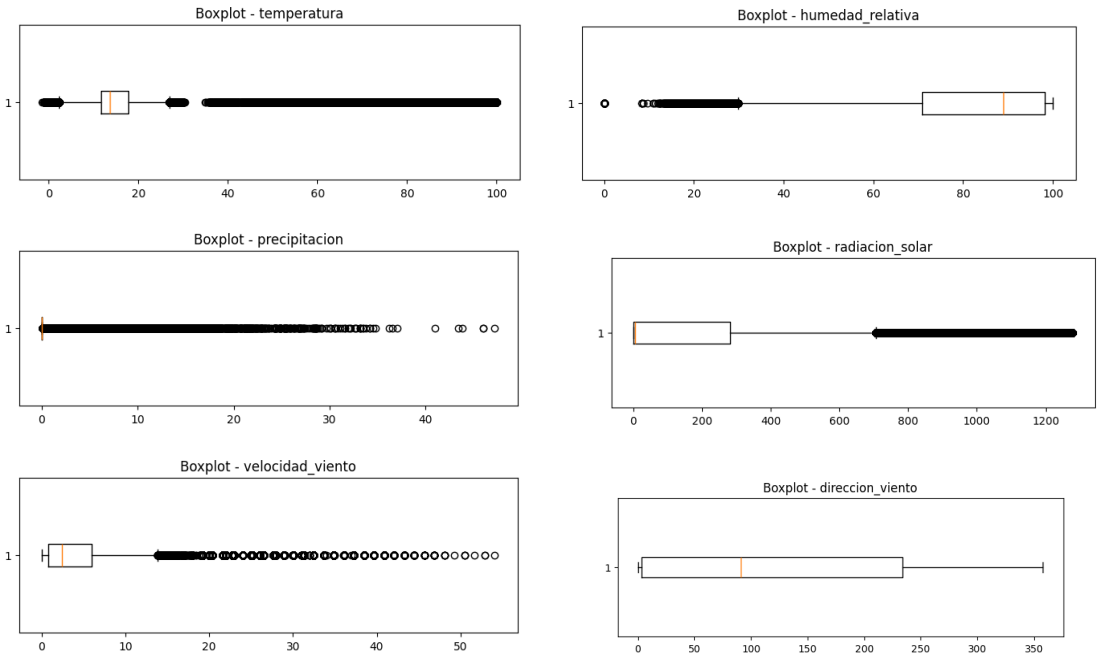


Figura 11:Diagramas de caja de las variables meteorológicas.

Los diagramas de caja muestran la presencia de valores atípicos en varias de las variables analizadas, especialmente en la precipitación, la radiación solar y la velocidad del viento. Estos valores corresponden a condiciones meteorológicas poco frecuentes, pero físicamente posibles, por lo que fueron conservados para el entrenamiento de los modelos predictivos, considerando que representan eventos reales registrados por las estaciones meteorológicas.

Correlación entre variables meteorológicas

Finalmente, se analizó la relación lineal entre las variables meteorológicas mediante una matriz de correlación.

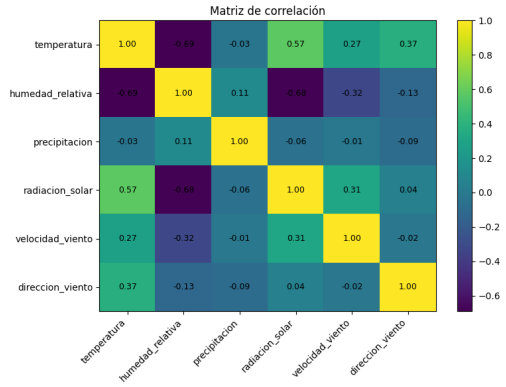


Figura 12:Matriz de correlación de las variables meteorológicas.

El análisis de correlación evidencia una fuerte correlación negativa entre la temperatura y la humedad relativa, indicando que a medida que aumenta la temperatura tiende a disminuir la humedad del ambiente, asimismo, se observa una correlación positiva alta entre la temperatura y la radiación solar, mientras que la humedad relativa presenta una relación negativa con esta última, en cambio la precipitación muestra correlaciones débiles con el resto de variables, lo que indica un comportamiento relativamente independiente dentro del conjunto de datos, finalmente, la velocidad y la dirección del viento presentan una correlación positiva moderada, asociada al comportamiento conjunto de ambas variables durante los eventos meteorológicos registrados.

Los resultados del análisis exploratorio permitieron caracterizar las condiciones meteorológicas utilizadas como variables predictoras y verificar que la información disponible era adecuada para el entrenamiento y evaluación de los modelos de aprendizaje automático desarrollados en la presente investigación.

Relación	**r de Pearson**
Temperatura – Humedad relativa	**-0,69**
Humedad relativa – Radiación solar	**-0,68**
Temperatura – Radiación solar	**0,57**
Radiación solar – Velocidad viento	0,31
Temperatura – Velocidad viento	0,27
Humedad relativa – Velocidad viento	-0,32

Tabla 8:Análisis exploratorio de la información meteorológica

La matriz de correlación de Pearson mostró que la relación lineal de mayor magnitud se presentó entre temperatura y humedad relativa, con un coeficiente de r = -0,69, indicando una relación inversa.

También se observó una correlación negativa entre humedad relativa y radiación solar (r = -0,68) y una correlación positiva moderada entre temperatura y radiación solar (r = 0,57).

La velocidad del viento presentó correlaciones positivas de menor magnitud con radiación solar **(**r = 0,31**)** y temperatura **(**r = 0,27**),** mientras que la precipitación presentó correlaciones lineales débiles con las demás variables.

2.3.1.3. Integración de información meteorológica e incendios forestales

Una vez depuradas y homologadas las bases de información, se realizó la integración de los registros históricos de incendios forestales con las observaciones meteorológicas, este procedimiento tuvo como finalidad asociar a cada evento las condiciones meteorológicas registradas por la estación más cercana espacialmente y con información disponible en el momento temporal más próximo al incendio.

Para este proceso se utilizaron 2.615 registros válidos de incendios forestales**,** previamente depurados, y la base meteorológica consolidada de 15.355.869 observaciones**.** La información meteorológica correspondió a 48 estaciones con coordenadas geográficas disponibles.

Asignación espacial de la estación meteorológica

La primera etapa consistió en determinar la estación meteorológica más cercana a cada incendio forestal, para lo cual se utilizaron las coordenadas geográficas de los eventos latitud y longitud y las coordenadas correspondientes a cada estación latitud_estacion y longitud_estacion.

La distancia entre cada incendio y las estaciones disponibles se calculó mediante la fórmula de Haversine, considerando un radio terrestre de 6.371,0088 km, para cada registro de incendio se seleccionó la estación que presentó la menor distancia geográfica.

Como resultado de este procedimiento fue posible asignar una estación meteorológica a la totalidad de los 2.615 incendios forestales, no se registraron eventos sin fecha y hora ni eventos a los que no fuera posible asignar espacialmente una estación.

Resultado de asignación espacial	Registros
Incendios disponibles	2.615
Incendios con fecha y hora válida	2.615
Incendios con estación asignada	2.615
Incendios sin estación asignada	0
Cobertura espacial	100%

Tabla 9: Resultado de la asignación espacial de estaciones meteorológicas
Fuente: Creación de autor, D. Tiupul (2026)

Este resultado permitió incorporar a cada evento las variables estacion_asignada y distancia_estacion_km, necesarias para realizar posteriormente la asociación temporal con las observaciones meteorológicas.

Integración temporal de los registros

Una vez asignada la estación más cercana, se procedió a relacionar cada incendio con el registro meteorológico temporalmente más próximo correspondiente a dicha estación.

Para ello se utilizó una unión temporal por proximidad mediante merge_asof, considerando como criterios simultáneos la estación asignada y la fecha y hora del evento. La búsqueda se realizó en ambas direcciones temporales, seleccionando la observación meteorológica más cercana al momento del incendio.

La tolerancia final establecida para la asociación fue de ±3 horas, de esta manera, únicamente se consideraron coincidencias meteorológicas cuya diferencia temporal respecto al incendio se encontrara dentro de dicho intervalo.

Es importante mencionar la Selección de la tolerancia temporal inicialmente se evaluó una ventana temporal de ±1 hora para asociar cada incendio con la observación meteorológica más cercana de la estación previamente asignada, con este criterio se obtuvo una coincidencia meteorológica aproximada del 68,95 % de los registros, luego se amplió la tolerancia máxima a ±3 horas como criterio final de integración, obteniéndose 1.804 coincidencias de 2.615 incendios, equivalentes al 68,99 %, la ampliación de la ventana no produjo un incremento sustancial en la cobertura, lo que evidencia que la mayoría de las coincidencias disponibles ya se encontraban próximas temporalmente al evento, la ventana de ±3 horas se mantuvo como límite máximo para permitir la recuperación de observaciones meteorológicas cercanas cuando no existía una medición exactamente coincidente con la hora del incendio, evitando al mismo tiempo asociar registros excesivamente alejados temporalmente.

Después de aplicar este procedimiento, la base integrada mantuvo los 2.615 registros originales de incendios forestales, evitando la eliminación de eventos que no contaban con una observación meteorológica dentro de la ventana temporal establecida.

De los 2.615 incendios analizados, 1.804 (68,99 %) pudieron asociarse con un registro meteorológico dentro de la ventana temporal establecida. En los 811 casos restantes (31,01 %) no se encontró una observación meteorológica que cumpliera con este criterio.

Resultado de integración temporal	Registros	Porcentaje
Incendios totales	2.615	100,00%
Con coincidencia meteorológica	1.804	68,99%
Sin coincidencia meteorológica	811	31,01%

Tabla 10: Resultado de la integración temporal entre incendios y meteorología
Nota. Se consideró una tolerancia temporal máxima de ±3 horas. Elaboración propia.

Cobertura temporal por año

La disponibilidad de coincidencias meteorológicas presentó diferencias entre los años analizados.

Año	Incendios	Coincidencias meteorológicas	Cobertura
2021	992	760	76,61%
2022	375	277	73,87%
2023	868	521	60,02%
2024	334	209	62,57%
2025	40	34	85,00%
2026	6	2	33,33%

Tabla 11: Cobertura de la integración meteorológica por año

Los resultados muestran que la disponibilidad de registros meteorológicos cercanos a los incendios varió durante el período analizado. La mayor cobertura se obtuvo en 2025, con un 85,00 %. Para los años utilizados en el desarrollo y evaluación principal de los modelos, la cobertura fue de 76,61 % en 2021, 73,87 % en 2022, 60,02 % en 2023 y 62,57 % en 2024.

Disponibilidad de las variables meteorológicas después de la integración

La existencia de una coincidencia temporal no implicó necesariamente que todas las variables meteorológicas estuvieran disponibles, debido a las diferencias existentes entre las estaciones y sensores de las fuentes originales.

La precipitación presentó la mayor disponibilidad, con información en 1.403 registros 53,65 %, la temperatura estuvo disponible en 576 registros 22,03 %**,** la humedad relativa en 561 con un porcentaje de 21,45 %**,** la radiación solar y velocidad del viento en 548 registros cada una 20,96 %, mientras que la dirección del viento estuvo disponible en 139 observaciones 5,32 %**.**

Variable meteorológica	Registros con datos	Disponibilidad
Precipitación	1.403	53,65%
Temperatura	576	22,03%
Humedad relativa	561	21,45%
Radiación solar	548	20,96%
Velocidad del viento	548	20,96%
Dirección del viento	139	5,32%

Tabla 12: Disponibilidad de variables meteorológicas después de la integración
Nota. Los porcentajes se calcularon respecto de los 2.615 registros de incendios forestales.

Base resultante de la integración

Como resultado final se obtuvo la base denominada base_incendios_meteorologia, conformada por 2.615 registros y 34 variables.

Esta base incorporó en una misma estructura la información del evento, sus características temporales y territoriales, la estación meteorológica asignada, la distancia entre el incendio y dicha estación, la fecha y hora de la observación meteorológica seleccionada, las coordenadas de la estación y las variables meteorológicas disponibles.

Entre las variables incorporadas se encontraron:

- temperatura
- humedad relativa
- precipitación
- radiación solar
- velocidad del viento
- dirección del viento
- latitud y longitud de la estación
- estación asignada
- distancia a la estación
- diferencia temporal entre el incendio y la observación meteorológica

La base resultante fue almacenada en formatos Parquet y CSV para las etapas posteriores de análisis y preparación del conjunto destinado al modelado predictivo.

