Modelado predictivo de la ocurrencia de focos de incendios forestales en Colombia mediante técnicas de Machine Learning

Presentado para obtener el título de

MAGÍSTER EN MATEMÁTICAS APLICADAS Y CIENCIAS DE LA COMPUTACIÓN

Mateo Novoa Cardozo

Dirección:

Fabián Sánchez Salazar

Universidad del Rosario

Escuela de Ciencias e Ingeniería

Maestría en Matemáticas Aplicadas y Ciencias de la Computación

DEDICATORIA

Dedico este trabajo a mi familia, en especial a mi madre y a mis hermanas, por su apoyo y amor incondicional.

A Camilo, por su compañía y sus palabras de aliento en cada etapa del proceso.

A mi perrita Arya, que me acompañó silenciosamente durante todo el desarrollo, sobre todo en las noches en las que trasnochábamos juntos.

Y a todas las personas que han sido víctimas de los incendios forestales. Que sus historias nos inspiren a cuidar la vida y proteger la naturaleza.

RESUMEN

Los incendios forestales se han convertido en un riesgo creciente para los ecosistemas y las comunidades en Colombia, especialmente bajo condiciones de cambio climático y uso intensivo del suelo. En respuesta a esta problemática, este estudio desarrolla un modelo de predicción semanal de la ocurrencia de focos de incendios forestales, planteado como un problema de clasificación binaria mediante técnicas de *machine learning*.

Para su desarrollo, se emplean datos históricos del periodo 2010–2023 a nivel municipal en Colombia, integrando variables climáticas como humedad relativa, precipitación y temperatura, junto con registros históricos de incendios y características geográficas. Entre los modelos evaluados, *LightGBM* presentó el mejor desempeño, alcanzando un AUC de 0.85 y un *F1 score* de 76% en validación temporal.

Asimismo, el análisis de importancia de variables mostró que la temperatura máxima, la humedad relativa mínima y la precipitación máxima, con rezagos entre una y tres semanas, fueron los predictores más relevantes. En conjunto, estos resultados muestran que la información climática puede ayudar a anticipar incendios forestales con una semana de antelación, ofreciendo una herramienta útil para la prevención y la gestión del riesgo en Colombia.

ABSTRACT

Wildfires have become an increasing risk to ecosystems and communities in Colombia, particularly under conditions of climate change and intensive land use. In response to this challenge, this study develops a model for the weekly prediction of wildfire hotspot occurrence, formulated as a binary classification problem using machine learning techniques.

The model is built using historical municipal-level data from Colombia covering the 2010–2023 period, integrating climatic variables such as relative humidity, precipitation, and temperature, together with historical wildfire records and geographic features. Among the evaluated models, LightGBM achieved the best performance, reaching an AUC of 0.85 and an F1 score of 76% under temporal validation.

In addition, the variable importance analysis showed that maximum temperature, minimum relative humidity, and maximum precipitation, with lags ranging from one to three weeks, were the most relevant predictors. Overall, these results show that climatic information can help anticipate wildfires one week in advance, providing a useful tool for prevention and risk management in Colombia.

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

Capítulo 1
INTRODUCCIÓN

Los incendios forestales representan una de las mayores amenazas ambientales en Colombia. En los últimos años, se ha visto el aumento significativo en frecuencia y volumen de hectáreas quemadas afectando grandes áreas de bosques, poniendo en riesgo la biodiversidad y el bienestar de comunidades rurales. Aunque factores como las sequías, las altas temperaturas y el cambio climático crean condiciones propicias para que el fuego se propague [1], diversos estudios muestran que la mayoría de los incendios tienen origen en actividades humanas, como quemas agrícolas o descuidos. En este sentido, el clima no suele ser la causa directa, sino un factor que intensifica el problema [2]. Por eso, los incendios deben entenderse como el resultado de la combinación entre acciones humanas y condiciones ambientales, por lo que es necesario analizarlos considerando todos sus factores.

Colombia, por su ubicación geográfica y diversidad de ecosistemas, presenta una alta vulnerabilidad ante este tipo de eventos. A pesar de contar con más del 50 % de su territorio cubierto por bosques, las estrategias de monitoreo, prevención y respuesta frente a los incendios forestales siguen siendo limitadas [3] [4]. Específicamente en regiones como la Amazonía y la Orinoquía, donde los resultados de este estudio evidencian una escasa cobertura de estaciones meteorológicas, lo que dificulta el monitoreo climático y la planificación de acciones preventivas.

El desarrollo de herramientas que permitan anticipar estos eventos con mayor precisión se ha convertido en una necesidad estratégica. En este escenario, el uso de técnicas de *machine learning* surge como una alternativa innovadora y eficaz, teniendo en cuenta que, estas metodologías permiten analizar grandes volúmenes de datos ambientales, climáticos y geográficos para identificar patrones y predecir con mayor precisión la ocurrencia de incendios.

Mientras que países como España [5], Vietnam [6] o Corea del Sur [7] ya han desarrollado modelos con resultados prometedores, En Colombia, el avance en este tema ha sido más limitado. Aunque ya existen algunas iniciativas para predecir incendios forestales usando *machine learning* y datos del clima, la mayoría aún son proyectos en etapa inicial o pruebas académicas. Entidades como el IDEAM cuentan con sistemas de monitoreo y alerta, pero todavía no integran completamente modelos predictivos avanzados. Por eso, estos desarrollos aún necesitan fortalecerse, validarse mejor y adaptarse a las condiciones propias del país [8], [9], [10].

Este trabajo responde a esa necesidad, desarrollando un modelo predictivo basado en *machine learning* para hacer predicciones sobre la ocurrencia de posibles focos de incendios forestales en Colombia. A partir de la integración de datos históricos de incendios, variables meteorológicas rezagadas (como temperatura, precipitación y humedad), así como características geográficas y temporales, se construyó un modelo capaz de generar predicciones semanales a nivel municipal. La metodología siguió seis pasos clave: recolección y preparación de datos, integración y construcción del conjunto de datos, análisis exploratorio, desarrollo y validación interna del modelo, validación externa con datos nuevos y definición del protocolo de aplicación.

Entre los principales resultados destacó el desempeño del modelo *LightGBM*, que alcanzó un *F1-score* de 76%, mostrando una buena capacidad para anticipar incendios forestales en la semana siguiente. Adicionalmente, el análisis exploratorio permitió detectar tendencias temporales y espaciales clave, así como limitaciones en la cobertura de datos, especialmente en el suroriente del país.

Para facilitar la comprensión del estudio, el documento se organiza en siete capítulos. El capítulo 2 presenta los objetivos del estudio, mientras que el capítulo 3 expone el problema de investigación y la justificación de la propuesta. En el capítulo 4 se revisa el marco teórico y el estado del arte relacionados con incendios forestales y técnicas de modelado predictivo. El capítulo 5 detalla la metodología empleada y el capítulo 6 discute los

principales resultados obtenidos. Finalmente, el capítulo 7 presenta las conclusiones y recomendaciones para futuras investigaciones.

Los resultados obtenidos evidencian que el modelo es una herramienta útil para anticipar la ocurrencia de incendios forestales a partir de variables climáticas, y resaltan el potencial del *machine learning* como apoyo a la gestión ambiental y la prevención del riesgo. En este sentido, el trabajo no se limita al desarrollo de un modelo predictivo, sino que plantea una base técnica con potencial de aplicación en sistemas de apoyo a la toma de decisiones orientados a la gestión de incendios forestales.

Capítulo 2
OBJETIVOS

1.1 Objetivo general

Modelar la ocurrencia semanal de focos de incendios forestales en Colombia mediante técnicas de *machine learning*, utilizando datos históricos de incendios y variables climáticas.

1.2 Objetivos específicos

- Analizar la relación entre las condiciones climáticas, meteorológicas y geográficas, incluyendo sus rezagos temporales, y la ocurrencia semanal de focos de incendios forestales en los municipios de Colombia.
- Comparar diferentes algoritmos *de machine learning* para identificar el modelo que mejor aproxime la ocurrencia semanal de focos de incendios forestales, utilizando métricas de desempeño y pruebas estadísticas.
- Demostrar la capacidad de generalización temporal del modelo predictivo seleccionado, analizando su estabilidad frente a cambios en la distribución temporal de los datos.

Capítulo 3
PROBLEMA Y JUSTIFICACIÓN

Los incendios forestales se han convertido en una amenaza creciente para los ecosistemas, la biodiversidad y las comunidades en Colombia. En los últimos años, su frecuencia y gravedad han aumentado, en gran parte debido a factores como el cambio climático y la expansión de actividades humanas, especialmente la agricultura y la ganadería. Esta situación no solo pone en riesgo la riqueza natural del país, sino también el bienestar de miles de personas que dependen de los recursos forestales para su subsistencia [11].

A nivel global, la pérdida de cobertura forestal es alarmante. Según un estudio de la Universidad de Maryland, desde 2001 más de 6 millones de hectáreas se pierden cada año debido a incendios, una superficie similar al tamaño de Croacia [12]. Asimismo, el cambio climático ha intensificado esta problemática, ya que estudios han demostrado que las olas de calor extremo son más frecuentes y prolongadas, creando condiciones ideales para incendios forestales y agravando sus impactos sobre ecosistemas y comunidades [13]. Sin embargo, el problema es aún más profundo: los incendios no solo son una consecuencia del cambio climático, sino también un factor que lo agrava. Dado que, los incendios al liberar grandes cantidades de carbono a la atmósfera contribuyen al calentamiento global, cerrando un ciclo peligroso que amenaza con volverse incontrolable [14].

América Latina, y Colombia en particular, se encuentran entre las regiones más afectadas por esta situación, Esto se debe no solo a su riqueza forestal, sino también al impacto de actividades humanas descontroladas, que han facilitado la expansión del fuego, afectando especialmente a las comunidades rurales que dependen de la tierra y sus recursos para sobrevivir. Según la Organización Panamericana de la Salud, no solo consumen bosques y pastizales, sino que ponen en riesgo la salud de las personas, el agua y los alimentos que consumen y su derecho a vivir en un ambiente seguro y digno [15].

La propagación del fuego está influenciada por múltiples factores. Aunque el oxígeno, el combustible y la ignición son esenciales, otros elementos como la humedad del suelo, el

viento y la topografía también juegan un papel crucial [16]. En Colombia, más de la mitad del territorio está cubierto por bosques, lo que representa una fuente significativa de combustible para los incendios, por lo tanto, esta situación combinada con la presión de actividades humanas crea un ambiente propicio para la ocurrencia de incendios. En este sentido, las áreas protegidas y los parques naturales, que albergan una gran diversidad biológica, son especialmente vulnerables a estos eventos, lo que evidencia la urgencia de fortalecer las estrategias de prevención, monitoreo y control del fuego en el país.

Otro aspecto crítico es cómo los incendios afectan el ciclo del agua, ya que destruyen la vegetación que regula naturalmente su flujo, reduciendo la capacidad del suelo para retener humedad y aumentando la escorrentía y la evaporación. Esto altera los procesos hidrológicos, disminuyendo la infiltración y el agua almacenada bajo tierra, lo que a su vez reduce la disponibilidad de agua en los ecosistemas afectados. Además, la erosión y el arrastre de cenizas hacia ríos y lagos pueden deteriorar la calidad del agua, generando impactos adicionales tanto para la fauna y la flora como para las comunidades que dependen de estos recursos [17].

Ante este panorama, se han creado herramientas de detección satelital como el sistema *NASA FIRMS (Fire Information for Resource Management System*) [18], que permiten ubicar focos de calor casi en tiempo real y han representado un gran avance en la vigilancia de incendios. Sin embargo, su alcance es limitado, ya que suelen actuar cuando el fuego ya ha comenzado y, además, pueden requerir recursos tecnológicos y logísticos que no siempre están disponibles en todos los territorios, lo que dificulta respuestas rápidas y efectivas [19].

Por eso cobra relevancia la búsqueda de alternativas que permitan anticiparse a este tipo de situaciones. En este sentido, el *machine learning* ha permitido avanzar en la predicción de incendios forestales, ya que facilita el análisis de grandes volúmenes de datos ambientales, meteorológicos y socioeconómicos, e identifica patrones asociados a su

ocurrencia [20]. Gracias a esto, es posible entender mejor cómo se relacionan las variables explicativas y mejorar la estimación del riesgo frente a enfoques tradicionales.

En línea con lo anterior, estudios realizados en algunos países de Europa han mostrado resultados prometedores en la predicción de incendios forestales. Por ejemplo, un trabajo desarrollado en España [5], enfocado en la Comunidad Autónoma de Andalucía, logró niveles altos de precisión mediante el uso de datos históricos, variables meteorológicas y técnicas de *machine learning*. Sin embargo, estos enfoques se han desarrollado en contextos específicos, por lo que su efectividad depende de su adaptación a las condiciones locales. Factores como la topografía, la vegetación, el uso del suelo y el clima varían entre regiones, lo que limita su aplicación directa en el contexto colombiano.

A pesar de estos avances en monitoreo y detección temprana, la gestión de los incendios forestales sigue siendo principalmente reactiva, ya que la mayoría de las herramientas disponibles permiten identificarlos cuando ya han ocurrido. Por esta razón, surge la necesidad de estimar de manera anticipada la probabilidad de que estos incendios ocurran a partir de información climática, geográfica e histórica. Desde esta perspectiva, el problema de investigación consiste en modelar la relación entre estos factores y la ocurrencia futura de incendios.

Para ello, es necesario definir una representación del fenómeno acorde con la información disponible. Aunque los incendios forestales ocurren de forma continua en el espacio y el tiempo, los datos disponibles corresponden a observaciones discretas, por lo que es necesario establecer una unidad de análisis común. Por ello, en este trabajo el problema se formula a nivel municipal y semanal, considerando como variable objetivo la ocurrencia de al menos un foco de incendio durante la semana siguiente. Esta formulación permite abordarlo como una tarea de clasificación probabilística supervisada mediante técnicas de *machine learning*, con el fin de apoyar la toma de decisiones preventivas, optimizar la asignación de recursos para el control de incendios y reducir sus impactos ambientales y sociales.

Capítulo 4
MARCO TEÓRICO Y ESTADO DEL ARTE

4. 1 Marco teórico

El desarrollo de un modelo predictivo para incendios forestales requiere integrar conocimientos tanto del fenómeno en estudio como de las herramientas utilizadas para su modelación. Por esta razón, en esta sección se presentan los fundamentos conceptuales relacionados con los incendios forestales, sus factores asociados y su comportamiento, así como los principales conceptos *de machine learning* aplicados a problemas de predicción. Esta combinación permite establecer una base teórica que sustenta el enfoque metodológico del estudio y facilita la comprensión de los resultados obtenidos.

4.1.1 Ecosistemas forestales en Colombia

Los bosques son ecosistemas complejos y dinámicos, conformados por una amplia diversidad de especies que interactúan de manera constante. Estos sistemas son autosostenibles y evolucionan a lo largo del tiempo mediante relaciones ecológicas. En Colombia, los bosques cubren aproximadamente el 53% del territorio nacional e incluyen ecosistemas como bosques andinos, secos y húmedos tropicales, de galería y manglares [21], [22]. Estos cumplen funciones esenciales, como la conservación de la biodiversidad, la regulación del ciclo del agua y la mitigación del cambio climático, además de proveer recursos fundamentales para las comunidades.

