



Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica
Escuela Profesional de Ingeniería Geográfica

**Desarrollo de un modelo de aprendizaje
ensamblado para la detección de quema de caña de
azúcar en la costa norte y centro del Perú durante
el período 2017 – 2022**

TESIS

Para optar el Título Profesional de Ingeniero Geógrafo

AUTOR

Jair Francisco FLORES FARFAN

ASESOR

Dr. Francisco Alejandro ALCANTARA BOZA

Lima, Perú

2024



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Flores, J. (2024). *Desarrollo de un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar en la costa norte y centro del Perú durante el período 2017 – 2022*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica, Escuela Profesional de Ingeniería Geográfica]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios:

Datos de autor	
Nombres y apellidos	Jair Francisco Flores Farfan
Tipo de documento de identidad	DNI
Número de documento de identidad	71992642
URL de ORCID	https://orcid.org/0009-0009-2959-1899
Datos de asesor	
Nombres y apellidos	Francisco Alejandro Alcantara Boza
Tipo de documento de identidad	DNI
Número de documento de identidad	27074721
URL de ORCID	https://orcid.org/0000-0001-9127-4450
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	José Luis Quispe Vilchez
Tipo de documento de identidad	DNI
Número de documento de identidad	08119070
Miembro del jurado 1	
Nombres y apellidos	Dante Horacio Loayza Alatrasta
Tipo de documento de identidad	DNI
Número de documento de identidad	07564915

Datos de investigación

Línea de investigación	C.0.3.16. Procesamiento de Imágenes y Visión por computadora
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	País: Perú Departamento: Piura Latitud: -4.929 Longitud: -80.854 Departamento: Lambayeque Latitud: -6.085 Longitud: -80.063 Departamento: La Libertad Latitud: -8.094 Longitud: -78.962 Departamento: Áncash Latitud: -9.146 Longitud: -78.281 Departamento: Lima Latitud: -10.674 Longitud: -77.823
Año o rango de años en que se realizó la investigación	Septiembre 2023 – Septiembre 2024
URL de disciplinas OCDE	Ciencias del medio ambiente https://purl.org/pe-repo/ocde/ford#1.05.08 Geociencias, Multidisciplinar https://purl.org/pe-repo/ocde/ford#1.05.01 Sensores remotos https://purl.org/pe-repo/ocde/ford#2.07.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE INGENIERÍA GEOLÓGICA, MINERA, METALÚRGICA Y GEOGRÁFICA

ESCUELA PROFESIONAL DE INGENIERÍA GEOGRÁFICA

ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR TÍTULO PROFESIONAL DE INGENIERO GEÓGRAFO

En el Salón de Grados y Títulos de la Escuela Profesional de Ingeniería Geográfica de la Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica de la Universidad Nacional Mayor de San Marcos, el jueves 19 de diciembre del 2024, siendo las 09:30 horas, en presencia de los Señores Docentes designados como Miembros del Jurado Calificador:

ING. JOSE LUIS QUISPE VILCHEZ

Presidente

ING. DANTE HORACIO LOAYZA ALATRISTA

Miembro

Reunidos en Acto Académico Público de Sustentación de la Tesis titulada: «**DESARROLLO DE UN MODELO DE APRENDIZAJE ENSAMBLADO PARA LA DETECCIÓN DE QUEMA DE CAÑA DE AZÚCAR EN LA COSTA NORTE Y CENTRO DEL PERÚ DURANTE EL PERÍODO 2017 - 2022**», para optar el Título Profesional de Ingeniero Geógrafo.

