ANÁLISIS DE LA DEFORESTACIÓN EN LA AMAZONÍA COLOMBIANA USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

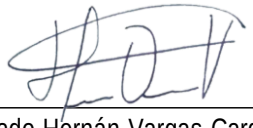
PAOLA ANDREA LEÓN ACOSTA

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Director Guillermo Andrés Otero

Jurado Valentina Corchuelo Guzmán



Jurado Hernán Vargas Cardona

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en ciencia de datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 06 de febrero de 2024

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 06 de febrero de 2024

Autor: PAOLA ANDREA LEÓN ACOSTA

Título del Trabajo de Grado: "ANÁLISIS DE LA DEFORESTACIÓN EN LA AMAZONÍA COLOMBIANA USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO"

Director: GUILLERMO ANDRÉS OTERO

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

_____

Firma del Director del Trabajo de Grado

Santiago de Cali, 06 de Febrero del 2024

Doctora
Gloría Inés Alvarez V.
Directora Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "ANÁLISIS DE LA DEFORESTACIÓN EN LA AMAZONÍA COLOMBIANA USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO", el cual fue realizado por el (los) estudiante (s) PAOLA ANDREA LEÓN ACOSTA con código (s) 8.975.326 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de GUILLERMO ANDRÉS OTERO.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Paola Andrea León Acosta

C.C. 1.022.384.363 de Bogotá D.C.

Guillermo Andrés Otero Martínez

C.C. 80.036.894 de Bogotá D.C.

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

Maestría en Ciencia de datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "ANÁLISIS DE LA DEFORESTACIÓN EN LA AMAZONÍA COLOMBIANA USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO."

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Conservación de recursos naturales
4. ESTUDIANTE (S): Paola Andrea León Acosta
5. CORREO ELECTRÓNICO: paolaleon12@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Calle 65 F # 80 d 09 - 3142077900
7. DIRECTOR: Guillermo Andrés Otero
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Externo
9. CORREO ELECTRÓNICO DEL DIRECTOR: guillermoterom@hotmail.com
10. CO-DIRECTOR(ES) (Si aplica):
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
12. OTROS GRUPOS O EMPRESAS:
13. PALABRAS CLAVE (al menos 5): Deforestación, aprendizaje automático, Google earth engine, imágenes satelitales, Amazonía colombiana, aprendizaje supervisado, aprendizaje no supervisado, redes neuronales artificiales.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Vida de ecosistemas terrestres
15. FECHA DE INICIO (Desarrollo del proyecto): 1/01/2023
16. RESUMEN (máximo 400 palabras).

Debido al alto impacto de la deforestación en el calentamiento global, el aumento de enfermedades zoonóticas y el riesgo de extinción de la biodiversidad, surge la necesidad de desarrollar nuevos enfoques para la medición y análisis de la deforestación que permitan a los gobiernos tener una mejor compresión de este fenómeno para centrar su atención y recursos a atender esta crisis ambiental en las zonas más vulnerables. Dada esta situación y considerando el amplio uso de los algoritmos de aprendizaje automático para analizar datos complejos como imágenes y textos, este proyecto tuvo como objetivo analizar el comportamiento de la deforestación en la Amazonía colombiana usando diferentes técnicas de aprendizaje automático con imágenes satelitales de Google earth engine, considerando estas metodologías como nuevas propuestas de medición en el análisis de la cobertura forestal. Posteriormente, se evaluaron estos modelos mediante métricas de evaluación, una vez seleccionado el modelo con mejor rendimiento, se identificaron las zonas con deforestación en las imágenes satelitales, y a partir de estos resultados se cuantificó y analizó el incremento de la perdida de bosques en un periodo determinado con el propósito de generar alertas de las zonas más vulnerables, y así brindar una herramienta que se pueda considerar como un insight para la formulación de planes de acción y políticas para la prevención y reforestación.

ANÁLISIS DE LA DEFORESTACIÓN EN LA AMAZONÍA COLOMBIANA USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO.

Paola Andrea León Acosta
Código 8.975.326

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

Director(a)
Guillermo Andrés Otero

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 6 DE 2023

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

El aumento de la deforestación en la Amazonía colombiana durante los últimos años tiene en alerta al gobierno y a las principales organizaciones internacionales encargadas de la conservación del medio ambiente, que consideran esta perdida desbocada de bosques como una crisis ambiental sin precedentes y que al llegar a cierto punto de deforestación ya no habrá retorno. Esta deforestación acelerada tiene repercusiones críticas en la biodiversidad, el calentamiento global, la aridez del terreno y las enfermedades zoonóticas. Específicamente, cuando se destruyen los bosques tropicales, una enorme cantidad de dióxido de carbono almacenado por la vegetación se libera a la atmósfera, lo que acelera el calentamiento global [1]. Ante esta situación surge la necesidad de desarrollar nuevos modelos de medición de la deforestación que permitan analizar dicho comportamiento, cuantificar e identificar las zonas más vulnerables en la Amazonía colombiana, para que así el gobierno pueda centrar su atención en la formulación de estrategias y recursos en ellas.

En este proyecto se identificaron las zonas donde se presenta deforestación en la Amazonía colombiana con técnicas de aprendizaje automático, las cuales se caracterizan por su gran utilidad en el análisis de datos más grandes y complejos como lo son las imágenes satelitales, que durante las últimas décadas han tomado mayor relevancia en el estudio científico de los componentes y características del espacio geográfico, para este proyecto en particular, nos interesó analizar la cobertura forestal. Adicionalmente, al desarrollar modelos de aprendizaje automático se pueden obtener resultados más precisos y que permiten calcular la tasa de deforestación en las imágenes satelitales de la Amazonía colombiana, es decir, cuantificar la pérdida de bosques. A partir de estos resultados, se desarrolló una herramienta visual que permite identificar las zonas con deforestación, y generar alertas de las zonas donde se presentó mayor deforestación en un periodo determinado, siendo una herramienta de fácil navegabilidad para la comprensión de los resultados. El desarrollo de este proyecto se considera como una aproximación inicial a un nuevo enfoque de medición de la deforestación en la Amazonía colombiana y un acercamiento a herramientas de análisis que son fundamentales para la formulación de políticas de contención y reforestación de bosques, que lo ha mencionado la ministra de ambiente y desarrollo sostenible como uno de los principales retos.

El presente documento se organiza de la siguiente manera: primero, se define el problema que se propone abarcar en este proyecto aplicado referente a la medición y monitoreo de la deforestación en la Amazonía colombiana junto con la formulación del problema. Segundo, se presenta el objetivo general y los objetivos específicos que se deben cumplir. Posteriormente, se presenta el marco teórico con los conceptos centrales de la deforestación, las imágenes satelitales y los algoritmos de aprendizaje automático, seguido de los antecedentes de proyectos con objetivos similares o en los que usaron técnicas de aprendizaje automático para analizar la dinámica de la deforestación. A continuación, se describen cada una de las actividades que se realizaron para el desarrollo de este proyecto. Finalmente, se exponen las conclusiones, los trabajos futuros y las referencias bibliográficas.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

La Amazonía comprende el 42% del territorio colombiano y se le conoce como el pulmón del mundo debido a su gran extensión selvática. Esta selva tropical suministra humedad a toda Sudamérica, influye en las lluvias de la región, contribuye a la estabilización del clima global y posee la mayor biodiversidad del mundo. Sin embargo, nunca en la historia la Amazonía había estado tan amenazada por el aumento crítico de la deforestación, a causa de la expansión de la agricultura, la ganadería, las concesiones mineras, los cultivos ilícitos, los incendios forestales y la tala ilegal. Estas son solo algunas de las presiones que tienen en riesgo a la selva más grande del mundo, y que tienen en peligro la integridad de los ecosistemas, las especies y las comunidades [1].

El proceso de deforestación en la Amazonía colombiana ha sido constante, si bien ha tenido picos históricos muy altos, también hubo años en los que bajó significativamente; sin embargo, en los últimos años ha venido subiendo de manera persistente, lo que se considera una crisis ambiental sin precedentes. Durante el primer semestre del 2022 la Amazonía Colombiana perdió más de 52 mil hectáreas, correspondiente a un aumento del 11% frente al 2021, esta deforestación se concentró en los departamentos de Meta, Caquetá, Guaviare y Putumayo [2]. Esta es una de las amenazas más grandes que tiene Colombia para su desarrollo sostenible y económico. De un modo general, este fenómeno vuelve desérticas las tierras, contribuye al calentamiento global con el aumento de la emisión de gases de efecto invernadero y causa la pérdida de biodiversidad, por eso se considera que el ser humano está causando la extinción masiva de especies.

Ante esta situación, la ministra de ambiente y desarrollo sostenible Susana Muhamad tiene entre sus retos enfrentar la deforestación, ya que señala que no hay una contención y que por el contrario está aumentando, además menciona que lo preocupante es que en unos años se podría ver afectado el flujo de agua que llega desde el amazonas a la región andina. Entre sus planes de acción esta implementar nuevas lógicas de medición que permitan identificar que se está quedando atrás y el impacto acumulativo que tiene la deforestación, hace énfasis en que no solo hay que frenar la deforestación, sino restaurar [3].

El problema que se abordó en este proyecto está relacionado con uno de los retos del ministerio de ambiente mencionado anteriormente: desarrollar nuevas lógicas de análisis y medición de la deforestación en la Amazonía colombiana a través de algoritmos de aprendizaje automático, que brindan resultados más precisos y que permiten analizar datos más complejos. Por tal razón, se

desarrolló un modelo de aprendizaje automático que identificara en las imágenes satelitales las zonas donde se presenta deforestación en la Amazonía colombiana, y a partir de estas se calculó la tasa de deforestación para analizar el aumento de la pérdida de bosques en un periodo determinado e identificar las zonas más vulnerables. De esta forma, se brinda un acercamiento a este fenómeno desde los datos para apoyar a las investigaciones de las facultades de ciencias naturales, y además de ser una aproximación inicial para proporcionar herramientas de análisis que son indispensables en la formulación de planes de acción y de políticas de conservación y sostenibilidad ambiental.

1.2. FORMULACIÓN DEL PROBLEMA

El problema que se abordó en el presente proyecto implicó responder los siguientes interrogantes: ¿Cómo evaluar la deforestación en la Amazonía colombiana usando técnicas de aprendizaje automático?, ¿De qué manera incorporar los modelos de aprendizaje automático para analizar la deforestación con imágenes satelitales?, ¿Cuáles técnicas de aprendizaje automático deben ser seleccionadas de tal forma que aporten adecuadamente a la solución del problema?, ¿Cómo seleccionar los datos requeridos para el análisis?, ¿Cómo evaluar el nivel de desempeño del modelo desarrollado?, ¿Cómo identificar las zonas más afectadas por la deforestación en la Amazonía colombiana? ¿Cómo realizar una herramienta de visualización con los resultados?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Analizar la deforestación en la Amazonía colombiana mediante técnicas de aprendizaje automático e imágenes satelitales de Google earth engine para cuantificar la pérdida de bosque y desarrollar nuevas herramientas fundamentales en la formulación de políticas de contención y reforestación.

2.2 OBJETIVOS ESPECÍFICOS

1. Realizar la minería de las imágenes satelitales de la Amazonía colombiana en Google earth engine y seleccionar las variables de análisis que se utilizarán.
2. Desarrollar los modelos de aprendizaje automático para identificar las zonas con deforestación.
3. Evaluar el desempeño de los modelos desarrollados con métricas estadísticas.
4. Analizar el aumento de la deforestación en un periodo determinado para identificar las zonas prioritarias.
5. Implementar una herramienta visual que permita consultar los resultados del análisis.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

A continuación, se presentan los temas que se relacionan con el desarrollo del proyecto, teniendo en cuenta que la temática fundamental es la deforestación, pero se utilizaron herramientas de ciencia de datos, tales como el procesamiento de imágenes satelitales, el modelado de algoritmos de aprendizaje supervisado y la validación.

3.1.1. DEFORESTACIÓN

La Amazonía es la mayor región de bosque tropical del planeta, y pierde cada año enormes extensiones de selva. El cambio en la cobertura y el uso de la tierra es un proceso generalizado, acelerado y significativo que puede traer consecuencias negativas para los seres humanos. Durante los últimos cincuenta años la transformación de los ecosistemas de la Amazonía colombiana, causada principalmente por procesos de deforestación y expansión de la frontera agropecuaria, ha ocasionado impactos sin precedentes en la biodiversidad, el clima y el ecosistema [4]. Entender las causas y el funcionamiento de estos procesos se ha convertido en uno de los principales objetivos de la investigación a nivel mundial.

3.1.1.1. PRINCIPALES AGENTES Y CAUSAS DIRECTAS

En la tabla 1 se mencionan los principales agentes y causas directas de la transformación de los bosques en el territorio nacional identificados por el IDEAM[1] [5].

TABLA 1.
PRINCIPALES AGENTES Y CAUSAS DIRECTAS

Variable/determinante	Agente relacionado
Expansión de la frontera agropecuaria (actividades lícitas e ilícitas).	Agricultores, Ganaderos, Actores armados.
Minería (efectos indirectos por construcción de vías de acceso).	Empresas mineras
Variables biofísicas (características de los suelos, clima, etc.).	No aplica.
Variables demográficas (crecimiento, densidad, estructura, etc.).	Agricultores, Ganaderos
Crecimiento de los precios de los commodities en los mercados internacionales	Agricultores, Ganaderos, Actores armados, Empresas mineras.

[1] Instituto de Hidrología, Meteorología y Estudios Ambientales

Mercado laboral	Agricultores, Ganaderos, Empresas mineras.
Políticas agrarias y de tierras (ausencia, incentivos perversos, etc.).	Agricultores, Ganaderos, Actores armados, Empresas mineras.
Tecnologías de la producción	Agricultores, Ganaderos, Actores armados, Empresas mineras.

3.1.1.2. CAUSAS SUBYACENTES

Las causas subyacentes o procesos sociales son factores que refuerzan las causas directas de la deforestación o degradación forestal. En el caso particular de Colombia, el IDEAM ha identificado los siguientes factores que agrupan complejas variables sociales, políticas, económicas y tecnológicas [5].

- Consolidación de la tendencia de urbanización, impulsada por la creciente industrialización en las ciudades principales.
- Saturación de tierras de pequeños propietarios en la región Andina, con el subsecuente incremento en la migración a las zonas de frontera de los bosques de tierras bajas de la Amazonía y las faldas de los Andes.
- Creciente conflicto armado reforzado por actividades económicas ilegales.
- Desarrollo y aumento estable del crecimiento del narcotráfico que ha invadido progresivamente las fronteras agrícolas.
- Incursión progresiva en los mercados internacionales con una economía dictada cada vez más por el entorno macroeconómico global.
- Políticas proteccionistas parcializadas a un número limitado de productos agropecuarios.
- Ausencia de una política fiscal que promueva el uso eficiente de la tierra.
- Distribución desigual de la tenencia de la tierra.

3.1.1.3. CONSECUENCIAS DE LA DEFORESTACIÓN EN EL AMAZONAS

La deforestación de la Amazonía tiene efectos adversos. Entre ellos, cabe destacar:

- Calentamiento global

De acuerdo con Greenpeace, en la década de los 90 la selva amazónica absorbía 2.000 millones de toneladas de CO2, una cifra que, en la actualidad, se ha reducido a la mitad. La subsiguiente acumulación de mayores cantidades de CO2 en la atmósfera contribuye al cambio climático, aumentando la temperatura del planeta a causa del efecto invernadero [6].

- Pérdida de biodiversidad

Se calcula que la selva amazónica alberga el 10 % de la fauna conocida, también un gran número de la desconocida, oculta entre su exuberante vegetación y el 20 % de la flora, más de 10.000 de

sus plantas contienen ingredientes para uso médico o cosmético. La destrucción de su hábitat las sitúa al borde de la extinción, impulsando la pérdida de biodiversidad [7].

- Enfermedades zoonóticas

Según un informe de WWF, el 70 % de las enfermedades humanas son producidas por la destrucción de la naturaleza. En el caso del Amazonas, siendo la mayor selva tropical del planeta, su progresiva deforestación puede provocar un considerable aumento de las enfermedades zoonóticas de origen animal con graves consecuencias sobre la salud humana [8].