4.1.2 Incendios forestales y factores asociados

Los incendios forestales son fuegos descontrolados que afectan extensas áreas de vegetación, generando impactos ambientales, sociales y económicos. Estos pueden originarse por causas naturales, como descargas eléctricas, o por actividades humanas [23], siendo estas últimas predominantes en contextos como el colombiano [24].

El análisis de estos eventos se centra en el concepto de foco de incendio, entendido como el punto donde se inicia o se detecta el fuego, es decir, la ubicación donde ocurre la primera

ignición del material combustible [25], el cual constituye la unidad de análisis en estudios predictivos. Su ocurrencia se explica mediante el triángulo del fuego, que establece que para que un incendio ocurra deben coincidir tres elementos: fuente de calor, combustible y oxígeno [26].

La ocurrencia y propagación del fuego dependen de múltiples factores. Entre los más relevantes se encuentran las variables climáticas, como la temperatura, la precipitación, la humedad relativa y el viento, que influyen en la sequedad del combustible [26]. A su vez, factores biofísicos como el tipo de vegetación y la cantidad de biomasa disponible determinan la disponibilidad de material combustible. Finalmente, las actividades humanas, como la expansión agrícola o la deforestación, incrementan significativamente la probabilidad de ignición.

Para efectos de este estudio, es importante diferenciar entre la ocurrencia, la detección y la propagación de un incendio forestal. La ocurrencia se refiere a la presencia del evento en un lugar y periodo determinados; la detección corresponde a su identificación mediante reportes, sensores o imágenes satelitales; y la propagación describe la forma en que el fuego se extiende una vez iniciado. Este trabajo se centra en la ocurrencia de incendios, entendido como la presencia o ausencia de al menos un foco de incendio en una unidad espacio-temporal definida.

Desde una perspectiva teórica, el estudio de los incendios forestales se sustenta en la comprensión de su comportamiento y su riesgo. El modelo de Rothermel establece que la propagación del fuego depende de la interacción entre combustible, condiciones meteorológicas y topografía, lo que constituye la base conceptual para la selección de variables en modelos predictivos [27].

Por su parte, la teoría del riesgo de incendios plantea que este fenómeno debe entenderse como una combinación entre la probabilidad de ocurrencia y las consecuencias potenciales

[28], lo que permite diferenciar entre modelos orientados a la predicción del evento y aquellos enfocados en su impacto.

4.1.3 Modelado predictivo de incendios forestales

Para los conceptos de *machine learning* se tomaron como referencia los libros [29], [30], [31] y [32].

- **Modelo de Clasificación de Machine Learning**

El modelado predictivo permite estimar la ocurrencia de eventos futuros a partir de información histórica. En el contexto de los incendios forestales, este problema puede formularse como una tarea de clasificación binaria, en la cual el objetivo es determinar si en una unidad espacio-temporal específica (por ejemplo, municipio–semana) ocurre o no un incendio.

Desde la perspectiva del *machine learning*, este enfoque corresponde a un problema de aprendizaje supervisado, en el que los modelos se entrenan a partir de datos etiquetados (es decir, hay una variable objetivo), aprendiendo una función que relaciona un conjunto de variables de entrada (features) con una variable objetivo (target). En este caso, la variable objetivo representa la ocurrencia del evento (1 = hubo incendio, 0 = no hubo incendio), mientras que las variables predictoras corresponden a factores climáticos, espaciales y temporales asociados a su ocurrencia.

El proceso de entrenamiento consiste en optimizar una función de pérdida que mide el error entre las predicciones del modelo y los valores reales, con el fin de mejorar su capacidad de generalización a nuevos datos. Esto es fundamental, ya que el objetivo no es solo ajustar el modelo a los datos históricos, sino lograr un buen desempeño en datos no observados previamente, evitando problemas como el sobreajuste (*overfitting*).

- **Modelos de Machine Learning**

Los modelos de *machine learning* utilizados en problemas de clasificación pueden pertenecer a diferentes familias, como modelos lineales, basados en árboles de decisión,

métodos de ensamble, etc. Estos últimos combinan múltiples modelos simples para mejorar el desempeño predictivo y la capacidad de generalización.

Entre los métodos de ensamble, los algoritmos de boosting, como Gradient Boosting y LightGBM, construyen modelos de manera secuencial, donde cada nuevo modelo se enfoca en corregir los errores del anterior. Este enfoque permite capturar relaciones no lineales y patrones complejos en los datos, lo que resulta especialmente útil en problemas ambientales como la predicción de incendios forestales, donde múltiples factores interactúan de manera simultánea.

Por otro lado, los métodos basados en Bootstrap (técnica de muestreo con reemplazo), como el bagging (Bootstrap Aggregating), generan múltiples subconjuntos de datos a partir de la muestra original mediante muestreo con reemplazo. Cada subconjunto se utiliza para entrenar un modelo independiente, y sus predicciones se combinan posteriormente, generalmente mediante votación o promedio. Este enfoque reduce la varianza del modelo y mejora su estabilidad, siendo especialmente efectivo en algoritmos como Random Forest. A diferencia del boosting, que busca reducir el sesgo enfocándose en los errores, el bagging se centra en disminuir la variabilidad de las predicciones, lo que contribuye a un mejor desempeño general del modelo.

La elección del modelo depende de su capacidad para adaptarse a la estructura de los datos, manejar relaciones no lineales y mantener un equilibrio entre sesgo y varianza, y esto se verá reflejado en las métricas de evaluación, que enunciaremos en la siguiente sección.

- **Métricas de Evaluación**

En los problemas de clasificación binaria, el desempeño de los modelos suele analizarse a partir de la matriz de confusión (ver ilustración 1), una herramienta que permite comparar las predicciones del modelo con los valores reales. Esta matriz organiza los resultados en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos,

proporcionando una base para el cálculo de métricas como la precisión, el *recall* y el *F1-score.*

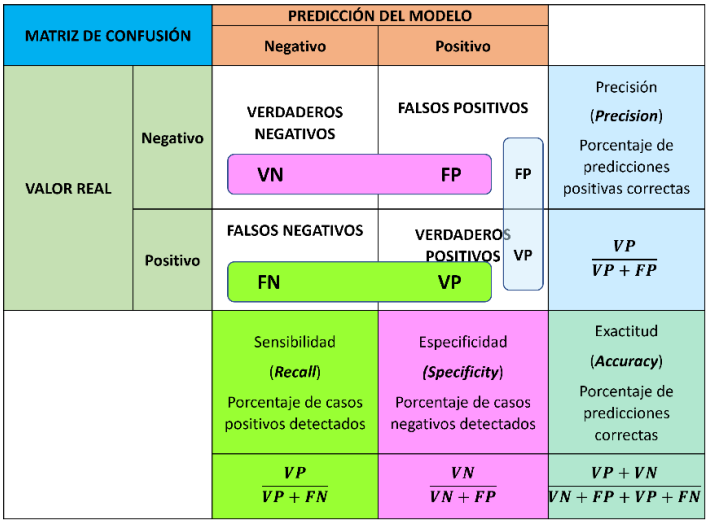

Ilustración 1 Métricas Matriz de Confusión

Fuente: https://github.com/Fabian830348/cursos/blob/master/Imagen/metricas.png

La evaluación del desempeño de los modelos de clasificación requiere el uso de métricas que permitan medir su capacidad para predecir correctamente la ocurrencia de eventos. En este contexto, no es suficiente utilizar únicamente la exactitud (*accuracy*), especialmente cuando existe desbalance entre clases.

La precisión (*precision*) mide la proporción de predicciones positivas correctas, mientras que la sensibilidad (*recall*) indica la capacidad del modelo para identificar correctamente los casos positivos. El F1-score combina ambas métricas en una sola medida, proporcionando un equilibrio entre precisión y *recall*, lo que resulta especialmente relevante en problemas donde es importante detectar eventos poco frecuentes, como los incendios forestales.

Estas métricas permiten evaluar el comportamiento del modelo desde diferentes perspectivas, facilitando la selección del algoritmo más adecuado para el problema planteado.

La curva ROC y el AUC

En problemas de predicción de incendios forestales, y en general en muchos escenarios reales, es común encontrar desbalance entre clases, donde la ocurrencia de incendios es menos frecuente que su ausencia. En estos casos, la evaluación del modelo no debe basarse únicamente en la exactitud (*accuracy*), sino en métricas más robustas como el *recall* y el F1-score, que permiten medir de forma más adecuada la capacidad del modelo para identificar correctamente los eventos de interés.

Adicionalmente, para comparar el desempeño de diferentes modelos, se utiliza la curva ROC (*Receiver Operating Characteristic*), una herramienta que evalúa el comportamiento de los modelos de clasificación binaria a distintos umbrales de decisión. Esta curva representa la relación entre la tasa de verdaderos positivos (*recall*) y la tasa de falsos positivos, permitiendo analizar la capacidad del modelo para discriminar entre clases (ver Ilustración 2).

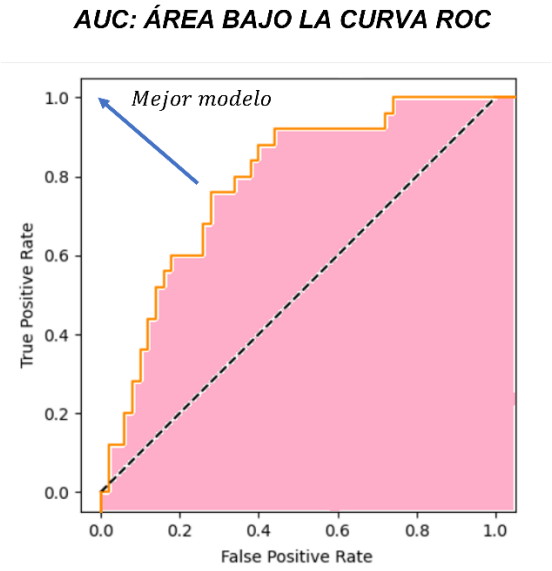

Ilustración 2. Curva ROC y AUV

Fuente: https://github.com/Fabian830348/cursos/blob/master/ROC/roc8.png

Como medida agregada de este comportamiento, se utiliza el área bajo la curva ROC (AUC), que resume en un solo valor la capacidad del modelo para diferenciar entre clases.

Valores cercanos a 1 indican una alta capacidad de discriminación, mientras que valores cercanos a 0.5 sugieren un desempeño similar al azar. Desde una perspectiva probabilística, puede interpretarse como la probabilidad de que el modelo asigne un mayor riesgo a un caso con incendio que a uno sin incendio. Como no depende de un umbral específico de clasificación, facilita la comparación entre modelos y es especialmente útil en problemas con clases desbalanceadas.

Finalmente, la interpretabilidad es un aspecto relevante en aplicaciones ambientales. El análisis de importancia de variables permite identificar cuáles factores tienen mayor influencia en la predicción, lo que no solo mejora la comprensión del modelo, sino que también aporta información útil sobre las condiciones que favorecen la ocurrencia de incendios forestales.

- **Test de McNemar para seleccionar el mejor modelo**

La prueba de **McNemar** es un método estadístico no paramétrico utilizado para comparar el desempeño de dos modelos de clasificación evaluados sobre el mismo conjunto de observaciones. Su objetivo es determinar si las diferencias observadas entre ambos modelos son estadísticamente significativas o si podrían atribuirse al azar. La prueba se basa en el análisis de los casos en los que los modelos discrepan en sus predicciones, construyendo una tabla de contingencia con los aciertos y errores de cada uno. En el contexto de *machine Learning*, el test de McNemar complementa métricas como Accuracy, Precision, Recall o F1-Score, ya que permite validar con rigor estadístico si un modelo realmente supera a otro. Si el valor *p* obtenido es menor al nivel de significancia establecido (generalmente 0,05), se concluye que existe evidencia suficiente para afirmar que el rendimiento de los modelos es diferente; de lo contrario, no se puede asegurar que uno sea superior al otro desde una perspectiva estadística.

La prueba de hipótesis que se plantea es:

$$\begin{cases} H_0 : & Ambos\ modelos\ tienen\ el\ mismo\ rendimiento \\ H_1 : & Existe\ diferencia\ significativa\ entre\ los\ modelos \end{cases}$$

- **El test de Kolmogorov Smirnov y el covariate shift**

La prueba de **Kolmogorov–Smirnov (KS)** es un método estadístico no paramétrico utilizado para comparar una distribución de datos observada con una distribución teórica de referencia, o para comparar dos distribuciones muestrales. En el contexto de análisis de datos y *machine Learning*, se emplea frecuentemente para evaluar si una variable sigue una distribución normal o para detectar cambios en la distribución de los datos entre diferentes conjuntos, como entrenamiento y prueba. La prueba calcula la máxima diferencia absoluta entre las funciones de distribución acumulada de los conjuntos comparados y, a partir de esta diferencia, determina si existen discrepancias estadísticamente significativas. Un valor de significancia (*p-value*) inferior al nivel establecido (generalmente 0,05) indica que las distribuciones son diferentes, mientras que un valor superior sugiere que no existe evidencia suficiente para rechazar la hipótesis de que ambas distribuciones son iguales. Debido a que no requiere supuestos sobre la forma de la distribución, el test KS es una herramienta ampliamente utilizada para la validación de datos y la detección de posibles fenómenos de *data drift* o *covariate shift* en modelos predictivos.

El **Covariate Shift** es un fenómeno que ocurre cuando la distribución de las variables predictoras (*features*) cambia entre el conjunto de entrenamiento y los datos utilizados posteriormente para validación o producción, mientras que la relación entre dichas variables y la variable objetivo permanece relativamente estable. En términos prácticos, esto significa que el modelo es entrenado con datos que presentan ciertas características, pero posteriormente debe realizar predicciones sobre una población con patrones diferentes. Este cambio puede afectar el rendimiento predictivo del modelo, ya que las observaciones futuras pueden no representar adecuadamente las condiciones aprendidas durante el entrenamiento. Por esta razón, en problemas con componente temporal es recomendable evaluar la presencia de *Covariate Shift* mediante pruebas estadísticas, comparación de distribuciones o análisis de *data drift*, con el fin de verificar la estabilidad

de los datos y garantizar que el modelo mantenga su capacidad de generalización a lo largo del tiempo.

4.1.4 Dimensión temporal y espacial en la predicción

Un aspecto especialmente relevante para esta investigación es la incorporación de dinámicas temporales en la modelación. La literatura reciente ha evidenciado que los incendios forestales no dependen únicamente de condiciones climáticas momentáneas, sino también de procesos acumulativos como sequías prolongadas, variaciones en la humedad del combustible y la persistencia de condiciones meteorológicas adversas, las cuales influyen directamente en la probabilidad de ignición y propagación del fuego [33], [34].

En este sentido, el presente estudio propone un enfoque que utiliza exclusivamente variables climáticas de semanas previas a la ocurrencia del incendio (hasta tres semanas de rezago), evitando el uso de información de la misma semana del incendio. Esta decisión metodológica busca prevenir la fuga de información (*data leakage*) y garantizar que el modelo opere bajo condiciones realistas de predicción.

Adicionalmente, aunque no se incorporan variables explícitas de uso del suelo o topografía, se incluyen variables como región natural y departamento, las cuales permiten capturar patrones espaciales latentes asociados a características ambientales no observadas directamente. Esta aproximación resulta especialmente relevante en contextos como el colombiano, donde la disponibilidad de datos es limitada y altamente heterogénea.

