Estimación de la tasa de recuperación de la vegetación tras incendios forestales mediante imágenes satelitales y machine learning

MILTON CARTAGENA MARTINEZ
GERMÁN DARÍO SÁENZ HERNÁNDEZ

Proyecto Aplicado para optar al título de Magíster en Ciencia de Datos

Directora
Yady Tatiana Solano Correa

Codirector:
Mario Milver Patiño Velasco

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
Febrero de 2026

FICHA RESUMEN

Estimación de la tasa de recuperación de la vegetación tras incendios forestales mediante imágenes satelitales y machine learning

1. ÁREA DE TRABAJO: Ciencia de Datos
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): Milton Cartagena / Germán Sáenz
4. CORREO ELECTRÓNICO: germansaenz@javerianacali.edu.co/ miltoncartagena@javerianacali.edu.co
5. DIRECCIÓN Y TELÉFONO: No aplica
6. DIRECTOR: Yady Tatiana Solano Correa
7. VINCULACIÓN DEL DIRECTOR: Pontificia Universidad Javeriana Cali
8. CORREO ELECTRÓNICO DEL DIRECTOR: tatiana.solano@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): Mario Milver Patiño Velasco
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Universidad del Cauca
11. OTROS GRUPOS O EMPRESAS: No aplica
12. PALABRAS CLAVE (al menos 5): Ciencia de datos, incendios, detección, cuantificación, imágenes satelitales, fotointerpretación.
13. FECHA DE INICIO: noviembre 2024
14. DURACIÓN ESTIMADA (En meses): 12 meses
15. RESUMEN: El presente proyecto tuvo como objetivo estimar la tasa de recuperación vegetal en áreas afectadas por incendios forestales en las regiones de Caquetá y Tolima, las cuales presentan condiciones climáticas diversas influenciadas por el fenómeno de El Niño-Oscilación del Sur (ENOS). Se desarrolló un modelo que utiliza imágenes satelitales de Sentinel-2, FACSAT-2 y de sensores Aerotransportados, empleando técnicas de entrenamiento supervisado y redes neuronales con el objetivo de detectar áreas afectadas por incendios y llevar a cabo análisis temporales, por lo tanto, se incorporaron variables climáticas relevantes en la recuperación vegetal, tales como la temperatura y la precipitación. Una vez detectada la zona de interés se aplicó una versión optimizada del algoritmo Gradient Boosting con histogramas (HGB) que permiten mejorar la eficiencia en la estimación de la recuperación vegetal en las zonas seleccionadas debido a su capacidad para manejar grandes volúmenes de datos, los resultados fueron visualizados en un tablero de dashboard de power BI para conocer los tiempos estimados de la tasa de recuperación en las dos zonas de estudio planteadas en este proyecto.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

Los incendios forestales representan una amenaza significativa en las regiones donde existe alta vegetación, sobre todo en áreas extensas y expuestas como en Colombia. En este tipo de zonas se presentan fenómenos que hacen que aumenten las precipitaciones o los tiempos secos, uno de estos factores es conocido como el fenómeno de El Niño (ENOS) que pone en riesgo la biodiversidad del país en la mayoría de sus regiones. La variabilidad climática y la intervención humana son componentes que generan gran impacto en las áreas de vegetación, el aumento de los incendios forestales crea un daño ambiental y social representativo. En este aspecto la tecnología es un aliado importante y fundamental para el desarrollo de actividades que permitan detectar las áreas vulnerables de manera eficiente [1].

En Colombia los incendios forestales son monitoreados por diferentes agencias entre los que se destacan la Unidad Nacional para la Gestión del Riesgo y Desastres - UNGRD, así como el Instituto de Hidrología, Meteorología y Estudios Ambientales – IDEAM, permitiendo que ambas entidades enfoquen sus esfuerzos en el planeamiento de estrategias y políticas ambientales para la preservación de los recursos naturales en zonas afectas por incendios. Teniendo en cuenta esta problemática y considerando al amplio uso de algoritmos de aprendizaje automático para efectuar análisis de grandes volúmenes de datos como imágenes, este proyecto tuvo como objetivo la estimación de la tasa de recuperación vegetal de zonas afectadas por incendios forestales, en los departamentos de Caquetá y Tolima, siendo estas regiones altamente afectadas por factores climáticos distintos debido a su geografía y ocurrencia de los fenómenos temporales conocidos como el Niño Oscilación del Sur (ENOS), en el cual se atenúan los periodos de lluvia y se intensifican los tiempos secos sin seguir un patrón común en las regiones del territorio colombiano, afectando puntualmente la tasa de recuperación de las zonas que ha sido afectadas por incendios forestales, por lo tanto, se propuso el desarrollo de un modelo que hace uso de imágenes satelitales de Sentinel-2, sensores aerotransportados y FACSAT-2 para la detección de las zonas de incendios empleando un entrenamiento supervisado y algoritmos de redes neuronales, este enfoque permite detectar el área de interés de nuestro trabajo y efectuar análisis temporal, una vez identificadas las áreas de incendios se procede a estructurar una base de datos que incluye datos de temperatura y precipitación, variables clave para la recuperación de la vegetación, y se propone una variante optimizada de Gradient Boosting que utiliza histogramas (HGB) para discretizar características continuas, haciendo el entrenamiento mucho más eficiente que permite estimar la tasa de recuperación de las zonas afectadas por incendios forestales.

En esta combinación tecnológica, las herramientas de visualización y análisis permiten contribuir a las acciones climáticas para preservar la vida terrestre y salvaguardar la vida de las comunidades afectadas. Generando un efecto positivo en el desarrollo ambiental, social y económico de nuestro País, siendo referentes de prevención y recuperación.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En la actualidad, los temas de biodiversidad y cambio climático tienen un alto impacto a nivel mundial en el desarrollo y gestión sostenible de los bosques y de la Amazonía, que son considerados el pulmón del mundo. De acuerdo con un artículo publicado por la revista Global Change Biology, se evidenció una disminución en la deforestación, pero un aumento de los incendios forestales en un 150% [2]. Sin embargo, no solo la Amazonía se ve afectada por la deforestación sino también otras zonas del país, debido a otros factores climáticos conocidos como el fenómeno de El Niño Oscilación del Sur (ENOS), en el cual se atenúan los periodos de lluvia y se intensifican los tiempos secos sin seguir un patrón común en las regiones Andina, Caribe y Pacífica [1], generando sequías extremas en el territorio colombiano, incrementando las posibilidades de incendios forestales, causando una amenaza constante para los ecosistemas, áreas rurales y urbanas, con impactos devastadores en la biodiversidad, las comunidades humanas, la calidad del aire y las economías locales.

Los incendios en Colombia son monitoreados por diferentes agencias entre los que se destacan la Unidad Nacional para la Gestión del Riesgo de Desastres – UNGRD, cuyo objetivo es: "dirigir la implementación de la gestión del riesgo de desastres, atendiendo las políticas de desarrollo sostenible, y coordina el funcionamiento y el desarrollo continuo del sistema nacional para la prevención y atención de desastres" [3], así como el Instituto de Hidrología, Meteorología y Estudios Ambientales – IDEAM cuya misión se enfoca en la producción de información confiable y oportuna sobre los recursos naturales y del medio ambiente [4], ambas entidades permiten la toma decisiones para influir directamente en el planteamiento de estrategias y políticas ambientales para la preservación de los recursos naturales.

En territorio colombiano los ecosistemas terrestres se caracterizan porque en su mayoría son propensos y sensibles al fuego, haciendo que los incendios forestales sean más recurrentes en nuestro territorio. Se estima que el 64% del territorio nacional corresponde al combustible más disponible que son los árboles, los cuales pueden durar hasta 100 horas incinerándose, además de poseer una gran cantidad de biomasa aérea de más de 150 toneladas por hectárea. Por otro lado, la susceptibilidad de estos ecosistemas difiere según la influencia del fuego, por lo que la susceptibilidad más concurrente en el país es muy alta con aproximadamente 1'100.000 km^2 ubicados principalmente en los biomas amazónicos, de la Orinoquía y los orobiomas bajo y alto de la región de los Andes [5].

Por lo tanto, es importante generar herramientas empleando imágenes satelitales, ciencia de datos y modelos de inteligencia artificial que permitan la identificación de zonas afectadas por incendios forestales, así como el seguimiento y monitoreo de la recuperación de las zonas vegetales post-incendio, de manera oportuna, a un bajo costo y sin arriesgar vidas humanas, con el fin de generar un impacto social, logrando de esta manera enfocar los esfuerzos de recuperación de estas áreas

en pro de mantener la lucha contra la desertificación, degradación de tierras y detener la pérdida de biodiversidad.

1.2 FORMULACIÓN DEL PROBLEMA

Pregunta de investigación: ¿Cómo se puede desarrollar un modelo para estimar la tasa de recuperación de vegetación tras incendios forestales, usando imágenes satelitales y técnicas de machine learning en el territorio colombiano?

Preguntas sistematización para formular el problema.

- ¿Cómo seleccionar y procesar imágenes satelitales y datos ambientales que representen adecuadamente las zonas afectadas por incendios forestales en Colombia para construir un conjunto de datos confiable y representativo?
- ¿Qué técnicas de machine learning son más efectivas para desarrollar un modelo que permita estimar la tasa de recuperación de la cobertura vegetal en zonas afectadas por incendios forestales?
- ¿Qué métodos de validación y métricas de rendimiento son los más apropiados para evaluar la precisión y efectividad del modelo en la estimación de la tasa de recuperación de la cobertura vegetal en diferentes ecosistemas afectados por incendios forestales?
- ¿Cómo se puede diseñar e implementar una herramienta de visualización que permita a los usuarios analizar e interpretar de manera sencilla los resultados relacionados con la tasa de recuperación de la vegetación en zonas afectadas?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo para la estimación de la tasa de recuperación de cobertura vegetal en zonas afectadas por incendios forestales en territorio colombiano, empleando imágenes satelitales y técnicas de machine learning.

2.2 OBJETIVOS ESPECÍFICOS

1. Construir un conjunto de datos representativo de imágenes satelitales y datos ambientales.
2. Implementar diferentes modelos para estimar la tasa de recuperación de zonas afectadas por incendios forestales.
3. Comparar los modelos propuestos considerando diferentes métricas para evaluar su precisión.
4. Desarrollar una herramienta de visualización con los resultados obtenidos.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1 Incendios forestales

Los incendios forestales son desastres naturales que pueden tener consecuencias devastadoras para el medio ambiente, las comunidades y la infraestructura de una región especifica. Las estrategias y modelos de prevención son cruciales para mitigar los impactos que generan dichos eventos [6]. Estos modelos normalmente emplean una combinación de datos meteorológicos, combustible y de actividad humana para predecir cual sería la probabilidad que un incendio se produzca en una zona específica [7]. Es importante resaltar que algunos modelos también tienen en cuenta el comportamiento del incendio una vez iniciado, considerando datos como velocidad de propagación y dirección del viento para estimar el área potencial que puede ser afectada, contar con esta información es crucial para la toma de decisiones en la planificación de una respuesta y el desarrollo de las estrategias para mitigar su impacto [8].

3.1.2 Detección de incendios

Los bosques juegan un papel relevante en la conservación del equilibrio ecológico de nuestros ecosistemas, sin embargo, en algunas ocasiones los incendios forestales pueden pasar desapercibidos y solo se tiene conocimiento de estos cuando ya se ha consumido un área considerable y extensa, lo cual dificulta las maniobras para control y extinción [9]. Los incendios forestales tienen consecuencias a largo plazo en los diferentes ambientes afectados, en el cambio climático y en la desaparición de especies de plantas y animales. La mayoría suelen presentarse en zonas aisladas, deshabitadas o con un mantenimiento inadecuado, donde la vegetación seca y marchita sirve de combustible para su propagación. Estos incendios pueden iniciarse por errores humanos, como la eliminación inadecuada de cigarrillos, o por causas naturales, como la intensa luz solar concentrada por cristales rotos [10].

Algunas técnicas de detección y control empleadas para la detección y monitoreo de incendios en zonas afectadas son; la vigilancia aérea y por satélite, red de observadores y los sensores con cámaras ópticas. Estas herramientas se pueden dividir en los siguiente: la notificación voluntaria y pública de incendios, así como aeronaves y trabajadores de campo en tierra, la quema controlada, la predicción meteorológica de incendios, las torres de vigilancia, la detección óptica de humo, los detectores de rayos, los detectores de infrarrojos, los aviones de vigilancia, los camiones de bomberos, redes de montañistas, guardaparques naturales y las notificaciones telefónicas, son algunos de los enfoques más populares para detectar incendios en una fase temprana.

3.1.3 Evaluación post-incendio

Para reducir el impacto de los incendios forestales, las autoridades locales y gubernamentales emplean diferentes técnicas de detección para el control y vigilancia de zonas, en Colombia el Ministerio de Ambiente, Vivienda y Desarrollo Territorial creó el plan nacional de prevención, control de incendios forestales y restauración de áreas afectada cuyo objetivo es: "Establecer los lineamientos de orden nacional para la prevención, control y restauración de las áreas afectadas por los incendios forestales, mitigando su impacto y fortaleciendo la organización nacional, regional y local con programas a corto (3 años), mediano (10 años) y largo plazo (25 años)" [11].

Una vez que el incendio se produce y deja afectaciones en las áreas devastadas, algunos de los planes de evaluación incluyen el despliegue de personal, equipo especializado, transporte aéreo o terrestre, con el fin de efectuar las evaluaciones y extraer información correspondiente al impacto que el incendio forestal tuvo sobre sobre un área específica, estos despliegues y movimientos normalmente son costosos y además peligroso para el personal involucrado, teniendo en cuenta que normalmente los incendios se producen en terrenos aislados dificultando el acceso a los mismos; por lo tanto, un monitoreo manual puede llegar a ser muy peligroso y con elevados costos [12].

3.1.4 Teledetección para el Monitoreo Ambiental

La teledetección es el área de las ciencias que busca obtener información sobre la Tierra mediante sensores remotos, los cuales detectan la radiación electromagnética reflejada o emitida por los objetos; estos sensores, ubicados en satélites o aviones, adquieren datos en diferentes bandas del espectro electromagnético, como el visible, infrarrojo o microondas [13].

La teledetección permite evaluar cambios en el uso del suelo, la deforestación, el monitoreo de emisiones de gases, monitoreo de suelos, el control de cuerpos de agua y otros fenómenos ambientales, su capacidad para ofrecer datos precisos y frecuentes en grandes extensiones hace de esta tecnología una herramienta crucial para la gestión ambiental y el análisis de cambios a nivel global [14].

3.1.5 Sensores Satelitales para teledetección

Los sensores satelitales que se encuentran en órbita se han convertido en una herramienta fundamental para recopilar información de la tierra, realizando diferentes técnicas derivadas de la teledetección es posible medir y obtener datos radiométricos para estimar de manera precisa heterogeneidad del paisaje y la diversidad de especies demostrando que es una herramienta poderosa y prometedora, ya que permite trabajar a una alta resolución espacial y temporal, logrando analizar sus cambios a través del tiempo, mitigando la dificultad asociada para recopilar datos en campo [15].

Por esta razón, existen sensores satelitales y constelaciones enfocadas a la teledetección de la tierra que permiten obtener, en algunos casos sin costo, imágenes que se pueden procesar para estudiar la degradación de los suelos, erosión, perdida de vegetación, pérdida de calidad del suelo, perdida de materia orgánica como incendios y deforestación entre otros. Los sensores más empleados para este tipo de estudio son los embarcados en satélites como: Landsat, Spot, Aster, EO, etc. (considerados de media a alta resolución espacial), los sensores de alta resolución temporal (embarcados en los satélites Modis, NOAA, etc.) y los sensores con radar (embarcados en los satélites Envisat, ERS, etc) [16].

Es importante resaltar que las capacidades de estos sensores en muchas aplicaciones son combinadas para obtener mejores resultados, el acceso a imágenes multiespectrales, térmicas o de radar no son adquiridas por la misma fuente o sensor, por lo tanto, la combinación de estas tecnologías son componentes críticos para la aplicación de la teledetección en la gestión forestal, los sensores multiespectrales permiten detectar radiación electromagnética en diferentes bandas espectrales como el espectro visible o el infrarrojo cercano, los sensores térmicos permiten medir la radiación térmica emitida por objetos facilitando la detección de diferencias de temperatura en la superficie terrestre, ahora bien, el acceso a imágenes multiespectrales, puede realizarse a través de diferentes fuentes, como satélites de observación de la tierra, aeronaves equipadas con sensores digitales o drones, dependiendo la aplicación se puede recurrir a la fuente, por ejemplo para trabajos a gran escala las imágenes satelitales son óptimas, mientras que los sensores a bordo de aeronaves y drones son ideales para obtener datos con mayor resolución [17].

3.1.5.1 Misión Sentinel 2

La misión Sentinel-2 consiste en una constelación de tres satélites (Sentinel-2A, 2B y 2C), como parte del programa Copernicus de la ESA, estos tres sensores capturan imágenes de tipo óptico y multiespectral con una cobertura global debido a su inclinación 98° y ofrece una revisita de 5 días. Cada uno de estos satélites está equipado con una carga útil MSI (MultiSpectral Instrument), capaz de registrar información en 13 bandas espectrales que incluyen el espectro visible y el infrarrojo de onda corta SIWR, las resoluciones que ofrece de 10m, 20m y 60m [18].

3.1.5.2 Misión FACSAT-2

La misión FACSAT-2 consiste en un sistema satelital, lanzado el 15 de abril de 2023 por la Fuerza Aeroespacial Colombiana y operado actualmente desde la ciudad de Cali, posee una órbita polar lo cual brinda acceso global y ofrece una revisita > 30 días sobre un punto especifico de territorio colombiano, está equipado con dos cargas útiles, la primera es una cámara óptica para observación de la tierra con una resolución de 4.75 m multiespectral con 8 bandas, que incluyen el espectro visible, NIR y Pancromático, la segunda carga útil es empleada para espectrometría y cuantificación

de gases de efecto invernadero.

3.1.6 Segmentación

La segmentación de imágenes satelitales representa un paso crucial en el análisis de datos, siendo particularmente desafiante debido a las texturas distintivas que complican los métodos de clasificación convencionales que operan a nivel de píxel. La tarea de segmentación implica la separación de los píxeles que componen los objetos presentes en la imagen. Para llevar a cabo esta labor, es imprescindible reunir una extensa cantidad de píxeles que compartan características homogéneas, con el objetivo de formar regiones de interés significativas. Posteriormente, se extraen de estas regiones descriptoras que facilitan la interpretación y reconocimiento de diferentes clases, ajustándose a la naturaleza específica del problema en cuestión [19].

Existen diversas metodologías para llevar a cabo una clasificación de clases, utilizando segmentación. La más frecuente se centra en documentar características específicas de los pixeles con el objetivo de determinar la categoría de un estudio específico, únicamente teniendo en cuenta lo visualmente registrado en cada uno de los pixeles. Sin embargo, también existen técnicas como la implementada por IDEAM, que utiliza diversas imágenes satelitales de sensores superpuestos para ajustar un thershold para definir los valores de reflectancia por región, con el objetivo de lograr una clasificación más precisa de la imagen objetivo, teniendo en cuenta factores como la calibración del sensor, el ángulo de la toma, la hora de la toma, entre otros [20].

3.1.7 Nubes y detección

Al trabajar con imágenes satelitales, uno de los principales fenómenos a tener en cuenta, es la nubosidad, ya que este tipo de factores afectan directamente la calidad de la señal observada y como consecuencia, la estimación de índices espectrales: NDVI o NBR. La presencia de sombras y nubes establece valores que no son representativos de la superficie terrestre, estos pueden ser interpretados con cambios reales en la vegetación o en la gravedad de la quema. Si estos píxeles que no corresponden, no se identifican y eliminan correctamente, van a generar error hacia las etapas que continúan en el análisis, generando sesgo en la caracterización del evento y la estimación de recuperación vegetal.

Con el propósito de abordar este problema, Sentinel-2 proporciona una capa de clasificación conocida como Scene Classification Layer (SCL). Esta se genera en el procesador Sen2Cor y asigna una clase a cada píxel en función de diversos factores, incluyendo agua, nubes, suelo desnudo, sombras, entre otros. Esta capa facilita la identificación eficaz de los píxeles vinculados a sombras y nubes, constituyendo el fundamento para la depuración de escenas, lo que permite una mayor precisión en la construcción de series temporales y el cálculo de índices espectrales.

Es importante aclarar que, al igual que con cualquier clasificador automático, el mapa SCL puede

presentar ciertas limitaciones en la identificación de nubes o algunas sombras proyectadas. Estas fluctuaciones son normales en las aplicaciones operativas y no obstaculizan su aplicación, dado que es imperativo tener una comprensión clara de su esencia y una interpretación de los resultados considerando estos márgenes de error. La aplicación del SCL como principal componente para el enmascaramiento de nubes y sombras se percibe como apropiada para el grado de detalle requerido y para alcanzar los objetivos del modelo de recuperación de la vegetación post-incendio [21].

3.1.8 Técnicas de detección de vegetación

En el análisis de incendios se pueden identificar tres fases diferentes para evaluación, la primera es la predicción sobre donde se podría iniciar un incendio, la segunda se enfoca en la predicción de la propagación del fuego una vez detectado y la tercera es la evaluación de la afectación del área después de consumido el fuego para analizar el índice de recuperación de esas zonas [10].

En todas estas fases es importante evaluar y conocer algunos indicadores que brindan información sobre el estado de la zona antes del incendio, durante el evento del conato y después de que ocurre el evento, las cuales son útiles para realizar estudios temporales del área afectada.

NVDI: el índice de vegetación diferencial normalizada o NDVI, es una variable que permite estimar el comportamiento de una vegetación sobre la base de la medición, con sensores remotos, de la intensidad de la radiación de ciertas bandas del espectro electromagnético que la misma emite o refleja. Normalmente es aplicado a las comunidades de plantas, vegetación o zonas de bosques, el índice arroja valores de intensidad del verdor de la zona, y da cuenta de la cantidad de vegetación presente en una superficie y su estado de salud o vigor vegetativo [22], la fórmula empleada para su cálculo es:

$$NDVI = \frac{NIR + RED}{NIR - RED} \tag{1}$$

Donde NIR es infrarrojo cercano y RED es la reflectancia en el espectro rojo visible.

NBR: el índice de tasa de quema o NBR, es empleado para resaltar las zonas afectadas por incendios y determinar la gravedad de estos. Las zonas quemadas presentan una alta reflexión en la banda del infrarrojo de onda corta cercano SWIR y una baja reflexión en el infrarrojo cercano NIR, la fórmula empleada para su cálculo es: [23].

$$NBR = \frac{NIR + SWIR}{NIR - SWIR} \tag{2}$$

Donde NIR es infrarrojo cercano y SWIR es el infrarrojo de onda corta.

3.1.9 Aprendizaje automático

En los últimos años se ha producido un aumento exponencial en el uso del aprendizaje automático [24] [25] [26] para una amplia gama de propósitos, desde la investigación hasta las aplicaciones prácticas, incluyendo la minería de textos, la detección de spam, la recomendación de vídeos, la categorización de imágenes y la recuperación de ideas multimedia [27]. El aprendizaje profundo (deep learning, DL) es un enfoque de aprendizaje automático (machine learning, ML) que se utiliza habitualmente en estos contextos [28]. El dominio de trabajo del DL es un subconjunto del ML y la inteligencia artificial (IA); por lo tanto, puede considerarse una función de la IA que imita la forma en que el cerebro humano procesa la información [24].