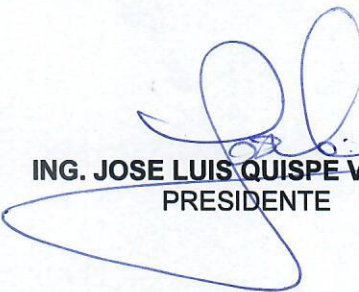
Expuesta la Tesis; los miembros del Jurado plantearon al Bachiller las preguntas pertinentes, que fueron absueltas por el tesista: *2. SATISFACCIÓN*

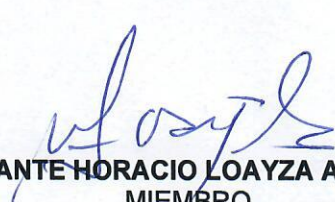
Concluida la sustentación de Tesis, el Jurado procedió a evaluar y calificar la calidad y sustentación en secreto, cuyo calificativo fue: *Diecisiete (17), Aprobado con mención Honrosa*

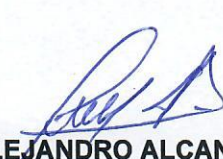
Habiendo sido aprobada la Sustentación de la Tesis por el Jurado Calificador, el presidente del Jurado recomienda que la Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica, otorgue el **TÍTULO PROFESIONAL DE INGENIERO GEÓGRAFO**, a Don **JAIR FRANCISCO FLORES FARFAN**

Siendo las *10.30* horas, se dio por concluido el acto académico, expidiéndose cinco (05) Actas Originales de la Sustentación de Tesis, firmadas por el Jurado Calificador.

Ciudad Universitaria, 19 de diciembre del 2024


ING. JOSE LUIS QUISPE VILCHEZ
PRESIDENTE


ING. DANTE HORACIO LOAYZA ALATRISTA
MIEMBRO


DR. FRANCISCO ALEJANDRO ALCANTARA BOZA
ASESOR DE TESIS



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Vicerrectorado de Investigación y Posgrado



CERTIFICADO DE SIMILITUD

Yo Francisco Alejandro Alcántara Boza, en mi condición de asesor acreditado con la Resolución Decanal N° 000358-2024-D-FIGMMG/UNMSM, de la tesis, cuyo título es **Desarrollo de un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar en la costa norte y centro del Perú durante el período 2017 - 2022**, presentado por el bachiller **Jair Francisco FLORES FARFAN** para optar el título de Ingeniero Geógrafo, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de **07 %** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título correspondiente.

Firma del Asesor: _____

DNI: 27074721

Asesor: Francisco Alejandro Alcántara Boza



*Dedico esta tesis a mi familia, por su amor
y apoyo incondicional que hicieron
posible este logro.*

Agradecimientos

Agradezco a todas las instituciones y personas que hicieron posible la realización de este trabajo de tesis.

En primer lugar, expreso mi gratitud al OEFA por proporcionar la problemática y los datos de validación necesarios para llevar a cabo este estudio.

Mi más profundo agradecimiento para Cesar Aybar, Fernando Prudencio y Roy Yali, cuya orientación y constante disposición para ayudarme fueron fundamentales a lo largo de este proceso.

Asimismo, agradezco especialmente a mi asesor de tesis, el Dr. Alejandro Alcántara Boza, por sus valiosos comentarios y sugerencias que contribuyeron a mejorar la calidad de esta investigación.

Quiero también agradecer a mis colegas Jeanett Valladares, Jose del Castillo y Julio Contreras por su invaluable ayuda en el etiquetado de las imágenes, un objetivo clave de esta investigación.

Finalmente, dedico un especial agradecimiento a mis padres Manuel e Irina y mis hermanos Fredd y Manuel, cuyo apoyo emocional y fortaleza fueron indispensables para completar con éxito esta tesis.

ÍNDICE

I	INTRODUCCIÓN	1
1.1	Introducción	1
1.2	Planteamiento del problema	2
1.2.1	Descripción de la problemática	2
1.2.2	Formulación del problema	4
1.2.2.1	Pregunta general.	4
1.2.2.2	Preguntas específicas.	4
1.3	Objetivos de la investigación	5
1.3.1	Objetivo general	5
1.3.2	Objetivos específicos	5
1.4	Importancia y alcance de la investigación	5
1.4.1	Importancia	5
1.4.2	Alcance	6
1.5	Limitaciones de la investigación	6
1.5.1	Limitaciones metodológicas	6
1.5.2	Limitaciones operacionales	7
1.5.3	Limitaciones en la validación	7
II	EL MARCO TEÓRICO	8
2.1	Antecedentes de la investigación	8
2.1.1	Antecedentes internacionales	8
2.1.2	Antecedentes nacionales	11
2.2	Bases teóricas	12
2.2.1	Detección de áreas quemadas mediante imágenes satelitales	12
2.2.1.1	Imágenes satelitales Sentinel-2.	13
2.2.1.2	Enfoques basados en reglas.	14
2.2.1.2.1	Índice de Quema Normalizado (NBR).	14