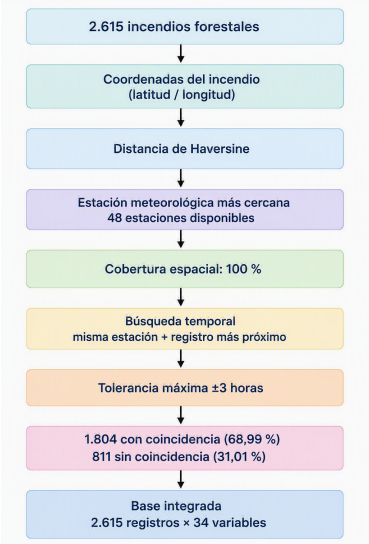


Figura 13:Proceso de integración espacial y temporal de los registros de incendios forestales con la información meteorológica.

2.3.1.4. Preparación de los datos para el modelado predictivo

Una vez integrada la información histórica de incendios forestales con los registros meteorológicos, se procedió a construir el conjunto de datos destinado al entrenamiento y evaluación de los modelos predictivos.

Inicialmente se utilizaron la base de incendios con información meteorológica, compuesta por 2.615 registros y 34 variables, y la base meteorológica consolidada, conformada por 15.355.869 registros y 14 variables.

Construcción de la información meteorológica horaria

Debido a que la información meteorológica original se encontraba disponible con diferentes frecuencias de registro, los datos fueron agregados a una escala horaria, utilizando como claves la fecha-hora y la estación meteorológica normalizada.

Para cada combinación de hora y estación se conservaron las coordenadas geográficas de la estación y se calcularon valores agregados de las variables meteorológicas, la temperatura, humedad relativa, precipitación, radiación solar y velocidad del viento fueron resumidas mediante la media, mientras que otras variables complementarias fueron procesadas de acuerdo con sus características.

Como resultado de este procedimiento se obtuvo una base meteorológica horaria de 1.676.114 observaciones y 11 variables**.**

Este proceso permitió disponer de una única observación representativa por estación y hora, reduciendo la diferencia de frecuencia existente entre las distintas fuentes meteorológicas.

Identificación temporal de los incendios forestales

Luego los registros de incendios forestales fueron transformados a la misma resolución temporal, la fecha y hora meteorológica asociada a cada incendio fue redondeada al inicio de la hora correspondiente y el nombre de la estación fue normalizado.

Después de eliminar combinaciones duplicadas de fecha-hora y estación, se identificaron 1.763 combinaciones únicas de hora y estación asociadas con incendios forestales**.**

A cada una de estas combinaciones se le asignó inicialmente el valor:

- Incendio=1

Luego esta información fue incorporada a la totalidad de la base meteorológica horaria mediante las variables fecha_hora_hora y estacion_normalizada.

Las observaciones meteorológicas que no presentaron correspondencia con una hora-estación asociada a un incendio fueron clasificadas como:

- Incendio=0

De esta manera se construyó una variable objetivo binaria destinada al problema de clasificación.

Construcción de la variable objetivo

Como resultado del etiquetado de las 1.676.114 observaciones meteorológicas horarias, se obtuvieron 1.674.353 registros correspondientes a la clase No incendio y 1.761 registros correspondientes a la clase Incendio**.**

Clase	Número de registros	Porcentaje
No incendio	1.674.353	99,89%
Incendio	1.761	0,11%
Total	1.676.114	100,00%

Tabla 13:Distribución de la variable objetivo en la base de modelado

La prevalencia observada de la clase positiva fue, por tanto, de apenas 0,1051 %**,** evidenciando un problema de clasificación con un desbalance extremo entre clases**.**

Selección de las variables predictoras

Para el desarrollo de los modelos se seleccionaron siete variables predictoras, correspondientes a cinco variables meteorológicas y dos variables espaciales asociadas con la ubicación de las estaciones.

Tipo	Variable
Meteorológica	Temperatura
Meteorológica	Humedad relativa
Meteorológica	Precipitación
Meteorológica	Radiación solar
Meteorológica	Velocidad del viento
Espacial	Latitud de la estación
Espacial	Longitud de la estación

Tabla 14:Variables predictoras utilizadas para el modelado

La base definitiva quedó estructurada con diez variables: fecha y hora, estación meteorológica normalizada, las siete variables predictoras y la variable objetivo incendio, por lo tanto la base final destinada al modelado presentó dimensiones de 1.676.114 registros y 10 variables la cual fue almacenada sin modificar la distribución original de la variable objetivo.

Tratamiento del desbalance

El gran desbalance identificado no fue corregido mediante submuestreo o sobremuestreo en esta etapa. A diferencia de la estrategia preliminar utilizada durante versiones iniciales del estudio, la base definitiva se conservó con su distribución real.

Esta decisión permitió evitar la construcción de un conjunto artificialmente balanceado antes de realizar la separación destinada al entrenamiento y evaluación.

El tratamiento del desbalance se realizó posteriormente exclusivamente durante el entrenamiento de los algoritmos, mediante mecanismos de ponderación de clases, mientras que los conjuntos utilizados para evaluación conservaron su distribución original.
Esta estrategia permitió evaluar los modelos bajo condiciones más próximas a la frecuencia real de ocurrencia de incendios forestales.

Manejo de valores faltantes

La preparación inicial evidenció presencia de valores faltantes en varias variables meteorológicas dentro de la base enriquecida de incendios, por ejemplo, en los 2.615 registros originales se identificaron 2.039 valores faltantes en temperatura, 2.054 en humedad relativa, 1.211 en precipitación y 2.067 tanto en radiación solar como en velocidad del viento.

Sin embargo, la imputación definitiva no se realizó sobre toda la base antes de dividir los períodos de entrenamiento y prueba, esta decisión evitó que información correspondiente a períodos posteriores influyera en los valores utilizados para completar registros históricos.

La imputación mediante la mediana se efectuó posteriormente dentro del proceso de modelado, ajustando el imputador exclusivamente con el conjunto de entrenamiento y aplicándolo después a los conjuntos de validación y prueba.

Este procedimiento permitió reducir el riesgo de fuga de información (*data leakage*).

Base final obtenida

La base final fue almacenada como base_modelo_clasificacion_sin_balancear.parquet con 1.676.114 registros, siete variables predictoras y una variable objetivo binaria.

En conclusión, el proceso de preparación permitió transformar los registros meteorológicos e históricos de incendios en una estructura adecuada para un problema de clasificación supervisada, conservando tres características fundamentales: la resolución temporal horaria, la localización espacial de las estaciones meteorológicas y la prevalencia real de la ocurrencia de incendios forestales.

2.3.1.5. Resultados de los modelos predictivos

Se desarrollaron y evaluaron tres algoritmos de clasificación supervisada: Regresión Logística, Random Forest y XGBoost, los tres modelos utilizaron las mismas siete variables predictoras, con el propósito de realizar una comparación bajo condiciones equivalentes.

La evaluación se realizó mediante una separación temporal, evitando una partición aleatoria de los registros, los datos correspondientes al período 2021–2023 fueron utilizados para el entrenamiento, mientras que el año 2024 se reservó como conjunto de prueba principal.

Los registros correspondientes a 2025–2026 fueron separados para una evaluación adicional posterior debido al cambio identificado en el criterio institucional de registro de incendios forestales a partir de 2025, por lo que no intervinieron en la evaluación principal de los modelos.

Distribución temporal de los conjuntos

El conjunto de entrenamiento quedó conformado por 964.455 registros, de los cuales 1.520 correspondieron a incendios forestales y 962.935 a condiciones sin incendio, la prevalencia de incendios en este período fue de 0,1576 %.

El conjunto de prueba correspondiente a 2024 estuvo compuesto por 314.318 observaciones, con 207 incendios forestales y 314.111 registros sin incendio, equivalente a una prevalencia de 0,0659 %.

Conjunto	Período	Registros	Incendios	No incendios	Prevalencia
Entrenamiento	2021–2023	964.455	1.520	962.935	0,16%
Prueba principal	2024	314.318	207	314.111	0,07%
Evaluación post-cambio	2025–2026	397.341	34	397.307	0,01%

Tabla 15:Distribución temporal de los conjuntos utilizados para el modelado
Fuente: Creación de autor, D. Tiupul (2026)

Preprocesamiento y tratamiento del desbalance

Los valores faltantes de las siete variables predictoras fueron tratados mediante imputación por la mediana, el imputador fue ajustado exclusivamente con los datos correspondientes al período de entrenamiento 2021–2023 y posteriormente aplicado, sin reajuste, a los conjuntos de 2024 y 2025–2026.

Después del procedimiento de imputación no permanecieron valores faltantes en ninguno de los conjuntos utilizados para el modelado.

El desbalance se trató exclusivamente durante el entrenamiento, conservando intacta la distribución real del conjunto de prueba de 2024. En el período 2021–2023 se identificaron 962.935 observaciones de no incendio frente a 1.520 observaciones de incendio, correspondiente a una relación aproximada de 633,51 registros negativos por cada registro positivo.

Para Regresión Logística y Random Forest se utilizó ponderación balanceada de clases mediante class_weight="balanced", mientras que XGBoost empleó scale_pos_weight, calculado a partir de la relación entre observaciones negativas y positivas del conjunto de entrenamiento, de esta manera no se aplicó submuestreo ni sobremuestreo al conjunto de prueba de 2024.

Resultados de la Regresión Logística

La Regresión Logística obtuvo una exactitud global (*Accuracy*) de 0,7056 y una exactitud balanceada (*Balanced Accuracy*) de 0,6838. En relación con la detección de incendios forestales, alcanzó un Recall de 0,6618, identificando correctamente 137 de los 207 incendios registrados durante 2024, equivalente al 66,18 %, el modelo no identificó 70 eventos, correspondientes al 33,82 %. El ROC-AUC obtenido fue de 0,7322, mientras que el PR-AUC alcanzó 0,004832.

La matriz de confusión presentó los siguientes resultados:

Resultado	Registros
Verdaderos negativos	221.661
Falsos positivos	92.450
Falsos negativos	70
Verdaderos positivos	137

Tabla 16:Matriz de confusión de la Regresión Logística sobre 2024

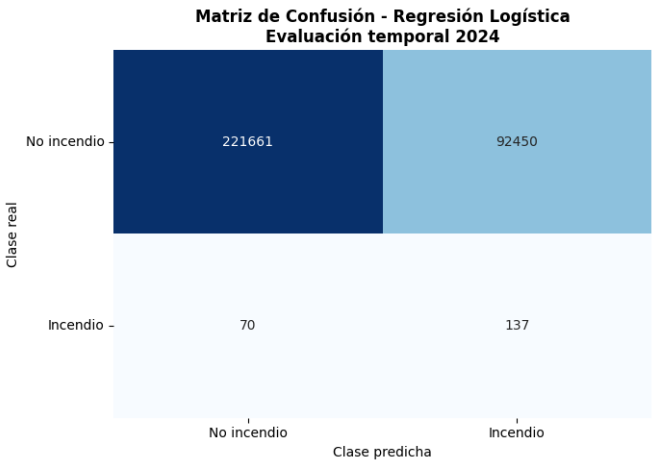


Figura 14:Matriz de confusión de la Regresión Logística en la evaluación temporal de 2024

Resultados de Random Forest

Random Forest presentó la mayor exactitud global de los tres modelos base, con un Accuracy de 0,8214, sin embargo, su *Balanced Accuracy* fue de 0,6354.

El modelo obtuvo un Recall de 0,4493 detectando correctamente 93 de los 207 incendios reales, equivalente al 44,93 %, por lo tanto 114 incendios no fueron identificados, correspondientes al 55,07 % de los eventos positivos del conjunto de prueba.

El ROC-AUC alcanzó 0,6825 y el PR-AUC fue de 0,001342.