En resumen, este marco teórico permite entender que los incendios forestales son un fenómeno complejo que no depende de un solo factor, sino de la combinación de varias condiciones. La literatura muestra que se han logrado avances importantes en su estudio, especialmente con el uso de *machine learning* y la inclusión de factores espaciales y temporales para mejorar las predicciones.

Sin embargo, todavía existen dificultades, principalmente relacionadas con la calidad y disponibilidad de los datos, así como con la aplicación de estos modelos en diferentes contextos. En países como Colombia, estos retos son mayores debido a la diversidad del territorio y a las limitaciones en la información disponible.

En el modelamiento propuesto de *machine learning* se utilizaron las siguientes técnicas:

- **Validación de Origen Rodante (Rolling Origin Validation)**

La validación de origen rodante es una técnica de evaluación utilizada en series de tiempo que respeta el orden cronológico de los datos. Consiste en entrenar el modelo con información histórica y evaluar su desempeño en períodos futuros, desplazando progresivamente la ventana temporal. Lo anterior mantiene la secuencia temporal de los datos, evita la fuga de información y permite evaluar la estabilidad del modelo a lo largo del tiempo

- **Validación Temporal**

La validación temporal consiste en entrenar un modelo utilizando datos históricos y evaluarlo con datos posteriores en el tiempo. Es decir, tiene como objetivo determinar si el modelo mantiene su capacidad predictiva cuando enfrenta información nueva.

4.2 Estado del arte

- **Evolución de enfoques en la predicción de incendios**

Los incendios forestales se han convertido en un problema cada vez más relevante debido a sus impactos sobre los ecosistemas y las comunidades. En las últimas décadas, los enfoques para su análisis y predicción han evolucionado de manera importante, y han estado impulsados por el avance en la disponibilidad de datos y el desarrollo de nuevas herramientas analíticas.

Inicialmente, se desarrollaron índices empíricos orientados a estimar el riesgo de incendios a partir de variables climáticas y del estado del combustible vegetal. Entre los primeros antecedentes se encuentran los índices empíricos desarrollados a inicios del siglo XX, como el propuesto por Munger (1916), considerados como algunos de los primeros intentos por relacionar condiciones de sequía con el riesgo de incendios [8]. Posteriormente, surgieron índices más estructurados como el Keetch-Byram Drought Index (KBDI), que permite estimar el déficit de humedad del suelo y su relación con la probabilidad de ignición del combustible [35]. Aunque estos enfoques fueron útiles para comprender el fenómeno, presentaban limitaciones al simplificar la dinámica real de los incendios.

Posteriormente, se desarrollaron modelos físicos como el propuesto por Rothermel (1972), que explican la propagación del fuego a partir de la interacción entre combustible, condiciones meteorológicas y topografía [27]. Si bien estos modelos permitieron una mejor comprensión del comportamiento del fuego, su aplicación práctica resulta compleja debido a la gran cantidad de variables requeridas.

Más adelante, se incorporaron modelos estadísticos tradicionales, como la regresión logística, que permiten estimar la probabilidad de ocurrencia de incendios a partir de variables ambientales [36]. Sin embargo, estos enfoques suelen asumir relaciones lineales, lo que limita su capacidad para representar la complejidad del fenómeno.

En años más recientes, el uso de *machine learning* en la predicción de incendios forestales ha aumentado considerablemente, impulsado por la disponibilidad de datos y el avance en la capacidad de procesamiento. Como resultado, estas técnicas han transformado su análisis, permitiendo modelar relaciones no lineales y aprovechar grandes volúmenes de información. Este avance se refleja en el desarrollo de distintos algoritmos y aplicaciones recientes donde modelos como Random Forest, Support Vector Machines, redes neuronales y métodos de ensamble han demostrado un mejor desempeño en la predicción, especialmente cuando se utilizan grandes conjuntos de datos climáticos y ambientales [37].

- **Aplicaciones Internacionales**

A nivel internacional, hay varios estudios que han demostrado el potencial del *machine learning* en la predicción de incendios forestales. En Vietnam, el estudio "*Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction*" evaluó distintos algoritmos, encontrando que modelos como *Random Forest* alcanzan valores de AUC cercanos o superiores a 0.90, evidenciando su capacidad para capturar patrones complejos [6].

De manera similar, Choi et al. (2024), investigadores de la Universidad Nacional de Kangwon, en Corea del Sur, propusieron un modelo basado en el ensamble de múltiples algoritmos utilizando variables meteorológicas a nivel nacional [7]. Los resultados mostraron que la combinación de modelos mejora la precisión de las predicciones y reduce los errores en comparación con enfoques individuales

Asimismo, Mohajane et al. [38] (2021), en el estudio Application of Remote Sensing and Machine Learning Algorithms for Forest Fire Mapping in a Mediterranean Area, desarrollado en Marruecos, evidencian que la integración de imágenes satelitales con algoritmos de *machine learning* permite identificar con mayor precisión las áreas afectadas y generar mapas de riesgo más confiables. En esta misma línea, Yang et al. [38](2021), en el estudio Predicting Forest Fire Using Remote Sensing Data and Machine Learning, desarrollado en la Universidad Nacional de Singapur, utilizando datos de Indonesia, demostraron que el uso de datos satelitales y sensores remotos mejora la detección temprana de incendios forestales, especialmente cuando se combinan con variables climáticas históricas.

Jain et al. (2020) realizaron una revisión de 300 estudios sobre aplicaciones de *Machine Learning* en la ciencia y gestión de incendios forestales, evidenciando un crecimiento significativo de estas técnicas durante las últimas décadas. Los autores identificaron que algoritmos como Random Forest, Redes Neuronales Artificiales, Máquinas de Vectores de Soporte y MaxEnt son los más utilizados para tareas de predicción, detección y evaluación

del riesgo de incendios. La revisión destaca que estas metodologías permiten modelar relaciones complejas y no lineales entre variables climáticas, topográficas y de vegetación. Asimismo, se concluye que el aprendizaje automático ofrece un gran potencial para mejorar la toma de decisiones en la gestión del fuego. Sin embargo, los autores enfatizan que la calidad de los datos y el conocimiento experto del dominio siguen siendo factores fundamentales para garantizar resultados confiables y aplicables en escenarios reales [39]

En conjunto, estos trabajos evidencian que el uso de *machine learning* permite mejorar significativamente la predicción de incendios forestales en diferentes contextos geográficos.

- **Aplicaciones en Latinoamérica**

En América Latina, el interés por la aplicación de técnicas de *machine learning* en la predicción de incendios forestales ha aumentado en los últimos años, en respuesta al incremento de eventos extremos asociados al cambio climático.

En Brasil, Freitas et al. (2025), en el estudio *Prediction of Forest Fire Susceptibility Using Machine Learning Tools in the Triunfo do Xingu Environmental Protection Area, Amazon, Brazil*, desarrollaron un modelo basado en *Random Forest* para identificar zonas con riesgo de incendios en la Amazonía, integrando variables ambientales, topográficas y socioeconómicas. Los resultados evidenciaron un alto desempeño predictivo y resaltaron la importancia de factores como la precipitación, el uso del suelo y la cercanía a zonas habitadas [40].

Asimismo, Silveira et al. (2020), en el estudio *Drivers of Fire Anomalies in the Brazilian Amazon: Lessons Learned from the 2019 Fire Crisis*, mostraron que una proporción significativa de los incendios en la Amazonía está asociada a actividades humanas como la expansión agrícola y la deforestación, lo que destaca la necesidad de considerar factores antrópicos en los modelos predictivos [41].

En México, González Martínez et al. (2024), en el estudio *Revisión de antecedentes para la predicción de incendios forestales mediante IA*, presentado en el Congreso Estudiantil de Inteligencia Artificial Aplicada a la Ingeniería y Tecnología, analizaron distintos enfoques basados en inteligencia artificial para la predicción de incendios forestales, concluyendo que los modelos que integran variables climáticas, datos históricos y análisis espacial presentan un mejor desempeño en la identificación de zonas de riesgo [42].

Estos estudios coinciden en que el uso de *machine learning* mejora la capacidad de predicción, especialmente cuando se dispone de información adecuada y se integran múltiples fuentes de datos.

Antecedentes en Colombia

En Colombia, la investigación en predicción de incendios forestales ha avanzado de forma progresiva en los últimos años. Ocampo-Zuleta y Beltrán-Vargas (2018) desarrollaron un modelo dinámico para analizar la ocurrencia de incendios en los Cerros Orientales de Bogotá, identificando una relación entre incendios, temperaturas elevadas y periodos de sequía [43].

Por otro lado, Barreto y Armenteras (2020) aplicaron el algoritmo *Random Forest* para modelar la probabilidad de incendios en los Llanos colombo-venezolanos, utilizando datos satelitales y variables ambientales. El modelo alcanzó una precisión del 94%, evidenciando el potencial del *machine learning* en contextos con limitada información meteorológica [44].

Más recientemente, Anzola, Fuentes y Rodríguez (2024) evaluaron modelos de *machine learning* para estimar el riesgo de incendios en Colombia, encontrando que algoritmos como Random Forest presentan un mejor desempeño en comparación con otros enfoques tradicionales [45]. Aunque estos estudios muestran avances importantes, la mayoría se han desarrollado en regiones específicas o con cobertura limitada, lo que restringe su aplicabilidad a nivel nacional.

En conjunto, los estudios revisados evidencian que el uso de técnicas de *machine learning* ha fortalecido la predicción de incendios forestales en distintos contextos. Sin embargo, persisten retos importantes, especialmente en países como Colombia, donde la disponibilidad de datos, la cobertura territorial y la validación en escenarios reales limitan la aplicación de estos modelos.

En este contexto, el presente trabajo propone un modelo de predicción semanal de la ocurrencia de focos de incendios forestales a nivel municipal en Colombia, utilizando variables climáticas rezagadas y validación temporal. Este enfoque busca aportar una herramienta con potencial de aplicación en la gestión del riesgo, adaptada a las condiciones del país y orientada a la anticipación de eventos.

Capítulo 5
METOLOGÍA

Este trabajo se enfocó en el desarrollo de un modelo predictivo basado en técnicas de *machine learning* para anticipar la ocurrencia semanal de los focos de incendios forestales en diferentes regiones de Colombia. A diferencia de otros estudios citados en el estado del arte, que analizan los incendios después de que ocurren o se centran en responder a ellos, esta propuesta busca anticiparse a su aparición y apoyar decisiones oportunas antes de que se presenten. Para ello, el modelo integró datos históricos de incendios y variables climáticas entre 2010 y 2023, lo que permitió identificar cambios ambientales significativos en los últimos años y generar predicciones que contribuyen a la gestión estratégica del riesgo. La metodología se estructuró en seis etapas principales: (1) recolección y preparación de datos, (2) integración de bases y construcción del conjunto de datos para modelamiento, (3) análisis exploratorio de datos, (4) modelado predictivo y validación interna, (5) validación externa con datos no utilizados en el entrenamiento y (6) definición del protocolo de aplicación del modelo.

1. Recolección y preparación de datos

Para el desarrollo del modelo predictivo, se recopilaron datos provenientes de distintas fuentes oficiales, incluyendo registros históricos de incendios forestales, información climática y datos geográficos a nivel municipal en Colombia.

Datos Históricos de Incendios: Se recopilaron registros históricos de incendios forestales en Colombia provenientes de la Unidad Nacional para la Gestión del Riesgo de Desastres (UNGRD) y del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Estas bases incluyen información relacionada con la ubicación del incendio (municipio y departamento), la fecha de ocurrencia, la extensión afectada por el incendio y las posibles causas reportadas.

Datos Climáticos: Se utilizaron registros climáticos históricos proporcionados por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), incluyendo las variables: temperatura del aire a 2 m de altura (°C), humedad relativa del aire a 2 m (%), precipitación acumulada (mm) y velocidad media del viento en intervalos de 10 minutos (m/s). Adicionalmente, las bases de datos contenían información de identificación y localización de cada estación de medición, como código de estación, municipio, departamento y coordenadas geográficas (latitud y longitud), lo que permitió asociar las variables climáticas a su contexto espacial.

Información Geográfica de Apoyo: Para organizar la información a nivel municipal, se utilizó la base DIVIPOLA-Códigos de municipios, disponible en la plataforma Datos Abiertos Colombia, la cual incluye el código de municipio (identificador oficial único), departamento, municipio y sus coordenadas geográficas (latitud y longitud). Esta base resultó necesaria, ya que la información de incendios no contaba originalmente con datos de ubicación geográfica. Mediante la llave territorial compuesta por departamento y municipio, fue posible asociar cada registro de incendio con sus coordenadas geográficas, lo que permitió incorporar la dimensión espacial al análisis y asegurar coherencia y precisión en la información utilizada.

Con el fin de garantizar la coherencia temporal, espacial y estadística de la información utilizada en el modelo predictivo, se realizó un proceso de preparación y transformación de datos. Esta etapa incluyó el tratamiento de las bases climáticas y de incendios forestales, la estandarización de identificadores geográficos y el ajuste de la escala temporal de análisis, permitiendo integrar fuentes diferentes en una estructura común a nivel municipio–semana.

Procesamiento y agregación de variables climáticas

Debido al gran volumen de información meteorológica del IDEAM, el procesamiento se realizó de forma escalonada para manejar adecuadamente las series históricas. En esta

etapa se llevó a cabo la limpieza de los datos, eliminando registros incompletos o inválidos y estandarizando la variable de fecha y hora.

Posteriormente, se generaron variables temporales como el Año ISO y la Semana ISO, utilizadas como referencia común para organizar la información en periodos semanales. Este sistema, basado en el estándar ISO-8601, define semanas completas de lunes a domingo y evita discontinuidades entre años, lo que permite mantener series temporales consistentes y comparables. Su uso facilitó alinear correctamente las variables climáticas con los registros de incendios, asegurando que ambas correspondieran a los mismos periodos de análisis.

A partir de esto, los datos diarios se agregaron a nivel semanal por estación meteorológica, calculando promedios, mínimos y máximos de variables como temperatura, humedad, viento y precipitación, junto con el número de observaciones por periodo.

Finalmente, la información se consolidó en una sola base, ajustando los promedios según la cantidad de datos disponibles, integrando la ubicación de cada estación y verificando que no existieran duplicados por estación, año y semana. Las variables utilizadas en el modelo se agrupan en componentes climáticos, temporales y estructurales, como se presenta a continuación.

Estandarización e integración espacial

Las bases de datos climáticas y de incendios presentaban inconsistencias en los nombres de municipios y departamentos, como diferencias ortográficas, uso de tildes y formatos de escritura. Para corregir esto, se realizó un proceso de normalización textual que incluyó la conversión a minúsculas, eliminación de tildes y caracteres no alfabéticos, así como la depuración de espacios y formatos.

Posteriormente, se utilizó un diccionario de equivalencias para unificar distintas formas de nombrar una misma entidad territorial y se trataron valores no válidos como datos faltantes, eliminando aquellos que no podían corregirse.

Con la información estandarizada, se construyó una llave territorial (departamento–municipio) que permitió integrar los datos con la base oficial DIVIPOLA, incorporando coordenadas geográficas a los registros de incendios. Este proceso fue clave para garantizar la coherencia espacial del análisis y evitar errores en la asociación con las variables climáticas

Preparación temporal y estructuración de la base de incendios

La base histórica de incendios presentaba inconsistencias en el registro de fechas, ya que la información estaba separada en columnas (día, mes en texto y año) y con formatos variados. Para corregirlo, se reconstruyó la fecha completa mediante la conversión de los meses a formato numérico, la corrección de años con dos dígitos y la validación de registros inválidos o incompletos. Las fechas resultantes se estandarizaron y se eliminaron aquellos registros con información temporal no confiable.