La red neuronal tradicional a partir de la cual se originó la DL se ha visto significativamente superada por su rendimiento superior. Además, el DL utiliza transformaciones y tecnologías de grafos en tándem para construir modelos de aprendizaje multicapa [29]. Recientemente se ha producido un aumento del interés por aplicar los algoritmos de aprendizaje automático en campos como el reconocimiento de imágenes, el reconocimiento óptico de caracteres, la predicción de precios, el filtrado de spam, la detección de fraudes, la sanidad, el transporte y muchos otros [30].

3.1.10 Modelos de Aprendizaje Automático para la Estimación de la Recuperación de la Vegetación Post-Incendio

En el contexto de la evaluación de la recuperación de la vegetación post-incendio, el uso de modelos de aprendizaje automático permite integrar información espectral, temporal y climática para capturar patrones complejos que no pueden ser representados mediante umbrales fijos o análisis descriptivos. En este proyecto, la metodología contempla la comparación de tres modelos de aprendizaje automático representativos: Regresión Logística (LR), Random Forest (RF) y HistGradientBoosting (HGB), los cuales responden a distintos niveles de complejidad, capacidad de generalización y representación de relaciones no lineales que se presentan a continuación:

3.1.10.1 Regresión Logística

La Regresión Logística se emplea como modelo base debido a su carácter probabilístico e interpretabilidad, permitiendo establecer una línea de referencia para el análisis de recuperación. Este modelo estima directamente la probabilidad de pertenencia a la clase positiva mediante la función logística (sigmoide), transformando una combinación lineal de variables explicativas en valores entre 0 y 1.

Matemáticamente, la probabilidad estimada puede expresarse como:

$$\boldsymbol{P(Y = 1|X)} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e} - (\boldsymbol{\beta} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\beta_{iX_i}}} \quad (3)$$

Donde X_i representan las variables predictoras (índices espectrales, variables climáticas y temporales) y β_i los parámetros estimados por máxima verosimilitud.

Si bien este modelo permite una interpretación directa del efecto de cada variable sobre la probabilidad de recuperación, asume relaciones lineales entre las variables explicativas y la respuesta, lo cual limita su capacidad para representar procesos ecológicos no lineales y efectos de interacción entre variables ambientales [31].

3.1.10.2 Random Forest

Es un modelo de tipo ensemble basado en la construcción de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios del conjunto de datos y de las variables predictoras. La predicción final corresponde al promedio de las probabilidades generadas por cada árbol individual, lo que reduce la varianza del modelo y mejora su capacidad de generalización.

Formalmente, el modelo puede representarse como:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x) \quad (4)$$

Donde $T_b(x)$ corresponde al árbol b entrenado mediante muestreo bootstrap y selección aleatoria de características.

Este enfoque permite capturar relaciones no lineales y efectos de interacción entre índices espectrales y condiciones climáticas. Además, su estructura reduce el riesgo de sobreajuste frente a modelos individuales de árbol. Sin embargo, aunque robusto, puede presentar limitaciones en la calibración fina de probabilidades en contextos de fuerte desbalance de clases [32].

3.1.10.3 Las Redes Neuronales Artificiales (RNA)

son modelos diseñados para manejar relaciones no lineales complejas. Utilizan una estructura de capas interconectadas para procesar información, lo que las hace muy versátiles. Estas redes pueden abordar desde tareas predictivas hasta análisis complejos de datos. El Aprendizaje Profundo (Deep

Learning) lleva las RNA un paso más allá, agregando múltiples capas que permiten identificar patrones más detallados y jerárquicos. Técnicas específicas como las redes convolucionales (CNN) y las redes recurrentes (RNN) son ideales para aplicaciones como el reconocimiento de imágenes y el análisis de datos secuenciales o series temporales [33].

3.1.10.4 HistGradientBoosting (HGB)

El modelo HistGradientBoosting (HGB) corresponde a una variante eficiente del algoritmo de Gradient Boosting, optimizada para grandes volúmenes de datos tabulares. A diferencia de Random Forest, que construye árboles de manera independiente, el boosting entrena árboles de forma secuencial, donde cada nuevo árbol corrige los errores residuales del modelo anterior mediante la optimización de una función de pérdida, El modelo puede representarse como:

$$\boldsymbol{F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)} \tag{5}$$

donde $h_m(x)$ representa el árbol ajustado al gradiente negativo de la función de pérdida y ν es el learning rate que controla la contribución de cada iteración. La versión basada en histogramas discretiza previamente las variables continuas en intervalos (bins), reduciendo el costo computacional y permitiendo una mayor eficiencia en escenarios con millones de observaciones, como ocurre en análisis a nivel de píxel.

El modelo HGB tiene las siguientes características:

- Capacidad para manejar desbalance de clases.
- Alto desempeño en métricas de priorización como PR-AUC.
- Capacidad para generar probabilidades bien calibradas (bajo Brier Score).
- Habilidad para capturar relaciones no lineales complejas entre severidad del incendio, condiciones climáticas y dinámica temporal.

Estas características lo convierten en un modelo adecuado para representar la recuperación vegetal como un proceso probabilístico continuo y no como una clasificación rígida binaria [34].

La inclusión de estos modelos en el marco teórico permite fundamentar metodológicamente la estrategia de modelamiento adoptada y establecer una conexión directa entre los datos de teledetección, los índices espectrales y la estimación probabilística de la recuperación.

3.1.11 Algoritmos combinados

Diversos estudios han demostrado que combinar diferentes algoritmos pueden llevar a resultados

más óptimos de segmentación y clasificación, esta combinación permite explotar de manera individual cada característica del algoritmo y extraer aquellas en la que es más fuerte realizando detecciones correctas de una clase especifica, en este sentido se puede afirmar que una variedad de opiniones y distintos aportes mejoran el proceso de toma de decisiones, este proceso es conocido como "aprendizaje hibrido", el cual mejora la precisión de un sistema facilitando la solución de problemas computacionales clave, como el reconocimiento facial o la detección de clases, los nuevos avances permiten a los usuarios aprovechar el poder de combinar algoritmos para dar soluciones cada vez más precisas a aplicaciones del mundo real [35].

3.1.12 Métricas para Evaluación

Las métricas de medición en aprendizaje automático dependen del tipo de problema que se está resolviendo (clasificación, regresión, agrupamiento, etc.), donde el objetivo es cuantificar la precisión, sensibilidad y predicción de los diferentes modelos, esto puede involucrar un entrenamiento empleando diversas técnicas tales como; regresión logística, Support Vector Machine, árbol de decisiones, Random Forest entre otros. También es posible emplear una combinación de los resultados que permitan a modelo generalizador efectuar predicciones finales con base en los datos de los modelos clusterizados para mejorar la precisión y los resultados obtenidos de manera individual [36].

- **Brier Score (BS**): Es una métrica de aprendizaje automático utilizada para evaluar el rendimiento de las probabilidades estimadas de un modelo, el cuál mide la diferencia cuadrática media entre las probabilidades predichas y los resultados reales (clases), proporcionando información sobre la calibración y la fiabilidad de un modelo predictivo [37].

$$BS = \frac{1}{n}\sum_{i=1}^{n}(Pi - yi)^2 \quad (6)$$

  Donde: n = Número total de observaciones, Pi = probabilidad predicha (0.0 a 1.0), yi = resultado observado (0 para falla y 1 para éxito).

- **Exactitud (accuracy):** se encarga de medir la proporción de las predicciones correctas, ya sean positivas o negativas respecto al total de observaciones, es común emplearlo con clases que están balanceadas [38].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

  Donde: TP = verdaderos positivos, TN = verdaderos negativos, FP = falsos positivos y FN = Falsos negativos.

- **Precisión (precision):** se encarga de medir la pureza de cada clase, es decir, que porcentaje de pixeles detectados como verdaderos positivos realmente lo son.

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

- **Sensibilidad (recall):** mide la capacidad del modelo para detectar de manera correcta los casos positivos, su empleo es común cuando se requiere minimizar falsos negativos [39].

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

- **F1-Score:** Es el promedio armónico entre la precisión y la sensibilidad (recall), ideal para conjuntos de datos desequilibrados [40].

$$F1 - Score = 2\,\frac{Precision \times Recall}{Precision+Recall} \quad (10)$$

- **Área bajo la curva ROC AUC:** es empleado para medir la capacidad del modelo de distinguir entre clases calculando el área bajo la curva de la tasa de verdaderos positivos vs la tasa falsos positivos, normalmente se emplea en presencia de clases desbalanceadas en clasificación binaria [38].

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (11)$$

Donde: TPR = tasa de verdaderos positivos y FPR = Tasa de falsos positivos

- **Área bajo la curva PR AUC:** cuantifica la capacidad del modelo para priorizar correctamente la clase positiva, calculando el área bajo la curva Precisión–Recall, definida por la relación entre precisión y sensibilidad (recall) a lo largo de distintos umbrales de decisión [41].

$$PR - AUC = \int_0^1 Precision\,(Recall)d(Recall) \quad (12)$$

3.1.13 Herramienta de visualización

Las herramientas de visualización son aquellas que permiten interactuar al usuario final con los datos, es realmente importante que durante su diseño se consideren diversos factores que permitan empatizar y entregar el mensaje de la mejor manera, los datos deben ser útiles y generar valor a quien los lee e interpreta, algunos aspectos relevantes a considerar son; etiquetas, texto alternativo,

interacción y accesibilidad en gráficos, color, contrastes, el objetivo es generar visualizaciones interactivas que permitan explorar los datos a diferentes usuarios, los cuales puedan tener acceso a Visualizadores, Tablas, Gráficos, Mapa de Calor, Repositorio de Imágenes entre otros [42].

Existen diferentes herramientas empleadas para la visualización de los datos, entre las que se destacan Power BI, Tableu, Google charts (Html5), Google looker Studio y lenguajes como Python cuyas librerías permiten efectuar visualizaciones personalizadas, todas las anteriores presentan ventajas y desventajas, por ejemplo, Power BI y Tableu que cuentan con un respaldo y soporte de Microsoft y SalesForce respectivamente, son interfaces intuitivas que poseen capacidades robustas para los procesos de extracción y transformación de datos de dato, sin embargo puede presentar altos costos asociados a licenciamiento, por otro lado existen herramientas libres como Google charts y Python que ofrecen una mayor flexibilidad de personalización, control de datos y una gran comunidad de soporte para desarrollar visualizaciones, sin embargo, estas herramientas de software libre requieren conocimientos previos de programación haciendo su curva de aprendizaje más inclinada para aplicar todas sus capacidades [43] [44].

3.2 ANTECEDENTES

3.2.1 Técnicas de teledetección para evaluar la recuperación de la vegetación tras un incendio [42]

El artículo “Remote sensing techniques to assess post-fire vegetation recovery” presenta diferentes técnicas de teledetección que permiten evaluar la recuperación de la vegetación después de un incendio, empleando imágenes satelitales para cubrir áreas extensas y diferentes ecosistemas para estudiar la recuperación de las zonas afectadas, utilizando diferentes índices y métricas para analizar la dinámica temporal de la cubierta vegetal, otra técnica empleada es la integración de diferentes sensores como el LiDAR, hiperespectral y térmicos para realizar un seguimiento más exhaustivo de la dinámica forestal, drones y análisis de series temporales también fueron identificadas, empleando datos para la detección de cambios y seguimiento de las perturbaciones forestales y los patrones de recuperación a través del tiempo [45].

Estas técnicas pueden contribuir colectivamente a entender mejor el objeto de estudio y como los ecosistemas responden a eventos de incendios, ofreciendo ventajas sobre las campañas tradicionales de recolección de información y datos, mejorando la relación costo beneficio, la cobertura y potencializando el monitoreo continuo de grandes áreas para recopilar datos consistentes y objetivos en el tiempo reduciendo los errores asociados a la recolección manual. Los resultados obtenidos estuvieron enfocados a establecer una relación entre los indicadores espectrales y la recuperación ecológica, la ventaja del uso de múltiples métodos de teledetección para proporcionar una evaluación más completa, identificación de vacíos en la temporalidad y recomendaciones para futuras investigaciones [45].

3.2.2 Monitoreo de la vegetación y su recuperación después de un incendio: caso de estudio Portugal [43]

Este artículo "Vegetation Monitoring and Post-Fire Recovery: A Case Study in the Centre Inland of Portugal" presenta un estudio en vegetaciones de árboles tipo pino y eucalipto, los cuales fueron afectados de manera significativa por incendios en Portugal, empleando sensores remotos la investigación enfoca sus técnicas en evaluar específicamente el índice normalizado de diferencia de vegetación – NDVI y el índice normalizado de tasa de quema - NBR, para evaluar el estado y salud de la cubierta vegetal después de un incendio, los valores de NDVI usados fueron los del año 2007 y desde 2020 a 2022, realizando un seguimiento para comprender la dinámica de la recuperación forestal de la zona y los cambios en la composición de los bosques considerando los valores de NBR para calcular la severidad de las quemaduras y diseñar estrategias de recuperación.

Este caso de estudio logró demostrar que el NDVI y el NBR contribuyen significativamente a enfocar los esfuerzos para la recuperación de estas zonas afectadas por incendios forestales de diferentes maneras, haciendo un seguimiento de la recuperación a lo largo del tiempo en distintos tipos de vegetación para comprender la rapidez y eficacia con la que se recuperan estas especies. La investigación también integró datos climáticos, empleando los índices espectrales junto con información de los niveles de precipitación proporcionando una comprensión más completa de los factores que puede correlacionarse con la recuperación después del incendio [46].

3.2.3 Monitoreo de Incendios Forestales y Dinámica de Vegetación con Teledetección [12]

El artículo "Forest Fire Spread Monitoring and Vegetation Dynamics Detection Based on Multi-Source Remote Sensing Images", presenta la combinación de imágenes de teledetección provenientes de diversas fuentes, junto con técnicas de aprendizaje automático, presentada en este estudio, tiene un gran potencial para desarrollar sistemas más eficientes y efectivos en el monitoreo y análisis de incendios forestales. Estas herramientas resultan particularmente útiles en regiones con alta vulnerabilidad ambiental, ya que permiten mejorar tanto la precisión como la rapidez en las acciones de respuesta frente a emergencias.

Por otro lado, la metodología basada en el uso de indicadores como el NBR y en el análisis de factores meteorológicos ofrece una base sólida que puede ser adaptada para estudios similares en otras áreas geográficas. Este enfoque técnico representa un recurso valioso para futuras investigaciones y aplicaciones orientadas a la gestión y prevención de desastres ecológicos [12].

3.2.4 Evaluación de Cambios y Recuperación de Vegetación Post-incendio con DL [36]

El artículo “Vegetation change detection and recovery assessment based on post-fire satellite imagery using deep learning”, presenta un enfoque en la combinación de manera efectiva de técnicas de aprendizaje profundo con el análisis detallado de series temporales, logrando una metodología flexible que se puede replicar en distintos escenarios de recuperación post-incendio. Este enfoque no solo es relevante en términos de precisión, sino que también tiene un enorme potencial para apoyar iniciativas enfocadas en la gestión ambiental y la respuesta ante desastres naturales. Además, su estructura modular permite adaptarlo a necesidades específicas, lo que lo convierte en una herramienta versátil tanto para investigadores como para responsables de políticas ambientales. Este tipo de innovación subraya la importancia de integrar tecnología avanzada en la solución de problemas críticos como la regeneración ecológica después de eventos [36].

4 CONSTRUCCIÓN CONJUNTO DE DATOS

4.1 METODOLOGÍA DE DESARROLLO

Para el desarrollo del siguiente trabajo se empleó el modelo CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es ampliamente utilizado debido a su enfoque iterativo y adaptable. Este modelo facilitó la orientación del proceso analítico, abarcando desde la comprensión del problema hasta la implementación de soluciones fundamentadas en datos, también en el desarrollo de algoritmos se adoptaron metodologías tipo MBSE (Model-Based Systems Engineering) y SCRUM.

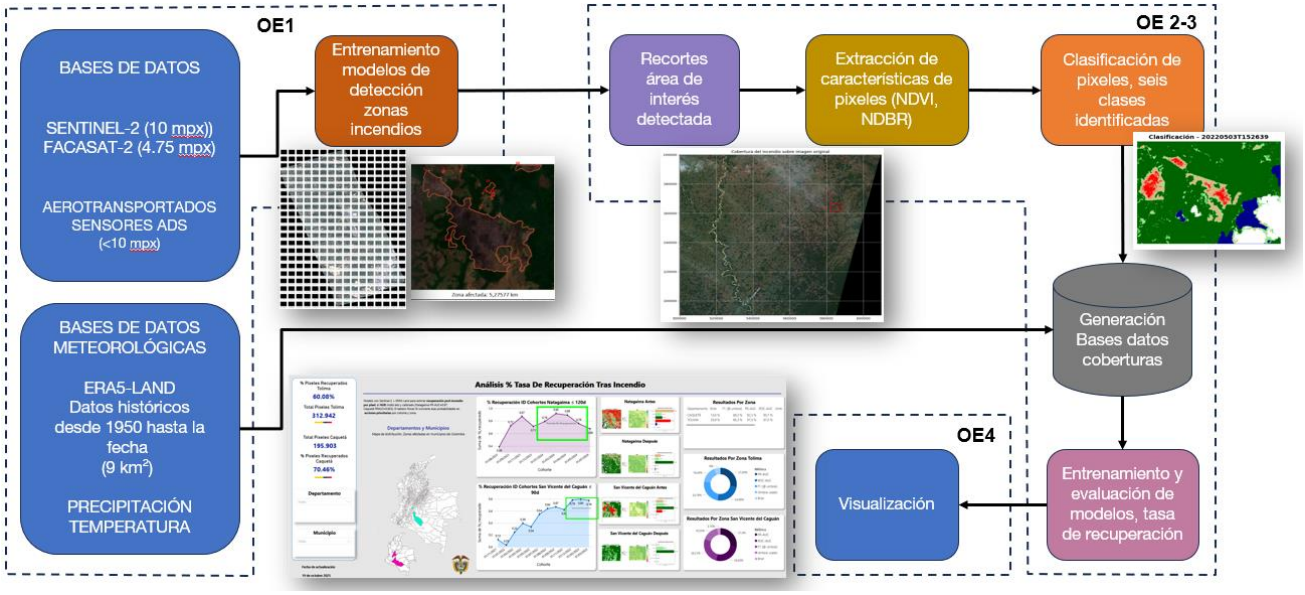

Figura 1. Flujo de trabajo metodología.

El proceso mencionado en la anterior figura, comienza con la definición del problema, que se basa en la necesidad de identificar de manera oportuna las áreas afectadas por incendios, con el fin de evaluar su recuperación posterior de manera objetiva, escalable y con una relación costo-beneficio que respalde la toma de decisiones en materia ambiental. Esta fase se traduce en objetivos analíticos precisos, los cuales se dividen en tres fases fundamentales: la primera consiste en el desarrollo de un algoritmo para la detección de incendios; la segunda se refiere a un algoritmo de clasificación multiclase; y la última se centra en un algoritmo destinado a evaluar la tasa de recuperación.

Posteriormente, se lleva a cabo la identificación y caracterización de los datos provenientes de las fuentes disponibles, que incluyen los sensores satelitales Sentinel-2A y FACSAT-2, así como los sensores digitales aerotransportados ADS. Esta etapa de la metodología también abarca la recopilación de datos históricos meteorológicos obtenidos de ERA5-land, con series que datan desde 1950. En este contexto, se realiza un análisis de la resolución espacial, temporal, cobertura y calidad radiométrica, lo que permite determinar la relevancia de los datos. Una vez recopilados, se procede a la fase de preparación de los datos, la cual es crítica dentro del flujo de trabajo. En esta fase, tras la

extracción de las características espectrales de los píxeles y los índices de NDVI/NDBR, se integran las variables climáticas, tales como la precipitación y la temperatura, generando así una base de datos estructurada de coberturas, lista para el entrenamiento. El resultado obtenido es un conjunto de datos consistente, especialmente alineado temporalmente y exento de ruido generado por estacionalidades prolongadas.

La fase de modelado consistió en el desarrollo de diversos procesos de entrenamiento de clasificación supervisada de píxeles, mediante la identificación de seis clases de cobertura en las áreas de interés, con el objetivo de llevar a cabo el entrenamiento para estimar la tasa de recuperación posterior a un incendio. Posteriormente, se realizó la evaluación de los modelos desde una perspectiva técnica y operativa, validando el desempeño de los clasificadores, así como la coherencia espacial y temporal de la recuperación estimada, y se llevó a cabo una comparación entre diferentes zonas, períodos y condiciones climáticas.

Finalmente, los resultados conducen a una fase de implementación y visualización, en la cual se ha desarrollado una herramienta interactiva de coberturas utilizando Power BI. Esta herramienta permite la visualización de los resultados de la tasa de recuperación en las regiones de Caquetá y Tolima, donde el valor del modelo seleccionado se traduce en información útil y aplicable para la toma de decisiones en materia ambiental y gestión de recursos.

4.2 DEFINICIÓN ZONAS DE INTERÉS

Los incendios en Colombia son monitoreados por diferentes agencias entre los que se destacan la Unidad Nacional para la Gestión del Riesgo de Desastres – UNGRD, cuyo objetivo es: “dirigir la implementación de la gestión del riesgo de desastres, atendiendo las políticas de desarrollo sostenible, y coordinar el funcionamiento y el desarrollo continuo del sistema nacional para la prevención y atención de desastres” [3], así como el Instituto de Hidrología, Meteorología y Estudios Ambientales – IDEAM cuya misión se enfoca en la producción de información confiable y oportuna sobre los recursos naturales y del medio ambiente [4], ambas entidades permiten la toma decisiones para influir directamente en el planteamiento de estrategias y políticas ambientales para la preservación de los recursos naturales.

De acuerdo con lo reportado por la UNGRD, en 2024 los incendios fueron los eventos más frecuentes en Colombia, donde se reportaron un total de 6.293 casos en los que se afectaron más de 216 mil hectáreas, la asistencia técnica para contrarrestar estos incendios forestales incluye la articulación de diferentes entidades a nivel Nacional, entre los que se destacan, Bomberos, PONALSAR, defensa civil, cruz roja, Fuerzas Militares entre otros, empelando 1.386 horas de vuelo y alrededor de 4.455 descargas de agua [47].

Para la selección del área de estudio se tomó la información histórica de incendios en Colombia reportadas por diferentes autoridades ambientales, este tablero de Dashboard (Figura 2) permitió

explorar los eventos reportados desde 2002 hasta 2023, esta información se filtró por año, departamento y municipios, lo cual fue de gran ayuda para identificar grandes zonas afectadas e iniciar la búsqueda de imágenes para descargar de Sentinel-2.