3.1.1.4. DEFORESTACIÓN EN EL DEPARTAMENTO DEL GUAVIARE

En Colombia, solo en 2017, se arrasaron 220.000 hectáreas de bosques. La cifra implica un aumento del 23% con relación al año anterior, y esta zona de la Amazonía fue la región más afectada concentrando el 65% del arrasamiento nacional. Al departamento del Guaviare, día a día le arrancan su verde. Su selva estratégica que conecta la Amazonía con la Orinoquía está siendo arrasada para convertir esos suelos en áridos campos de pastoreo al servicio de la ganadería. En los últimos cuatro años, ha tenido un crecimiento progresivo del fenómeno de la deforestación. Según el IDEAM, en 2014 se arrasaron 6.892 hectáreas, en 2015 desaparecieron 9.634, en 2016, la cifra llegó a 11.456 y en 2017 la deforestación arrasó con 38.221 hectáreas [9].

El importante incremento está relacionado con la salida de las FARC[2] como autoridad de facto en zonas del sur y oriente del país. Esa guerrilla no era una organización ambientalista ni nada por el estilo, pero sus lógicas de poder contuvieron por décadas el avance de los terratenientes y así se menguó el derribamiento de selvas y bosques. En los territorios de guerra había inestabilidad, lo que implica falta de garantías, de derechos, y eso contenía tanto a la gente como a las empresas, por tanto, no había motores de deforestación, simplemente porque los que habitaban esos territorios eran actores del conflicto.

Figura 1. Deforestación en el departamento del Guaviare
Fuente: Revista Semana- Artículo la selva a mordiscos

[2] Fuerzas armadas revolucionarias de Colombia.

Figura 2. Deforestación en el departamento del Guaviare
Fuente: Revista Semana- Artículo la selva a mordiscos

3.1.2. GOOGLE EARTH ENGINE E IMÁGENES SATELITALES

Google earth engine es una plataforma para el análisis científico y la visualización de conjuntos de datos geoespaciales para usuarios académicos, sin fines de lucro. Esta herramienta combina un catálogo de varios petabytes de imágenes satelitales, y también proporciona una API y otras herramientas para permitir el análisis de grandes conjuntos de datos, de esta forma, se puede analizar la cobertura forestal y de agua, el cambio de uso del suelo o evaluar la salud de los campos agrícolas, entre muchos otros análisis posibles.

3.1.2.1. CATÁLOGO DE DATOS DE EARTH ENGINE

Earth Engine aloja imágenes históricas terrestres desde hace más de cuarenta años. Se puede encontrar información de diferentes temáticas, asociadas a meteorología, campo geofísico como agricultura o cobertura terrestre, teledetección y dentro de este ámbito encontrar imágenes de satélites como Landsat o Sentinel [10].

LANDSAT Y SENTINEL

Los datos de Sentinel y Landsat son las dos fuentes más utilizadas para observar la Tierra, ya que proporcionan una mirada histórica de hasta 40 años y diferentes bandas espectrales aplicables para casi cualquier cosa: desde el análisis de cosechas hasta la detección de incendios y la vigilancia de glaciares. En específico, los datos del Sentinel-2A MultiSpectral Instrument (MSI) tienen bandas espectrales muy similares a Landsat 8 y 9 (excluyendo las bandas térmicas del sensor infrarrojo térmico (TIRS)).

- **IMÁGENES SATELITALES DE LANDSAT**

Los satélites LANDSAT 7 y 8 son los únicos que actualmente se encuentran activos y son

administrados por la NASA[3], en tanto que la producción y comercialización de las imágenes depende del Servicio Geológico de los Estados Unidos (USGS). La fecha de lanzamiento del satélite LANDSAT 8 fue en 2013 y fue diseñado para una vida útil de 5 años, tiene la capacidad de recolectar, así como transmitir hasta 532 imágenes por día. Se encuentra en una órbita Heliosincrónica, que significa que pasa siempre a la misma hora por un determinado lugar. Tiene visión de toda la superficie terrestre en un lapso de 15 días, y realiza 232 órbitas. Las imágenes LANDSAT 8 están compuestas por 10 bandas espectrales y 1 banda pancromática, que fueron elegidas especialmente para el monitoreo de la vegetación, para aplicaciones geológicas y para el estudio de los recursos naturales [11].

TABLA 2.
BANDAS SATÉLITE LANDSAT 8

BANDAS LANSAT 8	LONGITUD DE ONDA (μm)	RESOLUCIÓN (m)
Banda 1 - Coastal aerosol	0.43 - 0.45	30
Banda 2 - Blue	0.45 - 0.51	30
Banda 3 - Green	0.53 - 0.59	30
Banda 4 – Red	0.64 - 0.67	30
Banda 5 - Near Infrared (NIR)	0.85 - 0.88	30
Banda 6 - Short Wave Infrared 1	1.57 - 1.65	30
Banda 7 - Short Wave Infrared 2	2.11 - 2.29	30
Banda 8 - Panchromatic	0.50 - 0.68	15
Banda 9 – Cirrus	1.36 - 1.38	30
Banda 10 - TIRS 1	10.6 - 11.19	100
Banda 12 - TIRS 2	11.50 - 12.51	100

- **IMÁGENES SATELITALES DE SENTINEL**

Los Sentinel son una nueva flota de satélites diseñada específicamente para proporcionar los abundantes datos e imágenes de que se nutre el programa Copernicus, de la Comisión Europea. Este programa único de vigilancia medioambiental está cambiando drásticamente la forma en que gestionamos nuestro entorno, entendemos y abordamos los efectos del cambio climático y protegemos nuestra vida cotidiana.

Sentinel 2 lleva una innovadora cámara multiespectral de alta resolución, con 13 bandas espectrales que aportan una nueva perspectiva de la superficie terrestre y la vegetación. La misión se basa en una constelación de dos satélites idénticos en la misma órbita, separados por 180 grados, para lograr una cobertura y una descarga de datos óptimos. Cada cinco días los satélites

[3] National Space and Space Administration

cubrirán todas las superficies terrestres, grandes islas y aguas costeras. Sentinel-2A fue lanzada en junio de 2015 y Sentinel-2B en el primer trimestre de 2017 [12].

TABLA 3.
BANDAS SATÉLITE SENTINEL 2

BANDAS SENTINEL 2	LONGITUD DE ONDA (µm)	RESOLUCIÓN (m)
Banda 1 - Coastal aerosol	0.443	60
Banda 2 - Blue	0.49	10
Banda 3 - Green	0.56	10
Banda 4 – Red	0.665	10
Banda 5 - Vegetation Red Edge	0.705	20
Banda 6 - Vegetation Red Edge	0.74	20
Banda 7 - Vegetation Red Edge	0.783	20
Banda 8 – NIR	0.842	10
Banda 8A - Vegetation Red Edge	0.865	20
Banda 9 - Water vapor	0.945	60
Banda 10 - SWIR Cirrus	1.375	60
Banda 11 - SWIR	1.61	20
Banda 12 - SWIR	2.19	20

3.1.2.2. ÍNDICE DE VEGETACIÓN DIFERENCIAL NORMALIZADA (NDVI)

Es un indicador del vigor y densidad de la vegetación captada en una imagen de satélite, permite reconocer la presencia de vegetación en el territorio, reconocer ciertas estructuras vegetales, analizar series temporales de crecimiento de cultivos. El NDVI es probablemente uno de los índices de teledetección más comunes que existen, sus aplicaciones prácticas son increíblemente diversas, entre ellas el cuantificar las existencias forestales y ser utilizado como indicador de la sequía. Entre sus otros usos se encuentran la previsión de zonas de incendio y los mapas de desertificación. Este indicador se calcula con la diferencia entre dos bandas: la del rojo visible (RED) y la del infrarrojo cercano (NIR).

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

Este índice está definido por valores que van de -1.0 a 1.0, donde los valores negativos están formados principalmente por nubes, agua y nieve, y los valores negativos cercanos a cero están formados principalmente por rocas y suelo descubierto. Los valores muy pequeños de 0.1 o menos de la función NDVI corresponden a áreas sin rocas o arena. Los valores moderados de 0.2 a 0.3 representan arbustos y praderas, mientras que los valores grandes de 0.6 a 0.8 indican

bosques templados y tropicales [13].

3.1.3. TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Es una rama de la inteligencia artificial que, a través del uso de algoritmos tiene la capacidad de identificar patrones en datos masivos, y usarlos para elaborar predicciones o para llevar a cabo otros tipos de decisiones en un entorno de incertidumbre. Este tipo de aprendizaje permite a los computadores realizar tareas específicas de forma automática, ya que partirán de un conjunto de datos observados sobre los que se obtendrán reglas de clasificación o patrones de comportamiento, que serán aplicados sobre datos diferentes a aquellos utilizados para el análisis [14].

Existen múltiples técnicas de aprendizaje automático, dependiendo del tipo de información y del paradigma de aprendizaje que se utilice. La selección de la técnica dependerá del objetivo del modelo que se quiere construir. En general, se puede dividir en dos categorías, el aprendizaje supervisado y el no supervisado, y un método avanzado que abarca ambos, como lo son las redes neuronales.

3.1.3.1. APRENDIZAJE SUPERVISADO

La característica principal de este tipo de aprendizaje es que los algoritmos trabajan con datos etiquetados, es que decir que, en la base de datos que se utilizará para construir el modelo contiene información sobre la característica de estudio (variable de salida). El objetivo es entrenar el modelo para que a partir de unas variables explicativas (variables de entrada), les asigne la etiqueta de salida adecuada. Cuando la variable de salida es continua se habla de un problema de regresión, mientras cuando es nominal o discreta, se habla de un problema de clasificación [14].

Los algoritmos de clasificación básicamente consisten en que, para clasificar automáticamente una nueva muestra, se tiene en cuenta la información que pueda extraer de un conjunto de objetos disponibles divididos en clases (etiquetados) y la decisión de una regla de clasificación, algunos de los algoritmos que se pueden realizar en este tipo de problema son: regresión logística, clasificador Bayesiano ingenuo, arboles de decisiones, bosques aleatorios y máquinas de soporte vectorial.

3.1.3.1.1 BOSQUES ALEATORIOS [15]

Un bosque aleatorio es un algoritmo de aprendizaje automático, que combina la salida de varios árboles de decisiones para alcanzar un resultado único. Su facilidad de uso y su flexibilidad han impulsado su adopción, ya que permite manejar problemas de clasificación y regresión.

El algoritmo de bosque aleatorio está formado por un conjunto (ensemble) de árboles de decisión individuales, y cada árbol del conjunto se compone de una muestra de datos extraída de un conjunto de entrenamiento con sustitución, que se denomina muestra de programa de arranque. De esa muestra de entrenamiento, un tercio se establece aparte como datos de prueba, lo que se conoce como la muestra OOB (Out-Of-Bag). A continuación, se inyecta otra instancia de

aleatoriedad a través de la agregación autodocimante de características, lo que añade más diversidad al conjunto de datos y reduce la correlación entre los árboles de decisiones. Dependiendo del tipo de problema, la determinación de la predicción variará. Para una tarea de regresión, los árboles de decisiones individuales se promediarán y, para una tarea de clasificación, un voto mayoritario (por ejemplo, la variable categórica más frecuente) dará como resultado la clase pronosticada. Por último, se utiliza la muestra OOB para la validación cruzada, lo que finaliza la predicción.

3.1.3.1.2 MAQUINAS DE SOPORTE VECTORIAL [15]

Es un algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de clasificación y regresión, incluidas aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz.

- SVMs DE MÁRGENES DURAS

El objetivo del algoritmo SVM es encontrar el hiperplano que maximiza el margen de separación entre las clases, este margen se define como la suma de las distancias del hiperplano a su punto más cercano en cada clase y los puntos más cercanos al hiperplano de separación se denominan vectores de soporte. Por ejemplo, en la siguiente figura se reconocen tres vectores de soporte.

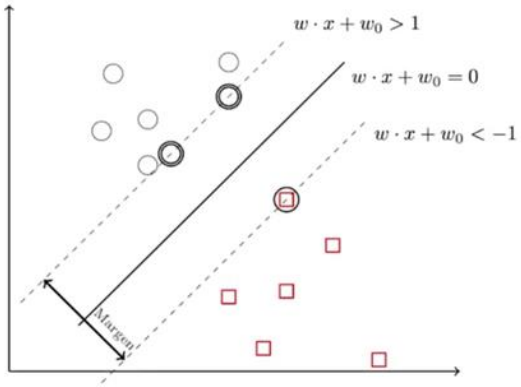


Figura 3. Ejemplo SVMs márgenes duras
Fuente: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Las SVMs determinan el hiperplano de separación a partir del siguiente problema de optimización cuadrática.

$$mín_{W,W_0} \frac{1}{2}\|W\|^2 \quad sujeto\ a \quad Y_n(W^T x_n + w_0) \geq 1, n = 1, \dots, N$$

Dado que el problema de optimización tiene restricciones, se usan multiplicadores de Lagrange; de esta forma, se configura el siguiente problema dual.

$$máx_a \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_n a_m y_n y_m x_n x_m\,, \quad sujeto\ a \quad a_n \geq 0, \quad \sum_{n=1}^{N} y_n a_n = 0$$

$$donde\ a = [a_1, \dots, a_N]$$

Una vez se determinan los valores del vector a y dada una muestra del conjunto de prueba x_*, se puede predecir su etiqueta como:

$$y(x_*) = \sum_{n=1}^{N} a_n y_n x_n x_* + w_0$$

Se nota que cuando el valor $a_n = 0$, la muestra x_n no participa en el cálculo de la predicción. Por el contrario, cuando $a_n > 0$, la muestra x_n se considera como un vector de soporte y contribuye al cálculo de las predicciones. De esta forma, el vector a funciona como un filtro, el cual usa, para el cálculo de las predicciones, únicamente las muestras más cercanas al hiperplano de decisión.

Por último, el parámetro w_0 es necesario para el cálculo de las predicciones. Este valor se determina a partir de la ecuación:

$$w_0 = \frac{1}{|S|} \sum_{n \in S} (y_n - \sum_{m \in S} a_m y_m x_n^T x_m)$$

Donde S es el conjunto de índices correspondientes a los vectores de soporte y $|S|$ representa el número de elementos en dicho conjunto.

- SVMs DE MÁRGENES SUAVES

La optimización de las SVMs con márgenes duras no tiene solución si los datos no son linealmente separables. En general, es difícil encontrar datos que cumplan esta característica y es más común encontrar datos, como los de la figura, donde la estructura es principalmente lineal, pero existen algunos datos que se traslapan.

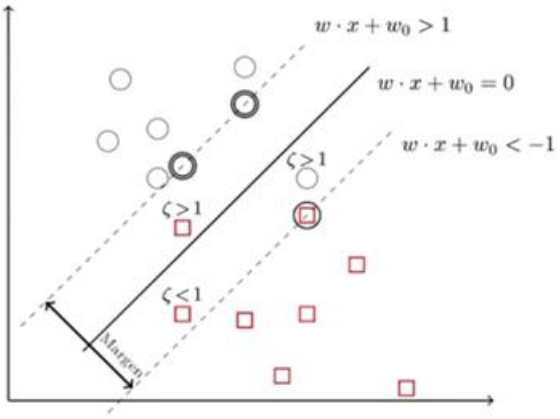


Figura 4. Ejemplo SVMs márgenes suaves
Fuente: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Una solución para esta condición es la inclusión de un conjunto de variables de holgura $\varsigma_n \geq 0$. De esta forma, las muestras con $\varsigma_n = 0$ están correctamente clasificadas y se encuentran fuera del margen. Para $0 < \varsigma_n \leq 1$, la muestra es clasificada correctamente, pero se encuentra dentro del margen. Finalmente $\varsigma_n > 1$ indica que la muestra está en el lado incorrecto del hiperplano y por lo tanto se clasificará de forma errónea.

Al incluir variables de holgura, el problema de optimización se puede escribir como:

$$mín_{W,W_0,\varsigma} \frac{1}{2}\|W\|^2 + C\sum_{n=1}^{N} \varsigma_n \quad sujeto\ a \quad Y_n(W^T x_n + w_0) \geq 1 - \varsigma_n, n = 1, \dots, N$$

Donde C es un hiperparámetro que controla cuántas muestras pueden violar el margen. Al igual que con el planteamiento original de las SVMs, se usan los multiplicadores de Lagrange para formular el siguiente problema dual.

$$máx_a \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_n a_m y_n y_m x_n x_m\,, \quad sujeto\ a \quad 0 \leq a_n \leq C, \quad \sum_{n=1}^{N} y_n a_n = 0$$

$$donde\ a = [a_1, \dots, a_N]$$

- SVMs PARA DATOS QUE NO SON LINEALMENTE SEPARABLES

Las formulaciones vistas hasta el momento se aplican para datos con estructuras lineales. sin embargo, es común encontrar datos con estructuras no lineales. Una alternativa es realizar una transformación no lineal φ de los datos a un espacio de mayor dimensión con el fin de mejorar su presentación.