A partir de estas fechas, se generaron variables temporales como Año ISO, Semana ISO, rangos semanales (inicio y fin) y mes, lo que permitió representar la información a nivel semanal y hacerla compatible con las variables climáticas.

Finalmente, se eliminaron duplicados y se agregaron los datos por municipio y semana, calculando el área total afectada, el número de incendios y las variables temporales correspondientes, junto con las coordenadas geográficas. El resultado fue una base consolidada a nivel municipio–semana, adecuada para el modelado y alineada con la estructura de los datos climáticos.

2. Integración de bases y construcción del conjunto de datos de entrenamiento

Tras finalizar los procesos de depuración, estandarización y agregación temporal de las bases climáticas y de incendios, se integraron para construir un conjunto de datos unificado a escala semanal. Dado el enfoque predictivo del estudio, las variables climáticas fueron consideradas como condiciones previas a la ocurrencia de incendios, evitando el uso de información futura que pudiera generar sesgos.

Las variables de precipitación, humedad y temperatura, previamente agregadas a nivel estación–semana, se integraron mediante uniones internas utilizando como claves el código de estación, el Año ISO y la Semana ISO, garantizando la consistencia de la información climática. Posteriormente, los datos fueron agregados a nivel municipio–semana, calculando promedios de las variables y un indicador del número de estaciones disponibles, como medida de cobertura.

La velocidad del viento, debido a su baja cobertura espacial y temporal, requirió imputación espacial mediante la asignación de valores de la estación más cercana. No obstante, tras pruebas preliminares, se evidenció que esta variable no mejoraba el desempeño del modelo e incluso introducía ruido, por lo que se decidió excluirla del conjunto final, aunque su tratamiento se documenta como parte del proceso metodológico.

Integración Integración espacio–temporal con la base de incendios

La integración entre las bases climáticas y los registros de incendios se diseñó buscando coherencia tanto temporal como espacial, teniendo en cuenta que provienen de sistemas distintos y con diferentes niveles de detalle.

Para la alineación temporal, se utilizaron las variables Año ISO y Semana ISO, lo que permitió organizar la información en periodos semanales consistentes (de lunes a domingo) y evitar inconsistencias entre años. Dado el enfoque predictivo del estudio, se evaluaron distintos rezagos temporales (1, 2 y 3 semanas) para representar las condiciones climáticas previas a la ocurrencia de incendios. Como resultado, se encontró que el rezago de hasta

tres semanas (t–3) ofrecía el mejor desempeño en las métricas del modelo, lo que sugiere que los incendios no responden únicamente a condiciones inmediatas, sino a procesos acumulativos en el tiempo.

En cuanto a la relación espacial, los incendios fueron georreferenciados a partir del centroide del municipio utilizando la base oficial DIVIPOLA, mientras que la información climática proviene de estaciones meteorológicas puntuales. Para vincular ambas fuentes, se probaron diferentes radios de influencia (10, 15, 20, 50, 80 y 100 km), encontrando que un radio de 100 km ofrecía el mejor desempeño predictivo. Esta decisión se basa en que los factores que favorecen la ocurrencia de incendios no se limitan a un punto específico, sino que suelen manifestarse en patrones regionales. Por ello, considerar un área de influencia más amplia permite reflejar de manera más realista las condiciones climáticas y ambientales que pueden contribuir a la generación de incendios [38].

La integración se realizó mediante un esquema de ponderación por distancia inversa (IDW), en el que cada incendio aporta a las unidades climáticas cercanas según su proximidad. A partir de este procedimiento se construyeron variables como el número de incendios cercanos, el área afectada y la distancia mínima al incendio más cercano, evitando registros duplicados y representando de manera más real cómo los incendios influyen en su entorno.

Hay que tener en cuenta que, usando este método de proximidad, algunos incendios que ocurren cerca de los límites de un municipio o departamento pueden contarse como si estuvieran en la zona vecina. Esto no es un error, sino una decisión metodológica para reflejar de manera más realista las condiciones ambientales, sin depender estrictamente de las fronteras administrativas.

Para representar la naturaleza cíclica de la variable SemanaDelAño, se aplicó una transformación armónica mediante las funciones seno y coseno: $Semana\_\sin = \sin(2\pi \cdot t/53)$ y $Semana\_\cos = \cos(2\pi \cdot t/53)$, donde t es la semana del año y 53

corresponde al número de semanas del calendario anual. Esta transformación proyecta cada observación sobre la circunferencia unitaria en $\mathbb{R}^2$, de manera que la distancia euclidiana refleja mejor la cercanía entre semanas en el ciclo temporal. Así, semanas próximas, incluyendo el cambio de año, permanecen cercanas en esta representación, evitando los saltos que genera una codificación lineal y permitiendo capturar mejor los patrones estacionales.

Por último, se construyó una variable objetivo binaria (hubo_incendio), que indica la presencia o ausencia de incendios en cada unidad municipio–semana. Es importante aclarar que las variables derivadas del proceso de integración espacial no se utilizaron como predictores dentro del modelo, con el fin de evitar fuga de información y mantener la coherencia del enfoque predictivo. A continuación, se describen las variables empleadas en el modelo, organizadas según su naturaleza climática, temporal y estructural.

Tipo de variable	**Variables**	**Descripción**
Variable objetivo	hubo_incendio	Indica la ocurrencia de incendio (1 = hubo incendio, 0 = no hubo incendio)
Estacionalidad (temporal)	Semana_sin, Semana_cos	Transformación cíclica de la semana del año para capturar patrones estacionales
Clima (rezagos)	Promedio_temperatura_lag1, Promedio_precipitacion_lag2, Promedio_precipitacion_lag3, Minimo_precipitacion_lag1, Minimo_precipitacion_lag2, Minimo_precipitacion_lag3,	Variables climáticas con rezagos temporales (1 a 3 semanas) que permiten capturar efectos acumulativos

	Maximo_precipitacion_lag1, Maximo_precipitacion_lag2, Maximo_precipitacion_lag3, Suma_total_precipitacion_lag1, Suma_total_precipitacion_lag2, Suma_total_precipitacion_lag3, Promedio_humedad_lag1, Promedio_humedad_lag2, Promedio_humedad_lag3, Minimo_humedad_lag1, Minimo_humedad_lag2, Minimo_humedad_lag3, Maximo_humedad_lag1, Maximo_humedad_lag2 Maximo_humedad_lag3, Minimo_temperatura_lag1, Minimo_temperatura_lag2, Minimo_temperatura_lag3, Maximo_temperatura_lag1, Maximo_temperatura_lag2, Maximo_temperatura_lag3	y condiciones previas a la ocurrencia de incendios
Variables categóricas	Departamento, Region, q_temp, q_hum, q_estaciones	Variables geográficas y de clasificación ambiental que permiten capturar patrones espaciales y heterogeneidad territorial

Tabla 1. Descripción de variables del modelo

3. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA) tuvo como objetivo comprender la estructura, distribución y comportamiento de las variables presentes en el conjunto de datos, así como identificar los patrones relevantes para el modelado predictivo de incendios forestales.

En primer lugar, se analizó la distribución espacial y temporal de los incendios forestales a nivel nacional. Luego se examinó la cobertura y variabilidad de las variables climáticas antes de la integración. Después, se evaluó cómo el proceso de integración afectó la estructura de los datos, asegurando consistencia y minimizando pérdida de información. Finalmente, se analizaron patrones de estacionalidad comparando los incendios con la precipitación, humedad y temperatura con un rezago de una semana, para identificar relaciones que respalden su uso en el modelo predictivo.

Este proceso permitió verificar la consistencia estadística del conjunto de datos resultante y entregó evidencia empírica sobre la influencia de las condiciones climáticas en la ocurrencia de incendios forestales.

Caracterización inicial de los incendios forestales

Como punto de partida del análisis exploratorio, se evaluó el comportamiento temporal del número de incendios forestales registrados en Colombia durante el periodo de estudio (2010-2023). Este análisis permitió identificar patrones interanuales del fenómeno y establecer una referencia inicial para evaluar posteriormente el impacto del proceso de integración de las variables climáticas.

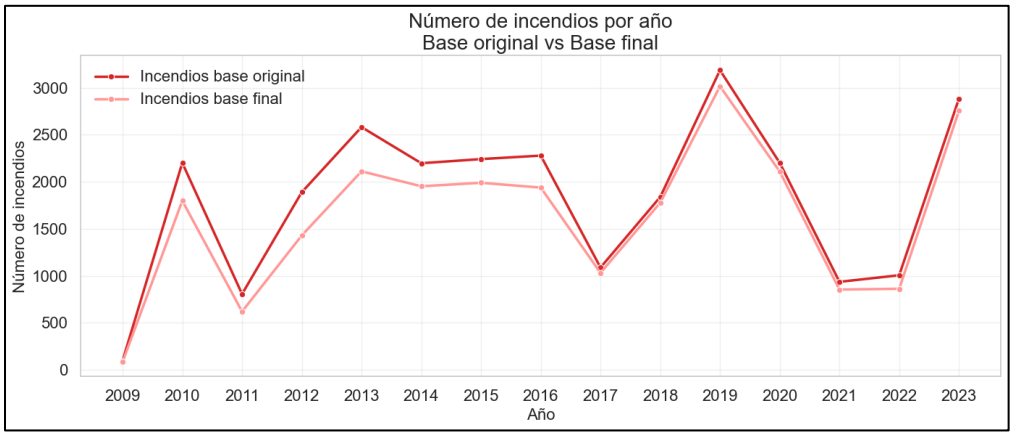

Ilustración 3. Comparativa de número de incendios por año

La ilustración 3 presenta el número total de incendios agregados por año, comparando la base original de incendios con la base final utilizada para el modelado, posterior a la integración con información climática.

En el análisis temporal se observan picos en 2010, 2013, 2016, 2019 y 2023. Algunos de estos años coinciden con periodos en los que Colombia experimentó condiciones más secas de lo habitual. Por ejemplo, en 2010 y 2016 el país estuvo bajo la influencia del fenómeno El Niño, lo que implicó menos lluvias y temperaturas más altas, condiciones que favorecieron la ocurrencia de incendios. De manera similar, en 2019 se registró un aumento de incendios en la región amazónica, y en 2023 se presentaron nuevamente condiciones secas asociadas al fenómeno El Niño.

En contraste, años como 2011 muestran niveles más bajos, lo cual coincide con el fenómeno La Niña, caracterizado por lluvias intensas que reducen la probabilidad de incendios. Para otros años, la variación parece responder a una combinación de factores climáticos y humanos. Aunque la base final tiene menos registros, debido a que, ahora solo se consideran aquellos con información climática disponible, el comportamiento general de los incendios se mantiene. Esto sugiere que, aunque se reduce la cantidad de datos, no se pierde la forma en que el fenómeno evoluciona en el tiempo. En otras palabras, la base

sigue representando bien la dinámica de los incendios forestales, lo que permite continuar con confianza hacia el análisis del impacto del proceso de integración de datos.

Tras analizar el comportamiento temporal de los incendios, se examinó su distribución espacial a nivel departamental. La ilustración 4 presenta los quince departamentos con mayor número de incendios, comparando los registros de la base original con los obtenidos después del proceso de integración de datos.

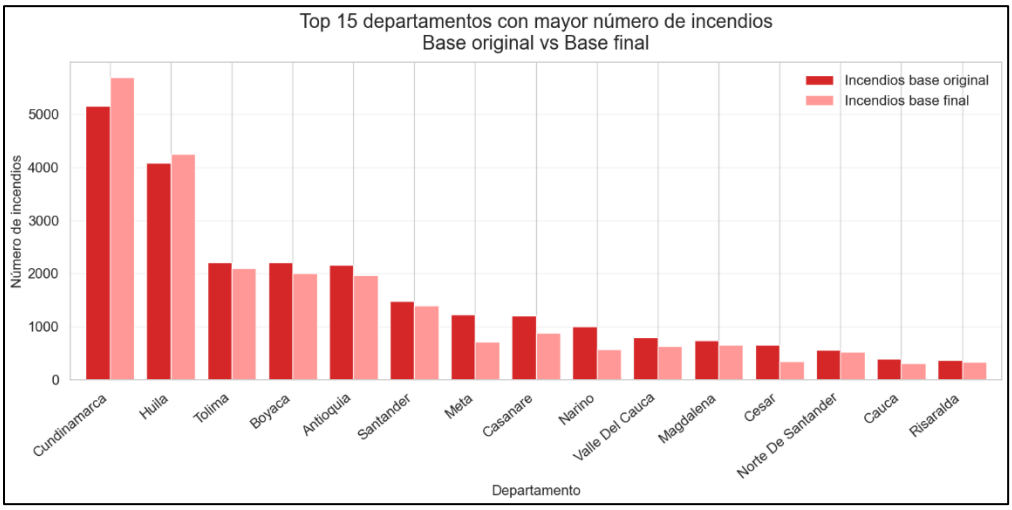

Ilustración 4. Comparativa de incendios por departamento

Los resultados muestran que Cundinamarca y Huila concentran el mayor número de incendios en ambos conjuntos de datos, seguidos por Tolima, Boyacá y Antioquia. En general, la distribución entre departamentos se mantiene similar, ya que las regiones con más incendios siguen siendo las mismas.

Sin embargo, es importante considerar que, debido a la forma en la que se integraron los datos, algunos incendios cercanos a límites departamentales fueron asociados al municipio más cercano con información climática, incluso si pertenecía a otro departamento. Esto puede generar pequeñas variaciones en los conteos, sin que represente cambios reales en la ocurrencia del fenómeno.

Además, estos departamentos comparten características en común, hacen parte de la región Andina, concentran gran parte de la población colombiana y de las actividades agropecuarias, y cuentan con mayor cobertura de estaciones meteorológicas. Todo esto favorece tanto a la ocurrencia como al registro de incendios. En conjunto, la base final conserva de forma adecuada la distribución espacial del fenómeno, lo que permite continuar con el análisis del impacto territorial.

En línea con lo anterior, la ilustración 5 muestra los 15 departamentos con mayor área quemada, comparando la base original con la base final tras la integración. Llama la atención que departamentos de la Orinoquía, como Vichada, Casanare, Meta y Arauca, tienen las mayores superficies afectadas. Sin embargo, en la base final se observa una gran reducción de las hectáreas quemadas en estas zonas, debido principalmente a la falta de estaciones climáticas en esta región.

A diferencia del área quemada, el número de incendios se mantiene muy similar entre ambas bases, lo que indica que la mayoría de los incendios sí fueron integrados. Sin embargo, la reducción en el área quemada sugiere que algunos de los incendios que no se integraron eran de gran extensión, lo que genera una subestimación del total de hectáreas en algunas regiones.

Aun así, como este estudio se centra en la ocurrencia de incendios, la estabilidad en el número de incendios permite mantener el patrón general entre departamentos y entender su distribución en el territorio.