2.2.1.2.2	Índice de Detección de Área Quemada (BADI).	15
2.2.1.3	Enfoques basados en aprendizaje automático.	15
2.2.1.3.1	Regresión logística binaria.	16
2.2.1.3.2	Entropía Cruzada Binaria (BCE).	18
2.2.1.3.3	Entropía Cruzada Binaria Ponderada (WBCE).	20
2.2.1.3.4	Función de pérdida de Dice.	20
2.2.1.3.5	WCE - Dice Loss.	21
2.2.1.3.6	Árbol de decisión.	21
2.2.1.3.7	Gradient Boosting Decision Tree (GBDT).	25
2.2.1.3.8	Algoritmo LightGBM.	26
2.2.2	Segmentación Semántica	30
2.2.2.1	Red Neuronal Convolutiva (CNN).	31
2.2.2.2	Arquitectura U-Net.	31
2.2.2.2.1	Capa convolutiva.	32
2.2.2.2.2	Padding y Stride.	33
2.2.2.2.3	Funciones de activación.	34
2.2.2.2.4	Capa de agrupación.	35
2.2.2.2.5	Capa de Deconvolución.	36
2.2.2.2.6	Función de costo.	37
2.2.2.2.7	Codificador y Decodificador.	38
2.2.2.3	Hiperparámetros del entrenamiento de una CNN.	39
2.2.3	Métricas de evaluación	40
2.2.3.1	Métricas basadas en F-measure.	40
2.2.3.2	Sensibilidad y Especificidad.	41
2.2.3.3	Precisión o Índice Rand.	41
2.2.3.4	Índice Kappa.	41
2.2.3.5	Curva ROC y Área Bajo la Curva (AUC).	42
2.2.4	Quema de caña de azúcar	43
2.2.4.1	Fenología de la caña de azúcar.	43

2.2.4.2	Firma espectral de cultivos de caña de azúcar.	44
2.2.4.3	Cosecha y quema de caña de azúcar.	45
2.2.4.4	Firma espectral de áreas quemadas.	47
2.3	Definición de Términos	48
2.3.1	Áreas con quema de caña de azúcar	48
2.3.2	Quema activa	48
2.3.3	Cosecha mecanizada en verde	48
2.3.4	Ingenio azucarero	48
2.3.5	Fundo, Predio y Campo Azucarero	49
2.3.6	Matriz de confusión en la segmentación binaria	49
2.3.7	Índice de Vegetación de Diferencia Normalizada (NDVI)	49
2.3.8	Índice de Agua de Diferencia Normalizada (NDWI)	50
III	HIPÓTESIS Y VARIABLES	51
3.1	Hipótesis	51
3.1.1	Hipótesis general	51
3.1.2	Hipótesis específicas	51
3.2	Variables	51
3.2.1	Variable independiente	51
3.2.2	Variable dependiente	51
3.3	Operacionalización de variables	52
IV	MATERIALES Y MÉTODOS	53
4.1	Área de estudio (Unidad de análisis)	53
4.2	Diseño de investigación	54
4.2.1	Tipo de investigación	54
4.2.2	Nivel de investigación	54
4.2.3	Enfoque de investigación	54
4.2.4	Diseño de investigación	54
4.3	Población y muestra	55