El teorema de Cover establece que, si a un conjunto de datos se le aplica una transformación no lineal a un espacio lo suficientemente grande, los datos transformados tienen alta probabilidad de ser linealmente separable. En este sentido, los datos transformados pueden ser clasificados a partir de rectas.

Las transformaciones no lineales se suelen realizar, de manera implícita, a partir de las funciones kernel k. De esta forma, las funciones Kernel mapean los datos originales a un espacio euclidiano, que en teoría, es de dimensión infinita. Una función Kernel es una representación del producto punto de los datos transformados. De esta forma,

$$\varphi(x_n)^T \varphi(x_m) = k(x_n, x_m)$$

Así, el problema dual para las SVMs con margen suave y con función kernel se escribe como:

$$máx_a \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_n a_m y_n y_m k(x_n, x_m)$$

De igual forma, para la ecuación predictiva tenemos:

$$y(x_*) = \sum_{n=1}^{N} a_n y_n k(x_n, x_*) + w_0$$

3.1.3.3. REDES NEURONALES ARTIFICIALES

Son un modelo computacional inspirado en el funcionamiento del cerebro humano, donde se busca proveer a las maquinas la capacidad de aprender de una forma similar a como lo hace un cerebro. Una red neuronal artificial está formada por neuronas artificiales, que son nodos que

reciben información externa o de otras neuronas, en general, cada neurona artificial está formada por: 1) un conjunto de entradas que son los enlaces por donde reciben información, 2) un conjunto de funciones de propagación, activación y transferencia y 3) la salida de la neurona que es el enlace por donde entrega el resultado al exterior de la red neuronal u otras neuronas [14].

3.1.3.3.1. REDES NEURONALES FULLY-CONNECTED [16]

La arquitectura de una red neuronal Fully-Connected o perceptrón multicapa se caracteriza porque tiene sus neuronas agrupadas en capas de diferentes niveles. Cada una de las capas está formada por un conjunto de neuronas y se distinguen tres tipos de capas diferentes: la capa de entrada, las capas ocultas y la capa de salida.

Las neuronas de la capa de entrada no actúan como neuronas propiamente dichas, sino que se encargan únicamente de recibir las señales o patrones del exterior y propagar dichas señales a todas las neuronas de la siguiente capa. La última capa actúa como salida de la red, proporcionando al exterior la respuesta de la red. Las neuronas de las capas ocultas realizan un procesamiento no lineal de los patrones recibidos.

Como se observa en la figura, las conexiones del perceptrón multicapa siempre están dirigidas hacia adelante, es decir, las neuronas de una capa se conectan con las neuronas de la siguiente capa, de ahí que reciban también el nombre de redes alimentadas hacia adelante o redes "feedforward".

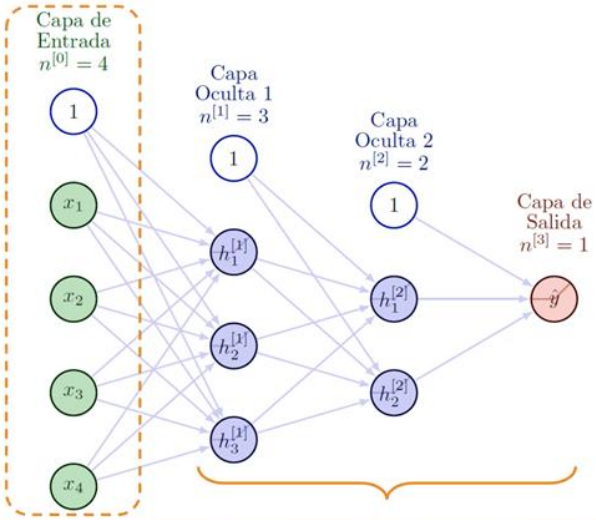


Figura 5. Ejemplo Redes neuronales Fully-Connected
Fuente: Elaboración propia

El Perceptrón multicapa define una relación entre las variables de entrada y las variables de salida de la red. Esta relación se obtiene propagando hacia adelante los valores de las variables de entrada. Para ello, cada neurona de la red procesa la información recibida por sus entradas y produce una respuesta o activación que se propaga, a través de las conexiones correspondientes, hacia las neuronas de la siguiente capa. A continuación, se muestran las expresiones para calcular las activaciones de las neuronas de la red. Sea un perceptrón multicapa con C capas, $C - 2$ capas

ocultas y n_c neuronas en la capa c, para $C = 1,2,..,C$. Sea $W^c = (w_{ij}^c)$ la matriz de pesos donde w_{ij}^c representa el peso de la conexión de la neurona i de la capa c para $c = 2, \dots, C$. Denotaremos a_i^c a la activación de la neurona i de la capa c. Estas activaciones se calculan del siguiente modo:

- Activación de las neuronas de la capa de entrada (a_i^1). Las neuronas de la capa de entrada se encargan de transmitir hacia la red las señales recibidas desde el exterior. Por tanto:

  $$a_i^1 = x_i \; para \; i = 1,2, \dots, n_1$$

  Donde $X = (x_1, x_2, \dots, x_{n1})$ representa el vector o patrón de entrada a la red.

- Activación de las neuronas de la capa oculta c (a_i^c). Las neuronas ocultas de la red procesan la información recibida aplicando la función de activación f a la suma de los productos de las activaciones que recibe por sus correspondientes pesos, es decir:

  $$a_i^c = f\left(\sum_{j=1}^{n_{c=1}} w_{ji}^{c-1} a_j^{c-1} + u_i^c\right) para \; i = 1,2, \dots, n_c \text{ y } c = 2,3, \dots, C-1$$

  Donde a_j^{c-1} son las activaciones de las neuronas de la capa $c-1$

- Activación de las neuronas de la capa de salida (a_i^C).. Al igual que en el caso anterior, la activación de estas neuronas viene dada por la función de activación f aplicada a la suma de los productos de las entradas que recibe por sus correspondientes pesos:

  $$y_i = a_i^c = f\left(\sum_{j=1}^{n_{C=1}} w_{ji}^{C-1} a_j^{C-1} + u_i^C\right) para \; i = 1,2, \dots, n_c$$

  Donde $Y = (y_1, y_2, \dots, y_{n_C})$ es el vector de salida de la red.

3.1.3.4. MÉTRICAS DE EVALUACIÓN

A continuación, se describen algunas de las métricas de evaluación más utilizados para los modelos de aprendizaje automático de clasificación [17].

- Matriz de confusión

Es una medida de rendimiento que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado, cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa las clases reales, de esta forma, permite ver qué tipos de aciertos y errores está teniendo el modelo.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos positivos (TP)	Falsos positivos (FP)
	Negativos	Falsos negativos (FN)	Verdaderos negativos (TN)

- Accuracy

La métrica accuracy representa el porcentaje total de valores correctamente clasificados, tanto positivos como negativos. Es recomendable utilizar esta métrica en problemas en los que los datos están balanceados.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Puntaje F1

El valor F1 se utiliza para combinar las medidas de precisión y recuperación (recall) en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la recuperación entre varios modelos, donde su puntaje F1 alcanza su valor en uno y su peor valor en cero.

$$Puntaje\ F1 = 2 \times \frac{Recall \times Precisión}{Recall + Precisión}$$

Recuperación: La métrica de recall, también conocida como el ratio de verdaderos positivos, es utilizada para saber cuántos valores positivos son correctamente clasificados.

$$Recall = \frac{TP}{TP + FN}$$

Precisión: La métrica de precisión es utilizada para poder saber qué porcentaje de valores que se han clasificado como positivos son realmente positivos.

$$Precisión = \frac{TP}{TP + FP}$$

3.2. ANTECEDENTES

A continuación, se describirán trabajos previos donde han abordado problemas similares al descrito en el presente trabajo, y donde se evidencia la versatilidad y rendimiento de los algoritmos de aprendizaje automático para analizar diferentes enfoques de este fenómeno.

- Artículo científico: Estimating deforestation using machine learning algorithms [18].

  En este artículo describen la deforestación como una de las principales causas del calentamiento global, y de allí expresan la necesidad de estimar las tasas de deforestación para pequeños estados insulares en desarrollo. Inicialmente realizan la segmentación sobre imágenes satelitales para determinar el porcentaje de áreas boscosas mediante algoritmos

de aprendizaje automático tradicionales como el clasificador Bayesiano ingenuo, máquinas de soporte vectorial, regresión logística, y algoritmos de aprendizaje profundo como redes neuronales convolucionales, con estos resultados calculan la tasa de cambio de la deforestación en un periodo de tiempo.
Los resultados demuestran que los algoritmos tienen un buen rendimiento en la clasificación y con resultados similares, pero el tiempo para entrenar la red neuronal fue significativamente mayor y también requirió mayor almacenamiento. Posteriormente, estos modelos se utilizaron para etiquetar imágenes de la isla de Trinidad, y se demostró que los modelos tenían clasificaciones similares, siendo el modelo del clasificador Bayesiano ingenuo el único atípico. De esta forma, se evidencia el buen desempeño de las diferentes técnicas de aprendizaje automático para analizar el comportamiento de la deforestación con imágenes satelitales, con lo cual, se puede justificar el uso de estas técnicas para el desarrollo de este proyecto.

- Artículo científico: Deforestation probability assessment using integrated machine learning algorithms of Eastern Himalayan foothills [19].

En este artículo exponen al parque nacional Jaldapara situado en las estribaciones del Himalaya oriental como una zona rica en biodiversidad, que se ha visto afectada por el aumento de la deforestación y recalcan la importancia de la protección de estas regiones para mitigar los efectos sobre el calentamiento global y la biodiversidad, es por ello que, el objetivo es identificar las zonas probables de deforestación en el parque y sus alrededores usando cinco algoritmos de aprendizaje automático (máquinas de soporte vectorial, clasificador Bayesiano ingenuo, bosques aleatorios, arboles de decisiones y redes neuronales artificiales), para realizar el proyecto se tuvieron en cuenta 11 variables que están relacionadas con la deforestación con un histórico de 30 años, algunas de estas variables son: altitud, densidad agrícola, distancia al rio, entre otros. Los resultados demuestran que el algoritmo que brinda mayor precisión es el modelo realizado con el algoritmo de máquinas de soporte vectorial, de acuerdo con este modelo se identificó que las secciones norte y media del parque se enfrentan a una alta tasa de deforestación debido a la invasión humana a gran escala, la caza furtiva y el tráfico de madera.
La principal diferencia que se puede identificar con el proyecto que se planea realizar, es que en este disponían de información histórica de variables adicionales que pueden explicar el comportamiento de la deforestación, y a partir de esto calcularon las zonas con mayor probabilidad de que suceda este evento. Al revisar este proyecto se puede destacar el valor agregado que brinda los modelos de aprendizaje automático para explicar este fenómeno, y como a partir de estos se busca generar herramientas para prevenir la deforestación debido a la criticidad de sus impactos.

- Artículo científico: Monitoring forest change in the amazon using multi-temporal remote sensing data and machine learning classification on google earth engine [20].

Inicialmente describen que la deforestación provoca diversas y profundas consecuencias para el medio ambiente y las especies, algunos de los efectos pueden estar relacionados con el cambio climático, la erosión del suelo, deslizamiento de tierra, etc. Considerando esto, mencionan que es importante el monitoreo oportuno y continuo de la dinámica forestal para seguir constantemente las políticas existentes y desarrollar nuevas medidas de mitigación. El objetivo de este proyecto es mapear y monitorear el cambio forestal de 2000 a 2019 y simular el futuro desarrollo forestal de una región de selva tropical ubicado en el estado de Pará, Brasil.

El desarrollo de este proyecto lo realizaron con el algoritmo de bosques aleatorios para clasificar la cobertura de la tierra para los periodos evaluados, y una vez obtenido esta clasificación utilizaron el módulo de QGIS para la evaluación del cambio de uso del suelo (MOLUSCE), el cual está diseñado para incorporar modelos ya probados y para modelar el uso potencial de la tierra y las transiciones forestales. Cabe destacar, que en este proyecto se propone un enfoque adicional donde utilizan dos herramientas para complementar el análisis propuesto, lo que hace que sea una proyección robusta ya que QGIS es un sistema especializado de información geográfica.

- Artículo científico: ForestNet: Classifying drivers of deforestation in Indonesia using deep learning on satellite imagery [21].

Los autores explican que identificar los procesos que conducen a la deforestación es fundamental para el desarrollo y la implementación de políticas específicas de conservación de bosques. Por esta razón, desarrollaron un modelo de aprendizaje profundo llamado ForestNet para clasificar los impulsores de la pérdida de bosques en Indonesia, un país con una de las tasas de deforestación más altas del mundo. Usando imágenes satelitales, ForestNet identifica los impulsores directos de la deforestación en imágenes de pérdida de bosques de cualquier tamaño.

Para realizar este proyecto seleccionaron un conjunto de datos de imágenes satelitales Landsat 8 de eventos conocidos de pérdida de bosques combinados con anotaciones de controladores expertos, usaron el conjunto de datos para entrenar y validar los modelos, a partir de los resultados, demostraron que ForestNet supera sustancialmente a otros enfoques de clasificación de controladores estándar. La principal diferencia con el proyecto que se planea realizar es que el objetivo de este era identificar los factores o causas que conllevan a la deforestación con imágenes satelitales etiquetadas, en cambio el trabajo

propuesto busca analizar el comportamiento de la deforestación en sí, para generar herramientas de monitoreo que aporten valor en la formulación de políticas ambientales; sin embargo, como se mencionó en los anteriores artículos se evidencia el uso de las técnicas de aprendizaje automático como un nuevo enfoque de análisis para este fenómeno.

4. MINERÍA DE DATOS

4.1. COMPRENSIÓN DE LAS IMÁGENES SATELITALES

Se revisó la documentación en la página oficial de Google Earth Engine y los datasets disponibles de los satélites SENTINEL y LANDSAT. De estas colecciones, se utilizaron las imágenes satelitales SURFACE REFLECTANCE de LANDSAT 8 OLI/TRS que tiene información disponible desde abril del 2013, y también se utilizaron las imágenes satelitales SURFACE REFLECTANCE de SENTINEL-2 MSI que tiene información disponible desde marzo del 2018.

4.2. DEFINICIÓN DE LA ZONA DE INTERÉS

En agosto de 2021 varios medios de comunicación del país como Semana y Caracol emitieron diversas noticias, artículos e inclusive un corto documental sobre la tragedia ambiental que devora velozmente los bosques del Guaviare [22] [9]. En este corto documental y noticias mostraron diversas imágenes, testimonios, causas y estadísticas de la alarmante perdida de bosques en la zona del Guaviare desde el 2017 hasta el 2020, producto de la apropiación de tierras, extensión de la ganadería y cultivos de uso ilícito. Por lo que, en este proyecto centramos nuestro interés en evidenciar el aumento de la deforestación mediante algoritmos de aprendizaje automático e imágenes satelitales de Google Earth Engine en el departamento del Guaviare durante este periodo de tiempo. Por lo anterior, la zona de interés que elegimos para el desarrollo de este trabajo tiene una cobertura total de $55{,}840 km^2$ perteneciente a la Amazonía colombiana.

Se sustentan las estadísticas y el comportamiento de la deforestación presentados por los medios de comunicación con el informe del monitoreo de deforestación entregado por el IDEAM para el año 2020, donde describen que la Amazonía es la región donde se presentó el mayor aumento de la superficie deforestada, un aumento de aproximadamente 11,000 hectáreas. En este mismo informe explican que el Guaviare es el tercer departamento después del Meta y Caquetá con mayor superficie deforestada en Colombia, con un área deforestada de 25,553 hectáreas en el 2020 [23]. De acuerdo con lo expuesto por los diferentes promotores ambientales, una de las principales causas del aumento acelerado de la deforestación en el departamento del Guaviare es el proceso de ganaderización, que a pesar de las diferentes iniciativas para buscar alternativas de manejo del bosque. La rentabilidad inmediata que permite el desarrollo ganadero, además de la consolidación de los procesos de apropiación indebida de baldíos y la facilidad de lavado de activos del narcotráfico, son un escenario propicio para esta actividad en la región.