Resultado	Registros
Verdaderos negativos	258.081
Falsos positivos	56.030
Falsos negativos	114
Verdaderos positivos	93

Tabla 17:Matriz de confusión de Random Forest sobre 2024

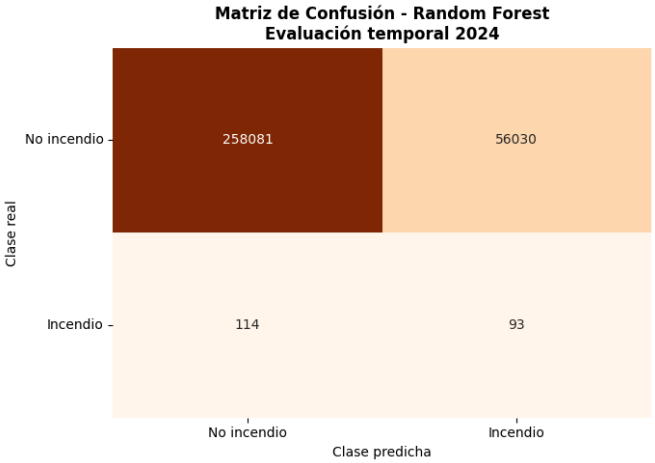


Figura 15:Matriz de confusión de Random Forest en la evaluación temporal de 2024

Resultados de XGBoost

El modelo XGBoost obtuvo un Accuracy de 0,7798 y un Balanced Accuracy de 0,7281, siendo este último el valor más alto entre los tres algoritmos evaluados, de la misma manera obtuvo un Recall de 0,6763 identificando correctamente 140 de los 207 incendios forestales registrados en 2024, equivalente al 67,63 % de los eventos positivos, los 67 incendios restantes no fueron identificados, equivalentes al 32,37, adicional XGBoost obtuvo un ROC-AUC de 0,8009, el valor más alto de los tres modelos base, y un PR-AUC de 0,002998**.**

Resultado	Registros
Verdaderos negativos	244.954
Falsos positivos	69.157
Falsos negativos	67
Verdaderos positivos	140

Tabla 18:Matriz de confusión de XGBoost sobre 2024

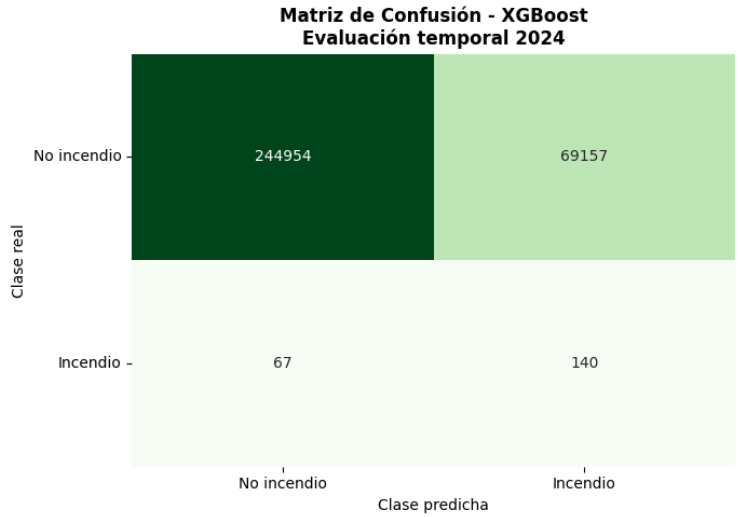


Figura 16:Matriz de confusión de XGBoost en la evaluación temporal de 2024

Comparación de los modelos predictivos

Los resultados obtenidos sobre el conjunto temporal de prueba correspondiente a 2024 se resumen en la siguiente tabla.

Métrica	Regresión Logística	Random Forest	XGBoost
Accuracy	0,7056	**0,8214**	0,7798
Balanced Accuracy	0,6838	0,6354	**0,7281**
Precision	0,00148	0,001657	**0,00202**
Recall	0,6618	0,4493	**0,6763**
F1-Score	0,002953	0,003302	**0,004029**
ROC-AUC	0,7322	0,6825	**0,8009**
PR-AUC	**0,004832**	0,001342	0,002998

Tabla 19:Comparación del desempeño de los modelos base sobre el conjunto temporal de prueba 2024

La comparación evidencia que ningún algoritmo presentó el mejor resultado en todas las métricas, Random Forest alcanzó el mayor Accuracy 0,8214, Regresión Logística obtuvo el mayor PR-AUC 0,004832, mientras que XGBoost presentó los valores más altos de Balanced Accuracy 0,7281, Precision 0,002020, Recall 0,6763, F1-Score 0,004029 y ROC-AUC 0,8009.

En términos de detección de los 207 incendios reales presentes en 2024, XGBoost identificó 140 eventos (67,63 %)**,** Regresión Logística 137 (66,18 %) y Random Forest 93 (44,93 %)**.**

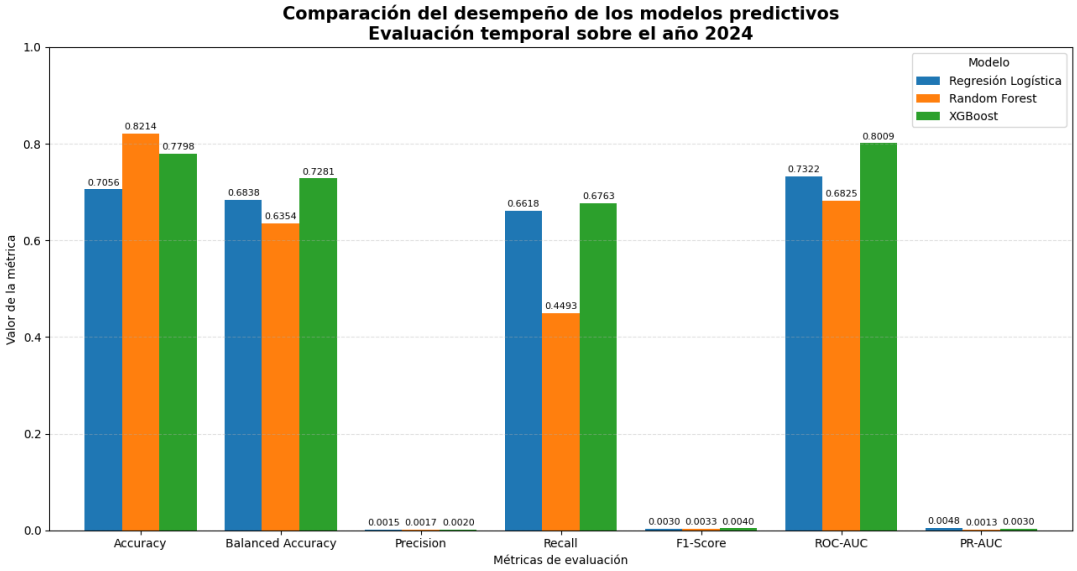


Figura 17:Comparación del desempeño de Regresión Logística, Random Forest y XGBoost en la evaluación temporal de 2024

Precisión y desbalance extremo

Los tres algoritmos presentaron valores reducidos de Precision y F1-Score especialmente para la clase incendio, este resultado debe presentarse junto con la prevalencia real del conjunto de prueba: únicamente 207 de las 314.318 observaciones de 2024 correspondieron a incendios equivalente al 0,0659 %**.**

Este punto será analizado con mayor profundidad en la sección 2.3.2 Análisis**,** debido a que el elevado número de falsos positivos constituye una de las principales limitaciones observadas en los modelos.

2.3.1.6. Optimización y selección del modelo predictivo

A partir de los resultados obtenidos con los modelos base, se realizó una etapa de optimización de Random Forest y XGBoost con el objetivo de determinar configuraciones de hiperparámetros que mejoraran la capacidad de identificación de incendios forestales.

La optimización mantuvo una estructura estrictamente temporal, los registros de 2021–2022 se utilizaron para el entrenamiento interno, 2023 se reservó para la selección de hiperparámetros y análisis del umbral de clasificación, y 2024 permaneció completamente separado como conjunto de prueba temporal final**.**

Los registros correspondientes a 2025–2026 no participaron en la optimización ni en la selección del modelo y fueron utilizados posteriormente como una evaluación adicional debido al cambio en el criterio institucional de registro de incendios forestales.

Como se explicó en el análisis exploratorio de los datos, este cambio correspondió a una modificación del criterio técnico de clasificación de los eventos y no únicamente a una variación en su frecuencia de ocurrencia, motivo por lo cual los registros de 2025–2026 fueron tratados como un período post-cambio y se mantuvieron separados del proceso principal de entrenamiento, optimización y selección del modelo.

Estrategia temporal para la optimización

La distribución utilizada fue la siguiente:

Conjunto	Período	Registros	Incendios	No incendios	Prevalencia
Entrenamiento interno	2021–2022	640.902	1.011	639.891	0,16%
Validación temporal	2023	323.553	509	323.044	0,16%
Prueba temporal final	2024	314.318	207	314.111	0,07%
Evaluación post-cambio	2025–2026	397.341	34	397.307	0,01%

Tabla 20:Distribución temporal utilizada durante la optimización y evaluación final

Esta estructura permitió seleccionar los hiperparámetros utilizando únicamente información disponible antes del período de prueba final, el año 2024 por tanto, no intervino en la selección de configuraciones ni en la elección del umbral de clasificación**.**

Optimización de Random Forest

Para Random Forest se evaluaron 24 combinaciones de hiperparámetros**,** considerando variaciones en número de árboles profundidad máxima número mínimo de muestras para dividir un nodo, número mínimo de muestras por hoja y cantidad de características consideradas en cada división.

La selección se realizó sobre el conjunto de validación temporal correspondiente a 2023, priorizando PR-AUC, seguida de Recall y Balanced Accuracy.

La configuración seleccionada fue:

Hiperparámetro	Valor seleccionado
n_estimators	200
max_depth	8
min_samples_split	2
min_samples_leaf	5
max_features	sqrt

Tabla 21:Hiperparámetros seleccionados para Random Forest

Con esta configuración, durante la validación temporal de 2023, Random Forest alcanzó un PR-AUC de 0,017990, ROC-AUC de 0,822540, Recall de 0,728880 y Balanced Accuracy de 0,726183.

Optimización de XGBoost

Para XGBoost se evaluaron 18 combinaciones, modificando el número de estimadores, profundidad máxima, tasa de aprendizaje, proporción de observaciones utilizada por árbol y proporción de variables utilizada durante la construcción de cada árbol.

La configuración con mejor resultado según el criterio establecido fue:

Hiperparámetro	Valor seleccionado
n_estimators	300
max_depth	4
learning_rate	0,03
subsample	0,8
colsample_bytree	0,8

Tabla 22:Hiperparámetros seleccionados para XGBoost

En la validación temporal de 2023, esta configuración obtuvo un PR-AUC de 0,020102, ROC-AUC de 0,845930, Recall de 0,772102 y Balanced Accuracy de 0,761166.

Estos resultados fueron superiores a los obtenidos por la configuración seleccionada de Random Forest en PR-AUC, ROC-AUC, Recall y Balanced Accuracy durante la validación temporal.

Evaluación del umbral de clasificación

Adicionalmente, se evaluó el comportamiento del XGBoost optimizado utilizando umbrales comprendidos entre 0,10 y 0,90, con incrementos de 0,05. Este análisis se realizó exclusivamente sobre el conjunto de validación correspondiente a 2023.

El cambio del umbral evidenció el compromiso existente entre la capacidad para detectar incendios y el número de falsos positivos. Los umbrales bajos permitieron incrementar considerablemente el Recall, pero generaron una mayor cantidad de falsas alarmas. En cambio, los umbrales elevados incrementaron la precisión, pero redujeron sustancialmente la proporción de incendios detectados.

Umbral	Precision	Recall	Balanced Accuracy	Incendios detectados
0,3	0,00329	0,915521	0,739267	91,55%
0,4	0,003913	0,833006	0,749451	83,30%
0,45	0,00452	0,801572	**0,76171**	80,16%
0,5	**0,004847**	**0,772102**	**0,761166**	**77,21%**
0,6	0,005724	0,687623	0,749719	68,76%
0,7	0,007159	0,59332	0,731833	59,33%
0,8	0,021009	0,239686	0,611044	23,97%

Tabla 23:Comportamiento del XGBoost optimizado según el umbral de clasificación en 2023

Finalmente, se conservó un umbral de 0,50, debido a que proporcionó un equilibrio entre sensibilidad y discriminación de ambas clases, con un Recall de 0,7721 y Balanced Accuracy de 0,7612, aunque el umbral de 0,45 presentó una Balanced Accuracy ligeramente superior (0,7617), la diferencia fue mínima y el valor de 0,50 mantuvo un comportamiento equilibrado sin requerir un desplazamiento adicional del punto de decisión.

Esta selección se realizó antes de observar los resultados de 2024, evitando ajustar el umbral en función del conjunto de prueba final.

Reentrenamiento de los modelos

Una vez seleccionados los hiperparámetros y el umbral mediante la información disponible hasta 2023, Random Forest y XGBoost fueron reentrenados utilizando conjuntamente los registros correspondientes a 2021–2023.