4.3.1	Población	55
4.3.2	Muestra	57
4.4	Procedimiento, técnicas e instrumentos de recolección de información	58
4.4.1	Pre-procesamiento de datos Sentinel-2 L2A	58
4.4.2	Detección temporal de áreas pseudoquemadas	58
4.4.3	Selección de imágenes con áreas quemadas	60
4.4.4	Etiquetado de imágenes	62
4.4.5	Generación del conjunto de datos para la detección de quemas	68
4.4.5.1	Pendiente.	68
4.4.5.2	Distancia a la cobertura agrícola.	69
4.4.6	Implementación del modelo CNN	69
4.4.7	Implementación del Gradient Boosting Decision Tree	70
4.4.8	Implementación del modelo Stacking	72
4.4.9	Validación del umbral de clasificación para el modelo Stacking	72
4.4.10	Inferencia del modelo Stacking en un área piloto	73
4.5	Análisis estadístico	75
4.5.1	Evaluación del modelo	75
V	RESULTADOS, DISCUSIÓN, CONCLUSIÓN Y RECOMENDACIONES	77
5.1	Resultados	77
5.1.1	Implementación de los modelos de aprendizaje automático	77
5.1.1.1	Configuración de la Red Neuronal Convolucional (CNN).	77
5.1.1.2	Resultados del modelo U-Net en la detección de áreas quemadas.	79
5.1.1.3	Configuración y resultados del Gradient Boosting Decision Tree (LightGBM).	80
5.1.1.4	Configuración y resultados del modelo Stacking.	83
5.1.2	Evaluación de los modelos	85
5.1.3	Validación del umbral de probabilidad del modelo Stacking.	85
5.1.4	Inferencia en el área piloto.	90

5.1.4.1	Marzo a junio de 2023.	90
5.1.4.2	Octubre a noviembre de 2023.	91
5.1.4.3	Marzo a abril de 2024.	92
5.2	Discusión	93
5.2.0.1	Discusión de la generación y etiquetado del conjunto de datos.	93
5.2.0.2	Discusión de los resultados de los modelos base.	95
5.2.0.3	Discusión de los resultados del modelo Stacking y su validación.	96
5.2.0.4	Discusión de la inferencia en el área piloto.	97
5.3	Conclusiones y Recomendaciones	98
5.3.1	Conclusiones	98
5.3.2	Recomendaciones	99
VI	REFERENCIAS Y ANEXOS	101
6.1	Referencias bibliográficas	101
6.2	Anexos	110

ÍNDICE DE FIGURAS

1	Función sigmoide utilizada en la regresión logística.	17
2	Función de pérdida de entropía cruzada binaria cuando la clase es positiva ($y = 1$)	19
3	Ejemplo de un árbol de decisión para un problema de clasificación de mamíferos.	22
4	Comparación de las medidas de impureza de Gini, entropía y error de clasificación.	24
5	Ilustración de cómo trabaja el Gradient Boosting Tree.	25
6	Algoritmo basado en histogramas.	27
7	Pseudocódigo del algoritmo Gradient-based One-Sided Sampling.	28
8	Pseudocódigo del método Exclusive Feature Bundling.	29
9	Ejemplo de cómo trabaja el Greedy Bundling - EFB.	30
10	Ejemplo de cómo trabaja el Merge Exclusive Features.	30
11	Ejemplo de segmentación semántica.	31
12	Ilustración de los primeros cinco pasos en una operación de convolución para una imagen de entrada $4 \times 4 \times 1$ y un núcleo de 2×2	32
13	Mapa de característica resultante después de una operación de convolución 2×2 .	33
14	Ejemplo de una operación de convolución utilizando un padding de 1 y un stride de 3.	34
15	Función de activación ReLU.	35
16	Ejemplo de una operación de max pooling utilizando un stride de 1 y un tamaño de filtro de 2×2	36
17	Ilustración de los pasos en una operación de deconvolución para una entrada de 2×2 y un núcleo de 2×2 con un stride de 1.	37
18	Arquitectura U-Net de un modelo de red neuronal convolucional.	39
19	Etapas del desarrollo de la caña de azúcar.	44
20	Firma espectral representativa de la caña de azúcar en la etapa de madurez. . .	45
21	Partes de la caña de azúcar.	46
22	Comparación de la respuesta espectral de la vegetación saludable y las áreas quemadas.	47