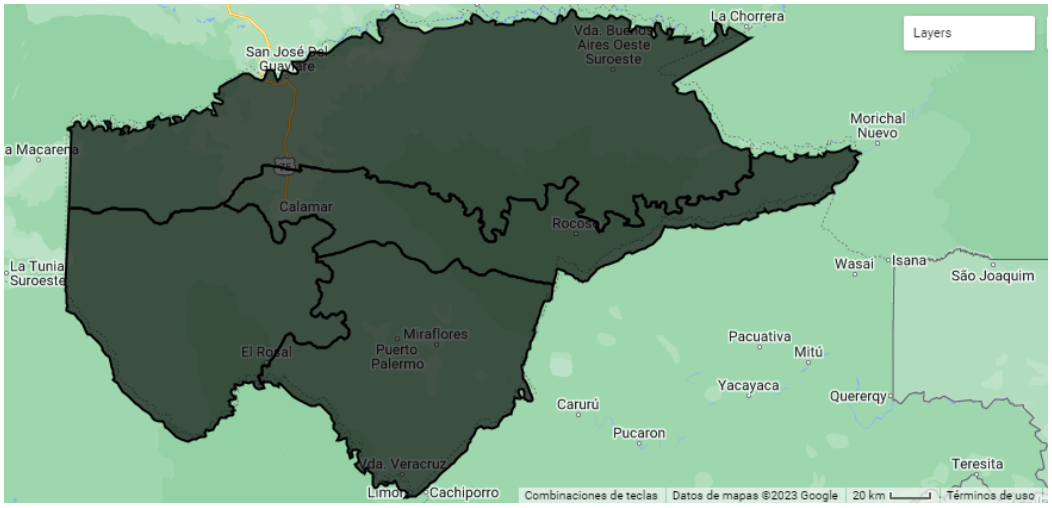


Figura 6. Zona de interés
Fuente: Google Earth Engine

4.3. CREACIÓN DE LA MÁSCARA DE NUBE

Uno de los principales retos al trabajar con imágenes satelitales es la presencia de nubes o cirros, lo que dificulta contar con información completa y exacta para realizar análisis y ciencia de datos sobre estas. Como parte fundamental e inicial de este análisis se realizó la limpieza y filtros de las imágenes, por lo que se crearon funciones para enmascarar las nubes en las imágenes satelitales de SENTINEL, y además se incluyeron filtros al momento de recolectar las imágenes de ambos satélites que cumplieran la condición que la proporción de nubes fuera menor del 10% para todos los periodos de análisis.

4.4. SELECCIÓN DE LOS PERIODOS DE ANÁLISIS

El objetivo del proyecto es evaluar el aumento de la deforestación en el periodo del 2017-2020 en el departamento del Guaviare. Por lo anterior, se evaluaron diferentes cortes de tiempo con el objetivo de tener imágenes satelitales de toda la zona de interés para fechas previas y posteriores a este periodo. Para contar con información completa, consistente y limpia de la zona de interés, se calculó la mediana de los valores espectrales de la colección de imágenes del satélite LANDSAT desde ene-14 hasta dic-15 considerado como el periodo previo de análisis, es importante recalcar que al calcular el promedio o mediana se reduce el efecto de la variación de los valores espectrales de las imágenes producto de las condiciones atmosféricas. También se realizó este mismo cálculo para la colección de las imágenes del satélite SENTINEL desde ene-21 hasta dic-21 consideradas dentro del periodo posterior de análisis. La razón principal de utilizar las imágenes de ambos satélites se debe a que, al momento de hacer las evaluaciones en diferentes periodos de tiempo, las imágenes de SENTINEL presentaban una mejor resolución, lo que es una característica primordial para realizar análisis de imágenes. Sin embargo, la colección de este satélite solamente está disponible en Google Earth Engine a partir de marzo del 2018.

4.5. DEFINICIÓN DE LOS DATOS DE ANÁLISIS

En base a lo expuesto anteriormente, se definen las 2 imágenes satelitales que serán el insumo principal para el desarrollo de este proyecto.

- **Zona de interés:** Departamento del Guaviare/Colombia
- **Primer periodo de análisis:** Imagen de la zona de interés calculada como la mediana de los valores espectrales de la colección de imágenes satelitales Landsat tomadas desde enero-2014 hasta diciembre-2015.

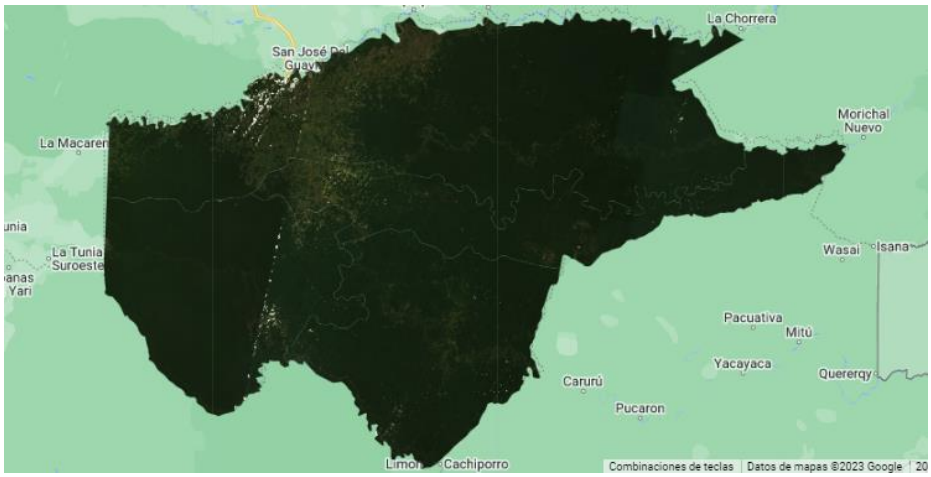


Figura 7. Imagen satelital Landsat ene-14 a dic-15
Fuente: Google Earth Engine

- **Segundo periodo de análisis:** Imagen de la zona de interés calculada como la mediana de los valores espectrales de la colección de imágenes satelitales Sentinel tomadas desde enero-2021 hasta diciembre-2021.

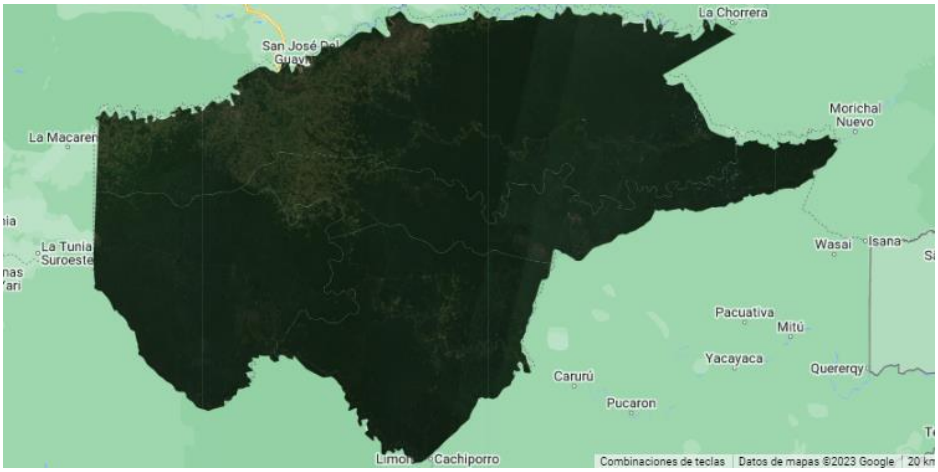


Figura 8. Imagen satelital Sentinel ene-21 a dic-21
Fuente: Google Earth Engine

4.6. CONSTRUCCIÓN DE INDICADORES Y FILTROS ADICIONALES

Sobre las imágenes satelitales ya seleccionadas se calculó el índice de vegetación de diferencia normalizada (NDVI). Este ha sido uno de los indicadores más utilizados en la observación remota desde su aparición en la década de los 70, y ayuda a diferenciar la vegetación de otros tipos de cobertura terrestre (artificial), también permite definir y visualizar las áreas con vegetación en el mapa.

A continuación, se presentan la presentación gráfica de los indicadores NDVI calculados sobre ambas imágenes satelitales. Este índice es adecuado para detectar y cuantificar la presencia de vegetación basado en cómo las plantas reflejan ciertos rangos del espectro electromagnético. De esta forma, permite conocer su estado actual, que luego podrá compararse con otra imagen temporal para observar su evolución en el tiempo.

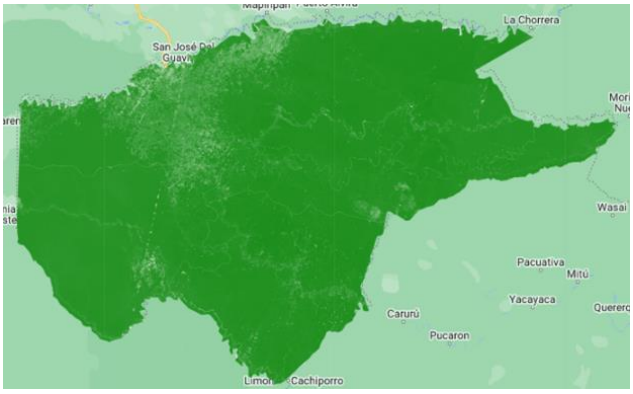


Figura 9. NDVI Departamento del Guaviare 2015
Fuente: Google Earth Engine

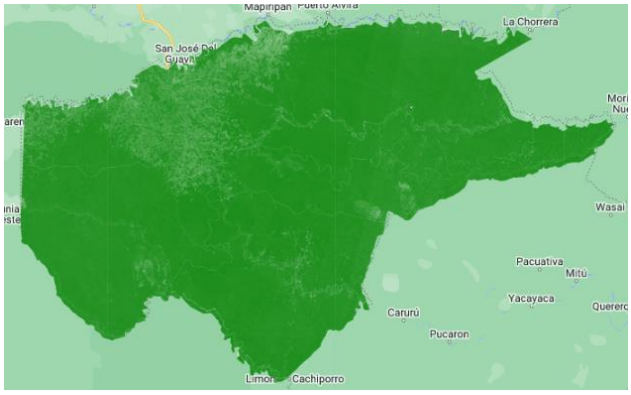


Figura 10. NDVI Departamento del Guaviare 2021
Fuente: Google Earth Engine

4.7. EXTRACCIÓN Y SELECCIÓN DE CARACTERÍSTICAS

- VARIABLES EXPLICATIVAS

De ambas imágenes satelitales se realizó la extracción de características de las bandas referentes a rojo, verde, azul, infrarrojo cercano y el indicador NDVI, y se hizo la estandarización de los valores de cada banda (a una resolución espacial de 30 metros) para la imagen completa del departamento del Guaviare de ambos periodos de análisis. Estas 5 características serán incluidas como las variables explicativas en el desarrollo del modelo.

- ETIQUETA

Con base a la revisión de los antecedentes, se decidió que el problema de este proyecto se abordaría desde el enfoque de aprendizaje supervisado, requiriendo contar con la etiqueta de las observaciones para el entrenamiento de los modelos. De esta forma, inicialmente se revisaron los catálogos de datos Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 [24] y ESA WorldCover 10m v200 [25] para la etiqueta de los datos. Sin embargo, se identificaron algunas

inconsistencias en las etiquetas de cobertura de tierra de estas colecciones, y que podrían generar algunos errores en la construcción de las bases de datos al tomar muestras aleatorias (pixeles) del mapa y de las etiquetas de estas colecciones. A continuación, se muestran algunos ejemplos de esta evaluación, donde se identifica que el principal riesgo es que se etiquete un pixel como forestación (etiqueta 1) cuando en la imagen satelital y en el NDVI se evidencia deforestación (etiqueta 0).

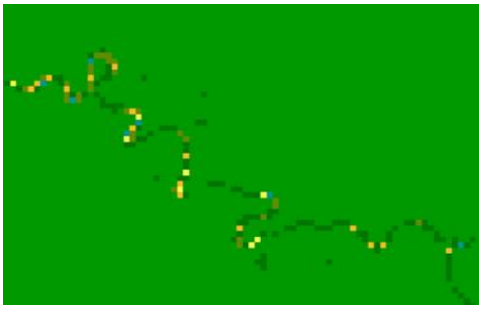
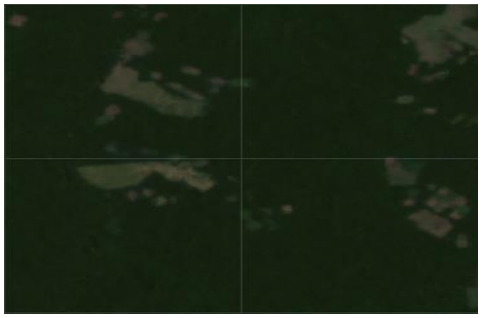
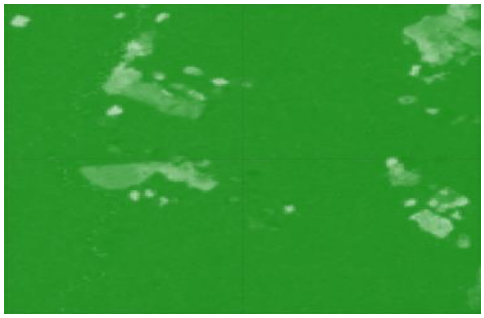
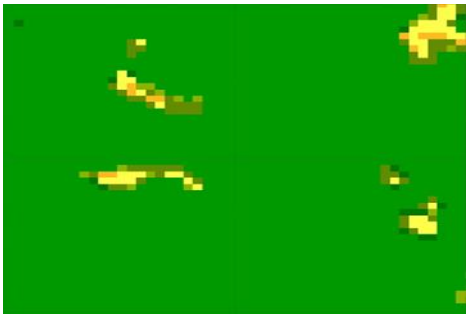
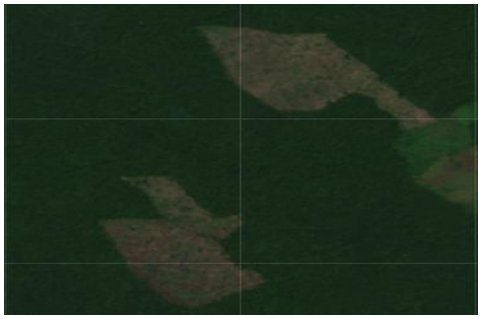
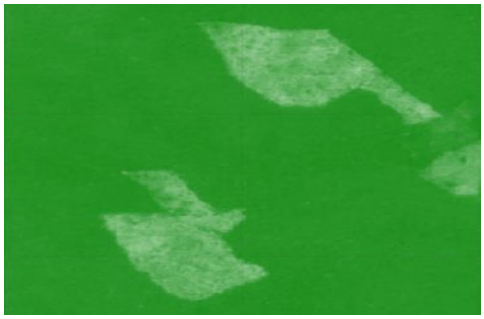
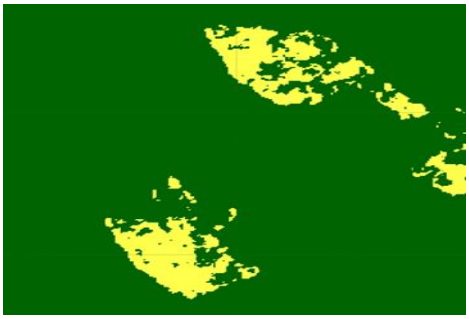
IMAGEN SATELITAL	NDVI	SUPERFICIE DE COBERTURA
		
		
		

Figura 11. Evaluación colecciones Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 y ESA WorldCover 10m v200.
Fuente: Google Earth Engine

Dado que el enfoque supervisado para el etiquetado de pixeles requiere que el usuario seleccione datos de entrenamiento representativos para cada una de las clases predefinidas [26], con el apoyo de un experto técnico se abordó un plan de acción para mitigar estos errores en la construcción de la base de datos y se decidió crear dos nuevas colecciones de polígonos de características usando como referencia las zonas en donde se identificará claramente zonas de forestación y deforestación (coberturas diferente a bosque) en el NDVI y en las colecciones disponibles en Google Earth Engine para cada periodo de análisis, considerando que en estos polígonos que se construyeran no se presentará esta disyuntiva en la etiqueta. Posteriormente, se validó y comparó la etiqueta de cada pixel de las colecciones creadas versus las etiquetas de las colecciones de Google Earth Engine, de esta validación no se presentó ninguna diferencia en las etiquetas de los pixeles y hubo una segunda revisión por parte del experto técnico.