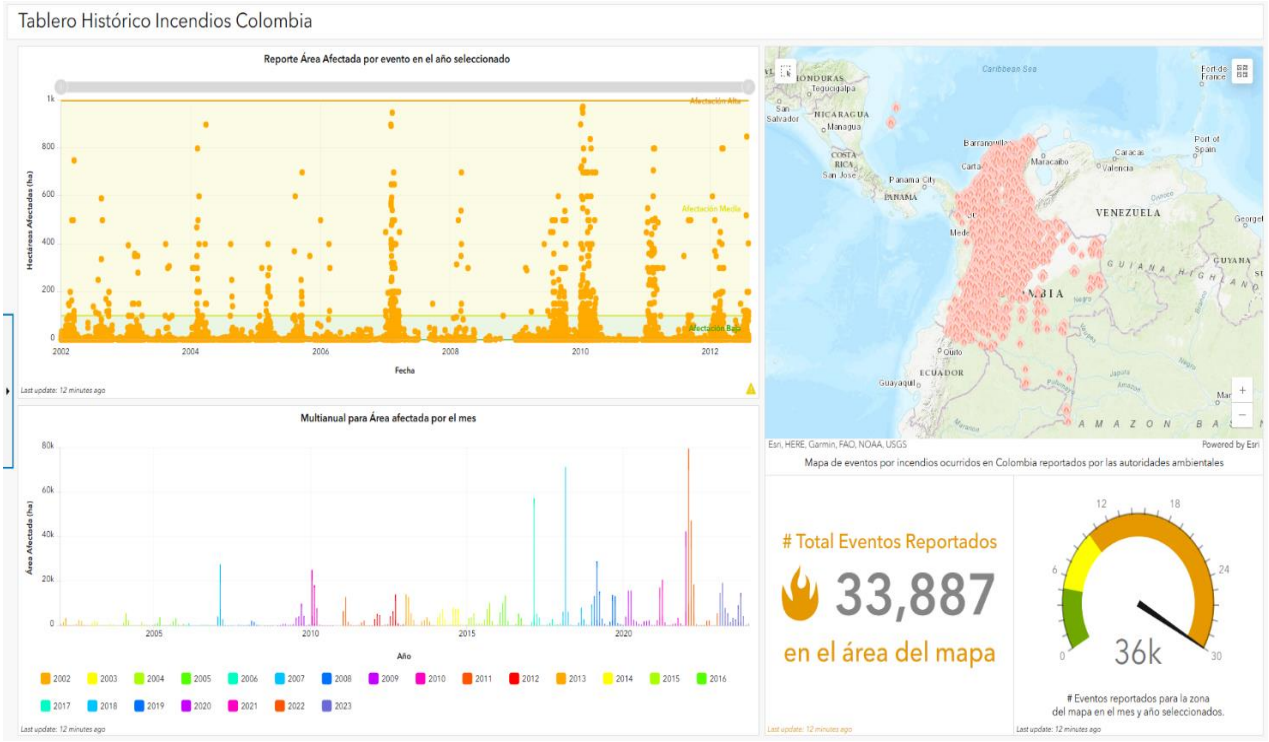

Figura 2. Dashboard – Tablero histórico de incendios en Colombia [48].

Se analizaron un total de 12 zonas de incendios forestales registrados en diferentes departamentos, en rojo se encuentran las zonas con alta densidad de nubes, en amarillo las de media densidad de nubes y verde las zonas que ofrecieron menor densidad de nubes y sombras, las cuales permitían un mejor análisis de temporalidad de acuerdo con las imágenes disponibles en Copernicus para la descarga. Para el análisis, se definieron dos zonas de interés: la primera en el departamento del Caquetá y la segunda en el departamento del Tolima, las cuales se presentan en la Tabla 1.

Tabla 1. Búsqueda de áreas de interés.

Fecha	Departamento	Municipio	Área_ha	Coordenadas
12/07/2023	VALLE DEL CAUCA	CALI	330	3.39909178/ -76.57632532000001
22/07/2021	VALLE DEL CAUCA	YOTOCO	200	3.90418281/ -76.39079706999999
14/08/2019	VALLE DEL CAUCA	YUMBO	175	3.59591367/ -76.5109879
13/08/2019	NARIÑO	POLICARPA	857	1.7353542/ -77.48134359000001
01/08/2023	NARIÑO	SANTACRUZ	650	1.2851799099999999/ -77.74457352
11/10/2023	HUILA	CAMPOALEGRE	1250	2.65868086/ -75.32941705

5/09/2019	HUILA	NEIVA	760	2.99336005/ -75.27236232
16/08/2023	HUILA	PALERMO	690	2.91449223/ -75.44065421
25/11/2019	HUILA	CAMPOALEGRE	300	2.65866668/ -75.32943803
15/06/2019	META	LA MACARENA	3000	2.16186392/-74.09488076000001
22/02/2022	CAQUETA	SAN VICENTE DEL CAGUÁN	408	1.6271/-741656
26/08/2023	TOLIMA	NATAGAIMA	50	3.7022/-75.1671

4.2.1 Zona departamento del Caquetá

De acuerdo con el boletín número 41 del cuarto trimestre de 2024, emitido por la subdirección de ecosistemas e información ambiental y el Sistema de Monitoreo de Bosques y Carbono (SMByC), se logró identificar que en la región de interés del incendio hay factores de deforestación que inciden en la causa de generación de incendios, este informe establece que el núcleo 4, se establece en el municipio de San Vicente del Cagúan, en las veredas del Camuya, Aguas Claras, Altagracia, altos del Yari y Caquetania, en donde las causas más comunes están asociadas a:

- Praderización para acaparamiento de tierras.
- Prácticas no sostenibles de ganadería extensiva.
- Infraestructura de transporte no planificada.
- Extracción de madera (tala ilegal).

Considerando lo anterior, se realiza la descarga de imágenes de este sector, en el cual se pueden encontrar grandes zonas de incendios, considerando que se realizan con conciencia para las actividades descritas anteriormente, la zona de interés cubre desde el 2021/11/16 hasta el 2023/03/21, la Tabla 2 muestra el análisis realizado a las 16 imágenes disponibles descargadas de Copernicus y en la Figura 3 se puede observar la ubicación del incendio registrado el 2022/02/22.

Tabla 2. Análisis de imágenes disponibles zona Caquetá.

Caquetá - Coordenadas 1.6271/-741656				
#	**Fecha**	**Apta**	**Observaciones**	**Peso Mb**
1	2021/11/16	SI	No nubosidad en el ROI	0.525
2	2021/11/19	NO	Nubosidad en el ROI	0.522
3	2022/01/03	NO	Nubosidad en el ROI	0.532
4	2022/01/08	SI	No nubosidad en el ROI	0.554
5	2022/02/22	SI	No nubosidad en el ROI (Fecha incendio)	0.526

Caquetá - Coordenadas 1.6271/-741656				
#	**Fecha**	**Apta**	**Observaciones**	**Peso Mb**
6	2022/03/24	SI	Se identifica bruma en una porción baja del ROI	0.546
7	2022/05/03	SI	Se identifica nubes sobre el ROI	0.594
8	2022/07/29	SI	No nubosidad en el ROI	0.538
9	2022/08/23	SI	No nubosidad en el ROI	0.523
10	2022/09/27	SI	No nubosidad en el ROI	0.535
11	2022/10/07	SI	Se identifica nubes/Sombras sobre una porción baja del ROI	0.556
12	2022/11/06	SI	Se identifica sombra sobre una porción baja del ROI	0.553
13	2022/12/19	SI	Se identifica nubes/Sombras sobre una porción baja del ROI	0.554
14	2023/02/02	NO	Nubosidad en el ROI	0.552
15	2023/02/19	SI	No nubosidad en el ROI	0.534
16	2023/03/21	SI	Se identifican unos pixeles no identificables en el ROI	0.535
	TOTAL	**13**	**Se identifican 13 imágenes para el análisis temporal**	

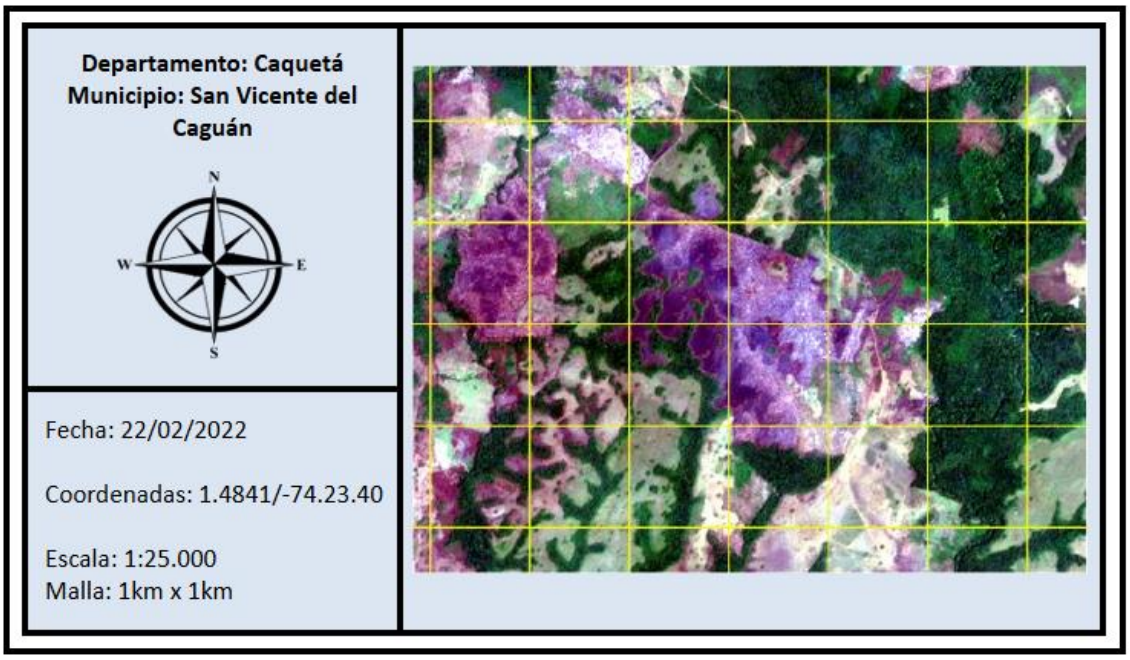

Figura 3. Zona de incendio departamento del Caquetá – ROI, Fuente: Diseño propio.

4.2.2 Zona Departamento del Tolima

La zona del sur del Tolima escogida para el estudio obedece a que este escenario es representativo para el mismo, esto debido a que esta zona es catalogada como una zona de alta frecuencia de incendios.

Adicional, la zona se encuentra dentro de la órbita 25 del satélite sentinel-2, lo que proporciona gran material de estudio, también cuenta con varias estaciones meteorológicas a su alrededor, donde debido a su biodiversidad nos proporciona bosques secos, bosques húmedos, cultivos y zonas con pastizales. Se descargaron una serie de imágenes que comprenden las fechas del 12 de junio de 2023 y el 16 de julio de 2024, las cuales cubren nuestra área de estudio localizada en el departamento del Tolima, Colombia, en municipios como Ibagué, Cajamarca, Rovira, San Luis y El Espinal. La Tabla 3 muestra el análisis realizado a las 10 imágenes disponibles descargadas de Copernicus y en la Figura 4 se puede observar la ubicación del incendio registrado el día 26-08-2023.

Tabla 3. Análisis de imágenes disponibles zona Tolima.

Tolima - Coordenadas 3.7022/-75.1671				
#	**Fecha**	**Apta**	**Observaciones**	**Peso Mb**
1	2023/06/12	SI	No nubosidad en el ROI	0.248
2	2023/08/26	SI	No nubosidad en el ROI (día incendio)	0.226
3	2023/09/30	SI	No nubosidad en el ROI	0.227
4	2023/12/09	SI	No nubosidad en el ROI	0.226
5	2023/11/14	SI	No nubosidad en el ROI	0.248
6	2024/01/23	SI	No nubosidad en el ROI	0.248
7	2024/03/18	SI	No nubosidad en el ROI	0.248
8	2024/04/07	SI	No nubosidad en el ROI	0.248
9	2024/05/07	SI	No nubosidad en el ROI	0.248
10	2024/07/17	SI	No nubosidad en el ROI	0.248
	TOTAL	**10**	**Se identifican 10 imágenes para el análisis temporal**	

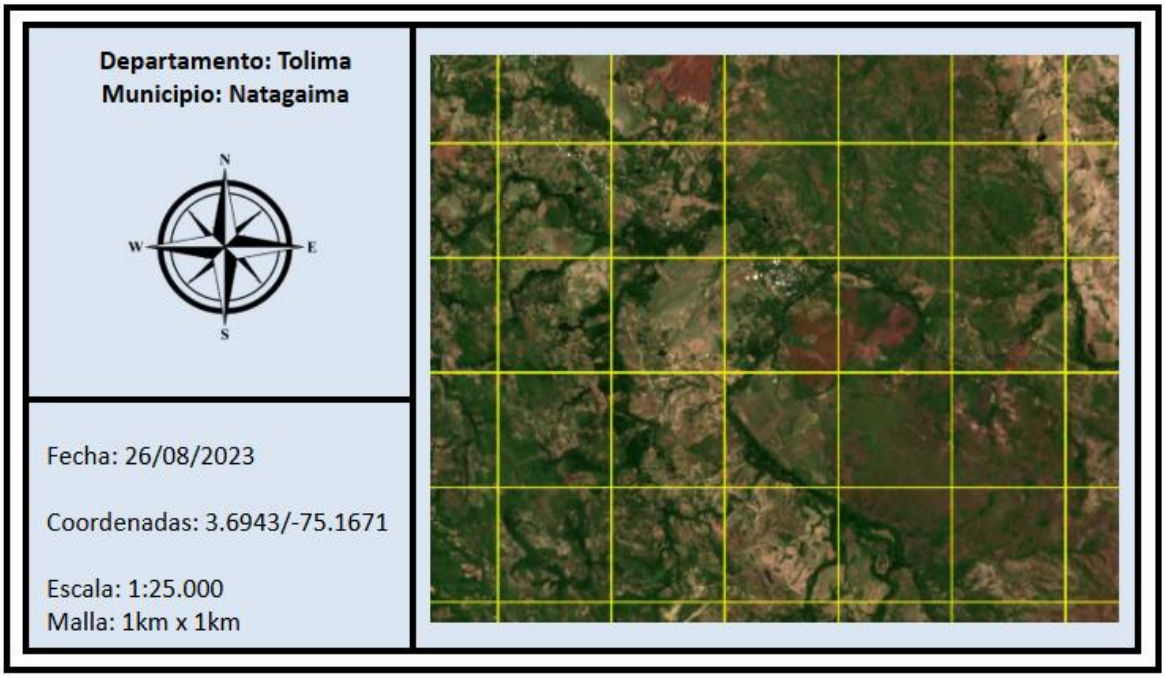

Figura 4. Zona de incendio departamento del Tolima – ROI, Fuente: Diseño propio.

4.3 CONSTRUCCIÓN DE BASE DE DATOS SATELITALES Y AMBIENTALES.

Con el propósito de lograr el primer objetivo, se implementaron una serie de etapas metodológicas que permiten la integración de datos provenientes de sensores remotos y datos meteorológicos, con el fin de iniciar el análisis de la recuperación de la vegetación afectada por incendios forestales.

4.3.1 Datos Satelitales.

Los datos satelitales son claves en este proyecto para la recolección de información de las zonas de interés sobre las cuales se desarrollará este trabajo, se revisaron diferentes plataformas para la descarga de datos, entre las que se encuentran Copernicus, Google Engine y Planet, se decidió trabajar con Copernicus debido a la facilidad de registro y acceso a imágenes de Colombia, las imágenes descargadas tienen un nivel de post-procesamiento de nivel 2A, que incluye correcciones geométricas, radiométricas y atmosféricas [49] [50].

Teniendo en cuenta que la fase inicial requiere el desarrollo de un modelo que permita identificar las zonas de incendios, se emplearon imágenes de diferentes sensores satelitales entre las que se destacan FACSAT-2, sensores aerotransportados digitales - ADS y Sentinel-2, para la construcción de la base datos, los cuales se presentan en la Tabla 4.

Tabla 4. Imágenes descargadas por sensor.

Sensor	Imágenes	Resolución m/px	Temporalidad	Acceso
SENTINEL 2	23	10 / 20	SI	SI
FACSAT-2	130	4.75	NO	SI
ADS	21504	<10	SI	Restringido

4.3.2 Preparación de los Datos

Una vez que se identificó la fuente de datos, se verifican las características de las mismas, como se indica en la Tabla 4, para el caso de FACSAT-2, el principal problema radicó en la carencia de temporalidad, considerando que no se cuentan con más de 3 imágenes de los puntos de interés, para las imágenes provenientes de sensores ADS su limitación se debe al uso restringido de las imágenes, por lo tanto, las imágenes de Sentinel-2 son la opción más acertada para efectuar un análisis de tasa de recuperación, las otras imágenes de los sensores restantes se emplearan para el entrenamiento y detección de zonas de incendios.

4.3.2.1 Sentinel-2

Una vez descargadas las zonas de interés para el área de Caquetá y Tolima, se debe trabajar en las imágenes para ajustarlas a una misma resolución de (10 m), teniendo en cuenta que se emplearon imágenes de Sentinel 2A, algunas de las bandas descargadas vienen originalmente en 20 m, por lo tanto, el código desarrollado permite realizar la interpolación espacial de las bandas de 20 m (B05, B06, B07, B8A, B11, B12) a 10 m, utilizando re-muestreo cúbico, recorte espacial del área del polígono y definición del área de interés para todas las bandas de 20m, el resultado fue una recolección de archivos .tif con nomenclatura uniforme por banda y por imagen, interpolada de 20 a 10 m.

Tabla 5. Total escenas por zona de interés.

ZONA	# imágenes	# bandas 10m	# bandas 20	Total bandas ZOI
Caquetá	13	04 (B,G,R,NIR)	06 (R1, R2, R3, NIR-a, SWIR1, SWIR2)	130 Recortes
Tolima	10	04 (B,G,R,NIR)	06 (R1, R2, R3, NIR-a, SWIR1, SWIR2)	100 Recortes

Una vez completado este proceso, se realiza el reescalado de pixeles, para cambiar la escala de cada imagen recortada y pasarlo a valores entre 0 y 1, como estrategia para reducir tamaños de almacenamiento y tiempos de procesamiento (multiplicando por el factor de cuantificación 0.0001).

4.3.2.2 Sensor FACSAT-2 y sistema aerotransportado digital

Se obtuvo acceso a un total de 130 imágenes del territorio colombiano a través de FACSAT-2, las cuales se centran en las regiones con mayor incidencia de incendios forestales. Sin embargo, esta cantidad de imágenes resulta insuficiente para el entrenamiento de un modelo destinado a la detección de áreas afectadas por incendios. En consecuencia, se opta por trabajar con un conjunto más amplio de 21,504 imágenes, de las cuales se logró acceder a 3,186 que fueron desclasificadas para los fines de este estudio.

La Figura 5 sintetiza el proceso de filtrado utilizado en la construcción final de la base de datos. En primer lugar, se aplicó un filtro que relaciona las imágenes a las cuales se otorgó acceso para este estudio. Posteriormente, se implementó un segundo filtro que identificó aquellas imágenes en las que se registraron zonas afectadas por incendios. Como resultado de este proceso, se obtuvo un total de 526 imágenes y 2630 aplicando técnicas de aumentación de datos para incrementar el tamaño del dataset, las cuales fueron utilizadas para el entrenamiento de los modelos de detección de incendios.

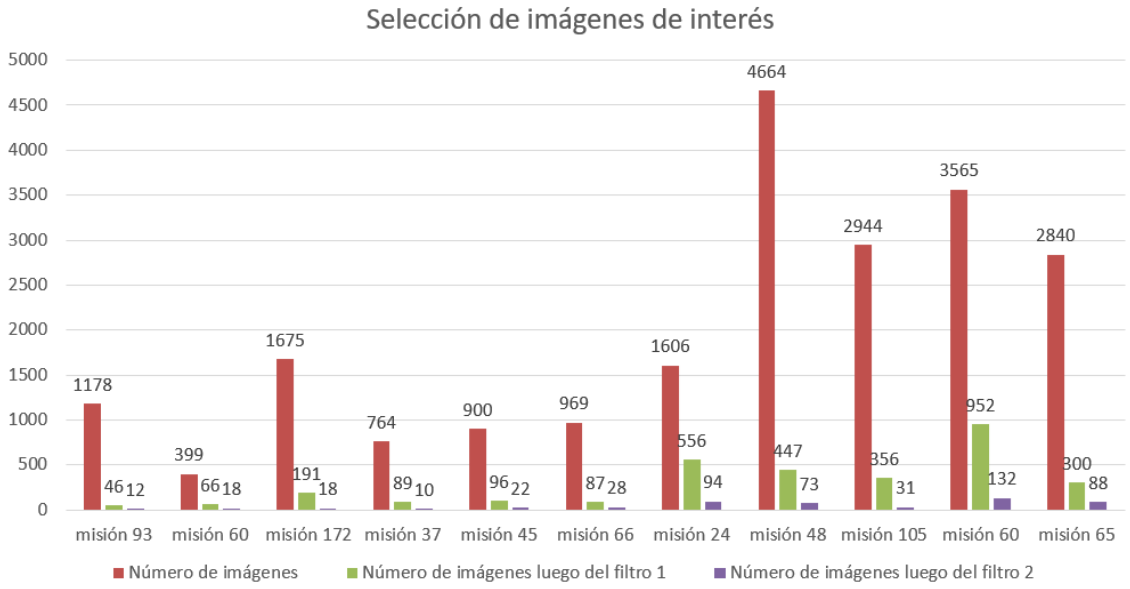

Figura 5. Selección de imágenes de interés, Fuente: Diseño propio.

4.3.3 Recolección de datos meteorológicos

Teniendo en cuenta la importancia de los datos meteorológicos para ser integrados a este estudio se realizó la verificación de datos con el portal del IDEAM, con el fin de identificar las estaciones meteorológicas más cercanas a la zona de estudio, priorizando aquellas localizadas en los departamentos previamente mencionados. Se recopilaron datos de las siguientes variables:

- Temperatura máxima (°C).
- Temperatura mínima (°C).
- Precipitación (mm).

El período de análisis abarcó desde noviembre de 2021 hasta marzo de 2023 para la primera zona, y desde junio de 2022 hasta julio de 2024, considerando:

- De uno a tres meses previo a los incendios.
- El período de impacto (momento del evento).
- Y el inicio del seguimiento a la recuperación post-incendio. (meses de recuperación).

La Tabla 6 y Figura 6 muestran las estaciones meteorológicas cercanas a los departamentos de estudio, donde se puede evidenciar que varias no tienen datos disponibles para los periodos de análisis [51].

Tabla 6. Ubicaciones estaciones meteorológicas zona Caquetá y Tolima.