Este procedimiento permitió aprovechar toda la información histórica disponible antes de la evaluación final, manteniendo 2024 completamente fuera del proceso de ajuste.

Evaluación temporal final sobre 2024

La evaluación final permitió comparar los modelos base y optimizados sobre los 314.318 registros de 2024, manteniendo la distribución real de las clases.

Modelo	Accuracy	Balanced Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Random Forest base	**0,821152**	0,635336	0,001655	0,449275	0,003298	0,682488	0,001342
Random Forest optimizado	0,718037	0,735818	0,001758	**0,753623**	0,003508	0,810807	0,002579
XGBoost base	0,779764	0,728081	**0,002020**	0,676329	**0,004029**	0,800906	0,002998
XGBoost optimizado	0,754802	**0,744558**	0,00197	0,7343	0,003929	**0,817874**	**0,004903**

Tabla 24:Comparación de los modelos base y optimizados en la evaluación temporal final de 2024

La optimización produjo mejoras particularmente importantes en los dos algoritmos. Random Forest incrementó su Recall de 0,4493 a 0,7536 y su ROC-AUC de 0,6825 a 0,8108, por otro lado XGBoost incrementó el Recall de 0,67630 a 0,7343, el ROC-AUC de 0,8009 a 0,8179 y el PR-AUC de 0,002998 a 0,004903.

Matrices de confusión

La comparación de las matrices de confusión permitió observar directamente la capacidad de detección de los 207 incendios reales registrados durante 2024.

Modelo	VN	FP	FN	VP	Detectados	No detectados
Random Forest base	258.010	56.101	114	93	44,93%	55,07%
Random Forest optimizado	225.536	88.575	**51**	**156**	**75,36%**	**24,64%**
XGBoost base	244.954	69.157	62	140	67,63%	32,37%
XGBoost optimizado	237.096	77.015	55	152	73,43%	26,57%

Tabla 25:Resultados de las matrices de confusión de los modelos evaluados en 2024

Random Forest optimizado presentó el mayor Recall identificando 156 de los 207 incendios (75,36 %). XGBoost optimizado identificó 152 incendios (73,43 %), dejando sin detectar 55 eventos, sin embargo, la selección final no se fundamenta únicamente en Recall.

Selección del modelo final

El modelo seleccionado fue XGBoost optimizado.

Esta selección se realizó considerando conjuntamente el comportamiento de las métricas sobre el conjunto temporal de prueba de 2024, aunque Random Forest optimizado presentó un Recall ligeramente superior (75,36 % frente a 73,43 %), XGBoost optimizado obtuvo mejores resultados en Balanced Accuracy (0,7446 frente a 0,7358), ROC-AUC (0,8179 frente a 0,8108) y PR-AUC (0,004903 frente a 0,002579).

En particular, PR-AUC fue considerada una métrica relevante debido a la extremadamente baja prevalencia de la clase incendio. Por tanto, XGBoost optimizado fue seleccionado como modelo final al presentar un comportamiento global más equilibrado, en lugar de seleccionar el algoritmo únicamente por Accuracy o Recall.

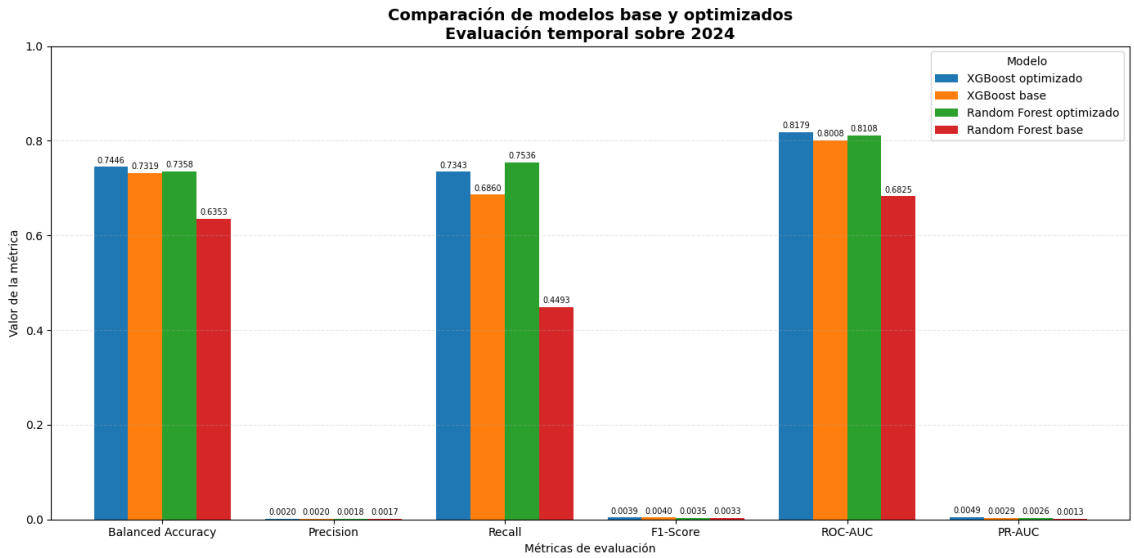


Figura 18:Comparación de los modelos base y optimizados durante la evaluación temporal de 2024

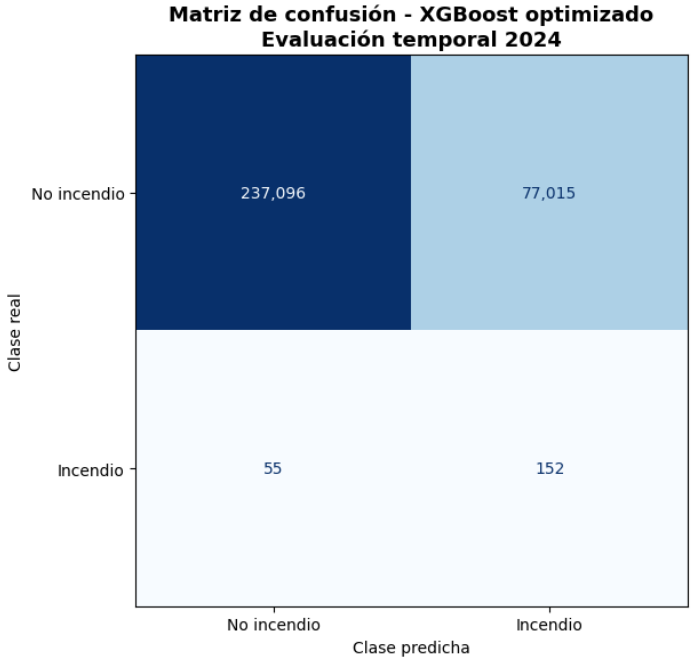


Figura 19:Matriz de confusión del modelo XGBoost optimizado sobre el conjunto temporal de prueba 2024

Análisis de sensibilidad del umbral de decisión

Debido al elevado número de falsos positivos observado en la evaluación temporal de 2024, se realizó un análisis adicional de sensibilidad para determinar el efecto de utilizar umbrales de clasificación más restrictivos en el XGBoost optimizado. Este análisis tuvo como objetivo examinar el compromiso entre la reducción de falsas predicciones positivas y la capacidad del modelo para identificar incendios forestales.

Es importante indicar que este procedimiento no se utilizó para seleccionar nuevamente el modelo ni modificar el umbral establecido durante la fase de validación, el umbral del modelo final fue determinado previamente utilizando el conjunto de validación de 2023, mientras que los diferentes umbrales aplicados sobre 2024 se analizaron únicamente como escenarios de sensibilidad para evaluar sus posibles implicaciones operativas.

Umbral	FP	FN	TP	Precision	Recall	Balanced Accuracy
0,5	77.015	55	152	0,00197	0,7343	0,744558
0,6	49.679	90	117	0,00235	0,565217	0,70353
0,7	43.509	95	112	0,002568	0,541063	0,701274
0,8	7.391	167	40	0,005383	0,193237	0,584853
0,9	3.722	176	31	0,00826	0,149758	0,568955
0,95	1.442	198	9	0,006203	0,043478	0,519444

Tabla 26:Elaboración propia a partir de las probabilidades generadas por el XGBoost optimizado sobre el conjunto temporal de prueba de 2024

Como se visualiza en la Tabla 26, el incremento del umbral produjo una reducción progresiva de los falsos positivos, acompañada de una disminución en la cantidad de incendios correctamente identificados. Con el umbral de 0,50 el modelo generó 77.015

falsos positivos e identificó correctamente 152 de los 207 incendios, alcanzando un Recall de 0,734300. Al incrementar el umbral a 0,70, los falsos positivos disminuyeron a 43.509 y los verdaderos positivos a 112, con un Recall de 0,541063.

La reducción fue más pronunciada con umbrales superiores. Con un umbral de 0,80, los falsos positivos disminuyeron a 7.391; sin embargo, el modelo identificó únicamente 40 incendios y el Recall descendió a 0,193237. Con un umbral de 0,90 se obtuvieron 3.722 falsos positivos y 31 verdaderos positivos, mientras que con 0,95 los falsos positivos se redujeron a 1.442, pero solamente se identificaron 9 de los 207 incendios registrados.

Los resultados muestran, por tanto, que el aumento del umbral reduce el número de predicciones positivas incorrectas, pero simultáneamente incrementa los falsos negativos y disminuye la capacidad de detección de incendios, este comportamiento evidencia un compromiso entre la reducción de falsas alarmas y la sensibilidad del modelo.

Evaluación adicional sobre el período 2025–2026

Finalmente, el modelo seleccionado fue aplicado al período 2025–2026 como una evaluación adicional independiente, este período no participó en el entrenamiento, optimización, selección de hiperparámetros, selección del umbral ni elección del modelo final.

Utilizando el mismo umbral de 0,50, el XGBoost optimizado obtuvo un Recall de 0,735294, ROC-AUC de 0,810148 y Balanced Accuracy de 0,731882, identificando correctamente 25 de los 34 incendios registrados. Sin embargo, la precisión descendió a 0,000232 y se produjeron 107.881 falsos positivos.

Estos resultados deben interpretarse con cautela debido a la prevalencia extremadamente baja de incendios durante 2025–2026 (0,0086 %) y al cambio en el criterio institucional de registro ocurrido a partir de 2025, por esta razón, este período se consideró una evaluación adicional post-cambio y no la prueba principal utilizada para seleccionar el modelo.

2.3.1.7. Interpretabilidad del modelo mediante SHAP

Una vez seleccionado XGBoost optimizado como modelo final, se aplicó la técnica SHAP (SHapley Additive exPlanations) para analizar la contribución de las variables predictoras en las estimaciones realizadas por el modelo.

El análisis se efectuó utilizando el modelo entrenado con los registros correspondientes al período 2021–2023, conformado por 964.455 observaciones y 1.520 incendios forestales, mientras que para la interpretación se utilizó una muestra reproducible del conjunto temporal de prueba correspondiente a 2024.

La muestra utilizada para el análisis SHAP estuvo compuesta por 5.207 registros, incluyendo los 207 incendios forestales disponibles en 2024 y una muestra aleatoria reproducible de 5.000 registros correspondientes a la clase No incendio, esta estrategia permitió conservar todos los eventos positivos disponibles durante el período de prueba y, al mismo tiempo, reducir el costo computacional del análisis de interpretabilidad.

Importancia global de las variables

La importancia global se determinó mediante el valor absoluto medio de las contribuciones SHAP de cada predictor, los resultados muestran diferencias claras en la participación de las siete variables utilizadas por el modelo.

Posición	Variable	SHAP medio absoluto
1	Longitud de la estación	1,186332
2	Latitud de la estación	0,646508
3	Humedad relativa	0,284675
4	Temperatura	0,224476
5	Precipitación	0,169948
6	Radiación solar	0,127871
7	Velocidad del viento	0,125958

Tabla 27:Importancia global de las variables predictoras según SHAP

La longitud de la estación presentó la mayor importancia global, con un valor SHAP medio absoluto de 1,186332, seguida de la latitud de la estación**,** con 0,646508.

Entre las variables meteorológicas, la de mayor influencia fue la humedad relativa, con 0,284675, seguida de la temperatura (0,224476) y la precipitación (0,169948), la radiación solar y la velocidad del viento presentaron valores inferiores de importancia global, con 0,127871 y 0,125958, respectivamente.