A continuación, se detalla la homologación de las etiquetas y las superficies para cada imagen satelital que se tuvieron en cuenta para la creación de las colecciones de características:

TABLA 4.
HOMOLOGACIÓN DE ETIQUETAS

Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 (2015)		
Valor	Descripción	Homologación
0	Desconocido	Deforestación
20	Arbustos. Plantas leñosas perennes de tallos persistentes y leñosos y sin tallo principal definido, de altura inferior a 5 m.	Deforestación
30	Vegetación herbácea. Plantas sin tallo ni brotes persistentes sobre el suelo. La cobertura de árboles y arbustos es inferior al 10%.	Deforestación
40	Vegetación/agricultura cultivada y gestionada.	Deforestación
50	Urbano/edificado.	Deforestación
60	Vegetación desnuda/escasa.	Deforestación
70	Nieve y hielo	Deforestación
80	Cuerpos de agua permanentes	Deforestación
90	Humedal herbáceo	Deforestación
100	Musgo y liquen	Deforestación
111	Bosque cerrado, hoja perenne de aguja. Copa de los árboles >70%	Forestación
112	Bosque cerrado, siempre verde de hoja ancha. Copa de los árboles >70 %	Forestación
113	Bosque cerrado, hoja acicular caducifolia. Copa de los árboles >70%	Forestación
114	Bosque cerrado, caducifolio de hoja ancha. Copa de los árboles	Forestación
115	Bosque cerrado, mixto.	Forestación
116	Bosque cerrado, que no coincide con ninguna de las otras definiciones.	Forestación
121	Bosque abierto, hoja perenne de aguja. Capa superior: árboles entre 15 y 70% y segunda capa: mezcla de arbustos y pastizales	Deforestación
122	Bosque abierto, siempre verde de hoja ancha. Capa superior: árboles 15-70% y segunda capa: mezcla de arbustos y pastizales	Deforestación
123	Bosque abierto, hoja acicular caducifolia. Capa superior: árboles 15-70% y segunda capa: mezcla de arbustos y pastizales	Deforestación
124	Bosque abierto, caducifolio de hoja ancha. Capa superior: árboles 15-70% y segunda capa: mezcla de arbustos y pastizales	Deforestación
125	Bosque abierto, mixto.	Deforestación
126	Bosque abierto, que no coincide con ninguna de las otras	Deforestación
200	Océanos, mares	Deforestación

ESA WorldCover 10m v200 (2021)		
Valor	Descripción	Homologación
10	Cobertura de árboles	Forestación
20	Matorral	Deforestación
30	Pradera	Deforestación
40	Tierras de cultivo	Deforestación
50	Construido	Deforestación
60	Vegetación desnuda/escasa	Deforestación
70	Nieve y hielo	Deforestación
80	Cuerpos de agua permanentes	Deforestación
90	Humedal herbáceo	Deforestación
95	Manglares	Deforestación
100	Musgo y liquen	Deforestación

- Polígonos de deforestación: En Google Earth Engine se creó una colección de características conformada por 220 polígonos de las zonas con deforestación de las imágenes satelitales seleccionadas previamente, estos polígonos equivalen a 76,989 pixeles. Al hacer esta selección de polígonos se verificó que se presentará deforestación en esa zona para ambas fechas de evaluación comparando con las clasificaciones de los catálogos de datos Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 y ESA WorldCover 10m v200, ya que esta colección se utilizó para el entrenamiento de los modelos para ambas fechas.

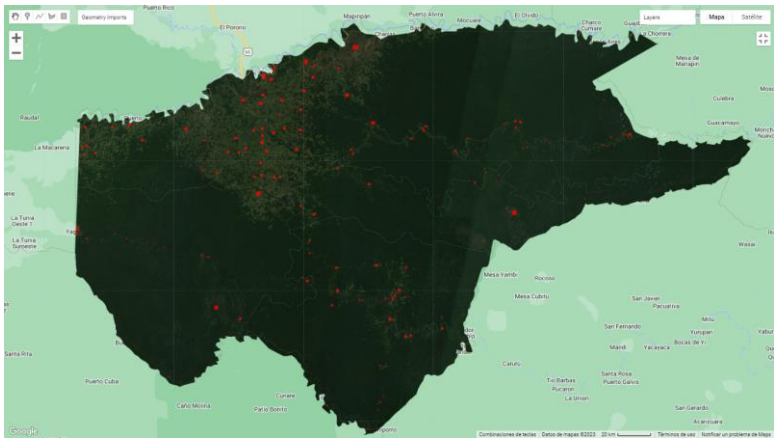


Figura 12. Polígonos de deforestación del Guaviare
Fuente: Elaboración propia en Google Earth Engine

- Polígonos de forestación: Al igual que en el anterior procedimiento, se creó una colección de características conformada por 220 polígonos de las zonas con forestación de las imágenes satelitales, que equivalen a 76,727 pixeles. Al hacer esta selección de polígonos se evaluó que se presentará forestación en esa zona para ambas fechas de evaluación comparando con las clasificaciones de los catálogos de datos Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 y ESA WorldCover 10m v200, ya que esta colección se utilizó para el entrenamiento de los modelos para ambas fechas.

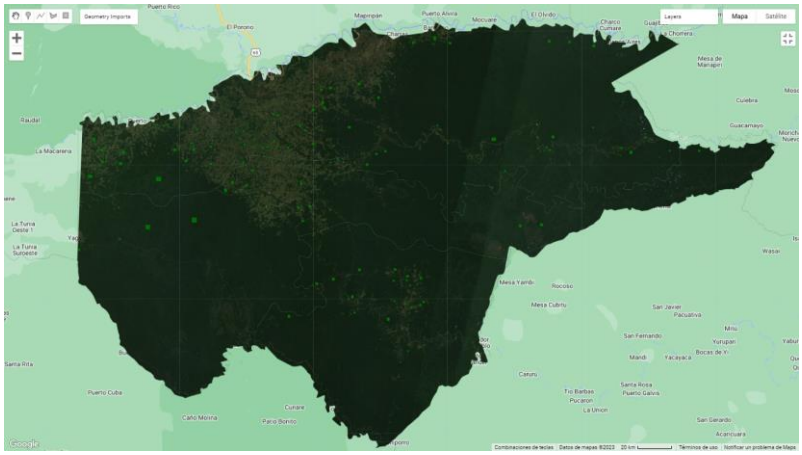


Figura 13. Polígonos de forestación del Guaviare
Fuente: Elaboración propia en Google Earth Engine

Una vez creadas las dos nuevas colecciones, se procedió a realizar un análisis descriptivo de las caracteristicas para cada periodo de evaluación, esto con el objetivo de asegurar que los datos están distribuidos adecuadamente y representan claramente cada clase, ya que los resultados de los modelos dependen directamente de la calidad de las muestras de entrenamiento [26]. En primer lugar, se realizó un histograma para cada clase y bandas seleccionadas para las muestras de cada periodo de análisis, de esto pudimos identificar 1) no hay valores extremos o sesgos en las distribuciones de los valores de reflectancia para cada clase, y 2) las distribuciones no se superponen sustancialmente entre clases.

En segundo lugar, con el objetivo de evaluar la separabilidad de las clases, se graficaron las medias de las signaturas espectrales y se calcularon las distancias Euclidiana, Mahalanobis, Bhattacharya y Jeffries–Matusita (JM) [27]. De forma general, de este análisis se identifica que las medias para las signaturas espectrales son diferentes entre clases para ambas imagenes, esta diferencia se amplifica principalmente en el indicador NDVI, también se ratifica la separabilidad de las clases para ambos periodos de analisis con distancias de Mhalanobis mayores a 3.1 y Bhattacharya a 2.1. Una vez verificados los supuestos de los datos, se procedió a continuar con el proyecto garantizando la calidad de las muestras.

5. DESARROLLO DEL MODELO

5.1. DEFINICIÓN DE LAS BASES DATOS PARA EL ENTRENAMIENTO Y PRUEBA DE LOS MODELOS

Luego de la creación de las colecciones de características para zonas con deforestación y forestación, se tuvieron un total de 440 polígonos para el desarrollo de este proyecto, que a una resolución espacial de 30 metros equivalen a 153,716 pixeles (observaciones). De cada colección, de forma aleatoria el 80% de los datos se utilizó para el entrenamiento y el 20% para el testing de los modelos. Es importante señalar que en ambas colecciones se hicieron polígonos de aproximadamente el mismo tamaño para contar con una base de datos de observaciones con deforestación y forestación balanceada.

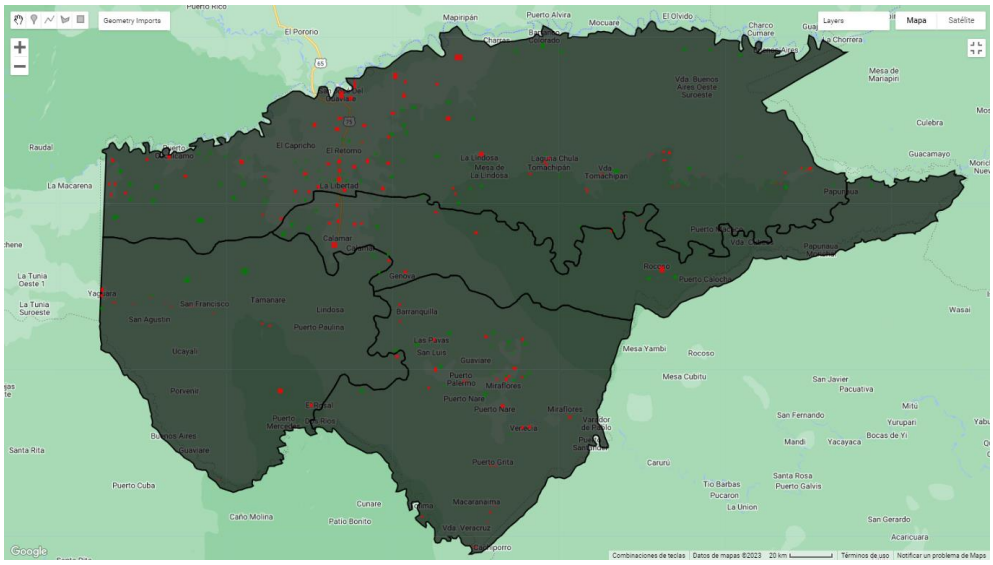


Figura 14. Base de datos de polígonos
Fuente: Elaboración propia en Google Earth Engine

5.2. SELECCIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Con base a la investigación de los antecedentes que se realizó, las técnicas de aprendizaje automático que se seleccionaron para identificar las zonas con deforestación y forestación en las imágenes satelitales de Landsat y Sentinel fueron 1) bosques aleatorios, 2) máquinas de soporte vectorial y 3) redes neuronales Fully connected explicados en el marco teórico, la justificación de la selección de estas técnicas radica en el buen desempeño de estos en los proyectos revisados en los antecedentes.

5.3. ENTRENAMIENTO Y DESARROLLO DE LOS MODELOS

Nuevamente se dividió la base de datos de entrenamiento en 80% - 20%, con el objetivo de realizar validación cruzada para la optimización de los hiperparámetros de los algoritmos de

bosques aleatorios, máquinas de soporte vectorial y redes neuronales fully-connected. Es decir, el 80% de la base de entrenamiento se utilizó para entrenar los modelos, y el 20% para evaluar cada modelo con sus respectivos hiperparámetros. De esta forma, se identificaron los hiperparámetros con los que se obtuvo un mejor accuracy en cada técnica de aprendizaje automático.

5.3.1. BOSQUES ALEATORIOS

- Se realizó la optimización de los hiperparámetros para el algoritmo de bosques aleatorios para las imágenes del satélite Landsat, para esto se realizó validación cruzada con las múltiples combinaciones de los hiperparámetros: número de árboles, fracción de submuestreo, número mínimo de observaciones en cada nodo y máximo de nodos. A continuación, se presenta la lista de los valores que se probaron para cada parámetro.

TABLA 5.
HIPERPARÁMETROS BOSQUES ALEATORIOS- LANDSAT

Number of trees	Bag fraction	Min leaf population	Max nodes
10	0.5	30	40
20	0.7	40	50
30	0.9	50	60
			70

- De acuerdo con la validación cruzada realizada en el punto anterior, la combinación de hiperparámetros con los que se obtuvo el mejor accuracy en la base de evaluación fue la siguiente:

TABLA 6.
RESULTADOS OPTIMIZACIÓN HIPERPARÁMETROS BOSQUES ALEATORIOS- LANDSAT

Hiperparametro	Valor
Number of trees	30
Bag fraction	0.9
Min leaf population	30
Max nodes	70
Accuracy	0.989205

- A continuación, se presenta la gráfica de los accuracies obtenidos en cada iteración, en total se realizaron 108 combinaciones de los hiperparámetros.

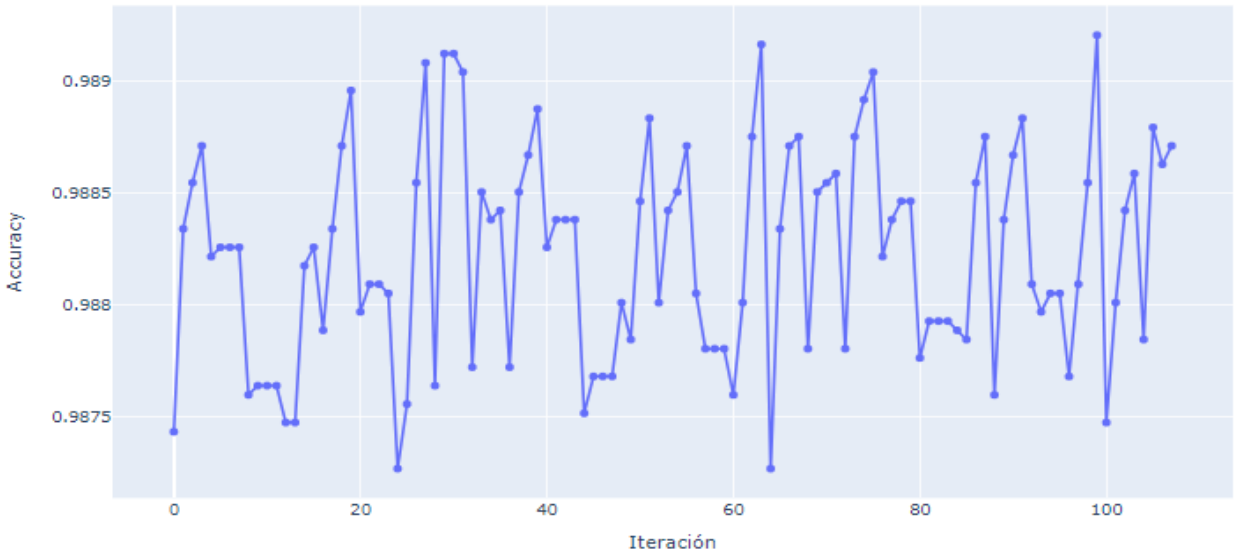


Figura 15. Resultados iteraciones Bosques aleatorios- Landsat
Fuente: Elaboración propia

- Se realizó la optimización de los hiperparámetros para el algoritmo de bosques aleatorios para las imágenes del satélite Sentinel, para esto se realizó validación cruzada con las múltiples combinaciones de los hiperparámetros: número de árboles, fracción de submuestreo, número mínimo de observaciones en cada nodo y máximo de nodos. A continuación, se presenta la lista de los valores que se probaron para cada parámetro.

TABLA 7.
HIPERPARÁMETROS BOSQUES ALEATORIOS- SENTINEL

Number of trees	Bag fraction	Min leaf population	Max nodes
10	0.5	30	40
20	0.7	40	50
30	0.9	50	60
			70

- De acuerdo con la validación cruzada realizada en el punto anterior, la combinación de hiperparámetros con los que se obtuvo el mejor accuracy en la base de evaluación fue la siguiente:

TABLA 8.
RESULTADOS OPTIMIZACIÓN HIPERPARÁMETROS BOSQUES ALEATORIOS- SENTINEL

Hiperparametro	Valor
Number of trees	10
Bag fraction	0.9
Min leaf population	30
Max nodes	40
Accuracy	0.997084

- A continuación, se presenta la gráfica de los accuracies obtenidos en cada iteración, en total se realizaron 108 combinaciones de los hiperparámetros.