Estación	Coordenada (LAT/LON)	Distancia al evento (KM)	Data IDEAM	Fecha estudio
Remolinos del Caguán	0.600 / -74.417	103	2001-03-01	2021/11/16 al 2023/03/2021
Cartagena del Chaira	1.350 / -74.846	69.8	2024-09-01	
Santa Rosa del Caguán	1.745 / -74.785	65.9	2024-05-01	
La macarena	2.341 / -73.890	98.7	2025-06-01	
San Vicente del Caguán	2.100 / -74.774	87.5	2019-03-01	
Natagaima	3.600 / -75.100	13.5	2010-11-01	2023/06/12 al 2024/07/16
Colache hacienda	3.766 / -75.237	10.8	2021-04-01	
Media Luna	3.775 / -75.120	10.1	2009-09-01	

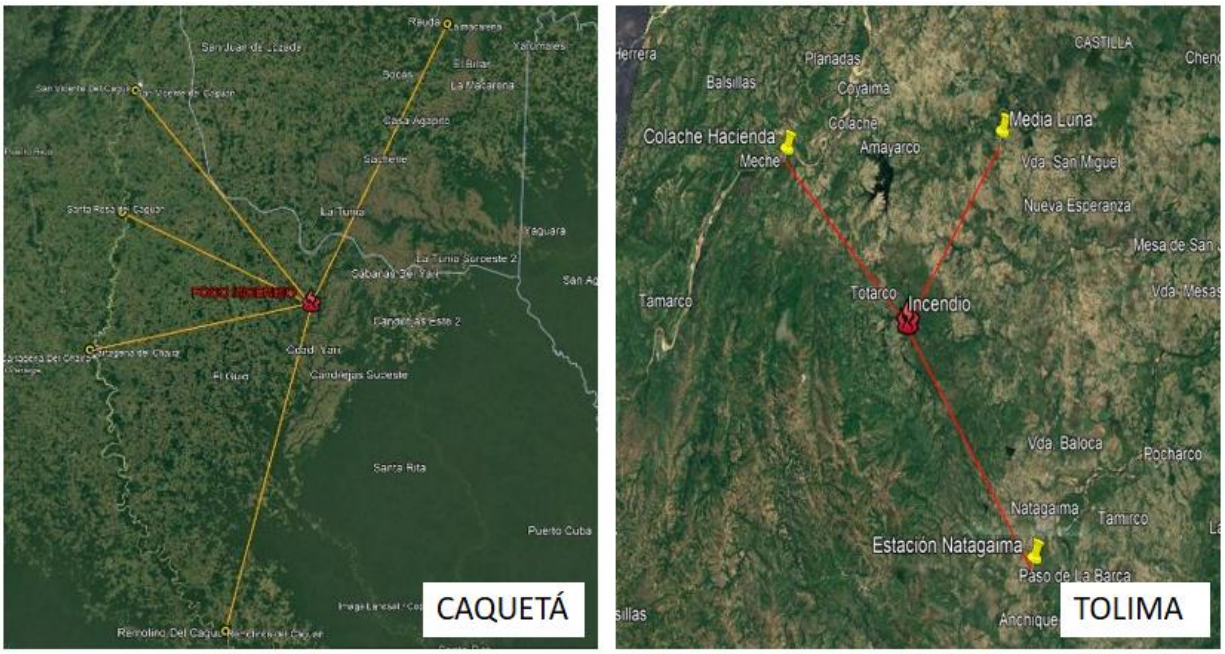

Figura 6. Distancias estaciones meteorológicas y conatos, fuente: Diseño propio.

Una vez analizada la información disponible, se logra evidenciar la falta de datos en las estaciones meteorológicas más cercanas a los incendios, por lo tanto, se decidió buscar datos de otras fuentes que permitieran obtener la temperatura y precipitación de las zonas de estudios con una buena resolución espacial. Se analizaron diferentes bases de datos históricas de clima como Worldclim y ERA5, se decidió usar la de ERA5-Land data set, la cual es de acceso público y contiene datos históricos desde 1950 hasta 5 días antes de la fecha actual con una resolución de 9 km². (se descartó Wordclim porque solo maneja datos históricos de 1970 hasta 2000).

4.3.4 Variables del clima y su preparación

Para la obtención de información climática, se empleó la fuente ERA5-Land, la cual es de acceso abierto y proporciona datos adecuados para las regiones de Tolima y Caquetá. Se seleccionaron los datos diarios correspondientes a la precipitación y a la temperatura (mínima, media y máxima). Para garantizar la adecuada preparación de los datos, se llevaron a cabo tres procesos específicos.

- Limpieza: se procedió a la estandarización de las unidades de medida en función de cada unidad, utilizando milímetros (mm) y grados Celsius (°C).
- Continuidad: se garantizó la integridad de los datos temporales, asegurando que no existieran días ausentes en el registro. En caso de que un día no estuviera documentado, se realizó una interpolación de la temperatura, mientras que la precipitación no registrada se consideró como cero.

- Información relevante: se consideraron de manera separada los acumulados y promedios climáticos en dos intervalos temporales; 7 días, que evidenciaban la variabilidad inmediata, y 30 días, que reflejaban un comportamiento más estable asociado a la escala mensual.

Posteriormente, se incorporaron estos datos a cada una de las fechas analizadas, de modo que cada píxel de la imagen se asoció con las variables climáticas pertinentes. Este enfoque facilitará la correlación entre los patrones territoriales y las variables climáticas de precipitación y temperatura.

4.4 GENERACIÓN DATASET PARA CIENCIA DE DATOS

Una vez que se han organizaron las imágenes a utilizar y sus correspondientes datos meteorológicos, se procedió a la elaboración de un conjunto de datos estructurado por escena. Cada escena se asoció a una fecha específica de las imágenes obtenidas mediante el satélite Sentinel-2, las cuales fueron transformadas en un archivo en formato Parquet (.parquet). Este formato facilito la optimización del almacenamiento y la velocidad de lectura en el manejo de grandes volúmenes de datos. Cada archivo contiene información estructurada en píxeles, que incluye la siguiente información:

Asignación Espacial: cada escena se vinculó con la información meteorológica proporcionada por ERA5, teniendo en cuenta la coordenada geográfica analizada con una resolución aproximada de 9 km, bajo la premisa de homogeneidad de las condiciones climáticas locales.

Asignación Temporal: para cada fecha de captura de imagen, se asignaron los registros diarios correspondientes de:

- Temperatura máxima, temperatura mínima, precipitación acumulada.
- x: coordenada este (Easting) en metros, sistema UTM.
- y: coordenada norte (Northing) en metros, sistema UTM.
- fecha de captura de la imagen (extraída del nombre del archivo).
- B02 – B12: valores de reflectancia limpia por cada una de las bandas seleccionadas:
  B02 (Azul), B03 (Verde), B04 (Rojo), B05–B07 (Red Edge), B08 (NIR), B08A (NIR angosto), B11 y B12 (SWIR), NDVI y NBR.
- Clase: Clasificación umbral (Nubes, Sombra, Incendio, Bosque, Pastizal y Suelo).

Adicionalmente, para cada escena se integraron todas las bandas con valores de reflectancia y máscaras de umbral en un único DataFrame con coordenadas x, y, bandas: banda_B02 hasta banda_B12, umbral_B02 hasta umbral_B12, se definió que los valores 0.0 de reflectancia o códigos no válidos que están fuera del rango, serían marcados como NaN, cada dataset fue guardado en formato .parquet, manteniendo valores válidos [44].

La generación del dataset implica la lectura de cada imagen procesada, la extracción de valores por

píxel y la consolidación en registros organizados. Se excluyó automáticamente cualquier píxel cuya máscara indicará la presencia de nubes, sombras o información no valida, la Tabla 7 presenta un resumen de las seis primeras variables y una escena para las zonas de estudio.

Tabla 7. Variables bases de datos zona Caquetá y Tolima.

DEPARTAMENTO TOLIMA									
clas	**fecha**	**lon**	**lat**	**B02**	**B04**	**B08**	**NDVI**	**NBR**	**clase**
clas_20231114	20231114	-75,2090	3,7472	0,1448	0,159	0,4532	0,4806	0,3021	5
clas_20231114	20231114	-75,2089	3,7472	0,1415	0,1478	0,4476	0,5035	0,3147	5
clas_20231114	20231114	-75,2088	3,7472	0,1458	0,1553	0,4299	0,4692	0,3083	5
clas_20231114	20231114	-75,2087	3,7472	0,1401	0,1444	0,434	0,5007	0,3177	5
clas_20231114	20231114	-75,2086	3,7472	0,136	0,1398	0,5107	0,5702	0,3916	4
clas_20231114	20231114	-75,2085	3,7472	0,1487	0,1542	0,5428	0,5575	0,4174	4
DEPARTAMENTO CAQUETÁ									
clas	**fecha**	**lat**	**lon**	**B02**	**B04**	**B08**	**NDVI**	**NBR**	**clase**
clas_20230612	20230612	1,4631	-74,2413	0,1097	0,1191	0,4348	0,5700	0,3585	4
clas_20230612	20230612	1,4631	-74,2413	0,139	0,1387	0,3988	0,4839	0,3942	4
clas_20230612	20230612	1,4631	-74,2413	0,1497	0,1379	0,462	0,5403	0,3025	4
clas_20230612	20230612	1,4631	-74,2413	0,1459	0,1353	0,434	0,5247	0,3147	4
clas_20230612	20230612	1,4631	-74,2412	0,1118	0,1205	0,4388	0,5691	0,3852	4
clas_20230612	20230612	1,4631	-74,2412	0,14	0,1387	0,4044	0,4892	0,3251	4

4.4.1 Universo de pixeles área de estudio

Se refiere a todos los píxeles del área de estudio, sin aplicar filtros: comprende aquellos que son afectados por el área de incendio, aquellos que se mantienen estables (no afectados) y también los que quedan sin información debido a la presencia de nubes o sombras en la escena, por lo que el NDVI se clasifica en nulo - NaN. La siguiente figura presenta la distribución de estos pixeles:

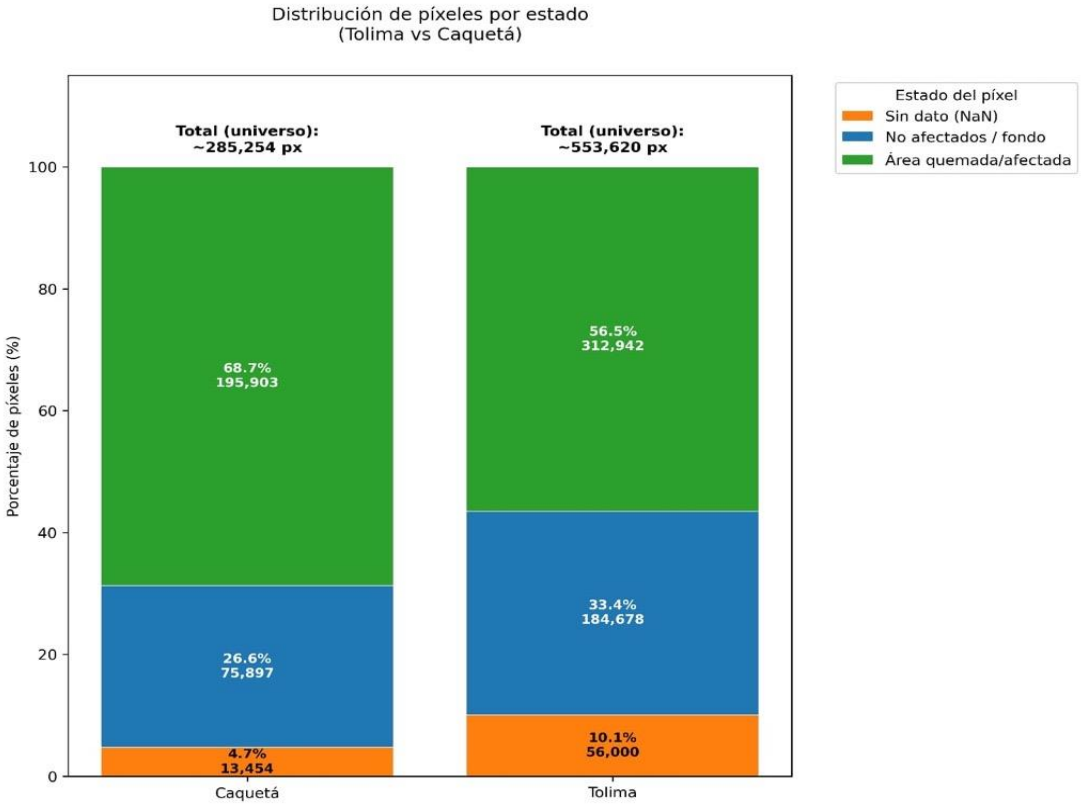

Figura 7. Universo total de pixeles zonas de interés

En Caquetá, una proporción significativa de la superficie presenta píxeles con indicación de afectación (69 %) en comparación con Tolima (57 %), lo que sugiere un evento de incendio más prevalente o predominante en el área examinada. Por el contrario, en Tolima se observa un porcentaje superior de píxeles sin información (~10 % en comparación con ~4,7 %). Esto se alinea con una mayor incidencia de nubes/sombras o dificultades en la observación, y también justifica la necesidad de implementar filtros más rigurosos para depurar el universo inicial.

4.4.2 Universo - episodios de incendio

Para la generación del dataset de "episodios de incendio", se seleccionan exclusivamente los píxeles afectados que cumplen los filtros de calidad (por ejemplo, consistencia temporal y eliminación de datos NaN problemáticos debido a nubes/sombras). Dicho conjunto, es el que verdaderamente sustenta el modelo. En este punto ya no quedan píxeles “de fondo” ni píxeles sin dato: es, básicamente, el universo final depurado de afectación sobre el que se entrena y evalúa el modelo.

5 IMPLEMENTACIÓN DE ALGORITMOS.

En esta sección se incluye el desarrollo de los tres algoritmos desarrollados durante el proyecto, el primero para la detección inicial de las zonas de incendios, el segundo para la clasificación del área de interés y el tercero para la estimación de la tasa de recuperación.

5.1 DETECCIÓN DE LAS ZONAS DE INCENDIO

El primer algoritmo fue diseñado para la detección de áreas afectadas por incendios. Para ello, se utilizó un conjunto de datos proveniente de los sensores ADS y se aplicó un enfoque de aprendizaje supervisado, que incluyó fotointerpretación por parte de un experto de dos clases, la primera las zonas de incendios asociadas a deforestación y la segunda a bosques, con el fin de detectar y segmentar adecuadamente las áreas de interés. Posteriormente, se llevó a cabo el entrenamiento de la red neuronal utilizando un conjunto de datos compuesto por 526 imágenes, al cual se le aplicaron técnicas de aumentación de datos (ver Figura 8), para incrementar el número de imágenes en el dataset, rotándolas 90° en sentido horario, efectuando patrones espejo y realizando zoom a las imágenes.

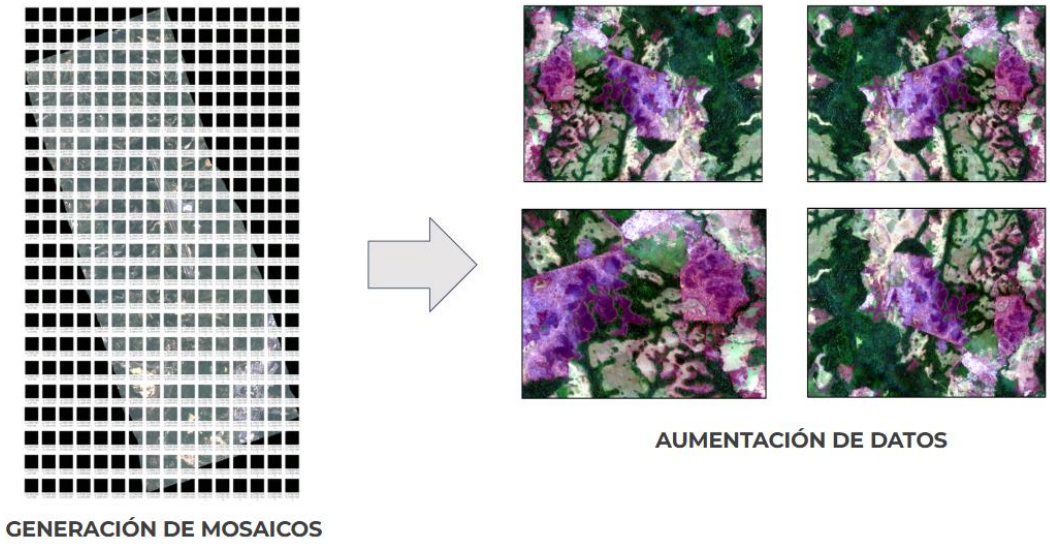

Figura 8. Generación mosaicos y aumentación de datos, fuente: Diseño propio.

Asimismo, se modificó el código encargado de la carga de datos ráster con el propósito de facilitar el manejo de rásteres de mayor tamaño, lo que amplía su aplicabilidad más allá de las dimensiones de las imágenes satelitales. El proceso de selección del modelo se encuentra estructurado dentro del flujo de trabajo representado en la Figura 9. Este marco permite la definición de un objetivo específico y la determinación de las características de la zona de estudio, facilitando así su posterior procesamiento y análisis.

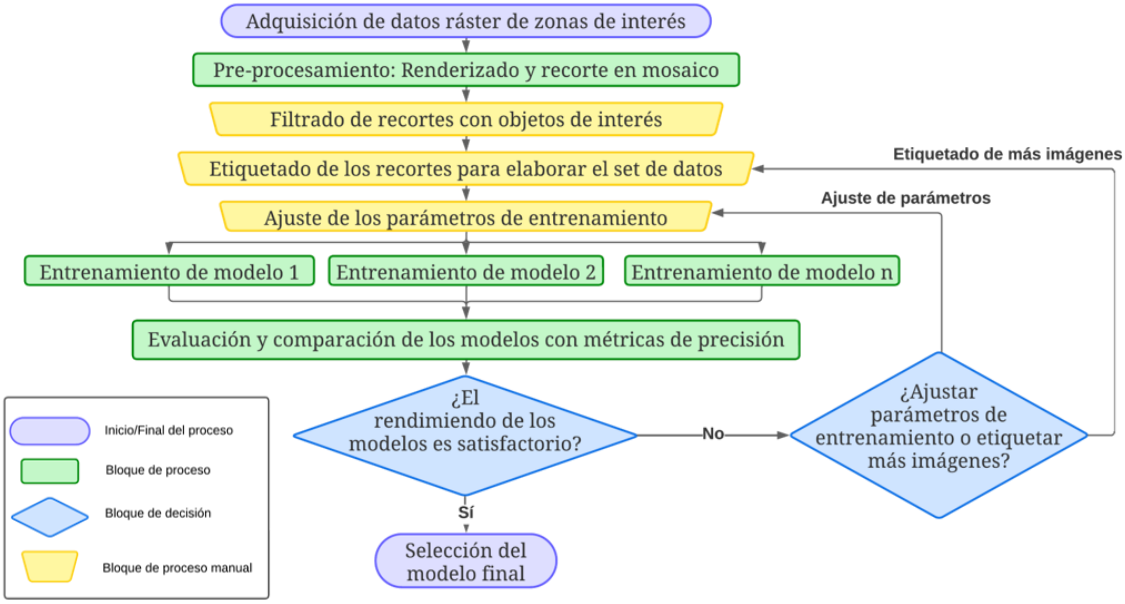

Figura 9. Flujo de trabajo selección modelo, fuente: Diseño propio.

Para el entrenamiento de los modelos se tuvieron en cuenta dos arquitecturas, Mask R-CNN y Yolov, de las cuales la última mostro mejor rendimiento en cuanto a velocidad y eficiencia, considerando que la detección de la zona afectada es el objetivo principal se escoge detección y eficiencia sobre segmentación y precisión, en la Figura 10 y Figura 11 se presentan los resultados de las versiones empleadas de Yolov y sus métricas.

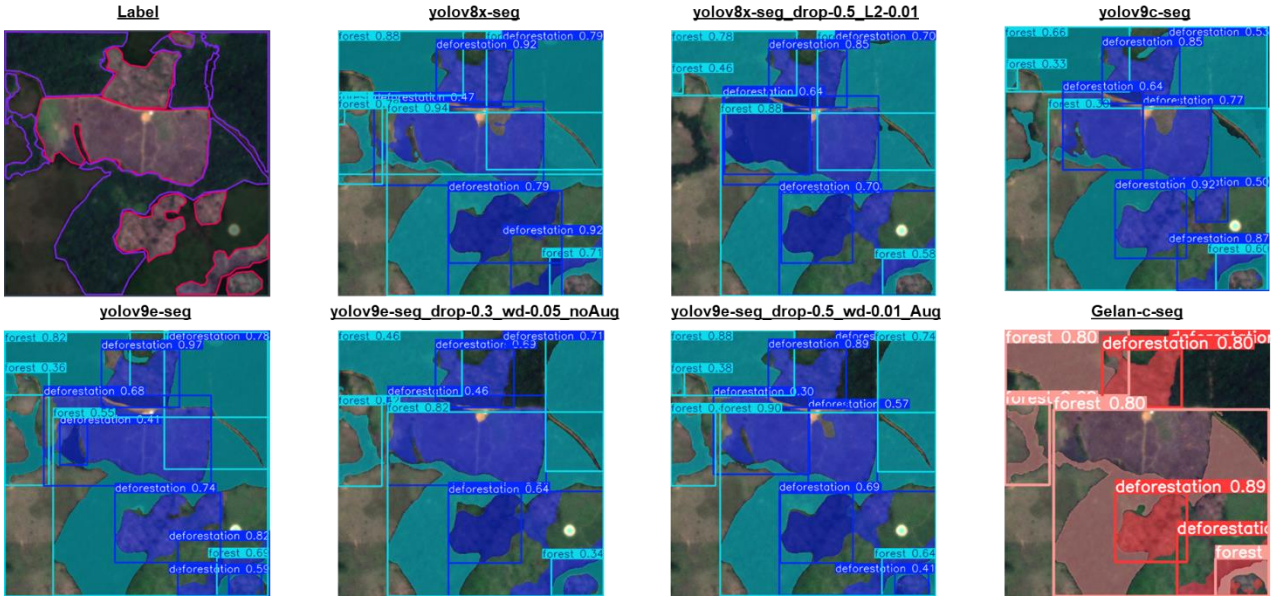

Figura 10. Algoritmos entrenados para detección de zonas deforestadas, Fuente: Diseño propio.

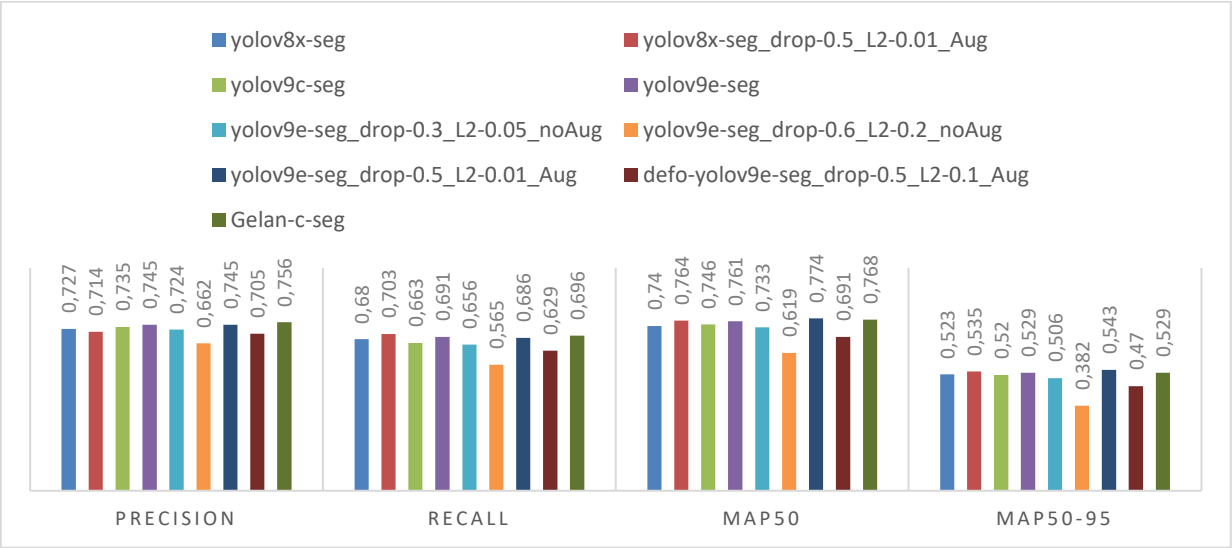

Figura 11. Validación Métricas segmentación por instancias para incendios, fuente: Diseño propio.