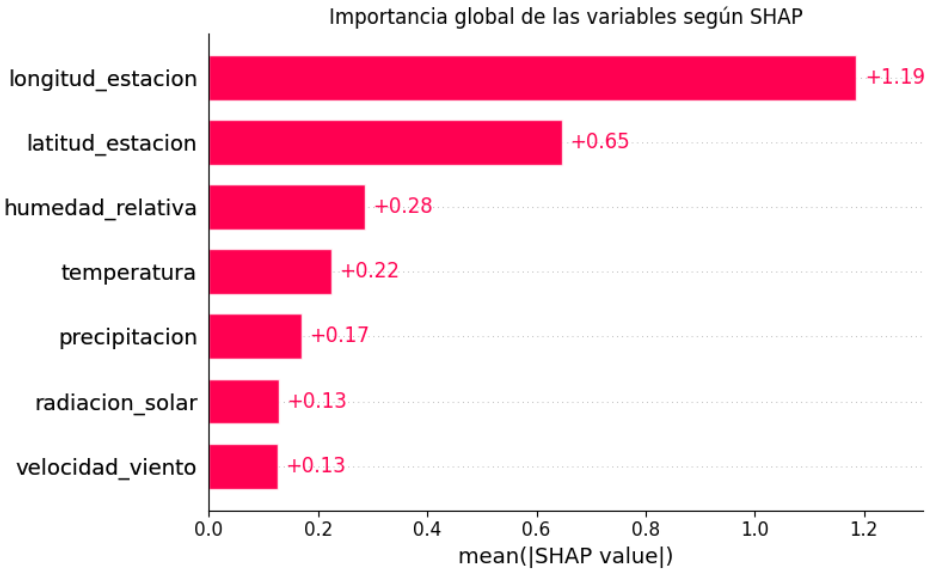


Figura 20:Importancia global de las variables predictoras según valores SHAP

SHAP Summary Plot

El gráfico resumen de SHAP permitió observar simultáneamente la importancia de las variables, la magnitud de sus contribuciones y la dirección del efecto sobre la salida del modelo.

Los valores SHAP positivos representan observaciones que desplazan la predicción del modelo hacia la clase Incendio, mientras que valores negativos desplazan la estimación hacia No incendio.

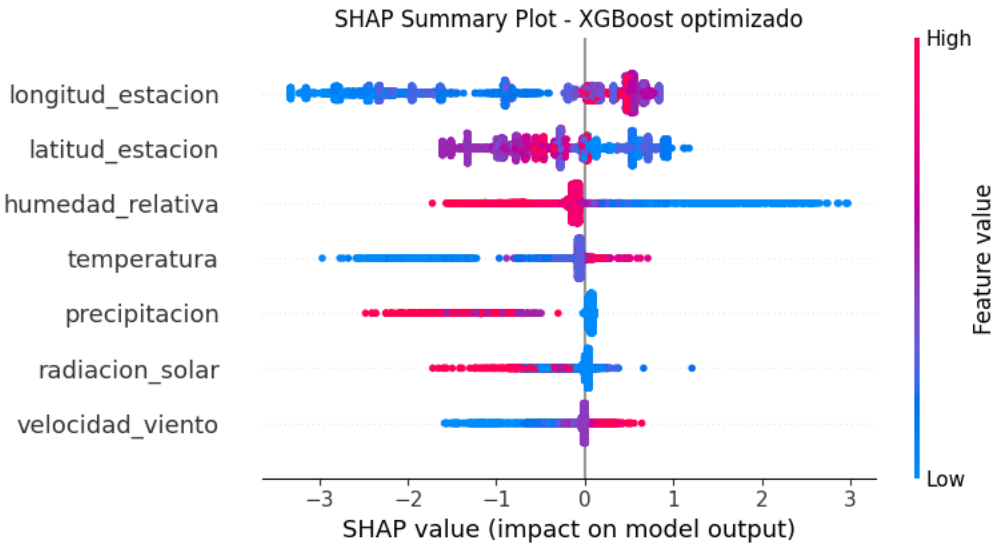


Figura 21:SHAP Summary Plot del modelo XGBoost optimizado

Componente espacial de las predicciones

Las variables de mayor influencia correspondieron a las coordenadas geográficas de las estaciones meteorológicas.

Los gráficos de dependencia mostraron que tanto longitud_estacion como latitud_estacion presentan contribuciones positivas y negativas dependiendo de la localización, las observaciones se concentran en valores discretos de coordenadas, debido a que múltiples registros meteorológicos comparten la ubicación de una misma estación.

Este resultado evidencia que el modelo utiliza de manera importante la localización de la estación meteorológica para diferenciar patrones predictivos entre distintos sectores geográficos.

Humedad relativa

La humedad relativa fue la variable meteorológica con mayor importancia según SHAP. El gráfico de dependencia mostró una relación predominantemente descendente, los valores bajos de humedad relativa presentaron contribuciones SHAP positivas, mientras que al incrementarse la humedad las contribuciones tendieron a disminuir y posteriormente a ubicarse en valores negativos.

En términos del comportamiento aprendido por XGBoost, los registros caracterizados por menor humedad relativa tienden a desplazar la predicción hacia la clase incendio, mientras que valores elevados contribuyen principalmente hacia la clase No incendio.

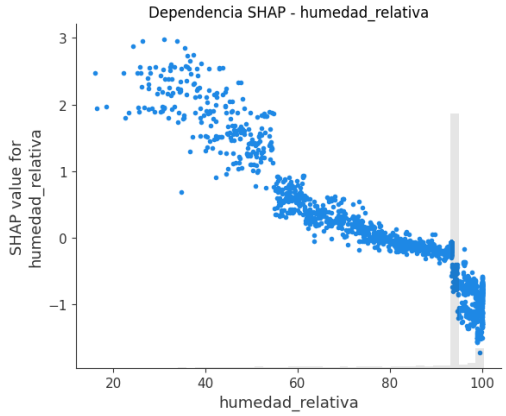


Figura 22:Dependencia SHAP de la humedad relativa

Temperatura

La temperatura ocupó la cuarta posición en importancia global, el gráfico de dependencia evidenció una relación no lineal, los valores de temperatura más bajos presentaron

contribuciones predominantemente negativas, mientras que conforme aumentó la temperatura los valores SHAP se aproximaron a cero y en determinados intervalos, presentaron contribuciones positivas.

Por tanto, la temperatura no presentó una relación estrictamente lineal con la salida del modelo, aunque se observó que las temperaturas bajas reducen considerablemente la estimación hacia incendio, mientras que temperaturas superiores dejan de ejercer ese efecto negativo.

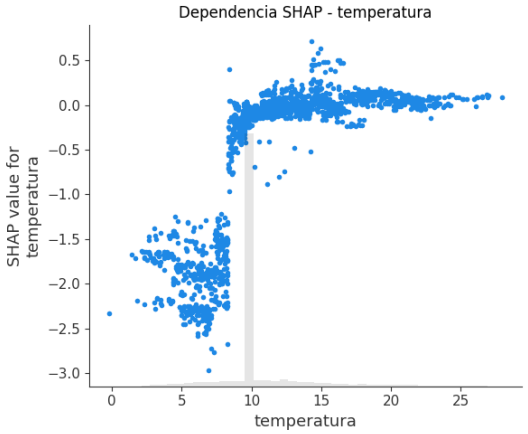


Figura 23:Dependencia SHAP de la temperatura.

Precipitación

La precipitación presentó una importancia SHAP media absoluta de 0,169948, la distribución de las observaciones mostró una fuerte concentración en valores próximos a cero, conforme la precipitación aumentó, las contribuciones SHAP fueron predominantemente negativas.

Este resultado indica que, dentro de la estructura aprendida por el modelo, la presencia de mayores niveles de precipitación tiende a desplazar la estimación hacia la clase No incendio.

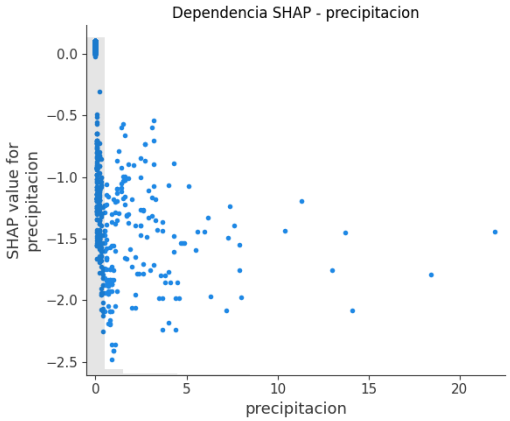


Figura 24:Dependencia SHAP de la precipitación

Radiación Solar

La radiación solar presentó una importancia SHAP media absoluta de 0,127871, el gráfico de dependencia mostró un comportamiento predominantemente no lineal, en varios rangos, el incremento de la radiación estuvo acompañado por una reducción de las contribuciones SHAP, llegando a observarse valores negativos para determinadas intensidades.

Este comportamiento debe interpretarse con cautela, debido a que SHAP refleja la contribución de la variable considerando simultáneamente las relaciones e interacciones aprendidas por el modelo con los demás predictores.

Por esta razón, los resultados no permiten afirmar que una mayor radiación solar reduzca causalmente la ocurrencia de incendios, sino únicamente que, dentro del modelo ajustado, determinados valores de radiación contribuyeron negativamente a la predicción.

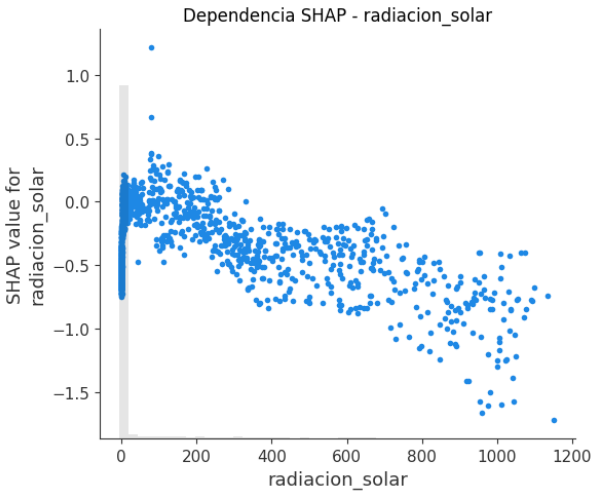


Figura 25:Dependencia SHAP de la radiación solar

Velocidad del viento

La velocidad del viento presentó la menor importancia global entre las siete variables, con un SHAP medio absoluto de 0,125958.

El gráfico de dependencia mostró diferencias en la dirección de las contribuciones, los valores reducidos de velocidad presentaron principalmente contribuciones negativas, mientras que valores mayores tendieron a desplazar progresivamente la salida hacia valores SHAP positivos.

El resultado evidencia que la velocidad del viento, aunque presenta una influencia global menor que otras variables, aporta información al modelo para diferenciar las condiciones asociadas a la predicción de incendios

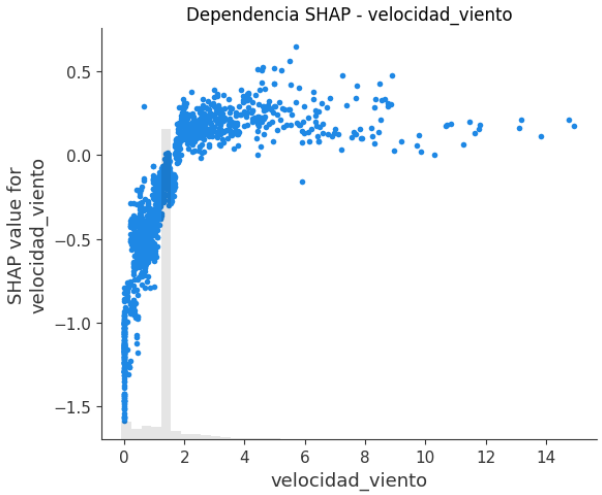


Figura 26:Dependencia SHAP de la velocidad del viento

Consistencia con la importancia interna de XGBoost

Como análisis complementario, la importancia SHAP fue comparada con la importancia interna calculada por XGBoost.

Variable	SHAP medio absoluto	Importancia XGBoost
Longitud estación	1,186332	0,395483
Latitud estación	0,646508	0,189726
Humedad relativa	0,284675	0,148525
Temperatura	0,224476	0,075385
Precipitación	0,169948	0,086019
Radiación solar	0,127871	0,056676
Velocidad viento	0,125958	0,048185

Tabla 28: Comparación de importancia SHAP e importancia interna de XGBoost

Ambos procedimientos ubicaron a longitud de la estación, latitud de la estación y humedad relativa como las tres variables de mayor relevancia para el modelo.