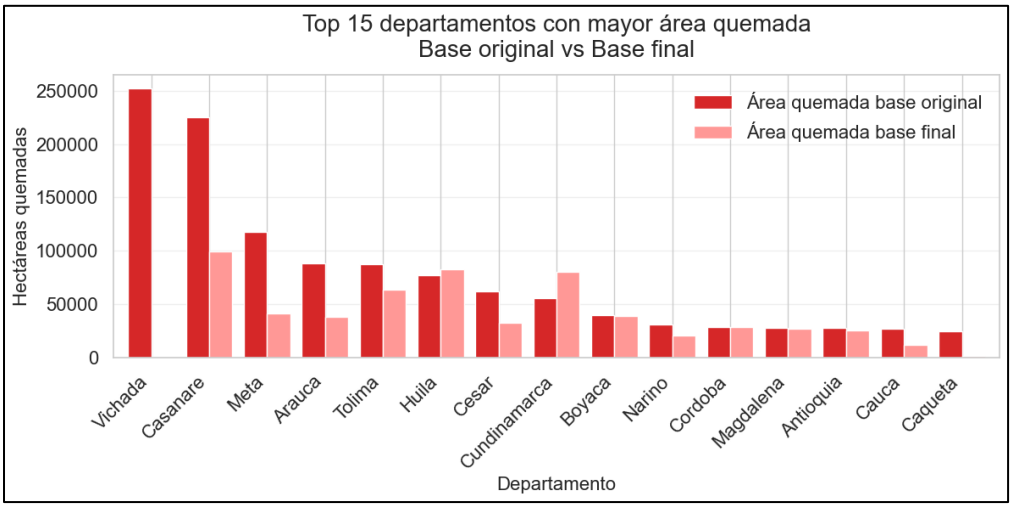

Ilustración 5. Comparativa de área quemada por departamento

Distribución de variables climáticas antes del proceso de integración

Con el fin de comprender el comportamiento climático previo a la integración de las bases de datos, se analizó la distribución de las principales variables meteorológicas registradas por las estaciones: temperatura, humedad relativa y precipitación. Estas variables fueron examinadas según las regiones naturales de Colombia (Caribe, Andina, Pacífica, Orinoquía, Amazonía e Insular) con el objetivo de identificar diferencias climáticas estructurales entre regiones antes del proceso de unión con la base de incendios forestales.

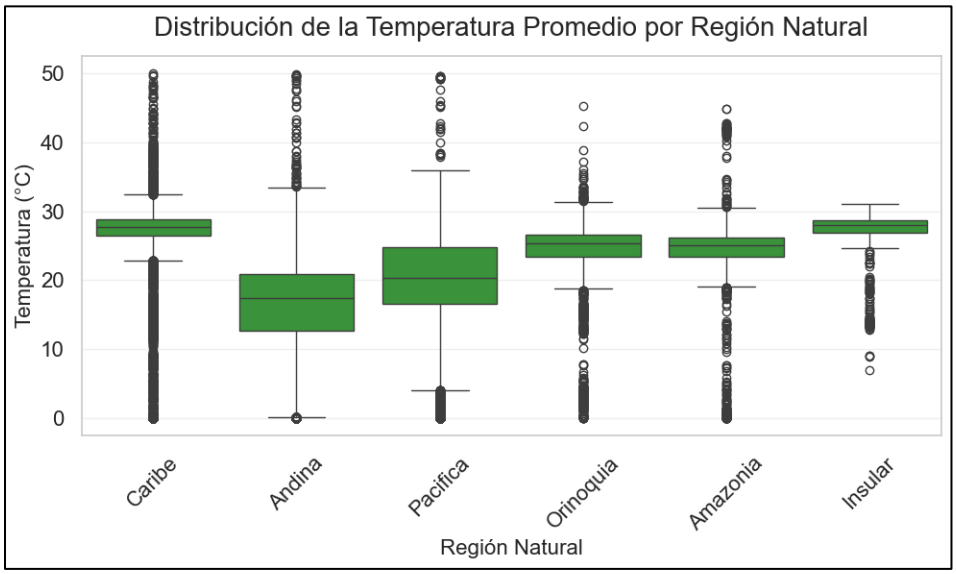

Ilustración 6. Distribución de temperatura por región natural

La Ilustración 6 muestra las diferencias en la temperatura entre las diferentes regiones del país. Las regiones Caribe e Insular presentaron las temperaturas más altas y estables a lo largo del periodo. De forma similar, las regiones de la Amazonía y la Orinoquía también registraron en promedio temperaturas altas, aunque ligeramente inferiores a las anteriores. En el caso de la Orinoquía, se observó además una mayor variación en los valores.

Por su parte, la región Pacífica mostró temperaturas intermedias con cierta dispersión, lo que indica cambios moderados en el tiempo. A diferencia de, la región Andina que presentó las temperaturas más bajas y la mayor variabilidad, reflejando la diversidad de condiciones dentro de la región.

En resumen, la mayoría de las regiones presentaron temperaturas altas, diferenciándose principalmente en qué tan estables o variables fueron, siendo la Región Andina la más variable frente a otras regiones más estables como Caribe e Insular.

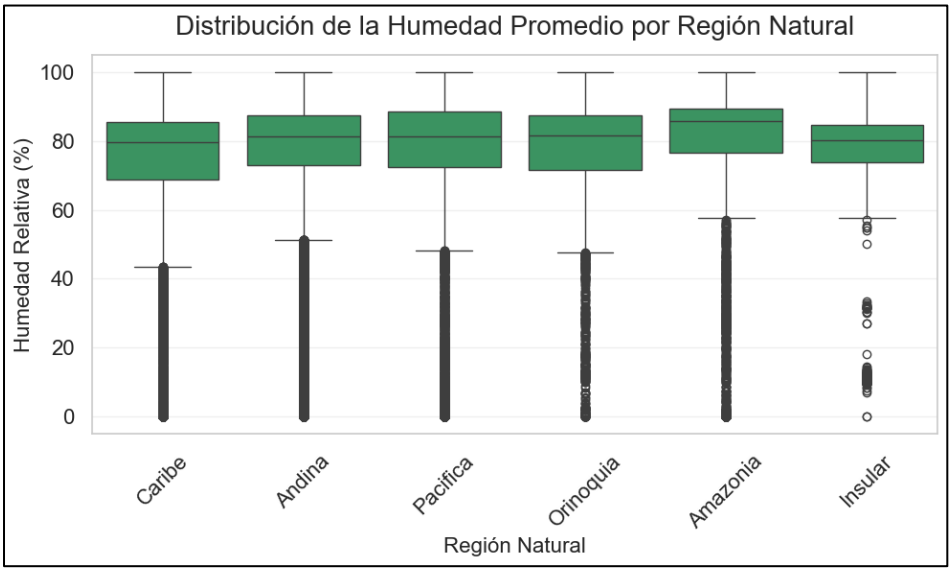

Ilustración 7. Distribución de humedad por región natural

En cuanto a la humedad relativa, la Ilustración 7 mostró valores altos en todas las regiones durante el periodo analizado. Sin embargo, las diferencias se observaron en la estabilidad de estos valores.

La Región Pacífica y la Región Amazonía presentaron los valores más altos y estables, lo que está directamente relacionado con sus altos niveles de precipitación y cobertura vegetal. La Región Andina también registró niveles elevados, pero con mayor variación, lo que refleja la diversidad climática dentro de la región.

Por su parte, la Región Orinoquía mostró una mayor dispersión y valores mínimos más bajos, lo que indica la presencia de periodos secos más marcados. Este patrón también se observó, aunque en menor medida, en la Región Caribe y la Región Insular.
En conjunto, aunque la humedad fue alta en todo el país, las diferencias en su estabilidad son clave, ya que las caídas en algunas regiones pueden generar condiciones más favorables para la ocurrencia de incendios.

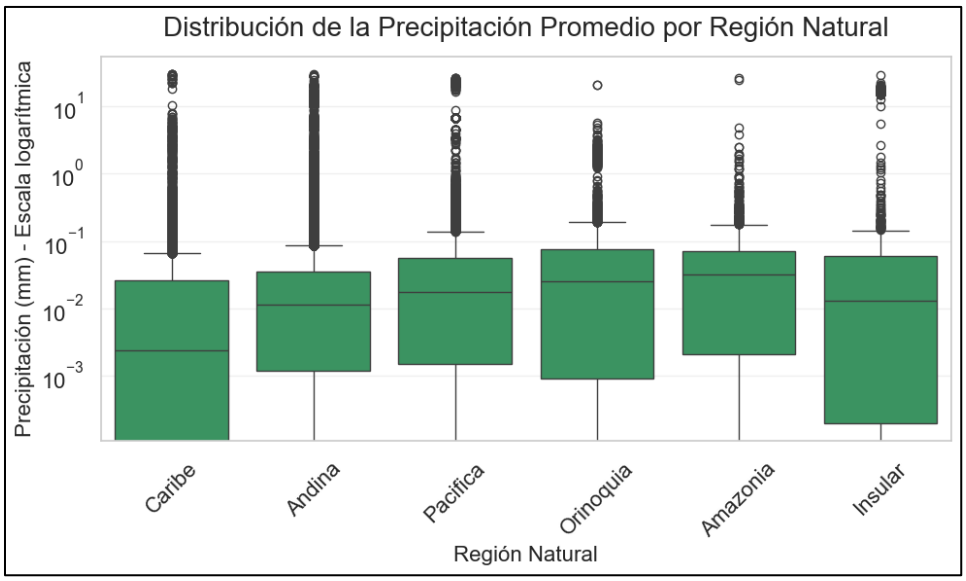

Ilustración 8. Distribución de precipitación por región natural

Respecto a la precipitación, la Ilustración 8 mostró diferencias más marcadas entre las regiones en comparación con las otras variables. La Región Pacífica y la Región Amazonía registraron los valores más altos, lo que confirma que son las zonas con mayor nivel de lluvias en el país.

La Región Andina y la Región Orinoquía presentaron niveles intermedios, aunque con una alta variabilidad, lo que indica cambios importantes en la cantidad de lluvia a lo largo del tiempo. Por su parte, la Región Caribe y la Región Insular mostraron valores más bajos en comparación con las demás regiones, junto con una mayor presencia de periodos con poca precipitación.

En general, la precipitación es la variable que más diferencia a las regiones. Mientras algunas mantienen lluvias constantes, otras tienen periodos secos más marcados. Estos cambios son importantes, ya que cuando la lluvia disminuye durante ciertos periodos, aumentan las condiciones que pueden favorecer la ocurrencia de incendios.

Impacto estructural del proceso de integración

Durante la integración de las bases climáticas se evaluó incluir la variable velocidad del viento. Sin embargo, su baja disponibilidad reducía considerablemente el número de registros, pasando de 86.076 a 56.017, lo que implicaba una pérdida cercana al 35 % de la información.

Aunque era posible aplicar métodos de imputación, esto habría requerido estimar una gran cantidad de datos, lo que podría introducir sesgos. Por esta razón, se decidió excluir esta variable y priorizar el uso de registros completos para temperatura, precipitación y humedad relativa.

Finalmente, se evaluó el impacto de la integración de las variables climáticas comparando sus estadísticas descriptivas antes y después del proceso, con el fin de identificar posibles cambios en su distribución.

Variable	Media Antes	Media Después	Std Antes	Std Después	Cambio % Media
Temperatura	19.08	19.97	7.12	6.27	4.66
Precipitación	0.09	0.07	0.93	0.48	-22.22
Humedad	75.54	76.67	21.19	18.24	1.5

Table 2 Resumen estadístico antes y después de la integración

Como se observa en la Tabla 2, los cambios entre los valores antes y después de la integración son en general reducidos, lo que sugiere que la transformación aplicada conservó adecuadamente la estructura original de los datos, tal como también se aprecia en las gráficas de estacionalidad, donde no se evidencian alteraciones en los patrones temporales.

La temperatura presentó un leve aumento en su media (4,66 %), manteniendo un comportamiento estacional consistente. La humedad mostró una variación mínima (1,5 %), lo que indica estabilidad tanto en su nivel promedio como en su dinámica temporal. En contraste, la precipitación presentó una disminución más marcada en términos relativos (-22,22 %); sin embargo, esta variación debe interpretarse considerando la baja magnitud absoluta de la variable, por lo que no implica un cambio sustancial en su comportamiento general. Adicionalmente, la disminución de la desviación estándar en las tres variables sugiere una ligera reducción en la dispersión de los datos después del proceso de integración, lo cual se refleja en las gráficas como series más estables, sin afectar los patrones estacionales ni la coherencia general de las variables climáticas.

En conjunto, estos resultados muestran que la base integrada mantuvo la consistencia estadística de las variables climáticas, brindando confianza sobre la calidad de la información para las etapas posteriores de modelado. Adicionalmente, se analizó la pérdida estructural de datos por departamento, considerando solo las semanas en las que coincidían temperatura, humedad y precipitación. Como se puede ver en la Ilustración 9, en la mayoría de los departamentos la pérdida fue menor al 40 %, lo que refleja una consistencia moderada entre las fuentes climáticas utilizadas.

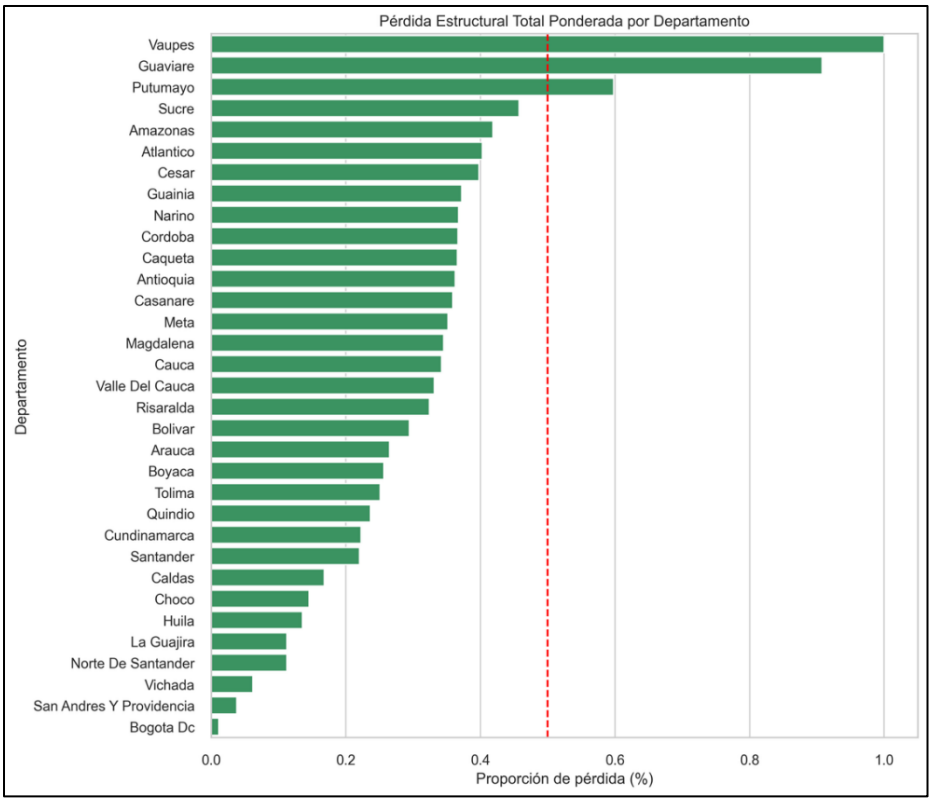

Ilustración 9. Perdida estructural ponderada por departamento

Como punto de referencia, se definió un umbral crítico del 50 %, ya que a partir de este punto se pierde al menos la mitad de la información original. Bajo este escenario, Vaupés, Guaviare, Putumayo y Guainía fueron los departamentos que concentraron la mayor pérdida de datos. Un aspecto común entre ellos es su territorio extenso, amplia cobertura de selva, poca población y una menor disponibilidad de estaciones meteorológicas, lo que dificulta contar con registros continuos y simultáneos de las variables climáticas.