En la figura anterior, se puede observar el resultado del entrenamiento con una precisión mayor al 70% de detección de la zona del incendio, se decide trabajar con la versión Yolov9e-seg_drop-0.5, la cual ofrece un Map50 (Mean Average precisión) de 0.774, lo cual sugiere que tiene el mejor promedio para todas las clases de objetos que se desean detectar, para este análisis en particular incendios y bosques.

5.1.1 Recorte de la zona de estudio

Una vez que se detectó la zona, se llevaron a cabo diversos tipos de recortes con el objetivo de reducir el área total a la correspondiente a nuestra área de interés. Este proceso incluye la georreferenciación de la zona de interés, extrayendo las coordenadas de las imágenes para ser visualizadas, tal como se ilustra en la Figura 12.

Posterior a la identificación del departamento y municipio donde ocurrió el evento del incendio, se establece un punto geográfico central utilizando las coordenadas correspondientes al píxel de cada imagen, lo cual representa el centro del área afectada por el incendio forestal. Este aspecto permite definir una referencia precisa para la construcción de la región de análisis de 6 km × 5 km, lo que asegura un enfoque localizado y representativo del fenómeno objeto de estudio.

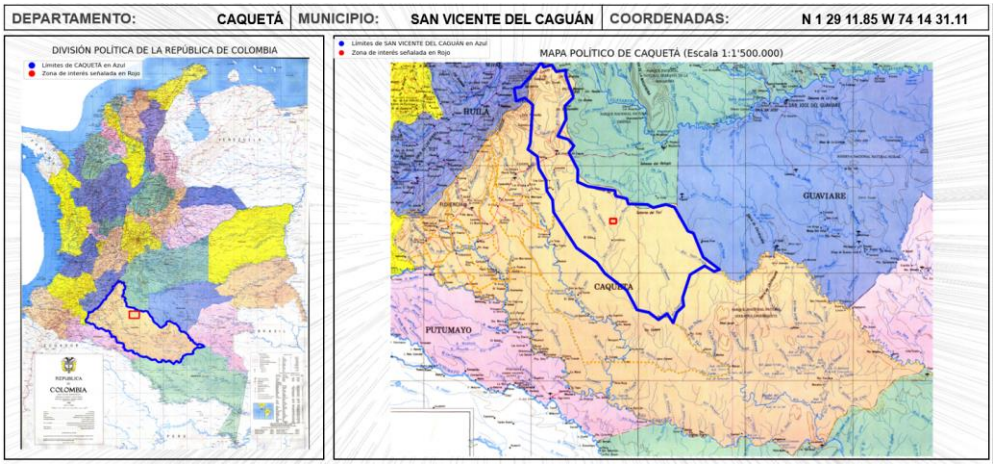

Figura 12. Georeferenciación y gráfica zona afectada, fuente: Diseño propio.

Este polígono funcionó como una máscara espacial para delimitar las imágenes satelitales de cada zona de estudio, lo que posibilitó restringir el análisis exclusivamente a los píxeles ubicados dentro del área afectada y sus alrededores. Esta alternativa optimiza la eficiencia del análisis y asegura la coherencia espacial de los datos que fundamentarán el modelo de recuperación post-incendio, el cual representa el objetivo principal de este estudio.

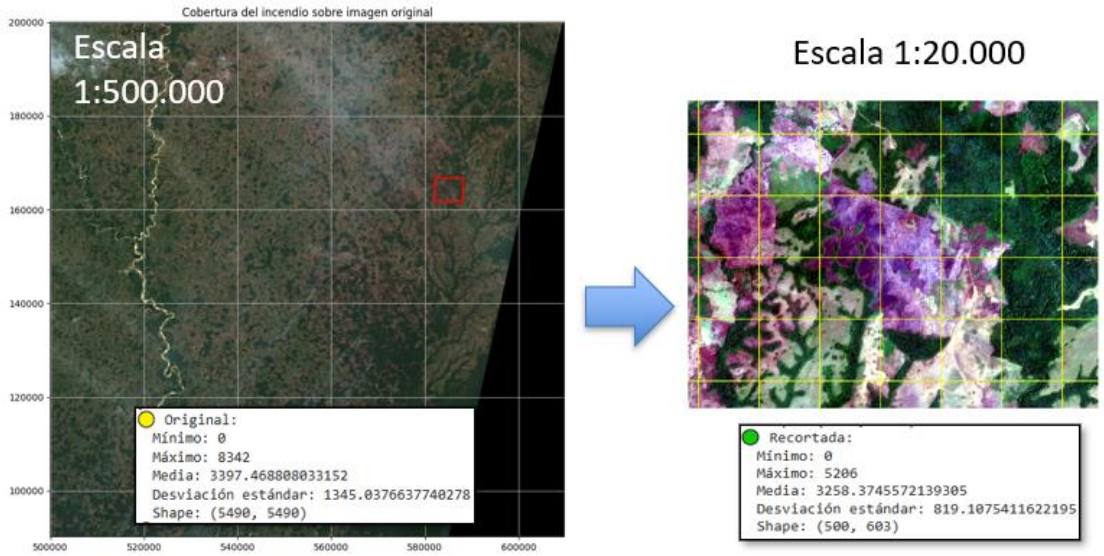

Figura 13. Proceso recorte área de interés 10m Sentinel 2, fuente: Diseño propio.

La Figura 13 facilita la verificación del recorte de la imagen original, enfocando la atención en el área de estudio. La disminución en la cantidad de píxeles es considerable. La misma dinámica fue aplicada a todas las bandas empleadas en el presente análisis para las imágenes con resoluciones de 10 m y 20 m.

El software desarrollado facilita la clasificación de todas las áreas afectadas por incendios dentro del ráster descargado, generando un informe para cada una de estas zonas. Este informe incluyo la

estimación del área afectada en kilómetros cuadrados, así como la ubicación geográfica correspondiente, tal como se presenta a continuación:

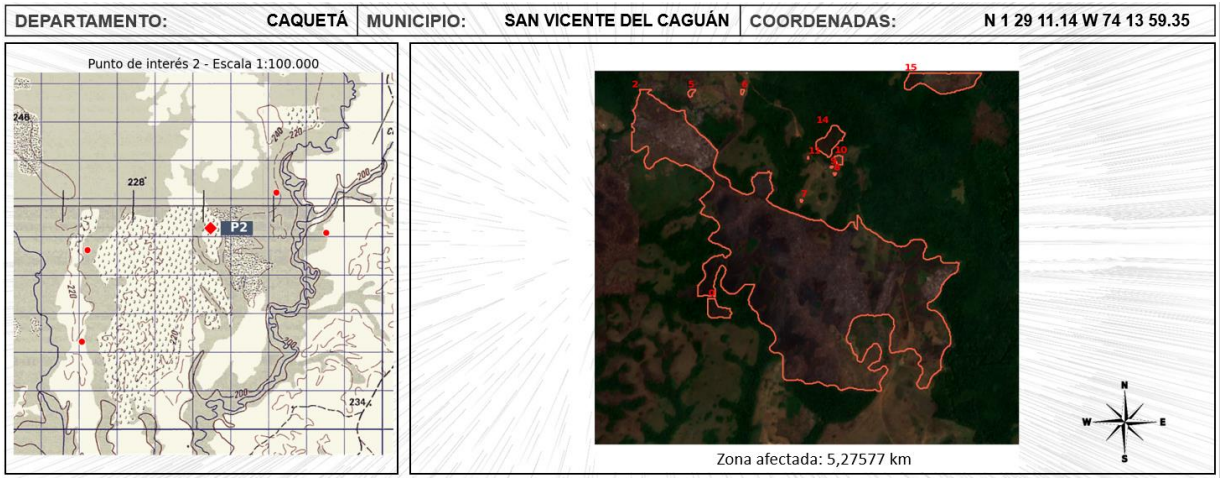

Figura 14. Análisis zona afectada por incendio ROI – Escala 1:15.000, fuente: Diseño propio.

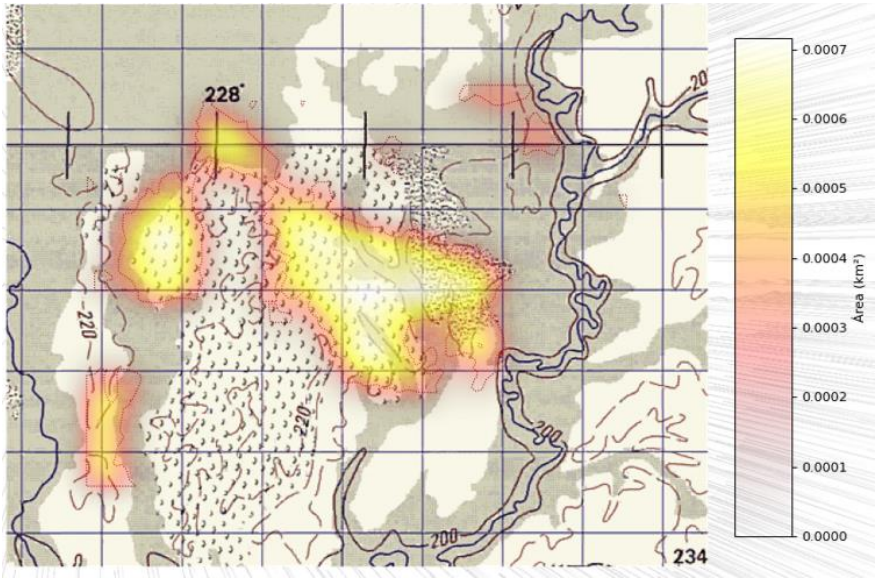

Figura 15. Detección zona afectada por el incendio, fuente: Diseño propio.

También es importante resaltar que una vez que se segmento el área de interés como la presentada en la Figura 14, el software desarrollado permitió generar mapas de calor dividiendo el territorio en celdas más pequeñas y calculando, en cada una, la cantidad total de área afectada en kilómetros cuadrados. Luego, estos valores se suavizaron para que la transición entre celdas fuera gradual y el resultado fuera una imagen continua. Finalmente, la barra de color tradujo los colores en valores de área, indicando cuántos km² corresponden a cada tono: colores más arriba en la barra indican zonas con mayor concentración de afectación, mientras que colores más abajo indican menor

concentración como se presenta en la Figura 15.

5.2 CLASIFICACIÓN DE CLASES

Una vez que se han recortado todas las escenas y se ha delimitado el área de interés (ROI), se procede al desarrollo del segundo algoritmo, el cual tiene como objetivo la clasificación de estas imágenes. Para llevar a cabo este proceso, es fundamental establecer los criterios de clasificación en función del objetivo principal, que es la tasa de recuperación. Para definir estas clases, se utiliza como referencia la clasificación propuesta por Copernicus [52], realizando una integración de algunas de estas categorías, tal como se detalla a continuación:

Tabla 8. Homologación etiquetas.

Valor	Descripción	Homologación	Clase
N/A	Se emplea mascara SCL = 8, 9, 10 y 11	Nubes	1
N/A	Mascara SCL = 3	Sombra nubes	2
N/A	NDVI <= 0.10	Incendios	3
125	Bosque abierto mixto	Bosques	4
30	Vegetación herbacea	Pastizales	5
60	Vegetación desnuda escasa	Suelo desnudo	6

Este proceso representa un aspecto fundamental en la clasificación de las diversas categorías asociadas a cada imagen y sus respectivas bandas. La clasificación inicial se realiza directamente sobre las imágenes, utilizando rangos de valores específicos (umbrales) que han sido validados a través de un análisis espectral de la zona. Estos umbrales se han tomado como referencia a partir de diversas fuentes bibliográficas, obtenidas mediante una revisión sistemática de los umbrales comunes para las clases en cuestión. Además, se considera la opinión de expertos del IDEAM, quienes señalan que dichos umbrales pueden no ser aplicables en todo el territorio nacional debido a diversos factores como la topografía, vapor de agua, estructura de la vegetación, ángulo del sol y ángulo del sensor. Por lo tanto, se propone la utilización y combinación de algoritmos automatizados para llevar a cabo la clasificación. Este enfoque garantiza una mayor flexibilidad y precisión en la clasificación de las diferentes clases, al adaptarse a las características específicas del área en cuestión.

5.2.1 Clasificación por umbrales

Una de las etapas más significativas en el procesamiento de imágenes satelitales es la introducción de nubes, sombras y otros componentes atmosféricos que pueden deteriorar los valores de reflectancia y ocasionar inexactitudes en la determinación de los índices espectrales o modelos requeridos para el objeto de estudio [53]. Las imágenes de Sentinel-2 vienen con una capa de

clasificación de escena (SCL) para detección de nubes ya preparada para clasificación, por lo tanto, estas pueden ser empleadas para identificar nubes dentro del área de estudio, sin embargo, en los análisis realizados se puede observar que no es muy buena identificando las nubes de la máscara SCL 9 (Cirrus delgados) como se puede apreciar en el caso 6 de la Figura 16, en donde en el recuadro amarillo no se realiza la correcta clasificación de la nube y su sombra.

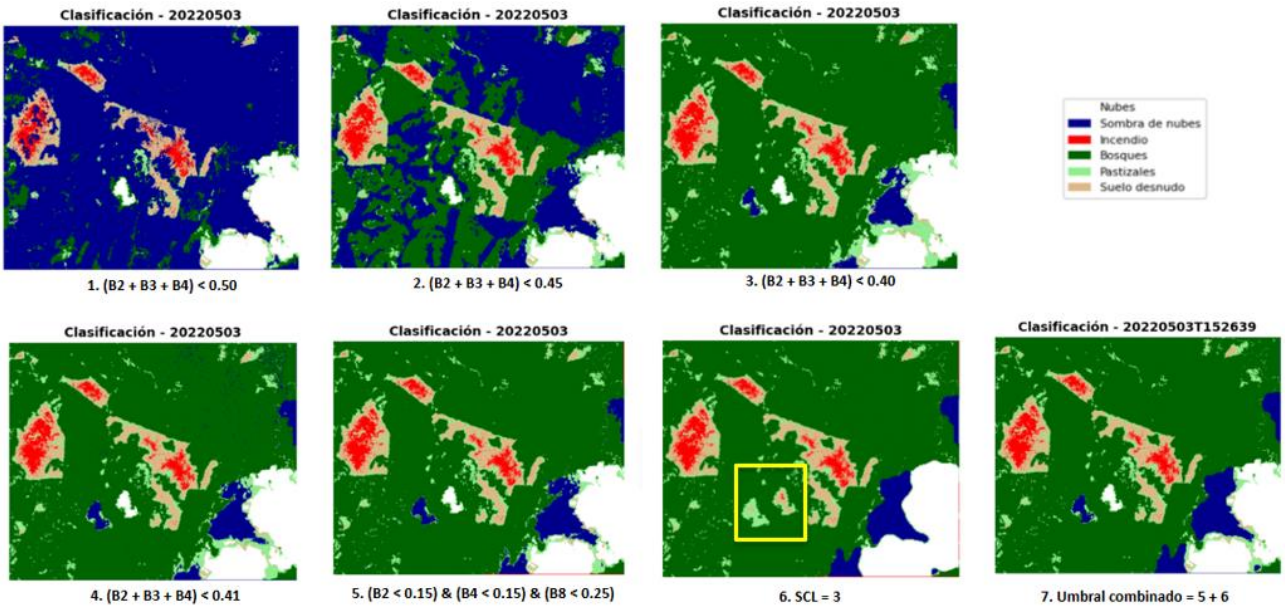

Figura 16. Proceso calibración por umbrales, fuente: Diseño propio.

La detección de nubes del tipo cirrus es un aspecto relevante para el presente estudio, dado que las imágenes obtenidas de Sentinel-2A incluyen escenas en las que se observa la presencia de nubes delgadas de este tipo. La falta de clasificación de estas nubes puede influir negativamente en los resultados, especialmente cuando se encuentran sobre alguna porción de la región de interés (ROI). Por lo tanto, se opta por emplear una combinación de umbrales junto con las bandas previamente clasificadas por S2A, así como la banda B2, que corresponde al espectro azul. La utilización de la banda B2 permite mejorar la detección de nubes tipo cirrus; al combinar esta información con la máscara de umbrales, se obtienen los resultados correspondientes al caso 7, los cuales se ilustran en la Figura 16.

Este paso inicial de clasificación por umbrales fue ejecutado para el resto de las escenas de las dos áreas de estudio Caquetá y Tolima con los umbrales que se relacionan en la Tabla 9, tomando como referencia la metodología del IDEAM empleada en la automatización de las alertas tempranas de deforestación como parte de uno de los componentes del Sistemas de Monitoreo de Bosques y Carbono [54], sin embargo, como se puede evidenciar en el caso 7 de clasificación de la Figura 16, no se obtuvieron buenos resultados para la discriminación de bosques y pastizales en la regiones de interés.

Tabla 9. Umbrales de clasificación [54].

ID	Clase	Umbral	Observaciones
1	Nubes	(B02 > 0.3) & SCL == 8,9,10 y 11	Identifica las nubes dentro de las imágenes, se emplean valores de reflectancia considerando que la máscara SCL 8, 9, 10 y 11, no tuvo éxito detectando patrones de nubes sobre el área de interés, con la reflectancia al tener valores altos en la banda 2 (Blue) se detectan las nubes con mejor precisión sobre todo aquellas que son más tenues.
2	Sombra de nubes	(B2 < 0.15) & (B4 < 0.13) & (B8 < 0.25) | (SCL==3)	Combina valores de reflectancia con la detección entregada por Sentinel en el SCL (Scene Classification Layer), para la sombra se decidió combinar con la capa SCL considerando que al emplear reflectancia + confirmación de SCL se logró mayor precisión que empleándolas de manera independientes.
3	Zona de incendio	(NDVI <= 0.10)	Emplea diferentes bandas para el cálculo del NDVI, valores muy bajos de NDVI significa vegetación muerta o áreas quemadas.
4	Vegetación - Bosques	(NDVI >= 0.35)	Emplea el NDVI para identificar vegetación densa y saludable, normalmente se suelen encontrar bosques saludables en valores de 0.3 a 0.8
5	Vegetación - Pastizales	(NDVI > 0.20) & (NDVI < 0.35)	Emplea el NDVI para identificar vegetación de media a baja densidad, la idea es lograr separar pastos o agricultura de bosques más densos.
6	Suelo desnudo	Cath-all	Incluye todo lo no clasificado anteriormente, para nuestra área objetivo son zonas con poca o ninguna vegetación.

De acuerdo con la anterior clasificación, posteriormente se generó una máscara binaria para cada imagen utilizando el siguiente proceso:

- 1 → Correspondían a áreas con valores de reflectancia válidos, asociadas a coberturas de bosques, pastizales, incendios o suelo desnudo, según la clasificación obtenida mediante umbrales físicos. Estos píxeles fueron considerados datos útiles y fiables para el estudio de la recuperación de la cobertura vegetal tras los incendios.
- 0 → Incluyen píxeles identificados como sombras, nubes o sin información (correspondientes a la clase 1 y 2 en la clasificación por umbrales). Estos valores enmascarados como NaN dentro del dataset para asegurar la calidad de los análisis posteriores.

Las máscaras se generaron mediante cálculos ejecutados en Python con rasterio y NumPy, y posteriormente se aplicaron a las imágenes de las áreas de interés. De esta forma, los valores contaminados quedaron eliminados del conjunto de datos, asegurando una alta calidad en los datos que fueron empleados para calcular la tasa de recuperación.

Estas máscaras fueron almacenadas en formato GeoTIFF, facilitando su integración al pipeline de análisis. La estandarización de resolución a 10 m permite que puedan ser aplicadas sobre cualquier banda reflectiva reescalada, sin necesidad de reproyección adicional.

5.2.2 Integración de algoritmos

Considerando la identificación de una deficiencia en la discriminación de bosques y pastizales (zonas con vegetación densa versus zonas con vegetación moderada) en la clasificación por umbrales, se opta por explorar la aproximación de los algoritmos utilizados por QGIS para el entrenamiento y clasificación de clases. En este contexto, se lleva a cabo una comparación entre los métodos semiautomatizados de QGIS y la clasificación basada en umbrales que se presentan en la Figura 17.

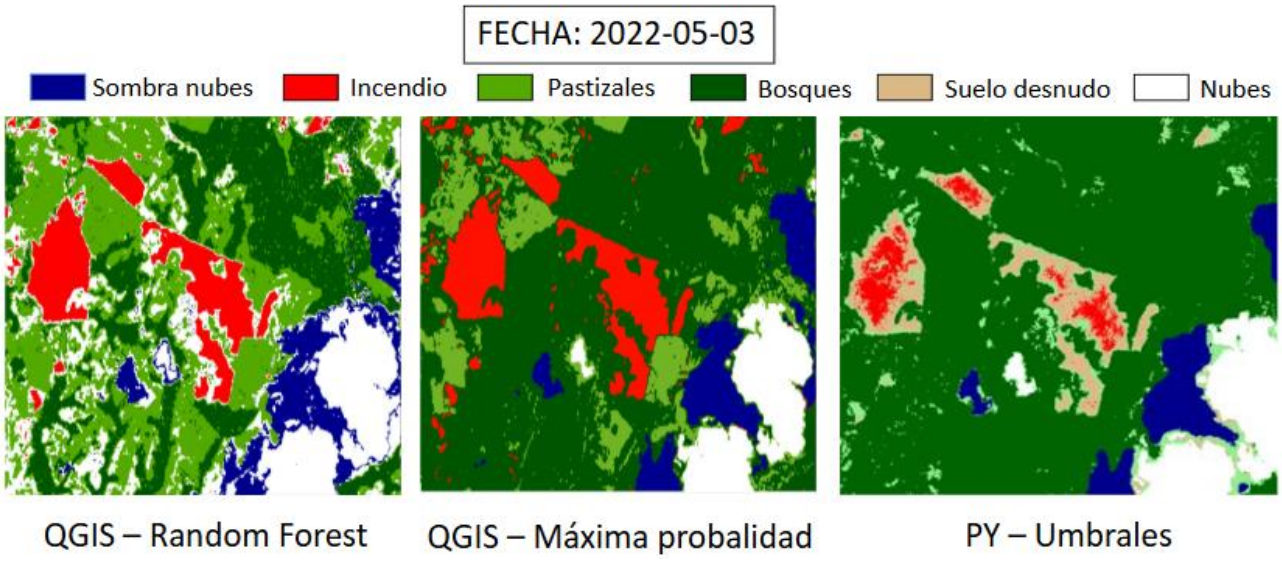

Figura 17. Comparación QGIS – Umbrales, fuente: Diseño propio.

Para el presente análisis, resulta fundamental establecer una distinción entre las clases de suelo desnudo y zonas afectadas por incendios, así como entre la vegetación densa, representada por los bosques, y la vegetación moderada, correspondiente a los pastizales. Esta diferenciación es relevante, dado que impacta en la estimación del modelo. En consecuencia, se opta por aplicar el modelo utilizando lo más eficaz de los tres algoritmos de clasificación disponibles, generando un conjunto de reglas básicas entre los píxeles para su comparación, con el objetivo de seleccionar aquel que ofrezca la mejor clasificación para el análisis, en la Tabla 10 se relaciona las ventajas y desventajas

encontradas en los modelos.