SHAP fue utilizado como mecanismo principal de interpretabilidad debido a que además de cuantificar la importancia global, permitió observar la dirección y distribución de las contribuciones individuales.

Resultado de la etapa de interpretabilidad

El análisis SHAP permitió establecer que el modelo XGBoost optimizado combina información espacial y meteorológica para generar sus predicciones.

Las variables con mayor influencia fueron las coordenadas geográficas de las estaciones meteorológicas, seguidas de la humedad relativa y la temperatura, entre las relaciones meteorológicas observadas, los niveles bajos de humedad presentaron contribuciones positivas hacia la predicción de incendio la precipitación elevada presentó contribuciones predominantemente negativas, y la velocidad del viento mostró una tendencia hacia contribuciones positivas en valores superiores.

Estos resultados permiten identificar las variables que mayor participación tuvieron en las decisiones del modelo y constituyen la base para su interpretación posterior en relación con los objetivos de investigación y los estudios previos.

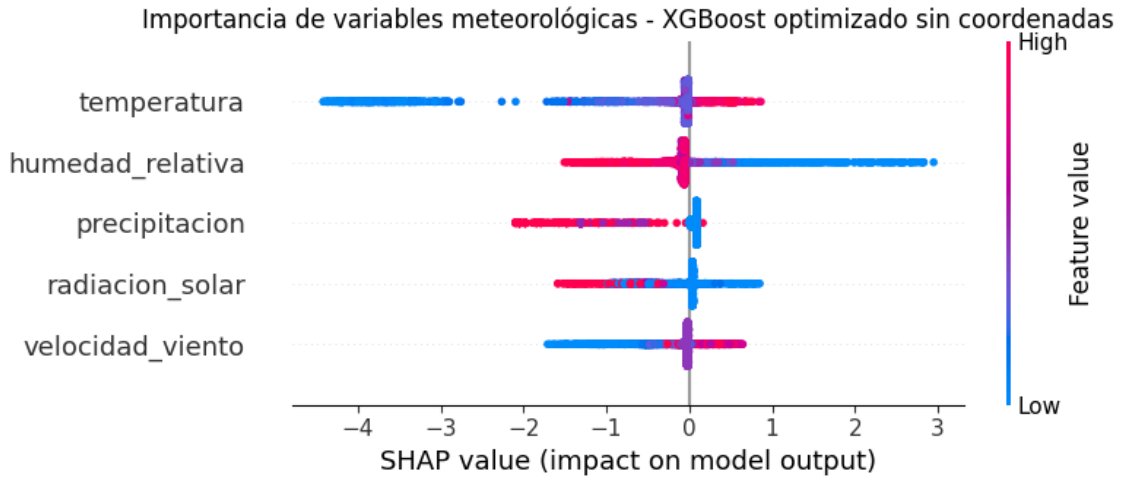
2.3.1.8 Evaluación del modelo sin componente espacial

Debido a la elevada importancia de latitud y longitud observada mediante SHAP, se realizó una prueba de ablación (Meyes et al., 2019), eliminando las dos variables para determinar el desempeño del modelo utilizando únicamente información meteorológica.

Métrica	Con coordenadas	Sin coordenadas
Umbral	0,5	0,6
TN	237.096	297.586
FP	77.015	16.525
FN	55	150
TP	152	57
Accuracy	0,754802	0,946949
Balanced Accuracy	0,744558	0,611377
Precision	0,00197	0,003437
Recall	0,7343	0,275362
F1-score	0,003929	0,00679
ROC-AUC	0,817874	0,694479
PR-AUC	0,004903	0,003347

Tabla 29:Comparación del XGBoost optimizado con y sin coordenadas geográficas en la prueba temporal de 2024

Con el propósito de identificar las variables que adquirieron mayor importancia una vez eliminado el componente espacial, se aplicó nuevamente el método SHAP al XGBoost optimizado sin coordenadas. El análisis se realizó considerando exclusivamente las cinco variables meteorológicas utilizadas por esta configuración: temperatura, humedad relativa, precipitación, radiación solar y velocidad del viento. SHAP permite estimar la contribución de las variables sobre las predicciones generadas por el modelo



Como se puede visualizar en el grafico de Shap, los valores altos de temperatura mostraron principalmente contribuciones hacia valores SHAP positivos, mientras que los

valores bajos presentaron contribuciones negativas, para la humedad relativa se observó un comportamiento diferente, debido a que los valores bajos se extendieron hacia contribuciones SHAP positivas, mientras que los valores altos se concentraron principalmente en contribuciones negativas, en la precipitación los valores altos mostraron predominantemente contribuciones negativas, la radiación solar y la velocidad del viento también participaron en las predicciones, aunque presentaron una importancia global menor respecto a las demás variables.

De esta manera la prueba realizada permitió obtener una segunda configuración del XGBoost optimizado basada exclusivamente en información meteorológica y determinar tanto su desempeño sobre el conjunto temporal de prueba de 2024 como la importancia de las variables utilizadas en sus predicciones.

2.3.2. Análisis

Los resultados obtenidos muestran que la predicción de incendios forestales en el Distrito Metropolitano de Quito presenta un desafío importante, principalmente debido a que estos eventos ocurren con poca frecuencia en comparación con la gran cantidad de registros meteorológicos disponibles, en la base general de modelado los incendios representaron apenas el 0,1051 % de las observaciones, mientras que en el conjunto temporal de prueba correspondiente a 2024 la prevalencia fue del 0,0659 %**,** con 207 incendios registrados entre 314.318 observaciones**.**

Esta diferencia entre la cantidad de registros con y sin incendio debe considerarse al momento de interpretar el desempeño de los modelos, en este tipo de escenarios, una exactitud (Accuracy) elevada puede estar influenciada principalmente por la correcta identificación de los numerosos casos en los que no ocurre un incendio y, por sí sola, no permite determinar qué tan bien se identifican los eventos de interés, por esta razón, la evaluación se complementó con métricas como Balanced Accuracy, Recall, ROC-AUC, PR-AUC, Precision y F1-Score**,** además del análisis de las matrices de confusión, permitiendo obtener una valoración más completa del comportamiento de los modelos.

2.3.2.1. Análisis comparativo del desempeño de los modelos predictivos

La comparación de los modelos predictivos permitió identificar diferencias importantes en la capacidad de Regresión Logística, Random Forest y XGBoost para detectar incendios forestales, en la interpretación de los resultados se visualiza un fuerte

desbalance de la variable objetivo, debido a que los incendios representaron únicamente el 0,1051 % de la base general de modelado, esta condición fue aún más marcada en el conjunto temporal de prueba correspondiente a 2024, donde se registraron 207 incendios entre 314.318 observaciones.

Debido al desbalance el desempeño de los algoritmos no puede interpretarse únicamente mediante Accuracy, puesto que una elevada exactitud global puede estar determinada principalmente por la correcta clasificación de la clase mayoritaria No incendio, por esta razón, se analizaron conjuntamente Balanced Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC y las matrices de confusión, prestando especial atención a la capacidad de identificación de la clase minoritaria.

Entre los modelos base, XGBoost presentó el comportamiento global más equilibrado, con un Balanced Accuracy de 0,7281, Recall de 0,6763 y ROC-AUC de 0,8009, identificando correctamente 140 de los 207 incendios registrados en 2024. La Regresión Logística alcanzó un Recall de 0,6618, correspondiente a 137 incendios detectados, mientras que Random Forest obtuvo un Recall de 0,4493, con 93 incendios correctamente identificados. Estos resultados evidencian que una mayor exactitud global no necesariamente implica una mayor capacidad para reconocer la clase incendio en conjuntos de datos altamente desbalanceados.

Posteriormente, la optimización de hiperparámetros mejoró el desempeño de Random Forest y XGBoost. Random Forest optimizado alcanzó un Recall de 0,7536, identificando 156 de los 207 incendios, mientras que XGBoost optimizado obtuvo un Balanced Accuracy de 0,7446, Recall de 0,7343, ROC-AUC de 0,8179 y PR-AUC de 0,004903, detectando correctamente 152 incendios. Aunque Random Forest optimizado presentó un Recall superior, XGBoost mostró un comportamiento conjunto más equilibrado en Balanced Accuracy, ROC-AUC y PR-AUC, motivo por el cual fue seleccionado como modelo final para continuar con el análisis de interpretabilidad mediante SHAP.

La capacidad de detección alcanzada por XGBoost debe analizarse conjuntamente con su Precision y su matriz de confusión. El modelo generó 77.015 falsos positivos frente a 152 verdaderos positivos, obteniendo una Precision de 0,001970. Por tanto, aunque logró identificar el 73,43 % de los incendios registrados en 2024, una proporción considerable de las observaciones clasificadas como positivas no correspondió a incendios reales. Este

comportamiento se encuentra estrechamente relacionado con la extremadamente baja prevalencia del evento y constituye una de las principales limitaciones para su aplicación operativa.

Desde una perspectiva práctica, los resultados muestran que existe un equilibrio entre la capacidad del modelo para detectar incendios y la cantidad de falsos positivos que genera, en la prevención de incendios forestales es importante identificar la mayor cantidad posible de eventos reales, sin embargo, el número de falsos positivos obtenido sigue siendo grande, por lo que no sería adecuado considerar cada predicción positiva como una alerta de incendio, el modelo podría utilizarse como una herramienta de apoyo para identificar condiciones o períodos que requieran mayor vigilancia, complementando sus resultados con otros criterios técnicos y fuentes de información, a futuro, será necesario continuar mejorando el modelo para reducir las falsas alarmas y fortalecer su utilidad.

Finalmente, los datos se separaron de forma temporal. La información de 2021–2023 se utilizó para el entrenamiento, optimización y selección del modelo, mientras que los registros de 2024 se reservaron para la prueba final. Esta separación permitió evaluar los modelos con datos de un período posterior sin utilizar información de 2024 durante el proceso de entrenamiento y ajuste.

Los resultados correspondientes al período 2025–2026 deben interpretarse considerando el cambio en el criterio técnico institucional de clasificación de incendios forestales aplicado a partir de 2025. Debido a que el modelo fue desarrollado principalmente con registros definidos bajo el criterio anterior, mientras que los eventos posteriores fueron clasificados bajo un criterio diferente, existe una ruptura en la definición de la variable objetivo, por lo tanto el desempeño obtenido durante 2025–2026 no es directamente comparable con el período 2021–2024 y se presenta únicamente como una evaluación adicional post-cambio.

En relación con la pregunta de investigación sobre el algoritmo que presentó mejores resultados, XGBoost optimizado mostró el desempeño más equilibrado entre los modelos evaluados. Sin embargo, su selección no se basó en una sola métrica, sino en el análisis conjunto de los resultados obtenidos. Aunque mostró una buena capacidad para detectar incendios, también presentó limitaciones relacionadas con la Precision y la generación de falsos positivos, motivo po la cual los resultados muestran que el modelo puede servir

como apoyo para la gestión del riesgo, pero aun requiere mejoras antes de considerar su uso como un sistema de alerta.

2.3.2.2. Análisis de la influencia de las variables meteorológicas y espaciales

El análisis mediante SHAP permitió identificar la influencia de las variables meteorológicas y espaciales en las predicciones realizadas por XGBoost, en el modelo optimizado que incorpora todas las variables, la longitud y la latitud presentaron la mayor importancia global, lo que permite evidenciar que el componente espacial aporta información relevante para distinguir las condiciones asociadas con la ocurrencia de incendios forestales, sin embargo estas variables deben interpretarse como indicadores de localización y de los patrones territoriales presentes en los datos históricos, y no como causas directas de la ocurrencia de incendios.

Dado que existe una elevada importancia de las coordenadas se planteó la necesidad de evaluar si el modelo dependía principalmente de la ubicación geográfica de las observaciones o si las variables meteorológicas conservaban capacidad predictiva de manera independiente, para analizar este aspecto se realizó una prueba de ablación, eliminando la longitud y latitud y optimizando nuevamente XGBoost utilizando únicamente las 5 variables meteorológicas que son temperatura, humedad relativa, precipitación, radiación solar y velocidad del viento, esta evolución permitió comparar el desempeño del modelo XGBoost optimizado con y sin información espacial bajo la misma estrategia de evaluación temporal.