En general, la integración mostró un buen resultado, aunque este hallazgo permite reconocer regiones donde la cobertura de datos sigue siendo más limitada y que deben interpretarse con mayor cuidado en el análisis espacial del modelo.

Patrones estacionales de incendios y efecto de rezagos climáticos

Como se puede ver en las Ilustraciones 8, 9 y 10, los incendios siguen un comportamiento mensual con una estacionalidad definida. Los valores más altos se concentran en enero, febrero, agosto y septiembre, mientras que los meses con menos incendios suelen ser abril, mayo, junio, octubre y noviembre. Este comportamiento permite reconocer dos momentos del año en los que los incendios aumentan de manera importante, lo cual coincide con temporadas secas en gran parte del país.

En la Ilustración 10, la temperatura con una semana de rezago presenta una relación positiva con los incendios, ya que los meses con mayores incendios suelen coincidir con temperaturas previas relativamente altas, especialmente al inicio y en la segunda mitad del año. Sin embargo, esta relación no es totalmente lineal, lo que indica que la temperatura por sí sola no explica el comportamiento de los incendios. Este efecto parece intensificarse en los periodos secos, cuando también disminuyen la humedad y la precipitación. Esto puede indicar que temperaturas elevadas en la semana previa ayudan a generar condiciones más favorables para la ocurrencia de incendios

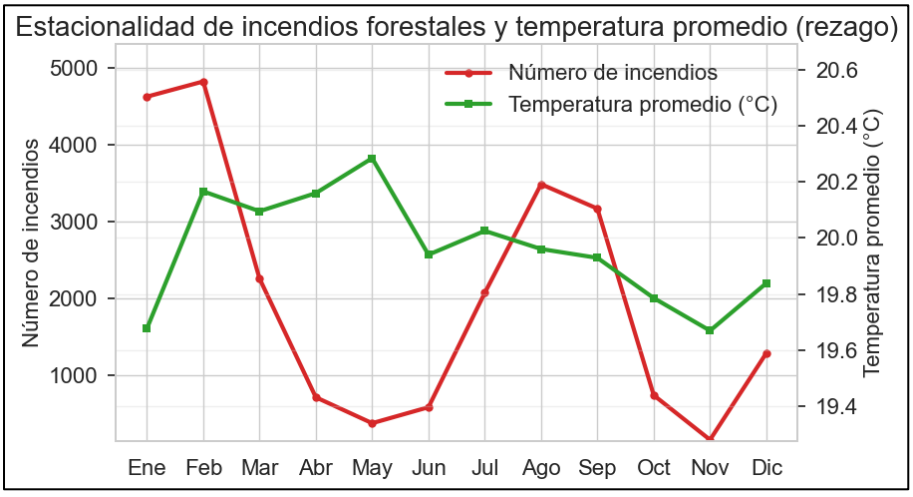

Ilustración 10. Relación mensual entre temperatura promedio (rezago) e incendios

Por su parte, la Ilustración 11 muestra que la precipitación tiene un comportamiento inverso frente a los incendios. Por ejemplo, en enero y febrero, cuando la lluvia de la semana anterior es menor, se registran los picos más altos de incendios. En cambio, en meses como abril, junio y noviembre, donde la precipitación aumenta, la cantidad de incendios disminuye de forma evidente. Esto sugiere que semanas previas con menos lluvia pueden incrementar la probabilidad de incendios.

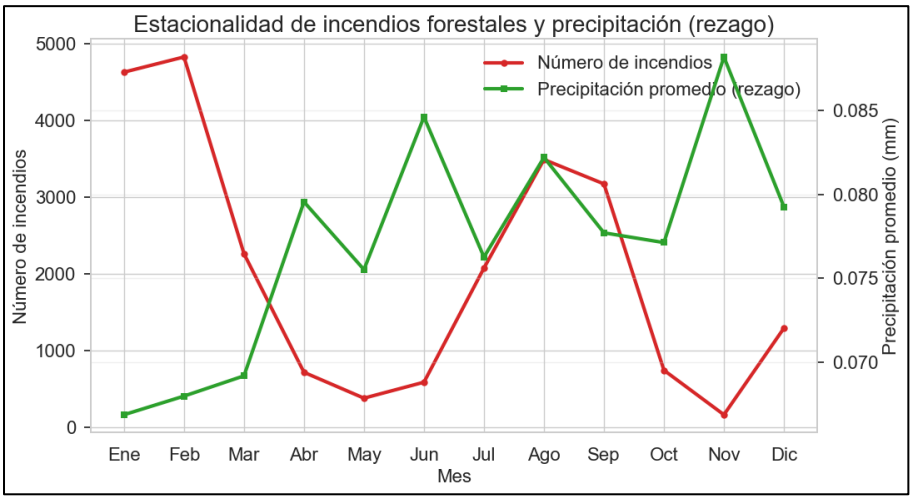

Ilustración 11. Relación mensual entre precipitación promedio (rezago) e incendios

Algo similar se observa en la Ilustración 12 con la humedad relativa, que también presenta una relación inversa con los incendios. Los mayores registros aparecen cuando la humedad de la semana anterior es más baja, mientras que al aumentar la humedad los incendios tienden a disminuir. Este comportamiento tiene sentido dentro de la dinámica del fuego, ya que niveles bajos de humedad favorecen el secado de la vegetación y hacen más probable la ocurrencia de incendios.

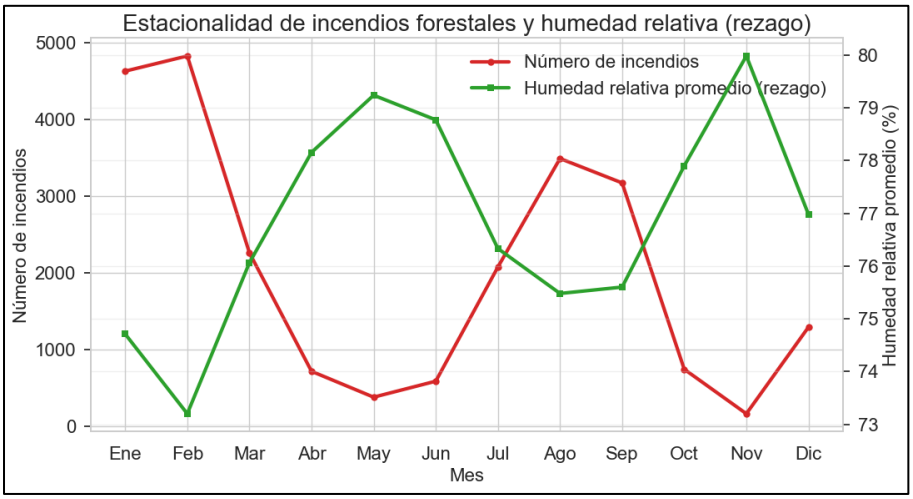

Ilustración 12. Relación mensual entre humedad promedio (rezago) e incendios

En conjunto, este análisis muestra que las variables climáticas de la semana previa aportan información valiosa para anticipar incendios. Esto respalda el uso de rezagos temporales dentro del modelo predictivo, ya que permiten capturar señales previas relevantes antes de que ocurra el incendio.

4. Modelado predictivo y validación interna

Para seleccionar los modelos de *machine learning* se utilizó PyCaret [46], una herramienta de código abierto en Python que permite manejar todo el flujo de modelado en un mismo entorno: desde la preparación de los datos, hasta el entrenamiento, la validación y la comparación de distintos algoritmos, incluyendo el ajuste de hiperparámetros. Esto facilitó probar varios modelos de manera ordenada y compararlos de forma objetiva para identificar cuál funcionaba mejor.

Además, PyCaret se basa en bibliotecas ampliamente reconocidas como *scikit-learn*, *LightGBM y XGBoost*, lo que permite evaluar distintos enfoques sin cambiar de entorno de trabajo. Este enfoque resulta muy útil para predecir incendios forestales, ya que variables climáticas, temporales y del municipio se combinan de manera compleja y no lineal. Por

esto, probar y comparar distintos modelos permitió identificar el más confiable para capturar estos patrones y generar predicciones precisas.

Variable objetivo del modelo

El modelo desarrollado en este estudio busca anticipar la ocurrencia de incendios forestales a partir de información climática e histórica. Para ello, se definió la variable hubo_incendio, que toma el valor 1 cuando se registra al menos un incendio en la semana y 0 cuando no ocurre ninguno.

Cada registro incluye cómo estuvo el clima en las semanas anteriores y si en la semana actual ocurrió o no un incendio. Para hacer las predicciones, se usó la información climática de hasta tres semanas antes, de manera que el modelo pueda identificar qué condiciones suelen aparecer antes de que se presente un incendio.

Este enfoque garantiza que el modelo solo use información disponible antes del incendio, evitando errores de fuga de información y permitiendo su uso en alertas tempranas. En resumen, el problema se plantea como una clasificación binaria, donde el modelo distingue entre contextos climáticos de mayor o menor probabilidad de incendios.

Balanceo del conjunto de datos

Antes de entrenar los modelos, se identificó un desbalance en la variable hubo_incendio, donde la clase positiva (1 = hubo incendio) contaba con 33,612 registros, mientras que la clase negativa (0 = no hubo incendio) tenía 52,464. Este desbalance puede sesgar el modelo hacia la clase mayoritaria y dificultar la detección de incendios. Para mitigarlo, se aplicó submuestreo aleatorio (*Random Undersampling)* sobre la clase negativa hasta igualarla con la positiva, obteniendo un conjunto balanceado de 67,224 registros (33,612 por clase) [47]. Este enfoque permite mejorar la capacidad del modelo para detectar eventos de interés, aunque implica la reducción de información disponible en la clase mayoritaria.

Adicionalmente, dado que se trabajó con variables climáticas, no se realizó imputación de valores faltantes, con el fin de evitar la incorporación de información artificial. Finalmente,

los datos fueron reorganizados de manera aleatoria para la construcción del conjunto de entrenamiento.

Configuración del entorno de modelado

El entrenamiento de los modelos se realizó con la función setup() de PyCaret, que permite configurar de forma unificada el proceso de preparación de datos y evaluación. Se definió como variable objetivo hubo_incendio, se aplicó normalización tipo *z-score* para estandarizar las variables y se eliminaron aquellas con alta correlación (mayor a 0.8) para evitar redundancias.

Además, se habilitó el uso de GPU para acelerar el entrenamiento y se utilizó validación cruzada estratificada con 10 particiones o *folds*, lo que permitió una evaluación más robusta de los modelos. Por último, se fijó una semilla (*session_id* 7727) para asegurar que los resultados puedan reproducirse.

Este proceso aseguró que todos los modelos fueran entrenados y evaluados bajo las mismas condiciones, permitiendo compararlos de manera uniforme.

Validación interna y selección de modelos

La selección de modelos se realizó usando validación cruzada con la función *compare_models*() de PyCaret. Primero, los datos se dividieron en dos partes: un conjunto de entrenamiento (*train*) con el 80% de los datos y un conjunto de prueba (*test*) con el restante 20%. El conjunto de entrenamiento se utilizó para construir y comparar los modelos, mientras que el conjunto de prueba se conservó únicamente para evaluar los modelos finales con datos que no fueron utilizados durante el entrenamiento.

Dentro del conjunto de entrenamiento se aplicó validación cruzada con diez particiones. Esto implicó que los modelos se entrenaran varias veces usando diferentes subconjuntos de datos y se validaran con la parte restante, lo que permitió obtener una evaluación más confiable de su desempeño.

Durante este proceso, PyCaret entrenó y comparó automáticamente distintos modelos de clasificación, como regresión logística, bosques aleatorios y métodos de *boosting*. Luego, los ordenó según su desempeño (*F1-score*) y seleccionó los cinco mejores. Finalmente, estos modelos se evaluaron sobre el conjunto de prueba, lo que permitió medir qué tan bien funcionaban con datos completamente nuevos, y a partir de estos resultados, se eligió el modelo que obtuvo simultáneamente los valores más altos de F1-score y AUC entre los modelos evaluados.

Sin embargo, la selección final no se basó únicamente en el F1-score y el AUC. También se consideró el estadístico de Kolmogorov–Smirnov (KS), ya que este permite evaluar la capacidad del modelo para diferenciar entre las clases de ocurrencia y no ocurrencia de incendios. De esta manera, se buscó elegir un modelo que no solo presentara un buen desempeño predictivo, sino que además mostrara una capacidad de discriminación significativamente superior respecto a las demás alternativas evaluadas.

5. Validación externa del modelo

Dado que el problema tiene una dimensión temporal, la evaluación del modelo elegido se realizó mediante un esquema de validación de origen rodante, en el cual el conjunto de entrenamiento se expande progresivamente en el tiempo y el modelo se evalúa sobre el periodo inmediatamente siguiente. Este enfoque permite simular un escenario real de predicción y analizar la capacidad de generalización del modelo en un contexto temporal no estacionario.

En este sentido, los datos se dividieron en tres iteraciones: entrenamiento hasta 2020 y prueba en 2021; entrenamiento hasta 2021 y prueba en 2022; y entrenamiento hasta 2022 y prueba en 2023. Para cada iteración se construyó un modelo independiente utilizando PyCaret, ajustando el preprocesamiento a la distribución específica de cada ventana temporal y generando predicciones sobre el conjunto de prueba correspondiente.

Adicionalmente, se aplicó la prueba de Kolmogorov–Smirnov (KS) como análisis complementario, con el objetivo de comparar la distribución de las probabilidades predichas entre los conjuntos de entrenamiento y prueba en cada iteración. Este enfoque permite evaluar si el comportamiento del modelo se mantiene estable en el tiempo o si existen diferencias significativas en la forma en que se distribuyen sus salidas probabilísticas entre periodos.

En este contexto, el test KS no se utilizó como una métrica de desempeño, sino como una herramienta diagnóstica para identificar posibles cambios en la distribución de los datos a lo largo del tiempo. Esto permite detectar indicios de covariate shift, es decir, variaciones en la distribución de las variables o en el comportamiento del modelo entre los periodos de entrenamiento y evaluación. De esta forma, el análisis complementó las métricas tradicionales como el F1-score y el AUC, aportando información adicional sobre la estabilidad del modelo en un entorno temporal no estacionario.

6. Definición del protocolo de aplicación del modelo

Para aplicar el modelo a nuevos datos, se definió un proceso de preparación que replica las transformaciones realizadas durante el entrenamiento. En primer lugar, se procesan los datos climáticos del IDEAM y se agregan a escala semanal para variables como temperatura, humedad y precipitación, calculando métricas como promedio, mínimo y máximo, y en el caso de la precipitación, la suma total.

Luego, se construyen variables derivadas como el mes y la semana del año, y se representa la estacionalidad mediante transformaciones seno y coseno. También se incorporan variables territoriales como departamento y región natural, y se crean categorías climáticas para temperatura y humedad, dividiendo sus valores en cuatro niveles (baja, media-baja, media-alta y alta) según su distribución en los datos.

Para realizar las predicciones, se utilizan las condiciones climáticas de la semana actual y las dos semanas anteriores, de manera que el modelo recibe la misma estructura de información con la que fue entrenado (tres semanas de contexto).