Tabla 10. Ventajas y desventajas modelos de clasificación.

Algoritmo	Ventajas	Desventajas
Random forest	Permite diferenciar entre vegetación densa (Bosques) y vegetación moderada (pastizales)	Parece confundir el suelo desnudo con la clase de nubes.
Máxima probabilidad	Buena detección de la clase nubes y sombras + empleo de SCL	Media diferenciación entre vegetación densa y moderada
Umbrales	Buena detección entre zona de incendio y suelo desnudo	Baja diferenciación entre vegetación densa y moderada.

Considerando la tabla anterior, se implementó una combinación de técnicas que permite aprovechar las ventajas de los tres modelos analizados. Este enfoque implica una comparación a nivel de píxeles, manteniendo la clase del modelo de referencia. Por ejemplo, si la clase de un píxel en análisis corresponde a vegetación densa, se validará la clasificación proporcionada por el modelo de random forest. En contraste, si la clasificación se refiere a la detección de incendios, se tomará como referencia el modelo PY-umbrales, el cual presenta un desempeño superior en la diferenciación de estos fenómenos. En la Figura 18 se presenta el flujo de clasificación de pixel al combinar los tres algoritmos.

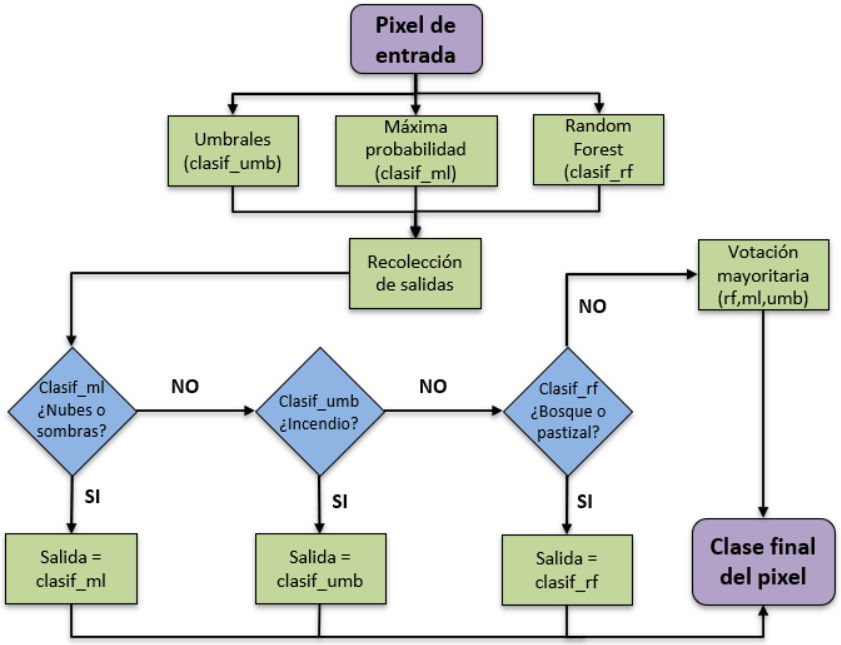

Figura 18. Combinación de modelos para clasificación (fuente: Diseño propio).

La combinación de algoritmos se aplica a todas las escenas de las áreas de interés, lo que facilita la obtención de resultados óptimos. Al incorporar una imagen de validación, la clasificación permite identificar de manera clara diferentes tipos de cobertura terrestre, tales como vegetación densa (bosques – Verde Oscuro), vegetación moderada (pastizales – Verde Claro), zonas afectadas por incendios (rojo) y suelo desnudo (tierra). Estas detecciones son pertinentes para el análisis de estimación. La Figura 19 ilustra la comparación de una imagen RGB y el resultado de su clasificación.

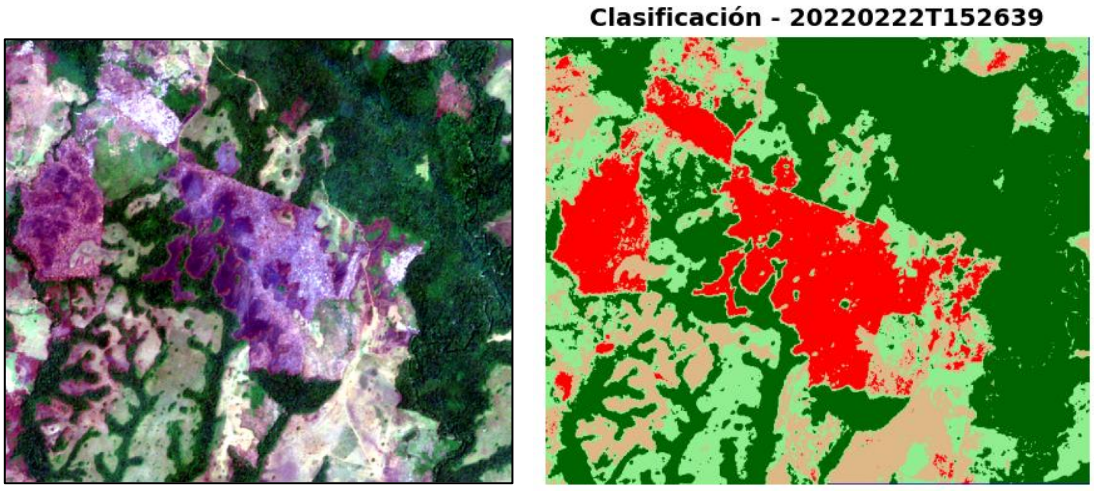

Figura 19. Resultados combinación de algoritmos, fuente: Diseño propio.

Adicionalmente, se logró desarrollar un modelo robusto y adecuadamente entrenado, alcanzando un promedio global de F1-Score de 0.906 para las seis clases evaluadas. En el caso de las nubes y sombras de nubes, se emplearon las capas SCL para su clasificación, combinadas con un algoritmo de máxima probabilidad, obteniendo un F1-Score de 0.935 para nubes y de 0.865 para sombras. Estos resultados indican una reflectancia característica bien capturada para las nubes. La detección de incendios alcanzó un F1-Score de 0.924, lo que sugiere una alta precisión; este modelo se caracteriza por ser conservador, priorizando la reducción de falsos positivos en comparación con los falsos negativos, con el objetivo de minimizar las falsas alarmas en el sistema de alerta temprana. En cuanto a los bosques, se obtuvo un F1-Score de 0.935, lo que indica que esta clase es la mejor equilibrada, con firmas espectrales bien definidas y contrastadas. Por otro lado, los pastizales presentan una mayor confusión Inter clase con el suelo desnudo, lo que podría estar relacionado con la presencia de pastizales secos que tienen una reflectancia similar a la del suelo desnudo, la Figura 20 presenta los resultados obtenidos de las diferentes métricas.

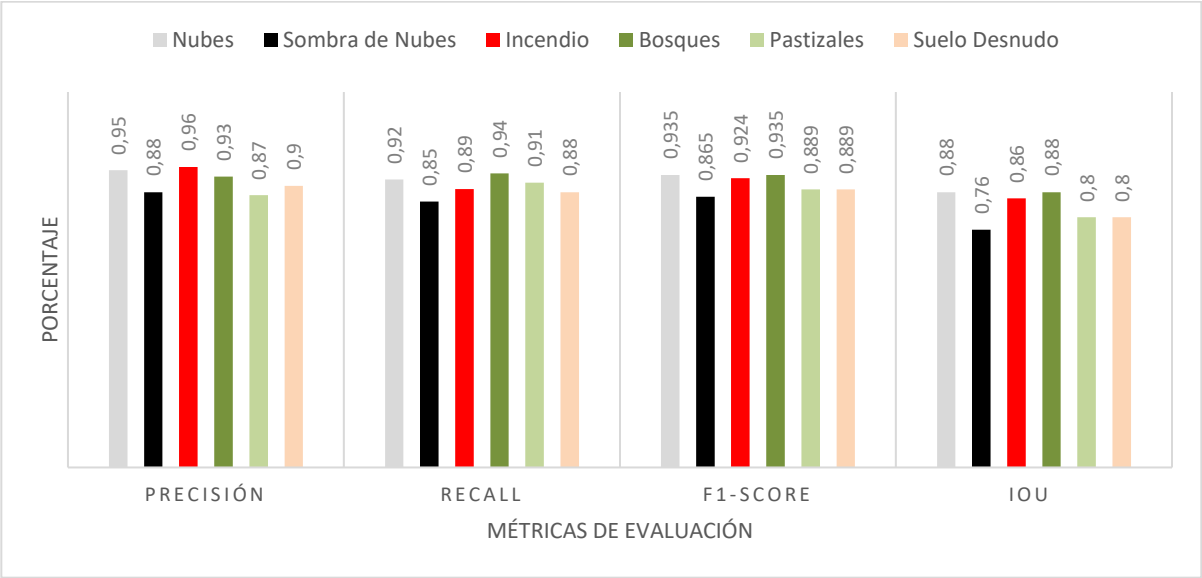

Figura 20. Resultados métricas de evaluación, fuente: Diseño propio.

5.3 ALGORITMO TASA DE RECUPERACIÓN

Esta sección describe cómo se implementaron los modelos de machine learning para estimar la tasa de recuperación de la vegetación en las zonas afectadas por incendios forestales, a partir del conjunto de datos descrito en el capítulo anterior. El objetivo es transformar la base de datos satelital y meteorológica en un sistema predictivo que otorgue, para cada píxel, una probabilidad de recuperación dentro de un horizonte operativo.

La implementación se basa en dos elementos clave del dataset:

- Clasificación basada en umbrales sobre las imágenes Sentinel-2, para que diferencie entre sombras, zonas quemadas, suelos desnudos, nubes, vegetación y pastos.
- Resumen organizado de variables climáticas: precipitaciones y temperatura calculadas en tiempos recientes, entre 7 y 30 días, asociadas a cada imagen y píxel para obtener el contexto atmosférico previo y posterior al incendio.

A partir de lo expuesto, se formuló un problema de clasificación binaria con el objetivo de estimar, para cada píxel la probabilidad de que se encuentre en estado de recuperación dentro de un horizonte temporal. Para ello, se utilizaron como insumos los índices espectrales y las condiciones climáticas recientes registradas en cada escena contenida en las bases de datos generadas. El desarrollo del modelo se llevó a cabo de acuerdo con las buenas prácticas en ciencia y análisis de datos. Este proceso incluyó la definición precisa de la variable objetivo, la selección y depuración de características relevantes, la evaluación de algoritmos candidatos, la formulación del esquema de validación y la optimización de los parámetros del modelo.

5.3.1 Variable objetivo

Se estableció una etiqueta binaria para cada píxel, asignando un valor de uno (1) en caso de que se evidencie recuperación y un valor de cero (0) si la zona no presenta signos de recuperación. Esta etiqueta permite determinar si la vegetación se recupera dentro del horizonte operativo definido para cada región de interés como se presenta a continuación:

- Natagaima: recuperación en menor o igual que 120 días.
- Caquetá: recuperación en menor o igual que 90 días.

La presente etiqueta se generó mediante la comparación del estado de la vegetación en función de la línea de tiempo correspondiente a la fecha del incendio y a las fechas posteriores. Para ello, se emplearon indicadores como ΔNDVI y ΔNBR, con el propósito de evaluar el nivel de recuperación entre píxeles, en relación con el estado previo al evento, así como para identificar aquellos píxeles que continúan en un estado de degradación.

5.3.2 Características relevantes

El vector correspondiente a cada píxel incorpora información derivada de tres componentes fundamentales. En primer lugar, se procede a la evaluación del estado de la vegetación y la severidad del incendio a través de índices como el NDVI y el NBR, así como de sus variaciones (ΔNDVI y ΔNBR). Esta evaluación también considera el número de días transcurridos desde la última condición estable de la vegetación y una clasificación de severidad (Alta/Media/Baja), la cual se determina mediante la aplicación de umbrales sobre los cambios observados. En segundo lugar, se integran variables climáticas, tales como la temperatura media, el clima reciente y el acumulado de precipitaciones en intervalos de 7 y 30 días. Estas variables son fundamentales para integrar la disponibilidad de precipitación y las condiciones térmicas que pueden favorecer o limitar el proceso de recuperación. Finalmente se incorporan características temporales y espaciales, tales como el día del año codificado de manera cíclica (doy_sin, doy_cos) y la zona geográfica (Natagaima o Caquetá). La combinación de estas características es fundamental para la representación de patrones estacionales y para la identificación de diferencias intrínsecas entre las áreas de estudio.

5.3.3 Partición por zona

Se elaboraron diferenciaciones para Natagaima y Caquetá, dado que los pronósticos de recuperación y las condiciones climáticas presentan variaciones significativas, este enfoque permite establecer métricas y umbrales de manera específica para cada región, con el objetivo de evitar la confusión de comportamientos que responden a distintos climas y patrones de incendios.

5.3.4 Selección de algoritmos candidatos

Se lleva a cabo una evaluación de modelos, utilizando como referencia los que se exponen a continuación:

La Regresión Logística (LR) y el modelo de Random Forest (RF) se constituyen como herramientas fundamentales en el análisis de problemas ambientales. Ambas metodologías ofrecen modelos robustos y son interpretables, lo que facilita su comprensión y aplicación en este ámbito. Su capacidad para proporcionar explicaciones claras sobre los resultados obtenidos contribuye a su utilidad en la investigación y la toma de decisiones en contextos ambientales.

El HistGradientBoostingClassifier (HGB) es un modelo fundamentado en histogramas que se adapta eficazmente a grandes volúmenes de datos. Este modelo tiene la capacidad de abordar el desbalance de clases y de identificar relaciones no lineales en los datos.

Se opta por utilizar el modelo HGB como la opción principal, dado su potencial para:

- Correcta priorización de los casos de recuperación en contextos de fuerte desbalance, evidenciada por un alto valor de PR-AUC.
- Se observa una distinción significativa a nivel global entre los pixeles recuperados y aquellos no recuperados, evidenciada por un valor elevado del área bajo la curva (ROC-AUC).
- La generación de probabilidades, caracterizadas por un bajo puntaje de Brier, se aproxima a los eventos que ocurren en la realidad. Estas probabilidades son fundamentales para la utilización de deciles de riesgo y la definición de umbrales operativos.

Los primeros resultados indican que HGB clasifica de manera adecuada los píxeles en función de su potencial de recuperación. Además, se observa que la segmentación es clara y que las probabilidades generadas son coherentes con el análisis de los datos. Por estas razones, se selecciona este modelo para las zonas de Natagaima y Caquetá.

5.3.5 Esquema de validación

Con el objetivo de asegurar que las métricas representen un desempeño realista y no estuvieran afectadas por sobreestimaciones, se estableció un esquema de validación que se centra en el control específico de la fuga de información, efectuando lo siguiente:

a. Segmentación 80/20 sin combinar píxeles ni escenas:

- El 80 % de los datos se empleó para el entrenamiento y el 20 % para realizar las respectivas pruebas.

- Se establece la prohibición del uso de escenas que compartan la misma fecha y el mismo identificador de píxel, tanto en el proceso de entrenamiento como en el de prueba. Esta medida tiene como finalidad mitigar el riesgo de que se memoricen patrones específicos asociados a una escena.

b. Validación cruzada temporal + LOGO (Leave-One-Group-Out) por escena:

- En el entrenamiento se implementó una validación cruzada donde cada fragmento hace referencia a un polígono.
- Cada bloque deja por fuera una escena completa, simulando el comportamiento del modelo frente a nuevas imágenes y reduciendo la varianza entre bloques (folds).
- La elección de configuraciones dio prelación a aquellas con baja dispersión de métricas entre bloques (folds), interpretadas como más fuertes y estables.

El esquema de validación presenta una estructura robusta, orientada hacia la aplicación operativa del modelo, con el propósito de generalizar nuevas fechas y escenas, más allá de lo observado durante el proceso de entrenamiento.

5.3.6 Optimización de hiperparámetros

En la subsección anterior, se expuso el proceso de validación para el empleo del modelo HGB, donde se llevó a cabo la segmentación inicial de los datos, asignando un 80 % para el conjunto de entrenamiento y un 20 % para el conjunto de prueba. Esta partición se realizó considerando las variables id_pixel y escena, con el objetivo de prevenir la fuga de información. Posteriormente, se realizó una búsqueda en rejilla sobre un conjunto de hiperparámetros con el propósito de identificar una configuración que proporcionara un equilibrio adecuado entre la estabilidad y el rendimiento del modelo.

La rejilla de valores (Tabla 11) se analizó con LOGO y validación cruzada temporal, dando prioridad a las configuraciones con mayor PR-AUC promedio, F1 más alto y consistente entre folds y el Brier Score bajo (Buena calibración).

Tabla 11. Grid de hiperparámetros evaluados para HGB.

Hiperparámetro	Valores
learning_rate	0.05, 0.06, 0.08
max_iter	400, 600
min_samples_leaf	80, 120
l2_regularization	0.5, 1.0
max_depth	6

Configuración seleccionada (modelo escogido), para ambas zonas se escogió la siguiente combinación (Tabla 12).

Tabla 12. Configuración óptima del modelo HGB para las zonas de Natagaima – Caquetá.

Hiperparámetro	Valor
learning_rate	0.06
max_iter	400
min_samples_leaf	80
l2_regularization	1.0
max_depth	6

5.3.7 Modelo de tasa de recuperación e índice de recuperación

El modelo seleccionado (HistGradientBoosting, HGB) genera como salida una probabilidad de recuperación por píxel, con valores continuos entre 0 y 1. Esta probabilidad representa la confianza del modelo en que un píxel afectado por incendio haya alcanzado condiciones compatibles con recuperación vegetal dentro del horizonte operativo definido para cada zona de estudio (≤120 días para Natagaima y ≤90 días para Caquetá).

Dado que la salida del modelo es probabilística, la clasificación operativa como “recuperado” o “no recuperado” requiere la definición de un umbral operativo específico para cada región, el cual permite convertir la probabilidad continua en una decisión binaria.

Natagaima (≤120 días): se estableció un umbral operativo de 0.573. Esto significa que un píxel se clasifica como recuperado (valor 1) únicamente si su probabilidad estimada es mayor o igual a 0.573; en caso contrario, se clasifica como no recuperado (valor 0).

Caquetá (≤90 días): se estableció un umbral operativo de 0.55. Esto significa que un píxel se clasifica como recuperado (valor 1) únicamente si su probabilidad estimada es mayor o igual a 0.55; en caso contrario, se clasifica como no recuperado (valor 0).

La formulación probabilística resulta coherente con el objetivo del proyecto, dado que la recuperación de la vegetación es un proceso progresivo y gradual, y no un fenómeno estrictamente binario. En este sentido, la salida del modelo permite capturar distintos niveles de avance en la recuperación post-incendio.

Con el fin de traducir esta salida probabilística en una medida interpretable y operativamente útil, se define el Índice de Recuperación (IR) directamente a partir de la probabilidad estimada por el modelo:

$$IRi = Pi \tag{13}$$

donde IR_i corresponde al índice de recuperación del píxel i y $p_i$es la probabilidad de recuperación estimada por el modelo HGB [55]. Bajo esta definición, valores cercanos a 1 indican una alta probabilidad de recuperación vegetal, mientras que valores cercanos a 0 representan baja probabilidad de recuperación.

Para facilitar el análisis a escalas superiores, el índice de recuperación se agrega espacial y temporalmente a nivel de municipio, cohorte temporal o región, mediante el promedio de los índices individuales:

$$IR_{Zona} = \frac{1}{N} \sum_{i=1}^{N} IRi \quad (14)$$

donde N corresponde al número de píxeles afectados en la zona analizada. Este valor agregado permite estimar una tasa de recuperación esperada, facilitando la comparación entre regiones y periodos temporales.

Adicionalmente, el índice de recuperación puede discretizarse mediante la definición de un umbral operativo, permitiendo clasificar los píxeles como recuperados o no recuperados. No obstante, el uso continuo del índice conserva mayor información y resulta más adecuado para procesos de priorización y análisis comparativo, mientras que el umbral se emplea principalmente con fines operativos, de reporte y visualización.

De esta manera, el modelo HGB no solo permite estimar la probabilidad de recuperación por píxel, sino que también proporciona un marco cuantitativo coherente para la generación de indicadores agregados, la comparación temporal entre cohortes y el soporte directo a la toma de decisiones en contextos de monitoreo y restauración vegetal.

5.4 TIEMPOS DE CÓMPUTO

Considerando que el proyecto implicó la implementación de tres algoritmos, se exponen a continuación los tiempos estimados de computación, software y hardware requerido, correspondientes a cada uno de los algoritmos desarrollados:

Tabla 13. Tiempos de cómputo de los modelos.

Algoritmo	Descripción	Tiempos	Software	Hardware
Detección de zonas de incendios	Empleado para detectar en una imagen satelital las	Raster 19x50 km = 0.5 min Raster 100 x 100 km =	Backend: Flask (Python 3). Base de datos: MongoDB. Frontend: Angular,	

	zonas asociadas a incendios forestales, el objetivo es identificar la zona de interés para estudio y clasificación (sección 5.1).	3.1 min Raster 60 x 400 km = 5.2 min	librerías Chart.js para visualización estadística y Mapbox GL JS para mapas interactivos. Frameworks:Tensorflow, Ultralytics. Seguridad: Autenticación basada en roles. SW requerido: Miniconda, Cuda Nvidia, Clonación de repositorio GIT	CPU: Intel Core i5 (8ª generación) o AMD equivalente Memoria RAM: 8 GB (mínimo) o superior Tarjeta Gráfica: 5 GB. Almacenamiento: 30GB o superior SSD Conectividad: Acceso a Internet estable durante la instalación.
Clasificación clases de la zona de interés	Se encarga de clasificar las seis clases disponibles de la zona de incendio detectada previamente (Sección 5.2).	Imagen 6 x 5 km = 0.3 seg (por imagen)	Framework: Google Colab Librerías: Shapely, geopandas, pyproj, rasterio, pyarrow, numpy, matplotlib	
Estimación de la tasa de recuperación	Modelo para la estimación de la tasa de recuperación vegetal (Sección 5.3), utilizando la partición 80/20 y la validación cruzada temporal + LOGO	Regresión lineal = 550 seg Random Forest = 670 seg HistGradientBoosting = 480 seg	Framework: Jupyter Librerias: numpy, pandas, rasterio, sklearn, tqdm, joblib, datetime, pathlib.	

6 COMPARATIVA DE MODELOS Y ANÁLISIS OPERATIVO DE RESULTADOS

En este capítulo se presenta un resumen comparativo del desempeño de los tres modelos analizados y evaluados: HistGradientBoosting (HGB), Random Forest (RF) y Regresión Logística (LR), en las dos zonas de estudio.

6.1 MÉTRICAS DE EVALUACIÓN

La métrica principal utilizada es el PR-AUC (Average Precision). Esta métrica prioriza el rendimiento en la clase positiva en contextos de desbalance, permitiendo un análisis de la eficacia del modelo en la concentración de recuperaciones reales, especialmente en los niveles altos de probabilidad.