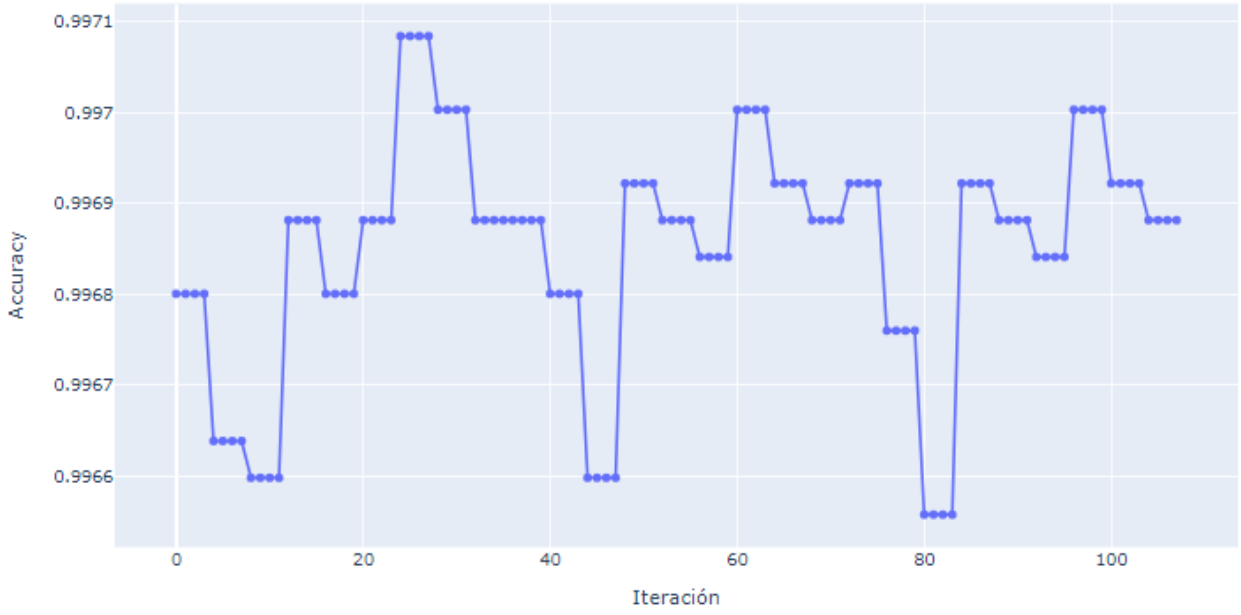


Figura 16. Resultados iteraciones Bosques aleatorios- Sentinel
Fuente: Elaboración propia

- Por último, se entrenaron los modelos finales del algoritmo de bosques aleatorios con los hiperparámetros seleccionados en la validación cruzada, y se evalúa el desempeño de estos modelos en las bases de datos de testeo, que se crearon en la actividad 5.1.

5.3.2. MAQUINAS DE SOPORTE VECTORIAL

- Se realizó la optimización de los hiperparámetros para el algoritmo de máquinas de soporte vectorial para las imágenes del satélite Landsat, para esto se realizó validación cruzada con las múltiples combinaciones de los hiperparámetros: tipo de Kernel, gamma y costo. A continuación, se presenta la lista de los valores que se probaron para cada parámetro.

TABLA 9.
HIPERPARÁMETROS MAQUINAS DE SOPORTE VECTORIAL- LANDSAT

Kernel type	Cost	Gamma
Linear	30	1
RBF	40	1.5
	50	2
	60	2.5
	70	3
	80	3.5
		4
		4.5
		5

- De acuerdo con la validación cruzada realizada en el punto anterior, la combinación de hiperparámetros con los que se obtuvo el mejor accuracy en la base de evaluación fue la siguiente:

TABLA 10.
RESULTADO OPTIMIZACIÓN HIPERPARÁMETROS MAQUINAS DE SOPORTE VECTORIAL- LANDSAT

Hiperparametro	Valor
Kernel type	RBF
Cost	30
Gamma	2.0
Accuracy	0.991595

- A continuación, se presenta la gráfica de los accuracies obtenidos en cada iteración, en total se realizaron 60 combinaciones de los hiperparámetros.

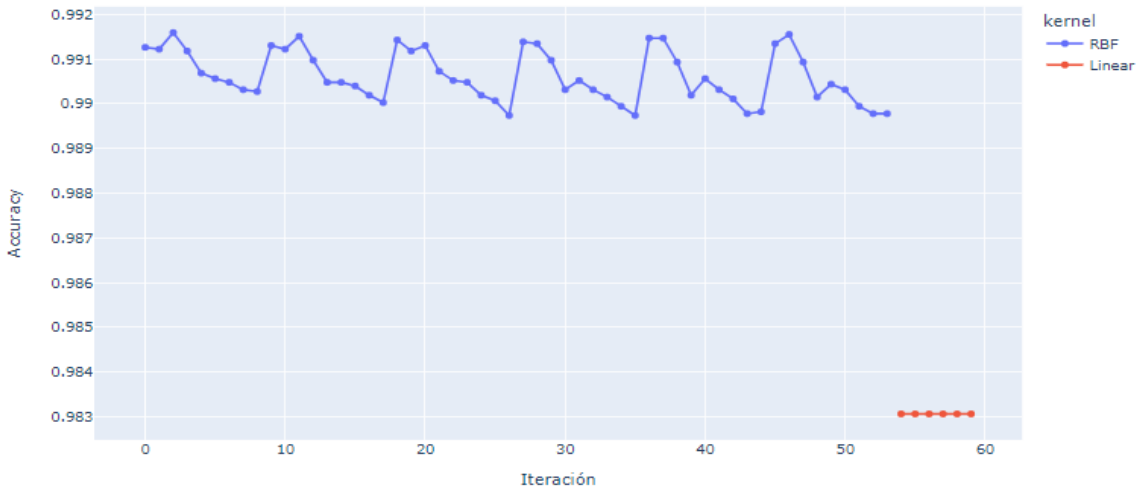


Figura 17. Resultados iteraciones Maquinas de soporte vectorial- Landsat
Fuente: Elaboración propia

- Se realizó la optimización de los hiperparámetros para el algoritmo de máquinas de soporte vectorial para las imágenes del satélite Sentinel, para esto se realizó validación cruzada con las múltiples combinaciones de los hiperparámetros: tipo de Kernel, gamma y costo. A continuación, se presenta la lista de los valores que se probaron para cada parámetro.

TABLA 11.
HIPERPARÁMETROS MAQUINAS DE SOPORTE VECTORIAL- SENTINEL

Kernel type	Cost	Gamma
Linear	30	1
RBF	40	1.5
	50	2
	60	2.5
	70	3
	80	3.5
		4
		4.5
		5

- De acuerdo con la validación cruzada realizada en el punto anterior, la combinación de hiperparámetros con los que se obtuvo el mejor accuracy en la base de evaluación fue la siguiente:

TABLA 12.
RESULTADO OPTIMIZACIÓN HIPERPARÁMETROS MAQUINAS DE SOPORTE VECTORIAL- SENTINEL

Hiperparametro	Valor
Kernel type	RBF
Cost	30
Gamma	1.0
Accuracy	0.997387

- A continuación, se presenta la gráfica de los accuracies obtenidos en cada iteración, en total se realizaron 60 combinaciones de los hiperparámetros.

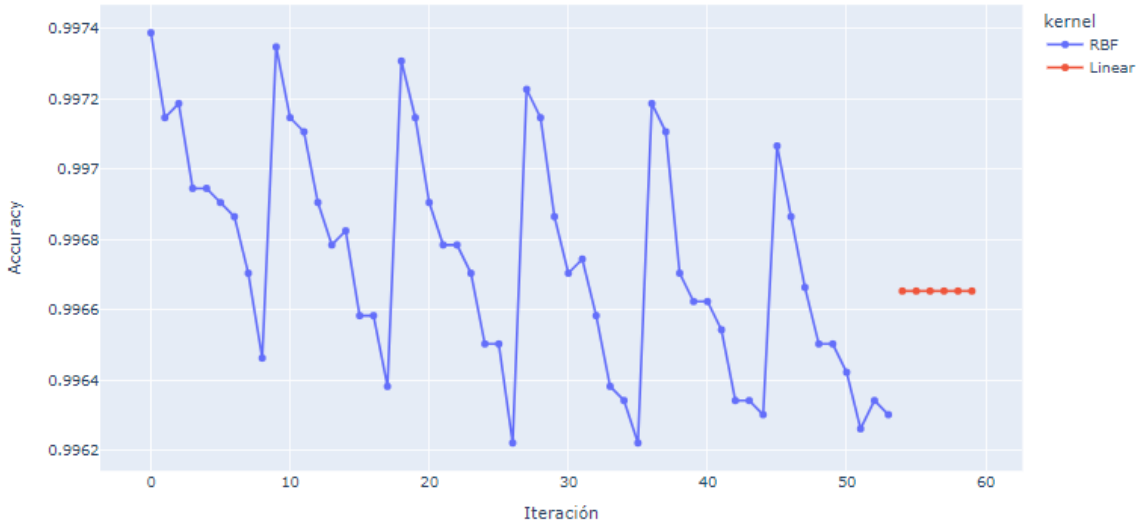


Figura 18. Resultados iteraciones Maquinas de soporte vectorial- Sentinel
Fuente: Elaboración propia

- Por último, se entrenaron los modelos finales del algoritmo de máquinas de soporte vectorial con los hiperparámetros obtenidos en la validación cruzada, y se evalúa el desempeño de estos modelos en las bases de datos de testeo, que se crearon en la actividad 5.1.

5.3.3. REDES NEURONALES

Imágenes Landsat

- Inicialmente se escoge la librería de Keras disponible en Python para el desarrollo de los modelos de redes neuronales Fully connected y se hace el cargue de los datos obtenidos de las imágenes Landsat de acuerdo con la estructura que requiere esta librería.

- Se realizó la optimización de los hiperparámetros numero de capas (2 y 3), número de unidades (164, 328 y 492) y tasa de Dropout (0.5 y 0.7) usando la base de entrenamiento y validación explicadas previamente. Además, se utilizó la función de activación RELU para las capas ocultas y la función Sigmoid para la capa de salida, en la compilación del modelo se utilizó el optimizador ADAM y la función de pérdida binary_crossentropy, por último, en el entramiento se utiliza 10 epochs y el batch size default de esta librería que es de 32.

- Después de realizar 10 iteraciones se obtuvo que el modelo con mejor accuracy y menor valor de pérdida tiene la siguiente estructura:

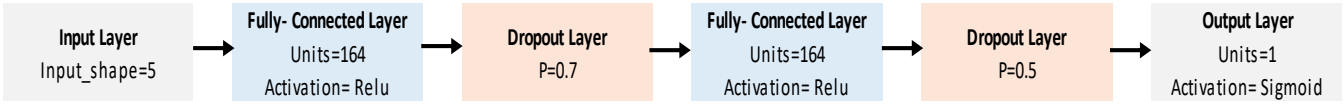


Figura 19. Red neuronal Fully connected – Landsat
Fuente: Elaboración propia

- En cuanto a las gráficas de rendimiento y perdida no se identifica sobreajuste del modelo, ni problemas de gradientes que explotan o desvanecen. Se identifica convergencia del modelo, obteniendo rendimientos similares tanto en la base de entrenamiento como de testeo en cada iteración, y obteniendo un accuracy de 0.987.

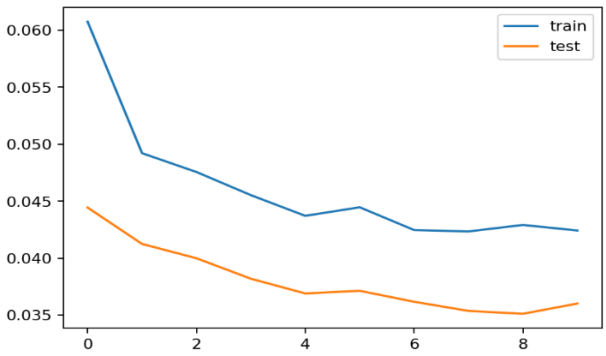


Figura 20. Función de costo Red neuronal
Fuente: Elaboración propia

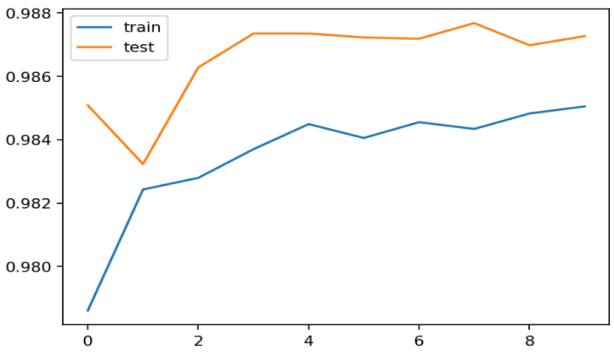


Figura 21. Accuracies Red neuronal
Fuente: Elaboración propia

- Por último, se evalúa el desempeño de este modelo en las bases de datos de testeo, que se crearon en la actividad 5.1.

Imágenes Sentinel

- Inicialmente se escoge la librería de Keras disponible en Python para el desarrollo de los modelos de redes neuronales Fully connected y se hace el cargue de los datos obtenidos de las imágenes Sentinel de acuerdo con la estructura que requiere esta librería.

- Se realizó la optimización de los hiperparámetros numero de capas (2 y 3), número de unidades (164, 328 y 492) y tasa de Dropout (0.5 y 0.7) usando la base de entrenamiento

y validación explicadas previamente. Además, se utiliza la función de activación RELU para las capas ocultas y la función Sigmoid para la capa de salida, en la compilación del modelo se utiliza el optimizador ADAM y la función de pérdida binary_crossentropy, por último, en el entramiento se utiliza 10 epochs y el batch size default de esta librería que es de 32.

- Después de realizar 10 iteraciones se obtuvo que el modelo con mejor accuracy y menor valor de pérdida tiene la siguiente estructura:

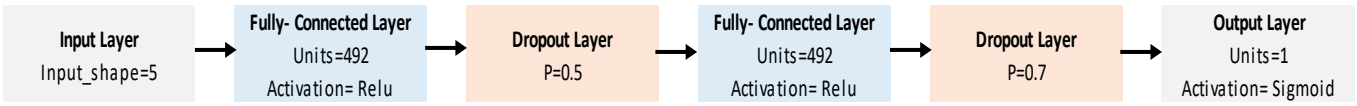


Figura 22. Red neuronal Fully connected – Sentinel

- En cuanto a las gráficas de rendimiento y perdida no se identifica sobreajuste del modelo, ni problemas de gradientes que explotan o desvanecen. Se identifica convergencia del modelo, obteniendo rendimientos similares tanto en la base de entrenamiento como de testeo en cada iteración, y obteniendo un accuracy de 0.988.

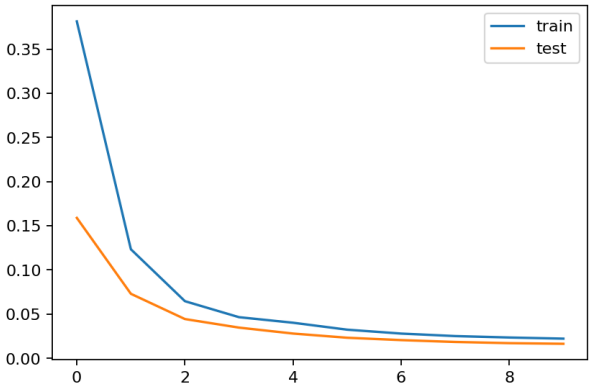


Figura 23. Función de costo Red neuronal
Fuente: Elaboración propia

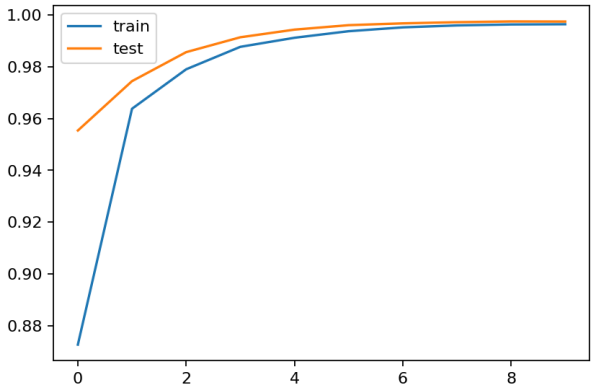


Figura 24. Accuracies Red neuronal
Fuente: Elaboración propia

- Por último, se evalúa el desempeño de este modelo en las bases de datos de testeo, que se crearon en la actividad 5.1.

6. EVALUACIÓN Y SELECCIÓN DEL MODELO

6.1. SELECCIÓN DE LOS CRITERIOS DE EVALUACIÓN

De acuerdo con los modelos seleccionados que corresponden a técnicas de aprendizaje supervisado y redes neuronales fully-connected, los criterios de evaluación que se seleccionaron fueron la matriz de confusión y el Accuracy explicados en el marco teórico y que están disponibles en las librerías de Google Earth Engine y Keras-Tensorflow.