Finalmente, el modelo se aplica mediante la función *predict_model()* de PyCaret, generando para cada combinación de municipio y semana dos resultados: una probabilidad de ocurrencia de incendios (*Score*) y una clasificación final (*Label*), obtenida al comparar esta probabilidad con el umbral definido. Estos resultados se exportan en un archivo de Excel para su posterior análisis.

Capítulo 6
RESULTADOS Y DISCUSIÓN

En esta sección se presentan los principales resultados del entrenamiento, validación y evaluación de los modelos de *machine learning* desarrollados para anticipar la ocurrencia de focos de incendios forestales en Colombia. Para seleccionar el mejor modelo se evaluaron los algoritmos generados mediante el proceso de experimentación automatizada realizado con PyCaret. En total se entrenaron 14 modelos de clasificación, de los cuales se analizaron los cinco con mejor desempeño. La Tabla 3 presenta una comparación de las principales métricas obtenidas por estos modelos en los conjuntos de entrenamiento y prueba.

Model	Training					Testing				
	Accuracy	AUC	Recall	Prec.	F1	Accuracy	AUC	Recall	Prec.	F1
Light Gradient Boosting Machine	76.47%	0.8495	76.36%	76.53%	76.44%	76.14%	0.8473	77.02%	75.69%	76.35%
Random Forest Classifier	75.42%	0.839	75.95%	75.16%	75.55%	75.32%	0.8363	76.22%	74.87%	75.54%
Extra Trees Classifier	75.56%	0.8399	75.53%	75.58%	75.55%	74.82%	0.8349	75.69%	74.39%	75.04%
Gradient Boosting Classifier	73.86%	0.8262	76.09%	72.84%	74.43%	73.78%	0.8229	76.61%	72.50%	74.50%
K Neighbors Classifier	72.25%	0.7908	72.67%	72.07%	72.37%	71.80%	0.7873	72.55%	71.48%	72.01%

Tabla 3. Desempeño comparativo de modelos de clasificación

El criterio principal para la selección del mejor modelo se basó en el desempeño del F1-score y el AUC, ya que ambas métricas permiten tener una visión más completa del rendimiento. El F1-score refleja el equilibrio entre la capacidad de detectar correctamente las semanas con incendios y la reducción de falsas alarmas, mientras que el AUC complementa este análisis al mostrar qué tan bien el modelo distingue entre la ocurrencia y no ocurrencia de incendios en distintos umbrales.

En este contexto, durante la validación externa se evaluaron distintos umbrales para convertir las probabilidades generadas por el modelo en predicciones. Los resultados

mostraron que un valor de 0.45 ofrecía el mejor equilibrio entre precisión y recall, alcanzando el mayor F1-score. En términos prácticos, este ajuste permitió identificar más semanas con incendios sin incrementar de forma considerable las falsas alarmas.

Bajo este criterio, el modelo con mejor desempeño fue *Light Gradient Boosting Machine* (LightGBM), que obtuvo un F1-score de 76% y un AUC de 0.85 en el conjunto de prueba. Estos resultados indican que el modelo no solo logra un buen balance entre precisión y sensibilidad, sino que además tiene una alta capacidad para distinguir correctamente entre la ocurrencia y la no ocurrencia de incendios.

Otros modelos evaluados, como *Random Forest Classifier* y *Extra Trees Classifier*, también presentaron un desempeño competitivo, con valores de F1 cercanos al 75% y AUC por encima de 0.83. Sin embargo, estos resultados se mantuvieron ligeramente por debajo del rendimiento alcanzado por el mejor modelo. En este sentido, LightGBM fue seleccionado como el modelo final del estudio.

Este resultado se complementa con la prueba de McNemar, la cual permite evaluar si las diferencias observadas en los errores de clasificación entre dos modelos son estadísticamente significativas o pueden atribuirse al azar. En este contexto, un valor p inferior a 0.05 indica que ambos modelos presentan un comportamiento significativamente diferente sobre el mismo conjunto de datos.

Los resultados muestran diferencias significativas entre LightGBM y los otros dos modelos evaluados ($p < 0.05$), lo que sugiere que las mejoras observadas en su desempeño no se deben únicamente a la variabilidad de la muestra (Tabla 4). Por el contrario, no se encontraron diferencias significativas entre Random Forest y Extra Trees ($p = 0.162$), indicando que ambos presentan un patrón de errores similar. Considerando además que LightGBM obtuvo los mayores valores de F1-score y AUC, la evidencia estadística respalda su selección como modelo final del estudio.

Comparación	Estadístico McNemar	p-value	Conclusión
LGBM vs Random Forest	319	0.000262928	Diferencia significativa
LGBM vs Extra Trees	346	0.023981081	Diferencia significativa
Random Forest vs Extra Trees	313	0.162025462	No significativa

Tabla 4. Comparación de mejores modelos mediante la prueba de McNemar.

Para complementar el análisis cuantitativo del desempeño del modelo seleccionado, se analizó la matriz de confusión correspondiente a las predicciones realizadas sobre el conjunto de prueba. Esta matriz permite observar de manera detallada el número de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN), proporcionando información clave sobre los tipos de error que el modelo comete.

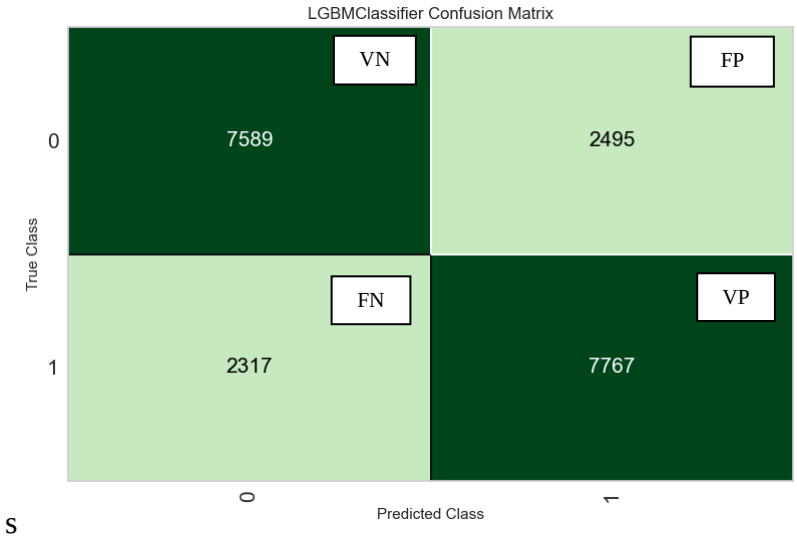

Ilustración 13. Matriz de confusión del modelo LightGBM

El modelo mostró un desempeño relativamente equilibrado, aunque la tasa de falsos negativos es del 23% y la de falsos positivos del 25%. Esto implica que aún se pueden omitir algunos incendios reales, lo cual es el error más crítico en contextos de alerta temprana, y también generar alertas que movilicen brigadas de bomberos hacia zonas

donde finalmente no ocurre un incendio, lo que sería un uso innecesario de recursos operativos.

En este contexto, ambos tipos de error tienen impactos diferentes, por lo que la evaluación no debe centrarse únicamente en el desempeño global, y sugiere ajustar el umbral de decisión del modelo para priorizar la detección de incendios reales, incluso si esto implica un aumento moderado de falsas alarmas, buscando un mejor balance entre detección temprana y uso eficiente de los recursos de respuesta.

Para evaluar el desempeño del modelo seleccionado en un contexto temporal, se aplicó una validación de origen rodante, ampliando progresivamente el conjunto de entrenamiento y utilizando como prueba el periodo siguiente.

Los resultados, presentados en la Tabla 5, muestran un comportamiento general estable a lo largo de las iteraciones. En la primera iteración (entrenamiento 2010–2020 y prueba 2021) se obtuvo un F1-score de 61.86% y un AUC de 0.81, mientras que en la segunda (entrenamiento 2010–2021 y prueba 2022) el F1-score descendió ligeramente a 60.47% y el AUC llegó a 0.79. En la tercera iteración (entrenamiento 2010–2022 y prueba 2023) , el modelo alcanzó su mejor desempeño, con un F1-score de 76.88%, una precisión de 77.24% y un AUC de 0.81.

En promedio, el modelo obtuvo un F1-score de 66.4% y mantuvo valores de AUC superiores al 0.79 en todos los escenarios, evidenciando una capacidad discriminativa estable. En conjunto, estos resultados demuestran que el modelo presenta una capacidad adecuada de generalización temporal y que mejora su desempeño al incorporar información más reciente.

Periodo de entrenamiento	Periodo de prueba	Accuracy	AUC	Recall	Prec.	F1
2010–2020	2021	69.84%	0.81	77.76%	51.36%	61.86%
2010–2021	2022	68.38%	0.79	74.44%	50.91%	60.47%
2010–2022	2023	73.42%	0.81	76.53%	77.24%	76.88%

Tabla 5. Resultados de la validación de origen rodante

Dado que el desempeño del modelo puede verse afectado por cambios en la distribución de los datos a lo largo del tiempo, se realizó un análisis de estabilidad mediante el test de Kolmogorov–Smirnov (KS), presentado en la Ilustración 16. Este test compara si las distribuciones de las probabilidades generadas por el modelo en el conjunto de entrenamiento y en el conjunto de prueba son similares o diferentes.

En este contexto, valores más altos del estadístico KS indican una mayor diferencia entre ambas distribuciones, mientras que valores más bajos reflejan mayor similitud y, por tanto, mayor estabilidad. Adicionalmente, valores p menores a 0.05 indican que estas diferencias son estadísticamente significativas.

Los resultados muestran diferencias significativas en todas las iteraciones ($p < 0.001$), lo que sugiere que la distribución de las predicciones cambia a lo largo del tiempo. En términos de magnitud, como se puede evidenciar en la Tabla 6, la mayor divergencia se observa en 2022 (KS = 0.194), seguida de 2023 (0.130), mientras que 2021 presenta la menor diferencia (0.075). Esto indica que el año 2022 es el periodo con mayor cambio en el comportamiento del modelo respecto a los datos de entrenamiento, lo que sugiere la presencia de variaciones temporales en el fenómeno analizado y evidencia de un entorno no estacionario, especialmente en ese año. Esto refuerza la necesidad de utilizar esquemas de validación temporal, como el origen rodante, para evaluar adecuadamente el desempeño del modelo seleccionado.

Iteración	**KS statistic**	**p-value**
2021	0.074868776	2.36871E-39
2022	0.193690158	2.1965E-251
2023	0.12957484	3.33513E-78

Tabla 6. Resultados de la prueba de Kolmogorov–Smirnov

Con el fin de comprender mejor los factores que influyen en las predicciones del modelo, se realizó un análisis de interpretabilidad mediante valores SHAP (SHapley Additive

Explanations). Esta técnica permitió identificar la contribución de cada variable y entender cómo el modelo toma sus decisiones.

Los resultados mostraron que variables como el departamento y la semana del año tienen una influencia importante en las predicciones, lo que evidencia que los incendios forestales en Colombia presentan patrones tanto geográficos como estacionales. Sin embargo, el hallazgo más relevante se observó en el comportamiento de las variables climáticas y sus rezagos temporales (Ver Ilustración 14).

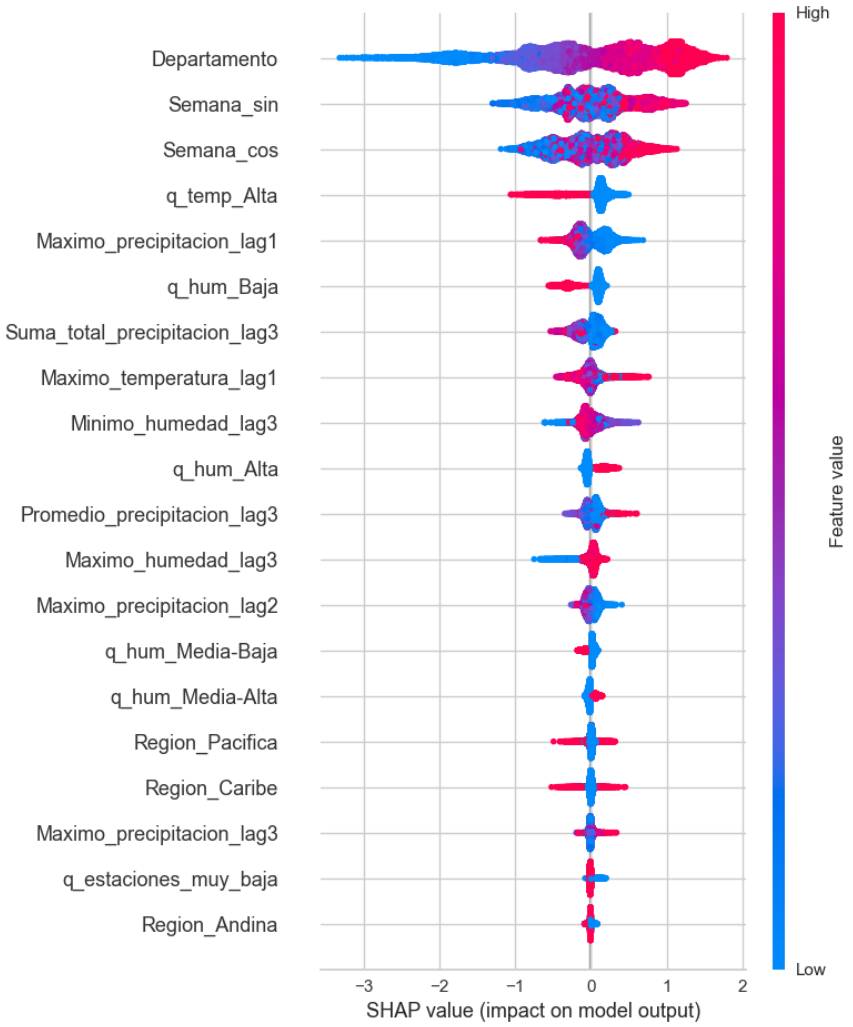

Ilustración 14. Importancia e impacto de las variables en las predicciones del modelo según valores SHAP

La humedad y la precipitación mostraron una mayor influencia cuando se evaluaron dos y tres semanas antes del periodo de predicción, mientras que la temperatura máxima tuvo un mayor impacto durante la semana inmediatamente anterior. Esto sugiere que las condiciones favorables para la ocurrencia de incendios no se generan de un momento a otro. Por el contrario, varias semanas con poca lluvia y baja humedad van secando progresivamente la vegetación y aumentando su susceptibilidad al fuego. Una vez que estas condiciones se han acumulado, un periodo reciente de altas temperaturas puede incrementar aún más el riesgo y favorecer la ocurrencia de incendios. En conjunto, estos resultados indican que el modelo captura patrones consistentes con la dinámica real de los incendios forestales, lo que aporta mayor confianza en la interpretación de sus predicciones.

En comparación con estudios previos desarrollados en Colombia, el modelo propuesto en esta investigación presenta una ventaja importante en términos de alcance espacial y temporal. Por un lado, el trabajo realizado en los Cerros Orientales de Bogotá se enfocó en una zona específica del país [43]. Por otro, el estudio de los Llanos colombo-venezolanos se desarrolló en una sola región y con información correspondiente al periodo 2015–2019 [44]. En contraste, el modelo planteado en este trabajo integra información de todo el territorio colombiano durante 12 años, lo que le permite representar una mayor variabilidad climática, geográfica y la amplia diversidad de ecosistemas presentes en Colombia.