6.1.1 Métricas complementarias

- **ROC-AUC:** mide la separación global para distinguir entre clases, independiente del umbral elegido.
- **Brier Score:** cuantifica el ajuste de las probabilidades definido como el error cuadrático medio entre score y etiqueta
- **F1-score al umbral operativo:** establece exactitud y recall, esto es importante para poder definir el punto de corte que convierte la probabilidad en decisión.
- **Análisis por fragmentos y Umbral operativo:** se establecieron umbrales operativos basados en F1, en Natagaima también en el criterio de Youden, esto permite descifrar el modelo como una regla sencilla se recupera o no se recupera.
- **Análisis de curvas ROC/PR** por escena y deciles de riesgo (lift), para convalidar que los píxeles con mayor puntuación concentran una fracción importante de las recuperaciones observadas, esta condición es necesaria para un ordenamiento por riesgo efectivo en campo.

Los resultados numéricos del modelo en el conjunto de prueba se sintetizan en la Tabla 14 para Natagaima (≤120 días) y la Tabla 15 para Caquetá (≤90 días), donde se resumen PR-AUC, ROC-AUC, F1 al umbral operativo, métricas de calibración y captura de recuperaciones en el top-10 %.

Tabla 14. Métricas consolidadas Natagaima - (≤120 días).

Métrica	Valor	Notas
Probabilidad media de recuperación ≤120d	0,644	Promedio de probabilidades sobre píxeles quemados
% recuperado (umbral Youden≈0.573)	0,601	Porcentaje etiquetado como recuperado (operativo)

Métrica	Valor	Notas
ROC-AUC	0,87	AUC ROC (validación OOF/hold-out)
PR-AUC	0,97	Average Precision (PR-AUC)
Recall @ umbral	0,8	Se prioriza sensibilidad en operación
Precisión @ umbral	0,94	Umbral seleccionado con Youden
Captura Top-10%	0,23	Concentración de eventos en el decil más alto

Tabla 15. Métricas consolidadas Caquetá - (≤90 días).

Métrica	Valor	Notas
Umbral Operativo	0.55	Seleccionado por estabilidad (OOF) y prioridad de recall
ROC-AUC (LOGO por escena)	≈0.60	Media aprox. (escenas sin clase única)
PR-AUC (LOGO por escena)	≈0.64	Media aprox.
F1 (LOGO por escena)	≈0.57	Media aprox.
ROC-AUC (fit all + cal)	0.907	Modelo final entrenado en observables
PR-AUC (fit all + cal)	0.923	Capacidad de priorización tras calibración
F1 (fit all + cal)	0.883	Desempeño al umbral operativo (0.55)
Brier score	0.106	Calibración de probabilidades
% positivo @ thr=0.55	~0.705	Promedio ponderado por escena (monitor)

6.2 DICCIONARIO DE MÉTRICAS Y VARIABLES DERIVADAS

El diccionario de datos especifica de manera clara las salidas del modelo, así como las variables derivadas que serán empleadas en los análisis operativos. Se presenta el nivel de agregación, tanto a nivel de píxel como de zona, así como la interpretación correspondiente de cada campo.

Tabla 16. Diccionario de datos.

Métrica	Campo sugerido	Nivel	Definición / Cálculo	Interpretación / Notas
Probabilidad (score)	Scores[score]	píxel	Salida del modelo HGB ∈[0,1].	Prob. estimada de recuperar en el horizonte (≤120d Nata / ≤90d Caque).
Etiqueta predicha	Scores[pred]	píxel	1 si score ≥ umbral; 0 en caso contrario.	Clasificación operativa. Umbral: **0.573** (Nata), **0.55** (Caque).
Umbral operativo	KPIs[threshold_used]	zona / cohorte	Valor de decisión para convertir score→clase.	Seleccionado por estabilidad y balance Precisión–Recall.
% Recuperado	(medida) %Recuperado	cohorte / zona	∑pred/N\sum pred / N∑pred/N	Tasa operativa al umbral. Usar **promedio ponderado** por n entre cohortes.
Probabilidad media	(medida) ProbMedia	cohorte / zona	avg(score)\text{avg}(score)avg(score)	Calibración temporal; comparar con %Recuperado.
N válidos	(medida) N válidos	cohorte / zona	Conteo de píxeles con valido=1 (sin nube/sombra).	Soporte de muestra para el análisis.
% válido	(medida) %Válido	cohorte / zona	N_va´lidos/N_totalN_válidos/N_totalN_va´lidos/N_total	Calidad de cobertura; filtrar cohortes con bajo % válido.
Precisión	(medida) Precisión	cohorte / zona	TP/(TP+FP)TP/(TP+FP)TP/(TP+FP)	Exactitud entre positivos predichos.
Recall (sensibilidad)	(medida) Recall	cohorte / zona	TP/(TP+FN)TP/(TP+FN)TP/(TP+FN)	Cobertura de verdaderos recuperados.
F1 @ umbral	(medida) F1	cohorte / zona	2PR/(P+R)2PR/(P+R)2PR/(P+R)	Balance Precisión–Recall al umbral.
Accuracy	(medida) Accuracy	cohorte / zona	(TP+TN)/(TP+FP+FN+TN)(TP+TN)/(TP+FP+FN+TN)(TP+TN)/(TP+FP+FN+TN)	Global; menos informativa ante desbalance.
PR-AUC	KPIs[prauc]	zona / modelo	Área bajo curva precisión–recobro.	**Prioriza** bien bajo desbalance. Nata≈**0.97**, Caque≈**0.923**.
ROC-AUC	KPIs[rocauc]	zona / modelo	Área bajo curva ROC.	Separación global. Nata≈**0.87**, Caque≈**0.907**.
Brier score	KPIs[brier]	zona / modelo	avg((score−label)2)\text{avg}((score - label)^2)avg((score−label)2)	**Calibración**: menor es mejor. Caque≈**0.106**.
Decil de	Scores[decil]	píxel	Ranking por score (1=top	Segmentación operativa.

Métrica	Campo sugerido	Nivel	Definición / Cálculo	Interpretación / Notas
riesgo			10%).	
Severidad	Scores[severidad]	píxel	Categoría (Alta/Media/Baja) según ΔNBR/ΔNDVI.	Útil para cortes de supervivencia y alertas.
Cohorte	Scores[cohorte]	píxel	Mes YYYY-MM del evento/escena.	Clave temporal para KPIs por cohorte.
Zona	Scores[zone]	píxel	Natagaima / Caquetá.	Filtro principal del tablero.

6.3 COMPARACIÓN DE MODELOS - NATAGAIMA

En la zona de recuperación de Natagaima, con un periodo de ≤120 días, se llevó a cabo un análisis comparativo de la evolución de los modelos HGB, RF y regresión logística, evaluando su desempeño a través de las métricas PR-AUC, ROC-AUC y F1. El modelo HGB se destaca en la comparación, presentando un PR-AUC de 0,970, un ROC-AUC de 0,870 y un F1 de 0,863. En contraste, el modelo de Random Forest (RF) se sitúa aproximadamente entre 3 y 5 puntos porcentuales por debajo de HGB en estas métricas, mientras que la regresión logística se encuentra entre 7 y 10 puntos porcentuales por debajo en las mismas características.

Esto evidencia una capacidad más alta en los casos recuperados (PR-AUC), ideal para el objetivo de recall y precisión en el umbral operativo (F1) y una separación global establecida entre clases (ROC-AUC). Estos resultados operativamente fortalecen a HGB como modelo de referencia para la zona de Natagaima, nos permite enfocar los recursos en los pixeles más críticos, el top -10% captura ≈23 % de las recuperaciones, mantiene consistencia con el umbral de referencia. La Figura 21 presenta los resultados obtenidos.

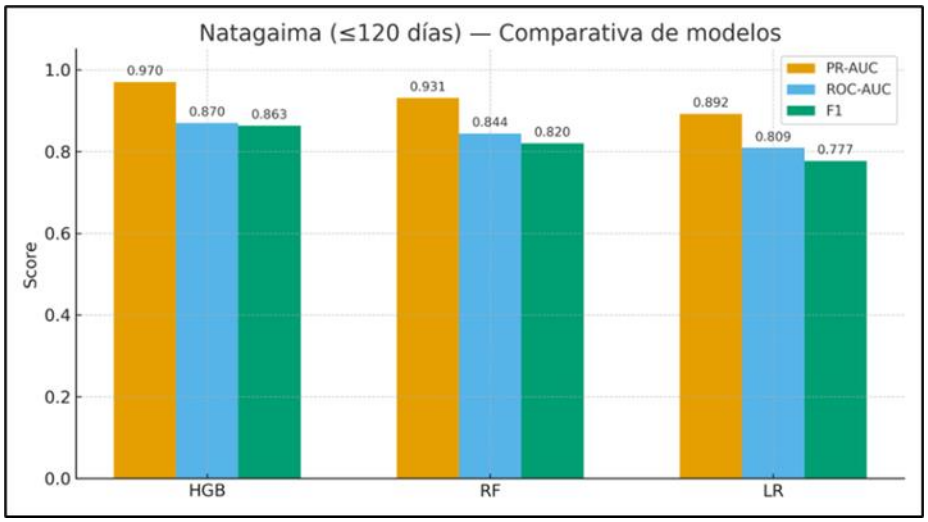

Figura 21. Resultados y comparación de modelos Natagaima.

6.3.1 Interpretación de Resultados

- El PR-AUC cercano a 0.97 hace referencia a una excelente capacidad de ordenamiento, los píxeles con mayor posibilidad estimada efectivamente se recuperan con una alta frecuencia
- El ROC-AUC de 0.87 este confirma una buena separación global entre no recuperados y recuperados.
- El F1 elevado (0.863) estima que el valor de referencia que se selecciono logra un balance adecuado entre recuperar en su mayoría de zonas y evitar un aumento de falsos positivos.
- La concentración de ~23 % de recuperaciones en el top-10 % sustenta el uso de este modelo para establecer el monitoreo y acciones de cierre en las cohortes temporales más críticos.

En términos operativos, estos resultados permiten priorizar meses y áreas donde la probabilidad de recuperación es menor o más incierta, alineando recursos de seguimiento y restauración con la información del modelo.

6.4 COMPARACIÓN DE MODELOS – CAQUETÁ

En el análisis correspondiente a la zona de Caquetá, con un horizonte temporal de ≤90 días, se llevó a cabo una comparación entre los modelos de HGF, RF y regresión logística, utilizando como métricas principales PR_AUC, ROC-AUC y F1. En este contexto, el modelo HGB demostró obtener los resultados más favorables, alcanzando un PR_AUC de 0,923, un ROC-AUC de 0,907 y un F1 de 0,883 al establecer un umbral de 0,55. En contraste, el modelo RF se situó por debajo de HGB en 5,4 y 6 puntos porcentuales en las métricas PR_AUC y ROC-AUC, respectivamente, mientras que la regresión logística se ubicó a 10, 7 y 10 puntos porcentuales por debajo en las mismas métricas.

Las diferencias indican que HGB se mantiene priorizando pixeles recuperables en un horizonte más viable, conserva un balance entre recall y precisión en el umbral operativo y nos presenta la separación global más estable entre clases. Adicional, el Brier ≈ 0,106 consolida la buena calibración de las probabilidades calculadas.

La Figura 22 representa esta comparación para Caquetá, y recalca la elección de HGB como modelo de referencia para focalizar el apoyo de recursos en el seguimiento y restauración en esta zona.

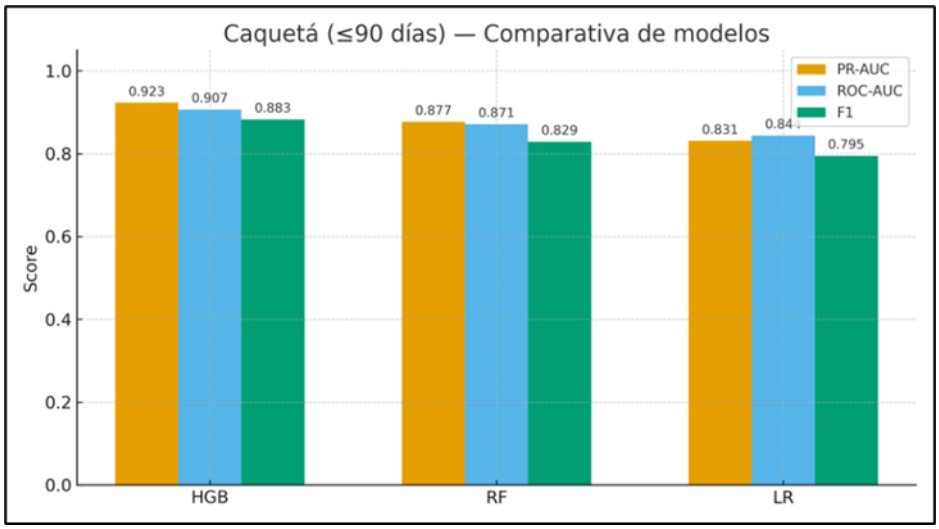

Figura 22. Resultados y comparación de modelos Caquetá.

6.4.1 Interpretación de Resultados

- El PR-AUC superior a 0.92, en un contexto de con un horizonte de recuperación más corto, se evidencia que el modelo mantiene una muy buena capacidad de dar prioridad a los píxeles recuperables.
- El ROC-AUC de 0.907 evidencia una separación global aún mejor que en la zona de Natagaima, a pesar de que el tiempo es más corto.
- El F1 de 0.883 al umbral 0.55 muestra un balance sólido entre recall y precisión, conveniente para las decisiones de operación a 90 días.
- El Brier de 0.106 confirma que las probabilidades estimadas son consistentes con las frecuencias observadas, condición importante para definir umbrales, alertas y deciles de riesgo.

6.5 ANÁLISIS OPERATIVO DEL MODELO

La curva presentada en la Figura 23 y la Figura 24 se elaboró mediante la aplicación del modelo de elección (HGB) a todos los píxeles válidos correspondientes a las regiones de Natagaima y Caquetá. Los resultados se agruparon de acuerdo con las cohortes mensuales (YYYY-MM).

Para cada cohorte, se determina, por un lado, el promedio de las probabilidades predichas por el modelo y, por otro, el porcentaje de píxeles que se recuperan de manera evidente en los días correspondientes a cada zona, utilizando el umbral operativo establecido en el Capítulo 5. Cada punto de la curva sintetiza, de manera mensual, la relación entre las predicciones del modelo y las evidencias observadas en los datos.

6.5.1 Zona Natagaima

En la Figura 23 se puede observar el comportamiento temporal del modelo, el cuál presenta picos evidentes en noviembre de 2023 y entre marzo y abril de 2024, los cuales se asocian con condiciones hídricas favorables, reflejadas en altos valores de precip_30d. Por otro lado, el descenso observado en enero de 2024 se correlaciona con una ventana seca y una menor disponibilidad de píxeles válidos. Asimismo, la disminución aparente en julio de 2024 puede interpretarse como resultado de la inclusión de una cohorte reciente, caracterizada por una ventana observacional más corta, lo que no sugiere un deterioro real en el desempeño del modelo. Ambas series presentan un desplazamiento consistente, lo que indica una adecuada calibración entre los puntajes del modelo y los resultados empíricos. Desde una perspectiva operativa, esta curva facilita la orientación de las tareas de monitoreo y restauración. En las áreas donde se presenta una probabilidad baja, se priorizan las verificaciones y el análisis de píxeles negativos. Por otro lado, en las zonas con una probabilidad alta, se centra el seguimiento en el cierre y la confirmación de resultados exitosos. El presente análisis se basa en la agregación de 312.942 píxeles válidos y en la aplicación de un umbral operativo fundamentado en el índice de Youden (aproximadamente 0.573). Este enfoque permite alcanzar un equilibrio adecuado entre la precisión (aproximadamente 0.94) y la sensibilidad (aproximadamente 0.80).

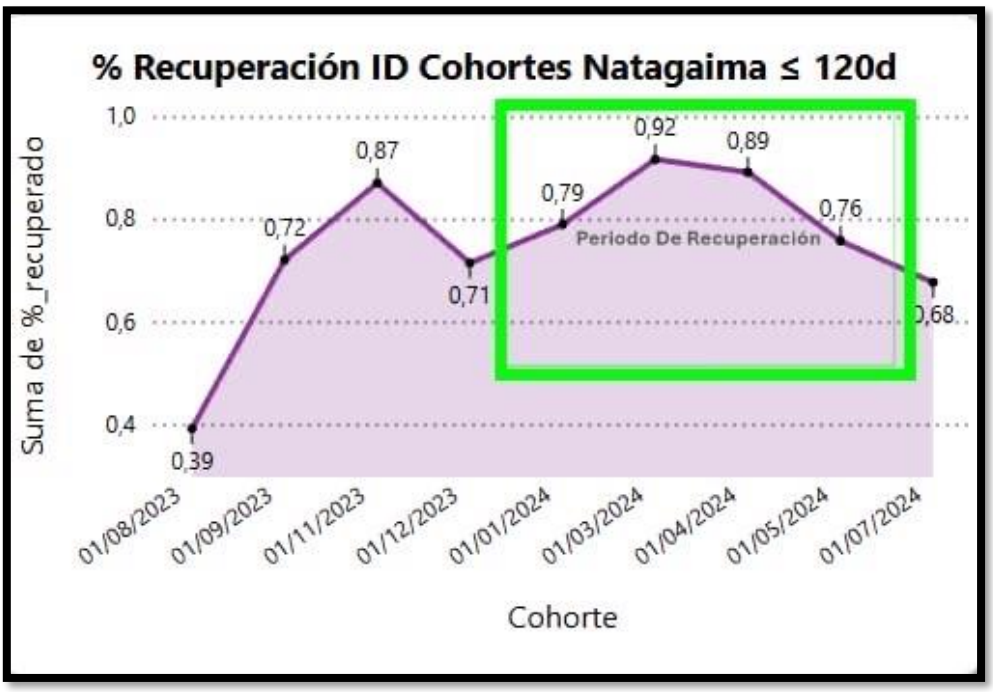

Figura 23. Probabilidad media de recuperación y porcentaje recuperado por cohorte - Natagaima.

6.5.2 Zona Caquetá

En la región de Caquetá, se puede observar en la Figura 24 que se identifican picos de recuperación durante los meses de febrero y marzo de 2023. En contraste, el valle observado en febrero del año 2022 se asocia a una cohorte temprana que presenta una señal menos pronunciada y un número limitado de píxeles válidos. Esta situación se ve condicionada por una ventana de observación reducida y episodios de mayor nubosidad. La probabilidad media y el porcentaje de recuperación estimado, utilizando un umbral operativo de 0.55, presentan una evolución casi paralela. Este fenómeno sugiere una adecuada calibración entre los puntajes generados por el modelo y los resultados observados. Desde la perspectiva de la lectura climática, los picos observados se correlacionan con condiciones hídricas más favorables, caracterizadas por una mayor precipitación reciente y temperaturas moderadas, lo cual es coherente con la importancia de la variable precip_30d. En contraste, el valle inicial indica un periodo seco, durante el cual se presenta una menor cantidad de datos utilizables, atribuible a la presencia de una máscara de nubes. Desde una perspectiva operativa, esta curva facilita la priorización de acciones. Los valores bajos de probabilidad demandan la verificación de zonas y la revisión de píxeles negativos, mientras que los valores altos indican la necesidad de realizar un seguimiento del cierre y la confirmación de áreas que han sido efectivamente recuperadas. El presente análisis se fundamenta en la agregación de 195.903 píxeles válidos, así como en el desempeño sólido del modelo en la región, evidenciado por métricas que incluyen un PR-AUC de aproximadamente 0.923, un ROC-AUC de aproximadamente 0.907, un F1 de aproximadamente 0.883 y un Brier Score de aproximadamente 0.106.

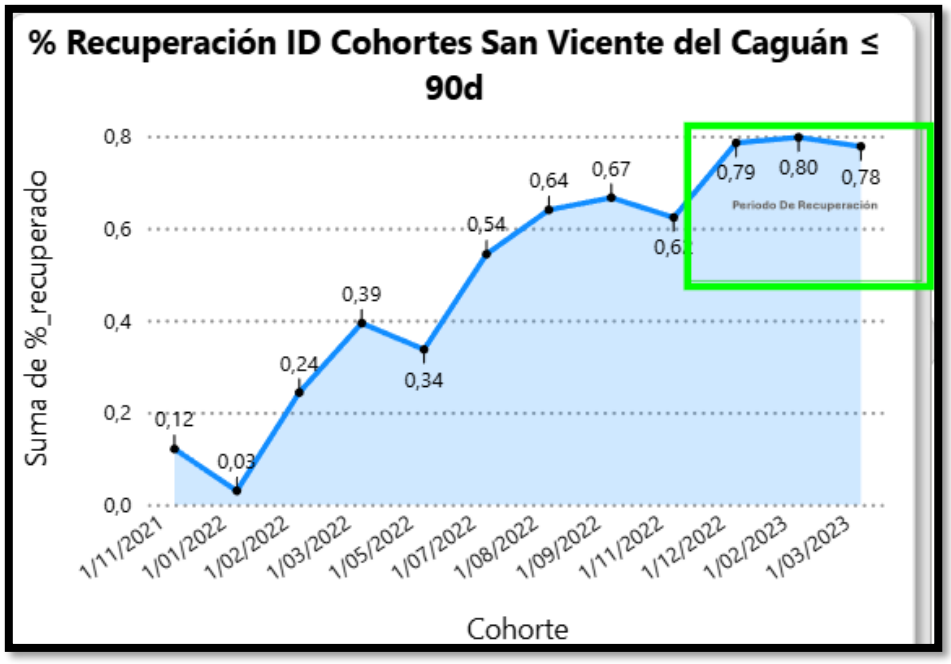

Figura 24. Probabilidad media de recuperación y porcentaje recuperado por cohorte - Caquetá.

6.6 SÍNTESIS COMPARATIVA DE RESULTADOS

En ambas áreas de investigación, el modelo HGB se establece como la opción más idónea para la estimación de la recuperación de la vegetación posterior a eventos de incendios. En Natagaima, se observa una notable capacidad de priorización, evidenciada por el PR-AUC más elevado. Por otro lado, en Caquetá, a pesar de contar con un horizonte más exigente, el umbral basado en ROC-AUC presenta la mejor separación global y un valor elevado de F1. Estos resultados respaldan la utilización de este enfoque para la elaboración del tablero de monitoreo y la generación de alertas clasificadas por deciles de riesgo. Las evidencias cuantitativas presentadas se correlacionan de manera directa con los objetivos específicos del proyecto, al validar un modelo que permite estimar la tasa de recuperación post-incendio, evaluar de manera rigurosa su desempeño y establecer las bases para la herramienta de visualización operativa. En conjunto, las métricas obtenidas indican que el modelo presenta una clasificación precisa de los eventos de recuperación. Además, proporciona probabilidades robustas y mantiene un equilibrio adecuado entre la maximización de la detección de áreas recuperadas y la limitación de falsos positivos, tanto en Natagaima (120 días) como en Caquetá (90 días).

7 HERRAMIENTA DE VISUALIZACIÓN

La herramienta de visualización fue desarrollada en Power BI, basándose en los resultados obtenidos en los capítulos 5 y 6. Esta herramienta se seleccionó debido a su curva de aprendizaje moderada y a sus capacidades avanzadas para los procesos de extracción, transformación y carga de datos (ETL). Entre las funcionalidades que ofrece se incluyen las probabilidades de recuperación por píxel, las agrupaciones mensuales por cohorte y las métricas de desempeño del modelo HGB.

El objetivo de la visualización es presentar en forma de síntesis, la tasa de recuperación estimada y observada en las dos zonas Natagaima y Caquetá, facilitando su uso operativo para priorización de acciones y monitoreo.