6.2. CÁLCULO DE LOS CRITERIOS DE EVALUACIÓN

A continuación, se presentan los resultados obtenidos en las métricas seleccionadas al evaluar cada modelo entrenado en las bases de datos de testeo.

- Métricas de evaluación para los modelos entrenados en la base de testeo para las imágenes del satélite Landsat

TABLA 13.
RESULTADOS CONSOLIDADOS CRITERIOS DE EVALUACIÓN- LANDSAT

	Bosques aleatorios	Máquinas de soporte vectorial	Redes neuronales Fully connected
Accuracy	0.987727	0.990892	0.986952

- Métricas de evaluación para los modelos entrenados en la base de testeo para las imágenes del satélite Sentinel

TABLA 14.
RESULTADOS CONSOLIDADOS CRITERIOS DE EVALUACIÓN- SENTINEL

	Bosques aleatorios	Máquinas de soporte vectorial	Redes neuronales Fully connected
Accuracy	0.997112	0.997089	0.997009

6.3. SELECCIÓN DEL MODELO CON EL MEJOR DESEMPEÑO

De acuerdo con la métrica de evaluación accuracy, se seleccionaron los modelos con mejor desempeño para cada base de datos, por lo siguiente, el modelo seleccionado para las imágenes Landsat fue el entrenado con el algoritmo de máquinas de soporte vectorial, en cuanto al modelo seleccionado para las imágenes satelitales Sentinel fue el entrenado con el algoritmo de bosques aleatorios.

6.4. ANÁLISIS DE LA DEFORESTACIÓN DE ACUERDO CON LOS RESULTADOS OBTENIDOS

Una vez seleccionados los modelos para cada periodo de observación, se realizó la predicción de los pixeles a una resolución espacial de 30 metros para las imágenes del satélite Landsat para el 2015 y las imágenes del satélite Sentinel para el 2021, generando así los mapas clasificados que se presentan a continuación. De forma visual, se identifica que los modelos desarrollados logran clasificar correctamente las zonas de deforestación (coberturas diferente a bosques) que se identificaban en las imágenes satelitales y en el indicador de NDVI, estos modelos también capturan el efecto de espina de pescado de la deforestación que se presenta principalmente en el municipio de San José del Guaviare. En una revisión detallada, se identifica que los modelos logran identificar y clasificar correctamente las zonas construidas, los cuerpos permanentes de agua, tierras de cultivo, vegetación escasa, entre otros.

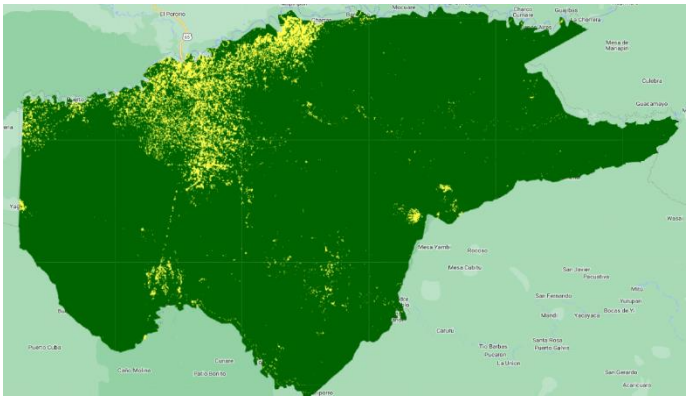


Figura 25. Resultado Mapa clasificado 2015 – Landsat
Fuente: Elaboración propia en Google Earth Engine

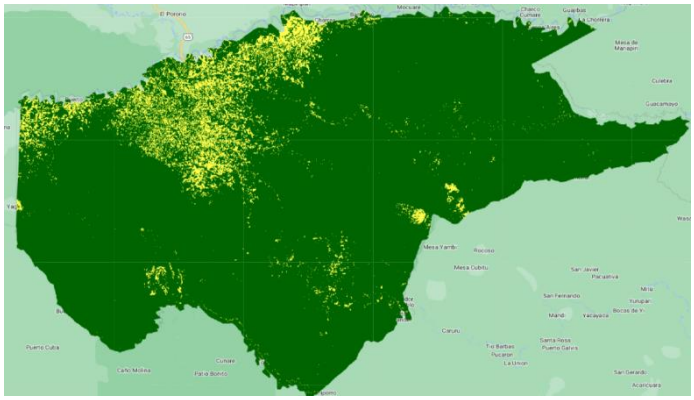


Figura 26. Resultado Mapa clasificado 2015 – Sentinel
Fuente: Elaboración propia en Google Earth Engine

Estos mapas calculados con zonas de deforestación y forestación para cada periodo de evaluación se procesaron para obtener un tercer mapa con las zonas donde hubo perdida de bosque durante este periodo de análisis, esto se realizó con funciones propias de la librería de Google Earth Engine, estos tres mapas son los resultados principales de este proyecto y serán el insumo inicial para realizar la herramienta visual y calcular los indicadores de deforestación para la zona de interés.

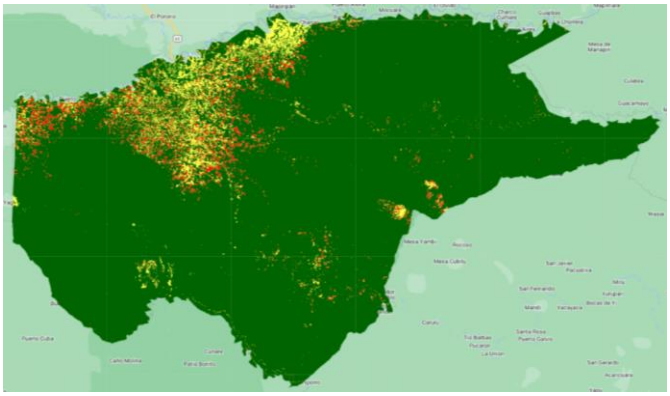


Figura 27. Resultado Mapa pérdida de bosque del 2015 al 2021
Fuente: Elaboración propia en Google Earth Engine

Adicionalmente, se realizó una validación externa comparando las etiquetas obtenidas con los modelos de clasificación versus las etiquetas de las colecciones de Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 y ESA WorldCover 10m v200 disponibles en Google Earth Engine, para esto se realizó muestreo estratificado de las zonas de deforestación (valor de 0) y forestación (valor de 1), para cada etiqueta se tomaron 16,000 pixeles aleatorios a una resolución espacial de 30 metros. De esta forma, se contaron con 32,000 observaciones para cada periodo de análisis y se realizó la misma homologación de etiquetas explicada en el numeral 4.7.

A continuación, se presentan los resultados de las matrices de confusión obtenidas de esta validación externa, de forma general se identifica que para ambos periodos hay una mayor divergencia en las etiquetas en los falsos negativos, es decir, cuando los modelos desarrollados en este proyecto clasifican los pixeles como zona de deforestación, pero las colecciones de Google consideran estos pixeles como coberturas forestales. Inicialmente podemos suponer que se debe, en primer lugar, a que las metodologías para estimar la superficie de cobertura son diferentes y, en segundo lugar, a las inconsistencias evaluadas anteriormente en los catálogos. Sin embargo, este proyecto no tiene como alcance profundizar en las metodologías de las colecciones de Google Earth Engine, ni evaluar que metodología tiene un mejor rendimiento, y tampoco se dispone de una base externa con las etiquetas definitivas para poder llevar a cabo dicho análisis.

Matriz de confusión periodo 2015

		Catálogos GEE	
		Deforestación	Forestación
Modelos desarrollados	Deforestación	12,484	3,516
	Forestación	232	15,768

Matriz de confusión periodo 2021

		Catálogos GEE	
		Deforestación	Forestación
Modelos desarrollados	Deforestación	12,834	3,166
	Forestación	16	15,984

Posteriormente, se realizó el cálculo de las métricas de evaluación Accuracy y F1-score para cada periodo de análisis, identificamos que hubo un alto porcentaje de similitud en las etiquetas de las muestras de los modelos desarrollados versus las etiquetas de los catálogos de Google Earth Engine, obteniendo acurracies superiores al 88% y F1-Scores superiores al 89%.

TABLA 15.
EVALUACIÓN EXTERNA

	2015	2021
Accuracy	0.882875	0.9005625
F1- Score	0.893776	0.909473

A continuación, se muestran algunos ejemplos de diferencias entre la clasificación de cobertura de los modelos desarrollados en este proyecto y las colecciones de Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 y ESA WorldCover 10m v200.

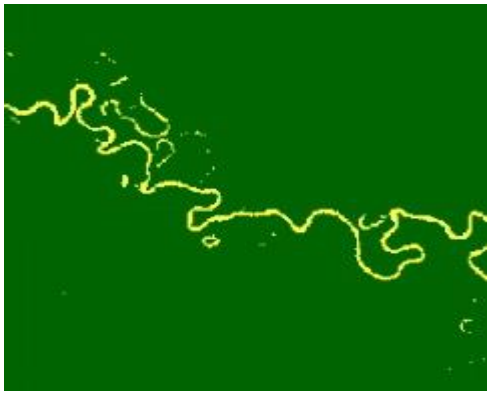
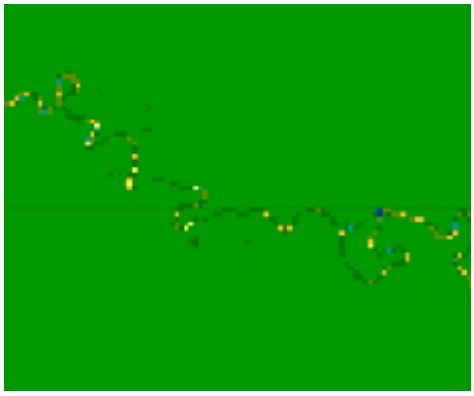
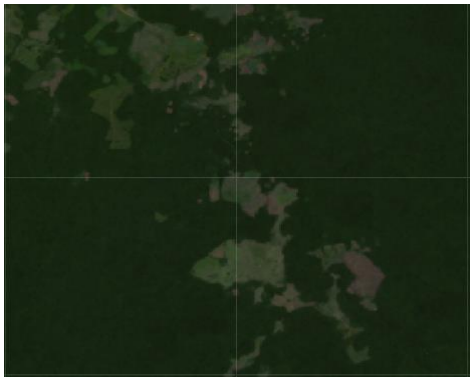
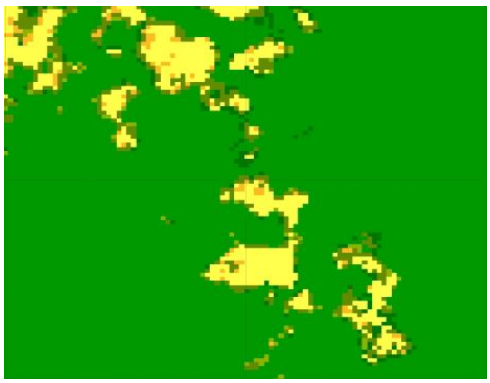
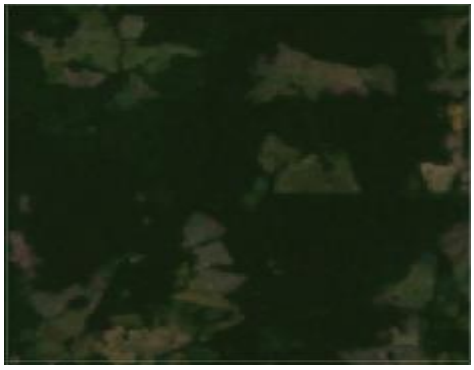
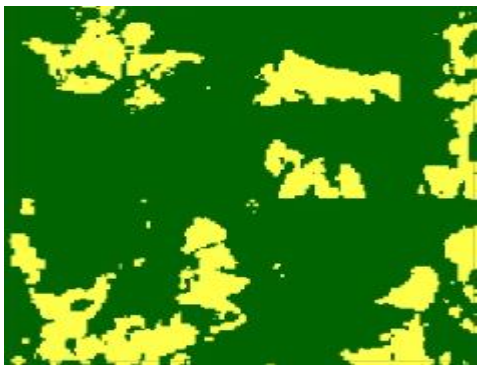
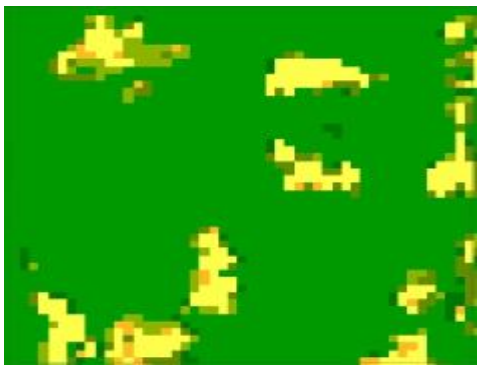
IMAGEN SATELITAL	MODELOS DESARROLLADOS	CATÁLOGOS EARTH ENGINE
		
		
		

Figura 28. Comparación modelos de aprendizaje automático y colecciones de Google Earth Engine
Fuente: Elaboración propia

7. ZONAS PRIORITARIAS

7.1. DELIMITACIÓN DE LAS ZONAS

Las zonas se delimitaron con base a los 4 municipios que hacen parte del departamento del Guaviare, siendo estos: El Calamar, El Retorno, Miraflores y San José del Guaviare.

7.2. CÁLCULO DE LOS INDICADORES DE DEFORESTACIÓN PARA CADA ZONA

Se realizó el cálculo de los indicadores de deforestación en base a los mapas calculados y creados previamente. De acuerdo con los resultados obtenidos, se decide ordenar los municipios con mayor alerta de deforestación conforme al área deforestada en el periodo de evaluación. De esta forma, se identifica que los municipios de San José del Guaviare y El retorno son los que tuvieron una mayor pérdida de bosque desde el 2015 hasta el 2021, siendo una perdida relativa superior al 5% del área total de cada municipio.

TABLA 16.
INDICADORES DE DEFORESTACIÓN

Zona	Área total (km2)	Área Deforestada 2015 (km2)	Área Deforestada 2021 (km2)	Bosque deforestado 2015-2021(km2)	% De perdida de bosque
Municipio El Calamar	11,907.87	407.15	552.26	145.11	1.2186%
Municipio El Retorno	8,775.38	836.73	1,284.1	447.37	5.0981%
Municipio Miraflores	11,458.46	393.08	527.51	134.43	1.1732%
Municipio San José del Guaviare	24,313.47	3,381.34	4,599.35	1,218.01	5.0096%
Departamento del Guaviare	**55,840.83**	**5,017.98**	**6,962.91**	**1,944.93**	**3.4830%**

8. HERRAMIENTA VISUAL

8.1. CONSOLIDAR LOS RESULTADOS DE LOS MAPAS Y EL SISTEMA DE ALERTAS

Desde Python se exportaron los mapas originales obtenidos en la colección de las imágenes satelitales de Landsat y Sentinel, y los mapas obtenidos luego de la clasificación de pixeles con los modelos de aprendizaje automático seleccionados (Formato GeoTIFF). A continuación, se enlistan los mapas que son insumo para el desarrollo de la herramienta visual.

- Mapa del departamento del Guaviare para el 2015, obtenido con las imágenes satelitales de Landsat.
- Mapa del departamento del Guaviare para el 2021, obtenido con las imágenes satelitales de Sentinel.
- Mapa del departamento del Guaviare para el 2015, con la clasificación de pixeles aplicando el modelo desarrollado con el algoritmo de máquinas de soporte vectorial.
- Mapa del departamento del Guaviare para el 2021, con la clasificación de pixeles aplicando el modelo desarrollado con el algoritmo de máquinas de Bosques aleatorios.
- Mapa del departamento del Guaviare con las áreas de perdida de bosque desde el 2015 hasta el 2021, este mapa fue calculado a partir de los dos mapas con las zonas de deforestación clasificadas y herramientas propias de la librería de Google Earth Engine.

8.2. SELECCIÓN DEL SOFTWARE A UTILIZAR

Para la selección del software se revisaron diferentes herramientas de visualización geoespaciales y los costos asociados al licenciamiento de estas, de esta forma, se escogió la herramienta de Earth Engine que funciona como servidor y permite realizar el cargue de los mapas desarrollados en este proyecto en formato GeoTIFF, además en la API de esta herramienta es posible desarrollar una aplicación en lenguaje de programación JavaScript con los resultados y realizar la publicación web.