Además, frente al estudio más reciente de la Universidad del Norte [45], una ventaja importante de esta investigación es el mayor tiempo de anticipación de la predicción. Mientras ese trabajo se apoya en variables satelitales muy cercanas al incendio ya detectable, el modelo propuesto utiliza información climática e histórica de las tres semanas previas para estimar la ocurrencia de incendios en la semana siguiente. Esto permite reconocer condiciones favorables antes de que el foco sea visible y brinda más tiempo de respuesta para fortalecer las acciones de prevención y gestión del riesgo a escala nacional.

Finalmente, aunque los resultados obtenidos son prometedores, es importante reconocer algunas limitaciones del estudio. Entre ellas se encuentra la falta de información histórica sobre variables como el uso del suelo y las prácticas agrícolas, así como la escasa cobertura de estaciones meteorológicas en varias regiones del país. Estas limitaciones pueden afectar la disponibilidad y calidad de los datos utilizados por el modelo, lo que podría reducir su capacidad de generalización en ciertos territorios, particularmente en regiones como la Orinoquía y la Amazonía.

A pesar de estas limitaciones, el enfoque propuesto muestra el alto potencial del *machine learning* como apoyo a la prevención de incendios forestales en Colombia, y abre la posibilidad de integrar este tipo de herramientas en futuros sistemas de alerta temprana y en estrategias de gestión del riesgo.

Capítulo 7
CONCLUSIONES

Este trabajo permitió desarrollar un modelo predictivo eficaz para la predicción de la ocurrencia de focos de incendios forestales en Colombia. Los resultados demuestran que es posible anticipar su ocurrencia en la semana siguiente a partir de variables meteorológicas rezagadas y patrones históricos de las tres semanas previas, incluso bajo la alta diversidad climática y de ecosistemas del país. El modelo *LightGBM*, seleccionado como el más eficiente, superó a varios de los algoritmos más utilizados en estudios previos y presentó el mejor desempeño general, destacándose por su equilibrio entre precisión, sensibilidad y capacidad de generalización.

El principal aporte de este trabajo no está solo en las métricas obtenidas, sino en la propuesta de un modelo adaptable al contexto colombiano, que puede servir como base para sistemas de alerta temprana a nivel nacional. Esta contribución es especialmente relevante en un contexto donde los incendios forestales tienen impactos crecientes sobre los ecosistemas y las comunidades.

Sin embargo, el modelo enfrenta limitaciones, especialmente relacionadas con la disponibilidad y calidad de datos en ciertas regiones. Esto sugiere la necesidad de fortalecer la infraestructura de monitoreo ambiental y de incorporar variables sociales, de uso del suelo y de comportamiento humano en futuras investigaciones.

En adelante, se recomienda avanzar hacia modelos que consideren no solo el clima y la historia de incendios, sino también el contexto territorial y socioeconómico junto con información proveniente de imágenes satelitales. Esto permitirá desarrollar herramientas predictivas más precisas y robustas, útiles para la planificación y la gestión del riesgo ambiental en Colombia.

REFERENCIAS

[1] J.-D. Rodríguez-Acuña, “Los incendios forestales en Colombia y su relación con la calidad del aire,” 2022. [Online]. Available: https://medioambiente.uexternado.edu.co/los-incendios-forestales-en-colombia-y-su-relacion-con-la-calidad-del-aire/ [Accessed: Apr. 25, 2026].

[2] United Nations, “¿Qué es el cambio climático?,” s.f. [Online]. Available: https://www.un.org/es/climatechange/what-is-climate-change [Accessed: Apr. 25, 2026].

[3] IDEAM, “Día Internacional de los Bosques 2026: Colombia avanza con el Sistema Nacional de Monitoreo Forestal,” 2026. [Online]. Available: https://www.ideam.gov.co/sala-de-prensa/noticia/dia-internacional-de-los-bosques-2026-colombia-avanza-con-el-sistema-nacional-de-monitoreo-forestal [Accessed: Apr. 27, 2026].

[4] A. N. D. Malpica, “Alerta por incendios forestales en Cundinamarca: CAR llama a los municipios a tomar medidas preventivas,” El Tiempo, 2026. [Online]. Available: https://www.eltiempo.com/bogota/alerta-por-incendios-forestales-en-cundinamarca-car-llama-a-los-municipios-a-tomar-medidas-preventivas-3519930 [Accessed: Mar. 31, 2026].

[5] J. Ruiz-Henestrosa, “Modelos de predicción de incendios forestales,” Universidad de Sevilla, 2024. [Online]. Available: https://hdl.handle.net/11441/165892 [Accessed: Apr. 23, 2026].

[6] B. T. Pham et al., “Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction,” Symmetry, vol. 12, no. 6, p. 1022, Jun. 2020, doi: 10.3390/sym12061022.

[7] S. Choi, M. Son, C. Kim, and B. Kim, “A Forest Fire Prediction Model Based on Meteorological Factors and the Multi-Model Ensemble Method,” Forests, vol. 15, no. 11, p. 1981, Nov. 2024, doi: 10.3390/f15111981.

[8] IDEAM, “Incendios de la cobertura vegetal,” s.f. [Online]. Available: https://www.ideam.gov.co/nuestra-entidad/ecosistemas-e-informacion-ambiental/incendios [Accessed: Mar. 31, 2026].

[9] C. Sarasty and J. Esteban, “Prototipo web para predicción y detección de incendios forestales en los cerros orientales de Bogotá, mediante una red de sensores e inteligencia artificial,” Feb. 2021. [Online]. Available: http://repository.unipiloto.edu.co/handle/20.500.12277/9883 [Accessed: Mar. 31, 2026].

[10] Universidad Sergio Arboleda, “Red de prevención y mitigación de incendios forestales: investigadores del IDEASA aportan a la innovación tecnológica a través de colaboración internacional para la gestión de riesgos,” 2025. [Online]. Available: https://www.usergioarboleda.edu.co/noticias/red-de-prevencion-y-mitigacion-de-incendios-forestales-investigadores-del-ideasa-aportan-a-la-innovacion-tecnologica-a-traves-de-colaboracion-internacional-para-la-gestion-de-riesgos/ [Accessed: Mar. 31, 2026].

[11] G. Archila, “Incendios forestales en Colombia: causas, consecuencias y aportes,” 2024. [Online]. Available: https://noticias.unad.edu.co/index.php/noticias-unad/incendios-forestales-en-colombia-causas-consecuencias-y-aportes [Accessed: Mar. 31, 2026].

[12] J. MacCarthy, J. Richter, S. Tyukavina, M. Weisse, and N. Harris, “Los últimos datos confirman: los incendios forestales están empeorando,” Dec. 2023. [Online]. Available: https://es.wri.org/insights/los-ultimos-datos-confirman-los-incendios-forestales-estan-empeorando [Accessed: Mar. 31, 2026].

[13] World Meteorological Organization, “Climate change and heatwaves,” 2023. [Online]. Available: https://public.wmo.int/content/climate-change-and-heatwaves [Accessed: Mar. 31, 2026].

[14] Organización Meteorológica Mundial, “Cambio climático, incendios forestales y contaminación atmosférica: un círculo vicioso,” 2025. [Online]. Available: https://wmo.int/es/news/media-centre/cambio-climatico-incendios-forestales-y-contaminacion-atmosferica-un-circulo-vicioso-que-acarrea [Accessed: Mar. 31, 2026].

[15] Organización Panamericana de la Salud, “Incendios forestales,” 2025. [Online]. Available: https://www.paho.org/es/temas/incendios-forestales [Accessed: Mar. 31, 2026].

[16] Argentina.gob.ar, “¿Cuáles son las variables y qué factores las afectan?,” 2018. [Online]. Available: https://www.argentina.gob.ar/seguridad/servicio-nacional-de-manejo-del-fuego/que-es-y-como-funciona-el-servicio-nacional-de-7 [Accessed: Mar. 31, 2026].

[17] National Geographic, “Los incendios forestales afectan el ciclo del agua: cómo impactan en la calidad de ese recurso,” 2024. [Online]. Available: https://www.nationalgeographicla.com/medio-ambiente/2024/09/los-incendios-forestales-afectan-el-ciclo-del-agua-como-impactan-en-la-calidad-de-ese-recurso [Accessed: Mar. 31, 2026].

[18] NASA Earth Science Data Systems, “FIRMS | NASA Earthdata,” 2024. [Online]. Available: https://www.earthdata.nasa.gov/data/tools/firms [Accessed: Mar. 31, 2026].

[19] D. Majumdar, “Decadal (2012–2023) account of spatio-temporal variability in satellite-detected biomass fires on Indian landmass and their fire radiative power,” Sci. Rep., vol. 15, Jul. 2025, doi: 10.1038/s41598-025-11200-w.

[20] A. Martínez Saucedo and P. E. Inchausti, “Predicción de incendios forestales mediante modelos de machine learning,” presented at CACIC, 2023. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/149568 [Accessed: Apr. 02, 2026].

[21] D. Rivera, “Cerca de 300 mil hectáreas de los ecosistemas colombianos son conservados y restaurados gracias a Masbosques y sus aliados,” 2023. [Online]. Available: https://masbosques.org/bosques-colombianos/ [Accessed: Apr. 25, 2026].

[22] WWF, “Colombia, la casa de los bosques,” 2025. [Online]. Available: https://www.wwf.org.co/?303630/Colombia-la-casa-de-los-bosques [Accessed: Jun. 25, 2025].

[23] D. M. J. S. Bowman et al., “Fire in the Earth System,” Science, vol. 324, no. 5926, pp. 481–484, Apr. 2009, doi: 10.1126/science.1163886.

[24] CAR, “¡Alto a las quemas en temporada de vientos! CAR advierte que el 90 % de incendios son por causas humanas,” 2025. [Online]. Available: https://www.car.gov.co/saladeprensa/alto-a-las-quemas-en-temporada-de-vientos-car-advierte-que-el-90-de-incendios-son-por-causas-humanas [Accessed: Aug. 24, 2025].

[25] Consejería de Medio Ambiente y Ordenación del Territorio, “Glosario de incendios forestales,” 2010. [Online]. Available: https://www.juntadeandalucia.es/medioambiente/web/participa_con_nosotros/campanas_comunicacion/infoca/2010/glosario.pdf

[26] D. S. Boshoff, “Understanding fire regimes: A biogeographical perspective,” Jàmbá J. Disaster Risk Stud., vol. 16, no. 1, p. 1673, Jul. 2024, doi: 10.4102/jamba.v16i1.1673.

[27] A. P. Rodríguez and M. L. Pérez, “Aplicación del modelo Rothermel a la gestión de riesgo de incendio en la región de la Chiquitanía, Santa Cruz Bolivia.”

[28] A. McEvoy, B. K. Kerns, and J. B. Kim, “Hazards of Risk: Identifying Plausible Community Wildfire Disasters in Low-Frequency Fire Regimes,” Forests, vol. 12, no. 7, p. 934, Jul. 2021, doi: 10.3390/f12070934.

[29] R. Ghali and M. A. Akhloufi, “Deep Learning Approaches for Wildland Fires Using Satellite Remote Sensing Data: Detection, Mapping, and Prediction,” Fire, vol. 6, no. 5, p. 192, May 2023, doi: 10.3390/fire6050192.

[30] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[31] T. M. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.

[32] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," Genet. Program. Evolvable Mach., vol. 19, no. 1, pp. 305–307, Jun. 2018, doi: 10.1007/s10710-017-9314-z.

[33] S. Yang, M. Lupascu, and K. S. Meel, "Predicting Forest Fire Using Remote Sensing Data And Machine Learning," Proc. AAAI Conf. Artif. Intell., vol. 35, no. 17, pp. 14983–14990, May 2021, doi: 10.1609/aaai.v35i17.17758.

[34] S. Sharma and P. Khanal, "Forest Fire Prediction: A Spatial Machine Learning and Neural Network Approach," Fire, vol. 7, no. 6, p. 205, Jun. 2024, doi: 10.3390/fire7060205.

[35] C. S. Gannon and N. C. Steinberg, "A global assessment of wildfire potential under climate change utilizing Keetch-Byram drought index and land cover classifications," Environ. Res. Commun., vol. 3, no. 3, p. 035002, Apr. 2021, doi: 10.1088/2515-7620/abd836.

[36] H. Singh et al., "A Comprehensive Review of Empirical and Dynamic Wildfire Simulators and Machine Learning Techniques used for the Prediction of Wildfire in Australia," Technol. Knowl. Learn., vol. 30, no. 2, pp. 935–968, Jun. 2025, doi: 10.1007/s10758-025-09839-5.

[37] M. K. Al-Bashiti and M. Z. Naser, "Machine learning for wildfire classification: Exploring blackbox, eXplainable, symbolic, and SMOTE methods," Nat. Hazards Res., vol. 2, no. 3, pp. 154–165, Sep. 2022, doi: 10.1016/j.nhres.2022.08.001.

[38] M. Mohajane et al., "Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area," Ecol. Indic., vol. 129, p. 107869, Oct. 2021, doi: 10.1016/j.ecolind.2021.107869.

[39] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor y M. D. Flannigan, "A review of machine learning applications in wildfire science and management," *Environmental Reviews*, vol. 28, no. 4, pp. 478–505, 2020. doi: 10.1139/er-2020-0019.

[40] K. M. Freitas et al., "Prediction of forest fire susceptibility using machine learning tools in the Triunfo do Xingu Environmental Protection Area, Amazon, Brazil," J. South Am. Earth Sci., vol. 153, p. 105366, Feb. 2025, doi: 10.1016/j.jsames.2025.105366.

[41] M. V. F. Silveira et al., “Drivers of Fire Anomalies in the Brazilian Amazon: Lessons Learned from the 2019 Fire Crisis,” Land, vol. 9, no. 12, p. 516, Dec. 2020, doi: 10.3390/land9120516.

[42] M. D. G. Martínez et al., “Revisión de antecedentes para la predicción de incendios forestales mediante IA,” 2024. [Online]. Available: https://virtual.cuautitlan.unam.mx/intar/memoriasceiaait/wp-content/uploads/sites/19/2024/12/36-Revision-de-antecedentes-para-la-Prediccion-de-Incendios-Forestales-mediante-IA-EDITADO.pdf

[43] K. Ocampo-Zuleta and J. Beltrán-Vargas, “Modelación dinámica de incendios forestales en los Cerros Orientales de Bogotá, Colombia,” Madera Bosques, vol. 24, no. 3, 2018, doi: 10.21829/myb.2018.2431662.

[44] J. S. Barreto and D. Armenteras, “Open Data and Machine Learning to Model the Occurrence of Fire in the Ecoregion of ‘Llanos Colombo–Venezolanos,’” Remote Sens., vol. 12, no. 23, p. 3921, 2020, doi: 10.3390/rs12233921.

[45] J. D. Anzola, L. D. Fuentes, and E. M. Rodríguez, “Desarrollo de un modelo de estimación para la prevención de incendios forestales en Colombia,” 2024. [Online]. Available: https://manglar.uninorte.edu.co/handle/10584/11968 [Accessed: Apr. 06, 2026].

[46] “PyCaret 3.0 | Docs,” s.f. [Online]. Available: https://pycaret.gitbook.io/docs [Accessed: Apr. 08, 2026].

[47] W. Chen et al., “A survey on imbalanced learning: latest research, applications and future directions,” Artif. Intell. Rev., vol. 57, no. 6, p. 137, May 2024, doi: 10.1007/s10462-024-10759-6.