Los resultados presentados en la Tabla 28 evidenciaron una disminución del desempeño al retirar las coordenadas, el ROC-AUC pasó de 0,817874 a 0,694479, mientras que la Balanced Accuracy disminuyó de 0,744558 a 0,611377, una diferencia significativa se evidenció en la detección de la clase incendio: el Recall disminuyó de 0,734300 a 0,275362, y los verdaderos positivos pasaron de 152 a 57, de la misma manera los falsos negativos aumentaron de 55 a 150, estos resultados demuestran que el componente espacial aporta información importante para la capacidad del modelo de identificar los incendios registrados en el conjunto temporal de prueba de 2024.

El modelo sin coordenadas presentó, por otra parte, una reducción de los falsos positivos de 77.015 a 16.525 y un incremento de la Precision de 0,001970 a 0,003437. Sin embargo, esta reducción debe interpretarse con precaución, debido a que el modelo con

coordenadas utilizó un umbral de clasificación de 0,50, mientras que el modelo sin coordenadas empleó un umbral de 0,60. Por tanto la disminución de falsos positivos no puede atribuirse únicamente a la eliminación de latitud y longitud, además es importante mencionar al obtener una reducción de falsas alarmas también se obtiene una pérdida considerable en la detección de incendios.

A pesar de esta disminución, el modelo sin coordenadas mantuvo un ROC-AUC de 0,694479, superior al comportamiento esperado de una clasificación sin capacidad discriminativa, este resultado indica que las variables meteorológicas conservan información útil para diferenciar las observaciones asociadas con incendios y no incendios, aunque con una capacidad menor que cuando se incorpora el componente espacial, por lo tanto los resultados no permiten sostener que XGBoost se limite exclusivamente a memorizar los lugares donde históricamente se han presentado incendios, sin embargo sí evidencian una dependencia relevante de la información espacial para alcanzar el desempeño observado en el modelo final.

Esta interpretación se complementa con el análisis SHAP realizado sobre el XGBoost optimizado sin coordenadas, al retirar latitud y longitud, la temperatura presentó la mayor importancia global, seguida por la humedad relativa, precipitación, radiación solar y velocidad del viento, evidenciándose que los valores altos de temperatura mostraron principalmente contribuciones hacia la predicción positiva, mientras que para la humedad relativa se observó una tendencia inversa, con valores bajos asociados con contribuciones positivas. Con estos resultados se logra evidenciar que en ausencia de información espacial, XGBoost continúa identificando patrones relacionados con las condiciones meteorológicas.

En conjunto, la prueba de ablación permite concluir que el modelo combina información espacial y meteorológica para realizar sus predicciones. Las coordenadas incrementan de manera importante su capacidad discriminativa y de detección, mientras que las variables meteorológicas mantienen un aporte independiente, aunque insuficiente para reproducir el desempeño alcanzado por el modelo completo. En consecuencia, la importancia de latitud y longitud debe entenderse como evidencia de la existencia de patrones espaciales asociados con los incendios históricos del DMQ.

2.3.2.3. Aporte del modelo a la predicción y prevención de incendios forestales

El modelo XGBoost optimizado se presenta como una potencial herramienta de apoyo para la identificación de condiciones asociadas con la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito. En la evaluación temporal correspondiente a 2024, el modelo identificó correctamente 152 de los 207 incendios registrados, equivalente al 73,43 %**,** con un Recall de 0,7343. Este resultado muestra que el modelo logró reconocer una proporción importante de los incendios forestales ocurridos en un período que no fue utilizado durante su entrenamiento.

Sin embargo, esta capacidad de detección debe analizarse conjuntamente con las demás métricas obtenidas, con el umbral de clasificación de 0,50, el modelo generó 77.015 falsos positivos y alcanzó una Precision de 0,001970, lo que evidencia que una proporción considerable de las observaciones clasificadas como positivas no correspondió a incendios reales. Este comportamiento representa una limitación importante para interpretar cada predicción positiva como una alerta individual y debe considerarse dentro del contexto del fuerte desbalance de clases y de la baja prevalencia de incendios en el conjunto temporal de prueba de 2024.

Con el propósito de analizar si un umbral de decisión más restrictivo podría reducir este volumen de falsas alertas, se realizó un análisis de sensibilidad sobre diferentes umbrales de clasificación. Los resultados mostraron que al incrementar el umbral a 0,70, los falsos positivos disminuyeron de 77.015 a 43.509, sin embargo los incendios correctamente identificados también disminuyeron de 152 a 112, y el Recall pasó de 0,7343 a 0,5411. Con un umbral de 0,80, los falsos positivos se redujeron considerablemente hasta 7.391, pero el modelo identificó únicamente 40 de los 207 incendios forestales**,** alcanzando un Recall de 0,1932. Con umbrales aún más restrictivos, la reducción de falsos positivos estuvo acompañada de una pérdida mayor en la capacidad de detección.

Estos resultados evidencian un compromiso entre la reducción de falsos positivos y la detección de incendios**.** En los escenarios evaluados no se observó un umbral que permitiera reducir sustancialmente las falsas alarmas sin disminuir de manera importante el Recall. Por esta razón, incrementar el umbral por sí solo no resuelve la principal limitación operativa del modelo, ya que una reducción considerable de falsas alertas implica dejar de identificar una mayor cantidad de incendios reales.

En consecuencia, el aporte práctico del modelo no debe interpretarse como la capacidad de generar de manera autónoma una alarma ante la posible ocurrencia de un incendio. Su utilidad se encuentra principalmente en el apoyo a procesos de priorización y vigilancia, utilizando sus predicciones como información complementaria para identificar condiciones, períodos o sectores que requieran mayor atención, en conjunto con otros criterios técnicos, ambientales y operativos disponibles para la toma de decisiones. De esta forma, el modelo podría contribuir a focalizar actividades de seguimiento y análisis, sin sustituir la valoración realizada por el personal responsable de la prevención y gestión de incendios forestales.

La incorporación de la interpretabilidad mediante SHAP complementa esta utilidad, debido a que permite identificar las variables que contribuyen en mayor medida a las predicciones, con estos resultados se lograron evidenciar una influencia relevante del componente espacial, adicional la prueba realizada sin coordenadas mostró que las variables meteorológicas conservaron capacidad discriminativa, destacándose principalmente la temperatura y la humedad relativa, de esta forma la salida del modelo puede complementarse con información sobre las condiciones que influyeron en la predicción y no limitarse únicamente a una clasificación de incendio o no incendio.

Desde el punto de vista de la prevención, los resultados constituyen una aproximación inicial para transformar información histórica, meteorológica y espacial en información predictiva orientada al apoyo de la toma de decisiones.

Para futuras versiones podrían incorporarse variables relacionadas con vegetación, topografía, uso del suelo, condiciones de sequedad, proximidad a áreas pobladas y otros factores ambientales o antrópicos, además de evaluar nuevas estrategias de modelado y calibración. Estas mejoras permitirían avanzar hacia modelos más robustos y con un equilibrio más adecuado entre la detección de incendios forestales y la generación de falsas alarmas.

2.3.2.4. Comparación de los resultados con estudios previos

Los resultados obtenidos en esta investigación presentan coincidencias con estudios previos que han aplicado técnicas de Machine Learning para la predicción de incendios forestales, en particular el desempeño alcanzado por XGBoost optimizado, con un ROC-AUC de 0,8179 y un Recall de 0,7343 se evidencia la capacidad de los algoritmos de

ensamble basados en árboles para identificar patrones complejos entre variables ambientales y espaciales, estos resultados deben analizarse considerando el fuerte desbalance de clases presente en el conjunto de datos y las características particulares del contexto geográfico del Distrito Metropolitano de Quito.

(Rodrigues et al., 2014) demostraron que la ocurrencia de incendios forestales presenta heterogeneidad espacial y que la influencia de los factores explicativos puede variar según la localización geográfica. Este hallazgo guarda relación con los resultados obtenidos mediante SHAP en la presente investigación donde la longitud y latitud de las estaciones meteorológicas fueron las variables de mayor influencia global, sin embargo las coordenadas utilizadas en este estudio deben interpretarse como indicadores de localización que permiten al modelo diferenciar patrones territoriales y no como factores causales directos de la ocurrencia de incendios, esta coincidencia evidencia la importancia de considerar el componente espacial en el análisis del riesgo de incendios forestales.

Los resultados también presentan similitudes con la investigación de Liao et al. (2025), desarrollada en Corea del Sur, en la cual se compararon Random Forest, XGBoost y redes neuronales artificiales utilizando variables climáticas, de vegetación, topográficas y socioambientales, en dicho estudio, XGBoost presentó el mejor desempeño, con un AUC de 0,76 y un F1-Score de 0,511, en la presente investigación, XGBoost también presentó el comportamiento global más equilibrado entre los modelos evaluados y alcanzó un ROC-AUC de 0,8179, sin embargo estos valores no deben compararse directamente como si correspondieran a condiciones experimentales equivalentes, debido a las diferencias existentes en las regiones de estudio, variables predictoras, distribución de las clases y estrategias de evaluación utilizadas.

Una diferencia relevante es que el modelo desarrollado para el Distrito Metropolitano de Quito se construyó principalmente a partir de variables meteorológicas y espaciales disponibles en las estaciones analizadas, mientras que investigaciones previas incorporaron una mayor diversidad de características relacionadas con vegetación, topografía, densidad poblacional y otros factores territoriales y humanos, esta diferencia es un aspecto importante para interpretar el desempeño obtenido y plantea una oportunidad para fortalecer futuras versiones del modelo mediante la incorporación de nuevas fuentes de información.

La aplicación de SHAP también permite establecer relaciones con investigaciones que han utilizado técnicas de Inteligencia Artificial Explicable para comprender los factores asociados con las predicciones de incendios forestales, en la presente investigación SHAP permitió determinar que entre las variables meteorológicas, la humedad relativa presentó la mayor influencia, seguida por la temperatura y la precipitación, asimismo los niveles bajos de humedad tendieron a contribuir hacia la predicción de incendio, mientras que valores elevados de precipitación presentaron principalmente contribuciones hacia la clase No incendio. Estos resultados muestran que el modelo aprendió patrones meteorológicos diferenciados y permiten complementar las métricas de desempeño con información sobre el comportamiento de las variables predictoras.

Otro aspecto que diferencia esta investigación de algunos estudios previos es la estrategia utilizada para la evaluación final, el modelo fue entrenado y optimizado utilizando información correspondiente al período 2021–2023, mientras que los registros de 2024 se mantuvieron como conjunto temporal de prueba, además, durante esta evaluación se conservó la distribución real de la clase incendio, cuya prevalencia en 2024 fue de apenas 0,0659 %. Esta condición representa un escenario particularmente exigente y contribuye a explicar la baja Precision y el elevado número de falsos positivos observados en la evaluación final.

En este sentido, aunque XGBoost optimizado identificó 152 de los 207 incendios registrados en 2024, equivalente a un Recall de 0,7343, también generó 77.015 falsos positivos y obtuvo una Precision de 0,001970. Estos resultados muestran que la capacidad de detección debe analizarse conjuntamente con las falsas alarmas generadas y refuerzan la necesidad de interpretar el modelo como una herramienta de apoyo para la vigilancia y priorización, más que como un mecanismo independiente para determinar la ocurrencia de incendios forestales.

En conjunto, la comparación con los estudios previos permite identificar coincidencias respecto a la utilidad de XGBoost para modelar relaciones complejas y a la relevancia de los componentes ambientales y espaciales, las diferencias metodológicas y en las variables utilizadas evidencian la necesidad de interpretar los resultados dentro de las características específicas de cada investigación, el principal aporte del presente estudio consiste en proporcionar evidencia basada en registros históricos y meteorológicos locales, aplicar una evaluación temporal sobre un período posterior e incorporar SHAP

para comprender las variables que participan en las predicciones del modelo en el contexto del Distrito Metropolitano de Quito.