A continuación, se presenta el resultado final de la visualización. La Figura 25 ilustra la vista principal del tablero. Se emplea una estructura y organización que facilita la distribución de los datos, proporcionando al usuario una experiencia inmersiva a través de la inclusión de texto alternativo y descripciones, así como un adecuado uso del color y contraste, y la interacción y accesibilidad gráfica. Todos los componentes se alimentan del mismo conjunto de datos consolidado sobre la recuperación post-incendio. Los indicadores resumen el total de los píxeles evaluados y el porcentaje de recuperación por zona. De acuerdo con el capítulo 6, los gráficos de líneas representan las curvas mensuales, mientras que los paneles de métricas evidencian los valores de PR-AUC, ROC-AUC, Brier y F1, los cuales se obtuvieron en la estimación del modelo.

En consecuencia, la herramienta de visualización, integra en una sola interfaz los resultados cuantitativos del modelo y su traducción a indicadores operativos.

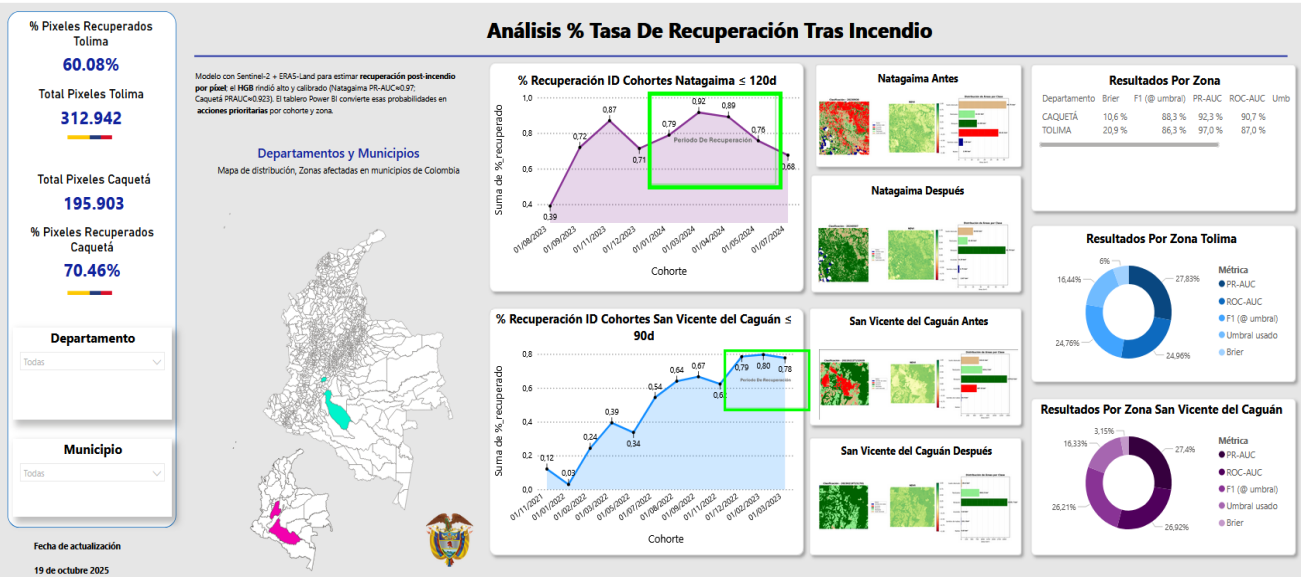

Figura 25. Tablero de visualización de la tasa de recuperación en Power BI.

El panel derecho proporciona acceso inmediato a datos de indicadores clave (KPIs), destacando el total de píxeles recuperados en cada una de las zonas, así como su correspondiente porcentaje. En esta sección, el usuario final cuenta con un filtro que permite seleccionar el área de interés correspondiente a los departamentos y municipios analizados.

En la sección central se presenta un mapa de Colombia que destaca la ubicación y posición de las zonas de interés. Este recurso visual facilita la observación del comportamiento y la tasa de recuperación mensual de cada área analizada.

Se incluyó una representación gráfica para cada resultado obtenido por zona, la cual ilustra la condición de la zona de incendio antes y después de la intervención en cada una de las áreas de interés. Es posible observar tanto la imagen correspondiente al día del incendio como la de un año posterior. La información se contrasta con el índice de vegetación NDVI con el propósito de validar las clasificaciones de las distintas clases, empleando su porcentaje correspondiente.

En la sección final, en el panel izquierdo, se presentan los resultados de las métricas utilizadas para cada uno de los modelos.

A continuación, se presenta el enlace de acceso a la herramienta de visualización desarrollada, la cual permite explorar los resultados del modelo, las métricas de desempeño y las curvas de priorización por zonas de estudio. Esta plataforma facilita el análisis operativo y la verificación de patrones espaciales y temporales asociados a los procesos de recuperación de la vegetación tras incendios

Herramienta de visualización para estimación de recuperación post-incendio [56]:

https://app.fabric.microsoft.com/view?r=eyJrIjoiMWU3Y2VkYWItNmU5OC00OGM3LTlkOGYtNTA5N2YyNTI2ZmM3IiwidCI6ImQ0OWRlNDMxLThlYzItNDYyNy05NWRjLWExYjA0MWJiYWIzMCIsImMiOjR9

8 CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

Los resultados de este proyecto confirman la capacidad del modelo seleccionado HGB para ser desplegado como un instrumento fiable en un sistema de apoyo a la toma de decisiones para la administración de zonas post-incendio. La evaluación permite resaltar no solo el rendimiento estadístico, sino también fortalezas complementarias en cada región que corroboran su aplicabilidad práctica.

Las métricas nos indican que el modelo da prioridad correcta, una proporción importante de las áreas que se regeneran y mantiene un equilibrio razonable entre errores y aciertos, por lo que tiene potencial para ser utilizado como herramienta de apoyo a la toma de decisiones en la administración de zonas post-incendio, siempre y cuando su desempeño se siga validando en operación.

En la zona de Natagaima en el departamento del Tolima, el modelo alcanzó su pico de eficiencia en la identificación y gestión de casos críticos con un PR-AUC de 0.970, ROC-AUC de 0.870 y F1 de 0.863 (umbral 0.573) en un horizonte de tiempo de (<120 días), de manera conjunta con el top-10% de los pixeles de mayor riesgo para capturar aproximadamente el 23% de las recuperaciones, se destaca una buena precisión para la focalización de esfuerzos. Esta “recuperación” de casos permite gestionar las intervenciones en las áreas donde el impacto será más significativo, asegurando de esta manera la eficiencia en la utilización de los recursos. La sincronía óptima (movimiento en tándem) entre la probabilidad promedio y el porcentaje de recuperación observado por cohortes constituyen una evaluación definitiva de una calibración óptima de la fiabilidad del modelo.

Caquetá es una zona con características diferentes, el horizonte de tiempo es más exigente (<90 días) y en un contexto ecológico de mayor complejidad, el modelo mantuvo un rendimiento equilibrado y robusto, demostrando un alto rendimiento con un PR-AUC de 0.923, ROC-AUC de 0.907, F1 de 0.883 (umbral 0.55) y Brier de 0.106, de manera general indica una capacidad superior para distinguir de manera global entre los pixeles recuperables y aquellos que no, una característica esencial para alertas tempranas por deforestación. Adicionalmente, un puntaje de Brier de 0.106 corrobora que las probabilidades propuestas están calibradas y reflejan de manera precisa la realidad observada, un aspecto relevante para que los responsables de las gestiones de intervención lo empleen como un apoyo a la toma de decisiones.

En conclusión, se puede deducir que el modelo desarrollado supera la simple exactitud estadística, en el municipio de Natagaima, proporciona el "panorama" más confiable y sólido, mientras que, en Caquetá, proporciona el "panorama" más fiable y robusto. Esta dualidad de fortalezas entre una priorización extrema en contraposición a una discriminación global confiable, afirma el modelo como una solución holística y escalable para el seguimiento de la recuperación post-incendios en diversos

contextos biogeográficos de Colombia.

8.2 TRABAJOS FUTUROS

Es relevante señalar que es posible llevar a cabo diversos trabajos futuros que complementen los resultados obtenidos y fortalezcan el sistema de estimación de la tasa de recuperación vegetal en el territorio colombiano. Para la clasificación por umbrales de las distintas clases, se sugiere integrar otros análisis en diversas zonas del país, con el objetivo de establecer una caracterización de la reflectancia que permita generar una fuente de datos en formato. JSON. Esta fuente facilitaría la extracción automática de las características del terreno, lo que posibilitaría la realización de análisis de manera automatizada, teniendo en cuenta los datos específicos de cada región del país para lograr una clasificación más precisa.

El siguiente paso podría centrarse en el desarrollo de un modelo espacio temporal que permita una interpretación más precisa del contexto. Este modelo debería integrar datos climáticos de alta resolución y considerar un mayor número de variables en los puntos de estudio, tales como la humedad, la radiación solar, la presión atmosférica, los vientos alisios, la nubosidad y las señales del fenómeno ENSO. Además, es fundamental incluir las características específicas de cada región, como los tipos de suelo, que podrían influir en el análisis del modelo.

Se sugiere que se desarrollen estudios relacionados no solo con la recuperación, sino también con un análisis más detallado de la severidad de los incendios, utilizando modelos como GBM/LightGBM. Este enfoque debería incluir una calibración que se ajuste a cada escenario, estación, nivel de severidad de la quema y umbrales dinámicos fundamentados en las pruebas más recientes. Asimismo, es posible evidenciar la incertidumbre asociada a cada estimación, así como detallar los resultados utilizando SHAP por cohorte. Este proceso puede culminar con la implementación de un enfoque de MLOps ligero, que incluya el monitoreo de cambios y la recalibración mensual, conectada de manera directa al tablero operativo.

9 BIBLIOGRAFÍA

[1] R. M.-F. A. Azcárate, «Meteorología, socioeconomía y gestión del riesgo de desastres del evento El Niño Oscilación del Sur en Colombia,» *Revista Mutis,* vol. vol. 6, nº no. 2, pp. pp. 95-109, 2016.

[2] M. P. Lizarazo, «INFOAMAZONIA,» 21 05 2024. [En línea]. Available: https://infoamazonia.org/es/2024/05/21/disminuyo-la-deforestacion-en-la-amazonia-pero-los-incendios-aumentaron-mas-del-150/. [Último acceso: 21 05 2024].

[3] UNGRD, «Unidad Nacional para la Gestión del Riesgo de Desastres,» UNGRD, 23 11 2024. [En línea]. Available: https://portal.gestiondelriesgo.gov.co/Paginas/Objetivos.aspx. [Último acceso: 23 11 2024].

[4] IDEAM, «Instituto de Hidrología, Meteorología y Estudios Ambientales,» IDEAM, 23 11 2024. [En línea]. Available: https://www.ideam.gov.co/transparencia/informacion-de-la-entidad/mision-vision-funciones-y-deberes. [Último acceso: 23 11 2024].

[5] F. H. B. T. F. G. A. M. M. R. D. P. C. E. P. R. Á. d. C. P. L. Dolors Armenteras, Incendios de la cobertura vegetal en Colombia, Cali: Universidad Autónoma de Occidente, 2011.

[6] e. a. S. Bova, «Wildfire prediction and management: A review of models and strategies,» *Forest Sci.,* vol. vol. 56, nº no. 4, pp. pag. 397-407, 2010.

[7] E. Chuvieco, «Satellite remote sensing of fires and burned areas,» *Geol. Nat. Hazards,* vol. vol. 45, nº no. 2, pp. pag. 855-872, 2002.

[8] e. a. N. Koutsias, «An overview of fire behavior prediction models,» *Environ. Model. Softw,* vol. vol. 20, nº no. 1, pp. pp. 3-15, 2005.

[9] D. R. K.C. Ryan, «Forest fires and the atmosphere,» *J. Geophys. Res.: Atmos,* 2001.

[10] P. R.-G. T. L. V. A. Sofia Giannakidou, «Leveraging the power of internet of things and artificial intelligence,» *Internet of Things,* vol. Vol. 26, pp. pp. 1-14, 2024.

[11] V. y. D. T. Ministerio de Ambiente, *lan nacional de prevención control de incendios forestales y restauración de áreas afectadas,* Bogotá D.C: Comisión Nacional Asesora para la Prevención y Mitigación de Incendios Forestales, 2002.

[12] Y. Z. W. M. L. B. W. a. X. Z. Tian, «Forest Fire Spread Monitoring and Vegetation Dynamics Detection Based on Multi-Source Remote Sensing Images,» *Remote Sensing,* vol. vol 14, nº no. 18, p. pp. 4431, 2022.

[13] T. Lillesand, R. Kiefer y J. W. Chipman, «REMOTE SENSING AND IMAGE INTERPRETATION,» *In Australian Journal of Geodesy,* vol. vol. 39, pp. pp. 1-39, 2015.

[14] A. F. &. K. L. Alqurashi, «Investigating the Use of Remote Sensing and GIS Techniques to Detect Land Use and Land Cover Change: A Review,» *Advances in Remote Sensing,* vol. vol. 2, nº no. 2, pp. 193-204, 2013.

[15] M. &. R. M. (. A. d. l. t. e. d. d. s. B. D. L. A. D. G. E. 2.-3. González, «González, M.E., & Rodríguez, M.P. (2013). Aplicaciones de la teledetección en degradación de suelos. Boletin De La Asociacion De Geografos Espanoles, 285-308.,» *González, M.E., & Rodríguez, M.P. (2013). Aplicaciones de la teledetección en degradación de suelos. Boletin De La Asociacion De Geografos Espanoles, 285-308.,* pp. pp 285-308, 2013.

[16] U. N. d. Colombia, *CARTOGRAFIA GEOLOGICA Y MODELAMIENTO ESTRUCTURAL DE LAS CUENCAS DE URABÁ Y SINÚ-SAN JACINTO A PARTIR DE LA INTERPRETACION DE IMÁGENES DE SENSORES REMOTOS*

Y MONITOREO SISMICO, Bogota: Agencia Nacional de Hidrocarburos, 2008.

[17] R. V. M. A. S. A. Navarro Cerrillo, «Sensores, acceso y procesado de imágenes multiespectrales y térmicas de interés forestal,» de *Geociencias aplicadas a la gestión forestal*, Universidad de Cordoba, 2024, p. 22.

[18] H. B. a. F. E.-S. A. E. Manurung, «Analysis of Drought at Terai Regions in Uttarakhand using Multiple Remotely Sensed Data,» *IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES),* pp. pp 1-7, 2022.

[19] L. Z. L. S. Philip Kinghorn, «A hierarchical and regional deep learning architecture for image description generation,» *Pattern Recognition Letters,* vol. 119, pp. pag. 77-85, 2019.

[20] M. Z. M. Y. Y. S. W. &. W. M. Fu, «LiDAR-based vehicle localization on the satellite image via a neural network,» *Robotics and Autonomous Systems,* p. pp 103519, 2020.

[21] K. G. T. R. L. S. P. &. M. d. S. J. R. Raiyani, «Sentinel-2 Image Scene Classification: A Comparison between Sen2Cor and a Machine Learning Approach,» *Remote Sensing,* vol. 13, nº 2, p. 300, 2021.

[22] D. M. V. H. P. R. a. M. V. U. V. Mayorga Arias, «USO DEL ÍNDICE NORMALIZADO DE VEGETACIÓN PARA LA ELABORACIÓN DE PLANOS DE CULTIVO.,» *Opuntia Brava,* vol. vol. 11, nº no. 2, pp. pp. 261-264, 2019.

[23] B. M. Meneses, «Vegetation Recovery Patterns in Burned Areas Assessed with Landsat 8 OLI Imagery and Environmental Biophysical Data,» *Fire,* vol. Vol. 4, nº no. 76, pp. pp. 2-20, 2021.

[24] I. H. Sarker, «Machine Learning: Algorithms, Real-World Applications and Research,» *SN Computer Science,* vol. vol. 2, nº no. 160, pp. pp. 5-8, 2021.

[25] M. N. S. S. Ke-Lin Du, «Neural Networks and Statistical Learning,» de *Fundamentals of Machine Learning*, Second ed., Springer, 2019, pp. pp. 21-63.

[26] Z. Q. Z. P. X. S. T. &. W. X. Zhao, «Object detection with deep learning: A review,» *IEEE transactions on neural networks and learning systems,* vol. vol. 30, nº no. 11, pp. pp. 3212-3232, 2019.

[27] R. P. T. &. S. K. Indrakumari, «Introduction to Deep Learning. Advanced Deep Learning for Engineers and Scientists: A Practical Approach,» de *Introduction to deep learning*, India, 2021, pp. pp. 1-22.

[28] S. D. C. G. a. A. L. R. J. Cintra, «Low-Complexity Approximate Convolutional Neural Networks,» *IEEE Transactions on Neural Networks and Learning Systems,* vol. vol. 29, nº no. 12, pp. pp. 5981-5992, 2018.

[29] A. S. a. A. Mahmood, «Review of Deep Learning Algorithms and Architectures,» *IEEE Access,* vol. vol. 7, pp. 53040-53065, 2019.

[30] L. W. J. R. a. T. L. J. Ker, «Deep Learning Applications in Medical Image Analysis,» *IEEE Access,* vol. vol. 6, pp. pp. 9375-9389, 2018.

[31] E. A. P. y. G. G. V. D. C. Montgomery, «Introduction to Linear Regression Analysis,» *Wiley Series in Probability and Statistics,* vol. 5th ed., 2012.

[32] W. G. L. P. W. H. Z. M. L. X. Sun Zhigang, «An improved random forest based on the classification accuracy and correlation measurement of decision trees,» *Expert Systems with Applications,* vol. 237, nº 0957-4174, 2024.

[33] T. Y. L. &. L. D. Liu, Change detection using deep learning approach with object-based image analysis. Remote Sensing Environ, Elsevier Inc, 2021.

[34] J. H. Friedman, «Greedy Function Approximation: A Gradient Boosting Machine,» *Annals of Statistics,*

vol. 29, nº 5, p. 1189–1232, 2010.

[35] A. J. F. M. A. T. Ferreira, «Ensemble Machine Learning: Methods and Applications,» de *Boosting Algorithms: A Review of Methods, Theory, and Applications*, New York, NY, Springer New York, 2012, pp. 35-85.

[36] R. V. K. Priya, «Vegetation change detection and recovery assessment based on post-fire satellite imagery using deep learning,» *Sci Rep,* vol. vol. 14, nº no. 12611, pp. pp. 1-23, 2024.

[37] A. A. S. S. S. a. T. H. Bradley, «Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score,» *Wea. Forecasting,* nº 23, pp. 992-1006, 2018.

[38] T. J. K. R. Rainio Oona, «Evaluation metrics and statistical tests for machine learning,» *Scientific Reports,* vol. 14, pp. 1-14, 2024.

[39] P. T. N. D. M. O. J. Miller Catriona, «A review of model evaluation metrics for machine learning in genetics and,» *Frontiers in bioinformatics,* vol. 4, 2024.

[40] R. R. P. A. R. M. K. Hoda Khoshvaght, «A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction,» *Journal of Environmental Chemical Engineering,* vol. 13, nº 6, pp. 1-10, 2025.

[41] T. S. y. M. Rehmsmeier, «The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,» *PLoS ONE,* vol. 10, nº 3, p. e0118432, 2015.

[42] T. Munzner, «Visualization Analysis and Design,» de *Chapter 3: Why: Task Abstraction*, Boca Raton, FL, AK Peters Visualization Series, 2015, pp. pág 44-57.

[43] R. Parthe, «Comparative Analysis of Data Visualization Tools: Power BI and Tableau,» *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT,* vol. 07, nº 11, pp. 1-2, 2023.

[44] M. K. Gokhale Leena, «Comparative Study of Data Visualization Tools,» de *International Conference on Engineering & Technology (ICET-20)*, Pune, India, 2020.

[45] R. M. D. B. A. Fernando Pérez-Cabello, «Remote sensing techniques to assess post-fire vegetation recovery,» *Current Opinion in Environmental Science & Health,* vol. vol. 21, pp. pp. 1-9, 2021.

[46] C. Alegria, «Vegetation Monitoring and Post-Fire Recovery: A Case Study in the Centre Inland of Portugal.,» *Sustainability,* vol. vol. 14, pp. pp. 1-31, 2022.

[47] U. N. P. L. G. D. R. D. DESASTRES, «UNGRD,» UNGRD, 24 12 2024. [En línea]. Available: https://portal.gestiondelriesgo.gov.co/Paginas/Noticias/2024/Incendios-forestales-inundaciones-y-movimientos-en-masa-las-emergencias-mas-frecuentes-en-Colombia-durante-2024.aspx#:~:text=Tras%20un%20balance%20entregado%20por,de%20216%20mil%20hect%C3%A1reas. [Último acceso: 10 04 2025].

[48] ESRI, «Mapa de eventos por incendios ocurridos en Colombia reportados por las autoridades ambientales.,» ESRI, 15 03 2024. [En línea]. Available: https://www.arcgis.com/apps/dashboards/e24388544f4a43e7920d2269fb0b1f9f. [Último acceso: 15 03 2024].

[49] A. E. Europea, «ESA,» 08 12 2024. [En línea]. Available: https://www.eoportal.org/satellite-missions/copernicus-sentinel-2#eop-quick-facts-section. [Último acceso: 12 06 2025].

[50] E. S. A. (. S.-2. M. –. L.-2. P. D. [. G. E. Copernicus., «Copernicus,» Copernicus, 2020. [En línea]. Available: https://sentiwiki.copernicus.eu/web/s2-applications. [Último acceso: 2025].

[51] IDEAM, «Catalogo Estaciones IDEAM,» IDEAM, 05 12 24. [En línea]. Available:

https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Catalogo-Estaciones-IDEAM/n6vw-vkfe. [Último acceso: 05 12 24].

[52] B. S. L. B. B. D. R. M. L. N.-E. T. L. L. &. A. T. Marcel Buchhorn, Copernicus Global Land Service: Land Cover 100m: version 3 Globe 2015-2019: Product User Manual, VITO, 2020.

[53] R. P. J. Lannacone, «Una revision del uso de imagenes Sentinel - 2 para el monitoreo de la cobertura boscosa a nivel global,» *Ingenieria y Competitividad,* vol. 25, nº 3, 2023.

[54] M. y. E. A. Instituto de Hidrología, «Sistema de Monitoreo de Bosques y Carbono - SMBYC,» IDEAM, 09 08 2017. [En línea]. Available: https://github.com/SMByC. [Último acceso: 21 07 2025].

[55] R. T. y. J. F. T. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2nd ed., New York: USA: Springer, 2019.

[56] «Herramienta de visualización para estimación de recuperación post-incendio,» 10 09 2025. [En línea]. Available: https://app.fabric.microsoft.com/view?r=eyJrIjoiMWU3Y2VkYWItNmU5OC00OGM3LTlkOGYtNTA5N2YyNTI2ZmM3IiwidCI6ImQ0OWRlNDMxLThlYzItNDYyNy05NWRjLWExYjA0MWJiYWIzMCIsImMiOjR9. [Último acceso: 08 12 2025].

[57] F. A. Colombiana, *Imagen satelital,* Cartagena, 2019.

[58] Z. H. L. Z. X. F. F. Z. Q. Y. L. F. Xubing Yang, «Preferred vector machine for forest fire detection,» *Pattern Recognition,* vol. Volume 143, 2023.