8.3. DESARROLLO DEL DASHBOARD CON LOS RESULTADOS

La herramienta de visualización se desarrolló en lenguaje JavaScript y se publicó en la siguiente página web:

https://ee-paolaleona12.projects.earthengine.app/view/analisis-de-la-deforestacion-en-el-guaviare

El dashboard está compuesto de 4 paneles que resumen de forma general el objetivo y los resultados obtenidos en este proyecto de forma gráfica. A continuación, se describe cada uno de ellos:

1. En el primer panel se da una breve introducción de la problemática de la deforestación en la Amazonía colombiana y el objetivo principal del proyecto abordado y del dashboard.
2. En el segundo panel se ven de forma gráfica los mapas resultantes de la clasificación de las zonas con deforestación y forestación en el departamento del Guaviare con los modelos de aprendizaje automático desarrollados y que obtuvieron el mejor rendimiento para cada periodo de análisis, también es posible visualizar las imágenes obtenidas de los satélites Landsat y Sentinel que fueron el insumo inicial.

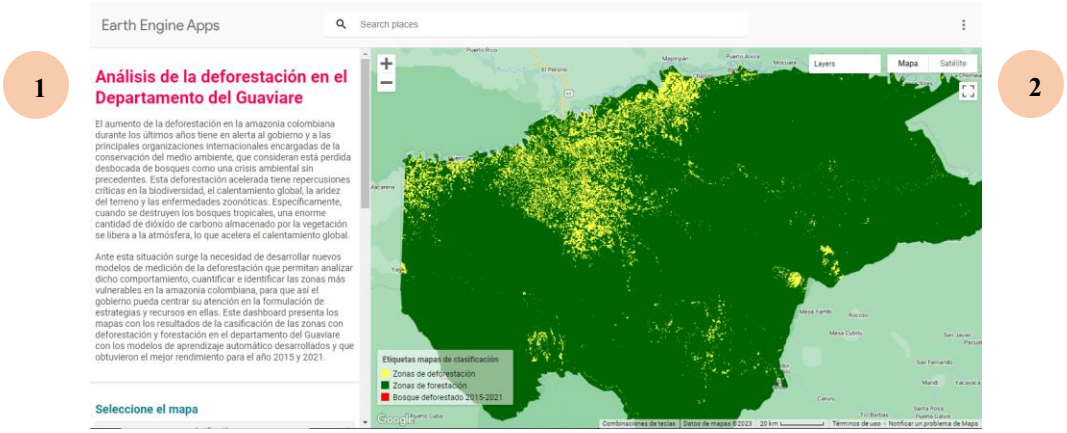


Figura 29. Aplicación web 1
Fuente: Elaboración propia

3. En el tercer panel es posible que el usuario interactúe con la aplicación seleccionando el mapa y la zona de interés. De esta forma, en el panel 2 se mostrará el mapa de acuerdo con los criterios seleccionados.
4. En el último panel se presentan los indicadores de deforestación calculados a partir de los mapas luego de la clasificación de pixeles y que también fueron presentados en este documento.

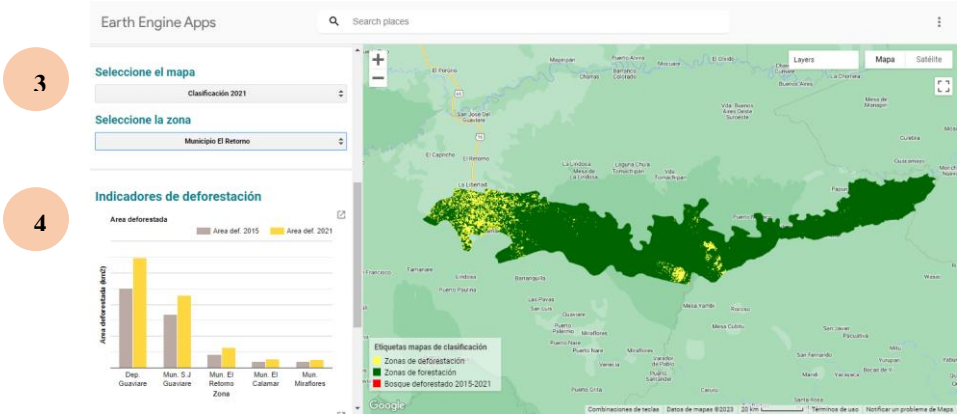


Figura 30. Aplicación web 2
Fuente: Elaboración propia

9. CONCLUSIONES Y TRABAJOS FUTUROS

9.1. CONCLUSIONES

Los modelos propuestos en este proyecto abordan este problema enfocado en la clasificación de pixeles de las imágenes satelitales en las clases de deforestación y forestación. Para el desarrollo de este proyecto, inicialmente se tuvieron que realizar diversas actividades en la minería de datos con el objetivo de contar con imágenes satelitales consistentes y completas de la zona de interés, ya que este tipo de datos se puede ver afectado por nubes y cirros que dificultarían sustancialmente los resultados del análisis, además se procedió a calcular el indicador de diferencia normalizada de vegetación y estandarizar los valores de las bandas que se consideran los principales insumos para el entrenamiento de los modelos. Así mismo, al validar diferentes colecciones de cobertura de tierra disponibles en Google Earth Engine, se identificaron varias inconsistencias en las etiquetas de estas colecciones y que generaría algunos errores al tomar muestras aleatorias para construir la base de datos, por esta razón, se decidió hacer un plan de acción creando dos colecciones nuevas con las características a evaluar: forestación y deforestación (cobertura diferente a bosque), garantizando la calidad de las muestras y validando la exactitud de las etiquetas a nivel de pixel, una vez creada la base de datos, se procedió a dividir en entrenamiento, validación y testeo.

Posteriormente, con la base de entrenamiento y validación, se entrenaron y optimizaron los hiperparámetros de los algoritmos de bosques aleatorios, máquinas de soporte vectorial y redes neuronales fully-connected, con estos modelos se procedió a evaluar su desempeño en la base de testeo. De forma general, se obtuvo un buen desempeño con todos los modelos para ambos periodos de tiempo, con accuracies superiores al 97%. Sin embargo, el mejor modelo para las imágenes del 2015 fue el entrenado con el algoritmo de máquinas de soporte vectorial, y para las imágenes del 2021 fue el entrenado con bosques aleatorios. Con los modelos seleccionados se realizó la clasificación de pixeles a una resolución espacial de 30 metros para la imagen completa de cada periodo de análisis, de forma visual, se identifica que los modelos desarrollados logran clasificar correctamente las zonas de deforestación que se identificaban en las imágenes satelitales y en el indicador de NDVI, como lo son cuerpos de agua permanente, praderas, tierras de cultivo, áreas construidas, entre otras. Además, estos modelos también capturan el fenómeno de espina de pescado de la deforestación que se presenta principalmente en el municipio de San José del Guaviare, y que ha sido el municipio con mayor deforestación.

Adicionalmente, se realizó una validación externa comparando las etiquetas obtenidas con los modelos de clasificación versus las etiquetas de las colecciones disponibles en Google Earth Engine, para esto se realizó muestreo estratificado y para cada etiqueta se tomaron 16,000 pixeles aleatorios a una resolución espacial de 30 metros. De forma general, se obtuvieron acurracies superiores a 88% y F1-Scores a 89%, se identifica que para ambos periodos hay una mayor divergencia en las etiquetas en los falsos negativos, es decir, cuando los modelos de aprendizaje supervisado clasifican los pixeles como zona de deforestación, pero las colecciones de Google Earth Engine etiquetan estos pixeles como coberturas forestales. Inicialmente podemos suponer que se debe a que las metodologías para estimar la superficie de cobertura son diferentes y a las inconsistencias evaluadas en los catálogos.

En base a los mapas creados en este proyecto, se procedió a calcular algunos indicadores de deforestación que nos pueden ayudar a cuantificar la perdida de bosque para el departamento del Guaviare y para cada municipio en el periodo de análisis. Por lo anterior, se identifica que los municipios de San José del Guaviare y El retorno son los que tuvieron una mayor pérdida de bosque desde el 2015 hasta el 2021, de 1,218 km2 y 447 km2 respectivamente, siendo una perdida relativa superior al 5% del área total de cada municipio. En cuanto al área deforestada para el departamento del Guaviare fue de 1,944 km2.

De acuerdo con el desarrollo y los resultados obtenidos en este proyecto, podemos evidenciar el buen desempeño de los modelos de aprendizaje automático supervisado para la clasificación de pixeles en las clases de forestación y deforestación (cobertura diferente a bosque), y en este caso en particular para analizar el comportamiento e incremento de la deforestación en la Amazonía colombiana con imágenes satelitales disponibles en la plataforma de Google Earth Engine. De esta forma, este proyecto aplicado brinda un acercamiento a este fenómeno desde los datos para apoyar a las investigaciones de conservación de recursos naturales de las facultades de ciencias, además de ser una aproximación inicial para proporcionar herramientas de análisis que son indispensables en la formulación de planes de acción y de políticas de conservación y sostenibilidad ambiental.

9.2. TRABAJOS FUTUROS

Como continuación de este trabajo y como en cualquier otro proyecto, existen diversas líneas de aplicación e investigación que quedan abiertas y en las que es posible continuar trabajando. Durante el desarrollo de este proyecto han surgido algunas líneas futuras que se han dejado abiertas y que se esperan atacar en un futuro: algunas de ellas, están más directamente relacionadas con este trabajo y son el resultado de cuestiones que han ido surgiendo durante la realización de este. Otras, son líneas más generales que, sin embargo, no son objeto de este proyecto; estas líneas pueden servir para retomarlas posteriormente o como opción a trabajos futuros para otros investigadores.

A continuación, se presentan algunos trabajos futuros que pueden desarrollarse como resultado de este trabajo o que, por exceder el alcance de este proyecto, no han podido ser tratados con la suficiente profundidad. Además, se sugieren algunos desarrollos específicos para apoyar y mejorar el modelo y metodología propuestos. Entre los posibles trabajos se destacan:

- Evaluar otras fuentes de imágenes satelitales que tengan una mejor resolución o inclusive donde exista información completa de más departamentos, para así evaluar el aumento de la deforestación incluyendo más territorio o en un mejor caso la totalidad de la Amazonía colombiana.
- Evaluar la posibilidad de realizar la conversión de los archivos a formatos abiertos, para que así, los datos sean compatibles con otras librerías de Python y poder realizar repeticiones del experimento con particiones hold-out aleatorias en la base de datos.
- Investigar, evaluar y entrenar modelos de clasificación como XGBOOST y procesos gaussianos.

- Desarrollo y entrenamiento de modelos de aprendizaje no supervisado, que no requieren que los datos se encuentren etiquetados. Con esto se podría evaluar y comparar los resultados con los mapas elaborados en este proyecto.
- Desarrollo y entrenamiento de redes neuronales convolucionales con el objetivo de realizar segmentación semántica de las imágenes satelitales. Para esto, en primer lugar, se debe realizar un procesamiento de las imágenes diferente al que se realizó en este proyecto, ya que se debe contar con máscaras de las imágenes para cada tipo de cobertura. Considerando que este tipo de redes neuronales son muy reconocidas en problemas de reconocimiento de imágenes y visión artificial.
- Evaluación de las metodologías de clasificación/segmentación de la cobertura de tierra de los catálogos disponibles en Google Earth Engine, con el objetivo de evaluar los pros, contras y rendimiento al comparar con modelos de aprendizaje automático.

11. REFERENCIAS BIBLIOGRÁFICAS

[1] Adelphi, WWF, Fundación ideas para la paz, «Un clima peligroso: Deforestación, cambio climatico y violencia contra los defensores ambientales en la Amazonía,» WWF Alemania, Berlín, 2021.

[2] K. Rodríguez, «La deforestación en Colombia subió 11 % en el primer semestre de 2022,» *Potafolio,* 16 Septiembre 2022.

[3] S. Muhamad, Interviewee, *En Colombia se han deforestado más de tres millones de hectáreas de bosque en las últimas dos décadas.* [Entrevista]. 7 Septiembre 2022.

[4] WWF Alemania, Fundación Ideas para la Paz (FIP), adelphi, «UN CLIMA PELIGROSO: Deforestación, cambio climático y violencia contra los defensores ambientales en la Amazonía colombiana,» Maro Ballach/WWF Alemania, Berlin, 2021.

[5] Instituto de Hidrología, Meteorología y Estudios Ambientales-IDEAM, «Análisis de tendencias y patrones espaciales de deforestación en Colombia,» Comité de Comunicaciones y Publicaciones del IDEAM, Bogotá D.C., 2011.

[6] Greenpeace, «Greenpeace: Bosques/Amazonas,» Greenpeace, Noviembre 2022. [En línea]. Available: https://es.greenpeace.org/es/trabajamos-en/bosques/amazonas/. [Último acceso: 13 Noviembre 2022].

[7] WWF, «WWF: Sobre la amazonía/El bioma amazónico,» WWF, Noviembre 2022. [En línea]. Available: https://wwf.panda.org/es/sobre_la_amazonia/. [Último acceso: 13 Noviembre 2022].

[8] L. Suárez , L. Rivera, M. Asunción, I. Pratesi, M. Galaverni y M. Antonelli, «Pérdida de naturaleza y pandemias. Un planeta sano por la salud de la humanidad,» WWF España, Madrid, 2020.

[9] REVISTA SEMANA, «GUAVIARE: La selva a mordiscos».*SEMANA.*

[10] Google, «Google Earth Engine,» Google, Noviembe 2022. [En línea]. Available: https://earthengine.google.com/. [Último acceso: 14 Noviembre 2022].

[11] NASA, «Landsat science NASA,» [En línea]. Available: https://landsat.gsfc.nasa.gov/satellites/landsat-8/. [Último acceso: 04 Abril 2023].

[12] Programme of the European Union, «Copernicus. Europe's eyes on Earth,» [En línea]. Available: https://www.copernicus.eu/es/sobre-copernicus/infraestructura/estos-son-nuestros-satelites. [Último acceso: 04 Abril 2023].

[13] ARCGIS, «ArcGIS Pro- Funciones de análisis,» [En línea]. Available: https://pro.arcgis.com/es/pro-app/latest/help/analysis/raster-functions/ndvi-function.htm. [Último acceso: 04 Abril 2023].

[14] S. Raschka y V. Mirjalili, Python machine learning, Birmingham: Packt Publishing, 2019.

[15] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, Inc, 2019.

[16] F. Chollet, Deep Learning with Python, Manning Publications Co., 2018.

[17] Scikit-learn's authors, «Scikit learn 3.3. Metrics and scoring: quantifying the quality of predictions,» Noviembre 2022. [En línea]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. [Último acceso: 20 Noviembre 2022].

[18] K. Nichols y P. Hosein, «Estimating deforestation using machine learning algorithms,» *Second international conference on intelligence data science technologies and applications (IDSTA),* pp. 82-87, 2021.

[19] S. Saha, S. Bhattacharjee, P. Kumar, N. Sengupta y B. Bera, «Deforestation probability assessment using integrated machine learning algorithms of Eastern Himalayan foothills,» *Resources, conservation & recycling advances,* vol. 14, nº 200077, 31 Marzo 2022.

[20] M. A. Brovelli, Y. Sun y V. Yordanov, «Monitoring forest change in the amazon using multi-temporal remote sensing data and machine learning classification on google earth engine,» *International journal of Geo-information,* vol. 9, nº 580, 2020.

[21] J. Irvin, N. Ramachandran, S. Johnson-Yu, S. Zhou, R. Rustowicz, K. Story, C. Elsworth y K. Austin, «ForestNet: Classifying drivers of deforestation in Indonesia using deep learning on satellite imagery,» Stanford ML Group, Vancouver, 2020.

[22] C. y. A. Expertos, Interviewee, *Tragedia ambiental devora velozmente los bosques del Guaviare y la Amazonía.* [Entrevista]. 1 Agosto 2021.

[23] IDEAM, «Resultados del monitoreo deforestadicón,» 2020.

[24] C. L. Service, «Earth Engine Data Catalog,» [En línea]. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_Landcover_100m_Proba-V-C3_Global. [Último acceso: 31 Octubre 2023].

[25] T. E. S. A. (ESA), «Earth Engine Data Catalog,» [En línea]. Available: https://developers.google.com/earth-engine/datasets/catalog/ESA_WorldCover_v200. [Último acceso: 31 Octubre 2023].

[26] P. M. Mather y B. TSO, Classification methods for remotely sensed data, Boca Raton, Florida: Chemical Rubber Company Press, 2009.

[27] R. A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, Academic Press, 2006.