Capítulo 3: Conclusiones y Recomendaciones

3.1. Conclusiones

- La presente investigación permitió desarrollar y evaluar un modelo predictivo de incendios forestales para el Distrito Metropolitano de Quito mediante la integración de registros históricos de incendios forestales, información meteorológica proveniente del CBDMQ y EPMAPS, el procesamiento, homologación e integración de estas fuentes permitió construir un conjunto de datos adecuado para la aplicación y evaluación de algoritmos de Machine Learning.
- La construcción de la variable objetivo evidenció un fuerte desbalance de clases, debido a que los registros asociados con incendios forestales representaron únicamente el 0,1051 % de la base general de modelado, esta característica determinó la necesidad de evaluar los algoritmos mediante métricas complementarias al Accuracy, principalmente Balanced Accuracy, Precision, Recall, F1-Score, ROC-AUC y PR-AUC, además de las matrices de confusión, los resultados demostraron que una elevada exactitud global no necesariamente representa una adecuada capacidad para identificar incendios forestales cuando existe una marcada diferencia entre la cantidad de observaciones de ambas clases.
- En relación con la comparación de los algoritmos de Machine Learning, se identificaron diferencias en el desempeño de Regresión Logística, Random Forest y XGBoost, después del proceso de optimización, XGBoost optimizado fue seleccionado como modelo final debido a su comportamiento conjunto en las métricas evaluadas, en el conjunto temporal de prueba correspondiente a 2024 alcanzó un Balanced Accuracy de 0,7446, Recall de 0,7343, ROC-AUC de 0,8179 y PR-AUC de 0,004903, identificando correctamente 152 de los 207 incendios registrados, equivalente al 73,43 %. Aunque Random Forest optimizado presentó un Recall ligeramente superior, XGBoost obtuvo un mejor comportamiento conjunto en Balanced Accuracy, ROC-AUC y PR-AUC.
- La separación temporal utilizada fortaleció el proceso de evaluación del modelo, los registros correspondientes al período 2021–2023 fueron utilizados para el entrenamiento y optimización, mientras que 2024 se mantuvo como conjunto de

prueba final, este procedimiento permitió evaluar la capacidad del modelo sobre información de un período posterior sin utilizar datos futuros durante el proceso de ajuste y conservando la distribución real de los incendios en el conjunto de evaluación.

- El análisis SHAP mostró que la longitud y latitud fueron las variables de mayor influencia en el XGBoost optimizado, sin embargo, la prueba de ablación evidenció que al eliminar las coordenadas el modelo mantuvo capacidad predictiva, aunque el ROC-AUC disminuyó de 0,817874 a 0,694479 y el Recall de 0,734300 a 0,275362. Estos resultados evidencian que el modelo presenta una dependencia relevante del componente espacial, pero no permiten sostener que sus predicciones se basen exclusivamente en la memorización de las zonas con alta ocurrencia de incendios forestales (hotspots históricos), ya que las variables meteorológicas conservaron capacidad discriminativa, en ausencia de las coordenadas, la temperatura y la humedad relativa fueron las variables de mayor influencia.
- El aporte del modelo a la prevención y toma de decisiones, los resultados evidencian una capacidad relevante para identificar una proporción importante de los incendios forestales registrados, sin embargo esta capacidad debe analizarse conjuntamente con las falsas alarmas generadas, en la evaluación temporal de 2024, XGBoost optimizado produjo 77.015 falsos positivos y obtuvo una Precision de 0,001970, aunque el modelo alcanzó un Recall de 0,7343, existe todavía una diferencia considerable entre la cantidad de predicciones positivas y los incendios efectivamente registrados.
- El modelo se presenta como una potencial herramienta complementaria para apoyar procesos de vigilancia, prevención y priorización territorial, permitiendo identificar condiciones que podrían requerir mayor atención y análisis, los resultados obtenidos pueden contribuir como un elemento adicional dentro del proceso de toma de decisiones, en conjunto con otros criterios técnicos, ambientales y operativos, para fortalecer una posible aplicación futura como herramienta de apoyo a la alerta de incendios forestales es necesario continuar mejorando su capacidad de discriminación y reduciendo la generación de falsos positivos.
- La presente investigación aporta una metodología reproducible para integrar información histórica, meteorológica y espacial, comparar diferentes algoritmos de Machine Learning, realizar una evaluación temporal e incorporar técnicas de Inteligencia Artificial Explicable mediante SHAP, los resultados obtenidos indican

una evidencia específica para las condiciones analizadas en el Distrito Metropolitano de Quito y constituyen una base para continuar desarrollando modelos predictivos orientados al fortalecimiento de la prevención y gestión del riesgo de incendios forestales.

3.2. Limitaciones

- Una de las principales limitaciones de la investigación corresponde al fuerte desbalance de la variable objetivo, la ocurrencia de incendios forestales representó una proporción muy reducida frente a las observaciones correspondientes a la clase No incendio, esta característica dificultó la discriminación de la clase minoritaria y se reflejó principalmente en los bajos valores de Precision y en la generación de un elevado número de falsos positivos.
- En la evaluación temporal correspondiente a 2024, el modelo XGBoost optimizado identificó correctamente 152 de los 207 incendios registrados, equivalente al 73,43 %, pero generó 77.015 falsos positivos y alcanzó una Precision de 0,001970, estos resultados evidencian, si bien el modelo presenta capacidad para detectar una proporción importante de los incendios todavía existe un amplio margen de mejora en la discriminación de las observaciones clasificadas como positivas, esta condición debe considerarse especialmente ante una posible aplicación del modelo como herramienta de apoyo para la generación de alertas.
- Otra limitación se relaciona con la disponibilidad y completitud de la información meteorológica, los registros provenientes del CBDMQ y EPMAPS presentaron diferencias en los períodos disponibles, frecuencia de medición y variables registradas, además no todas las estaciones meteorológicas disponían de información completa para las variables seleccionadas, por lo que fue necesario realizar procedimientos de homologación, integración y tratamiento de valores faltantes para construir la base utilizada durante el modelado.
- El modelo también estuvo condicionado por las variables disponibles en las fuentes de información utilizadas, la investigación incorporó temperatura, humedad relativa, precipitación, radiación solar, velocidad del viento y coordenadas geográficas de las estaciones meteorológicas, sin embargo no se contó con otras características que podrían aportar información adicional para representar las condiciones asociadas con los incendios forestales, la incorporación futura de variables relacionadas con vegetación, topografía, uso del suelo y factores antrópicos permitiría evaluar si una

mayor diversidad de información contribuye a mejorar la capacidad predictiva del modelo.

- Asimismo el modelo presentó una dependencia relevante del componente espacial, con la prueba de ablación se demostró que al eliminar la longitud y latitud, disminuyeron el ROC-AUC, la Balanced Accuracy y el Recall, evidenciando que las coordenadas aportan información importante al desempeño predictivo. Aunque el modelo sin coordenadas mantuvo capacidad discriminativa mediante las variables meteorológicas, la influencia del componente espacial puede limitar la generalización del modelo hacia zonas con características territoriales diferentes a las representadas en los datos de entrenamiento.
- La utilización de 2024 como conjunto temporal independiente permitió evaluar el modelo sobre un período posterior al utilizado durante el entrenamiento y optimización, los resultados corresponden a las condiciones y fuentes de información disponibles durante el período analizado, la evaluación con nuevos períodos temporales permitiría determinar si el comportamiento observado se mantiene y contribuiría a fortalecer la capacidad de generalización del modelo.
- El cambio en el criterio técnico institucional para la clasificación de incendios forestales aplicado a partir de 2025 es una limitación para la comparabilidad temporal de los datos, antes de este período la clasificación se encontraba asociada principalmente con una superficie afectada igual o superior a 100 m^2, mientras que desde 2025 se consideran las características del material involucrado en la combustión y no únicamente el metraje afectado, esta modificación genera una ruptura en la definición de la variable objetivo y debe considerarse en futuras actualizaciones, entrenamientos y evaluaciones del modelo.

3.3. Recomendaciones

- Se recomienda continuar fortaleciendo el modelo mediante la incorporación de nuevas variables ambientales, territoriales y antrópicas, como cobertura vegetal, tipo y humedad del combustible, topografía, pendiente, uso del suelo y factores relacionados con la actividad humana, la incorporación de estas nuevas variables permitiría evaluar si una mayor diversidad de información contribuye a mejorar la capacidad del modelo para diferenciar las condiciones asociadas con la ocurrencia y no ocurrencia de incendios forestales.

- Dado que se tiene un gran desbalance de clases identificado en la investigación y el elevado número de falsos positivos obtenido durante la evaluación temporal, se recomienda evaluar en futuros trabajos estrategias adicionales orientadas a mejorar la discriminación de la clase incendio, estas pueden incluir el análisis de diferentes umbrales de clasificación métodos para el tratamiento de clases desbalanceadas y la incorporación de nuevas variables predictoras, estas estrategias deberán evaluarse mediante conjuntos temporales independientes para determinar su efecto sobre el equilibrio entre la detección de incendios y la generación de falsos positivos.
- Se recomienda mantener la evaluación del modelo mediante diferentes métricas de desempeño y no únicamente mediante Accuracy, con el propósito de obtener una valoración más completa de su capacidad predictiva, para lo cual resulta conveniente analizar conjuntamente Balanced Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC y las matrices de confusión, poniendo mayor atención tanto a los incendios correctamente identificados como a la cantidad de falsos positivos generados.
- Se recomienda realizar nuevas evaluaciones temporales a medida que se disponga de información de períodos posteriores, esto permitirá determinar si el comportamiento del modelo se mantiene en el tiempo y contribuirá a evaluar su capacidad de generalización antes de considerar una aplicación operativa de mayor alcance.
- Debido al cambio en el criterio institucional utilizado para el registro de incendios forestales a partir de 2025, se recomienda documentar y mantener criterios claramente definidos para la construcción de la variable objetivo en futuros períodos, esto permitirá mejorar la comparabilidad temporal de los registros y facilitar la actualización y evaluación posterior del modelo.
- Para una posible aplicación institucional, se recomienda utilizar inicialmente los resultados del modelo como información complementaria en las actividades de vigilancia, prevención y priorización territorial. Esta información puede combinarse con criterios técnicos, ambientales y operativos para apoyar la toma de decisiones. Además, se recomienda continuar evaluando el modelo con nuevos datos para reducir la cantidad de falsos positivos.
- Se recomienda mantener el uso de técnicas de Inteligencia Artificial Explicable, como SHAP, en futuras versiones del modelo. Esto permitirá analizar si la importancia de las variables cambia cuando se incorporen nuevos datos y facilitará la interpretación de los resultados obtenidos.

4. Referencias

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

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barrett, K., Blanco, G., Cheung, W. W. L., Connors, S., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., … Ha, M. (with Lee, H.). (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge-Vorsatz, C. Xiao, N. Yassaa, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, … C. Péan, Eds.; First). Intergovernmental Panel on Climate Change (IPCC). https://doi.org/10.59327/IPCC/AR6-9789291691647

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. SPSS Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (Sixth edition). SAGE.

Food and Agriculture Organization of the United Nations. (2024). *Fire Management: Voluntary Guidelines: Principles and Strategic Actions*. Food and Agriculture Organization of the United Nations (FAO). https://openknowledge.fao.org/items/86091f4f-c0af-49f7-b6d6-a2e3efac3cb8

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

Jiawei Han, J. P., & Hanghang Tong. (2023). *Data Mining*. Elsevier. https://doi.org/10.1016/C2013-0-18660-6

Liao, B., Zhou, T., Liu, Y., Li, M., & Zhang, T. (2025). Tackling the Wildfire Prediction Challenge: An Explainable Artificial Intelligence (XAI) Model Combining Extreme Gradient Boosting (XGBoost) with SHapley Additive exPlanations (SHAP) for Enhanced Interpretability and Accuracy. *Forests*, *16*(4), 689. https://doi.org/10.3390/f16040689

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation Studies in Artificial Neural Networks* (Versión 2). arXiv. https://doi.org/10.48550/ARXIV.1901.08644

Ríos Hervas, J. G. (2023). Causas y origen de los incendios forestales en el Ecuador. *InnDev*, *2*(4), 11-21. https://doi.org/10.69583/inndev.v2n4.2023.79

Rodrigues, M., De La Riva, J., & Fotheringham, S. (2014). Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography*, *48*, 52-63. https://doi.org/10.1016/j.apgeog.2014.01.011

Zhou, T., Liao, B., Liu, Y., Li, M., & Zhang, T. (2025). Tackling the Wildfire Prediction Challenge: An Explainable Artificial Intelligence (XAI) Model Combining Extreme Gradient Boosting (XGBoost) with SHapley Additive exPlanations (SHAP) for Enhanced Interpretability and Accuracy. *Forests*, *16*(4), 689. https://doi.org/10.3390/f16040689

5. Anexos

5.1 Código de programación

Link de drive de Código Fuente del modelo

https://drive.google.com/file/d/1GSSvLQrd2KHd5FRPSgF0bBMV_Lttv5Vn/view?usp=sharing