23	Ubicación de los principales ingenios azucareros del país que conforman el gremio PERUCANÑA.	53
24	Áreas de cultivo de caña de azúcar.	55
25	Áreas con reporte de quema de caña de azúcar (ROI's).	57
26	Esquema metodológico para el pre-procesamiento de datos Sentinel-2	58
27	Esquema metodológico para la detección pseudoquemados.	59
28	Histograma de áreas pseudoquemadas detectadas (en hectáreas) en las 74 ROI's.	60
29	Combinaciones RGB utilizadas para la selección de imágenes con áreas quemadas.	62
30	Interfaz de etiquetado en IRIS incorporando las combinaciones de banda y una imagen de referencia.	63
31	Ejemplo de etiquetado de imágenes utilizando el método “Scribble”.	64
32	Visualización en IRIS según el criterio de visualización espectral para una imagen en la ROI_0097_04.	66
33	Visualización en IRIS según el criterio de contraste para una imagen en la ROI_0020_01.	66
34	Visualización en IRIS según el criterio de foco de calor para una imagen en la ROI_0084_01.	67
35	Visualización en IRIS según el criterio de ubicación y contexto espacial para una imagen en la ROI_0011_01.	67
36	Visualización en IRIS según el criterio de delimitación de umbral para una imagen en la ROI_0094_02.	68
37	Arquitectura U-Net para la segmentación semántica.	70
38	Esquema de la validación cruzada en diez pliegues.	71
39	Modelo LightGBM para cada pliegue en la validación cruzada.	71
40	Esquema del modelo de ensamblado de stacking.	72
41	Cronograma de actividades para la implementa de medidas de mitigación en el PAMA de Laredo S.A.A.	74
42	Área piloto seleccionada para la inferencia del modelo.	75

43	Resultados del modelo U-Net.	80
44	Importancia de las características en el modelo LightGBM.	82
45	Resultados del modelo LightGBM.	83
46	Resultados del modelo ensamblado.	84
47	Curva ROC para el modelo ensamblado.	86
48	Comparación de umbrales de probabilidad para la emergencia 0016.	87
49	Comparación de umbrales de probabilidad para la emergencia 0376.	88
50	Comparación de umbrales de probabilidad para la emergencia 0068.	89
51	Comparación de umbrales de probabilidad para la emergencia 0456.	90
52	Resultados de la inferencia en el área piloto (Marzo a Junio de 2023).	91
53	Resultados de la inferencia en el área piloto (Octubre a Noviembre de 2023).	92
54	Resultados de la inferencia en el área piloto (Marzo a Abril de 2024).	93
55	Ejemplo de etiquetado de áreas quemadas a nivel mundial.	94
56	Ejemplo de distintos estados del área de cultivo de caña de azúcar durante la época de cosecha.	95
57	Ejemplo de la detección de áreas quemadas en la costa norte y centro del Perú.	97
58	Resumen de las ROI's.	112

ÍNDICE DE TABLAS

1	Bandas espectrales de Sentinel-2 disponibles en Planetary Computer.	13
2	Cuadro de Operacionalización de variables	52
3	Principales ingenios azucareros en la costa norte y centro del Perú.	56
4	Descripción de las combinaciones RGB para diferentes tipos de áreas.	61
5	Criterios para el protocolo de generación de etiquetas de quema.	65
6	Configuración del modelo U-Net	77
7	Configuración del entrenamiento del modelo UNet	78
8	Principales parámetros para el entrenamiento del modelo GBDT	80
9	Configuración del modelo de regresión logística	83
10	Evaluación de los modelos en porcentaje (%).	85
11	Matriz de consistencia.	110

CAPÍTULO I

INTRODUCCIÓN

1.1. Introducción

En la costa norte y centro del Perú, la práctica de quemar los residuos de la caña de azúcar ha generado una creciente preocupación debido a sus impactos negativos en el suelo, el medio ambiente y la salud de la población circundante. Esta tradicional práctica agrícola libera grandes cantidades de gases de efecto invernadero y partículas contaminantes, deteriorando la calidad del aire y afectando la salud de las personas que viven en las zonas cercanas a estos campos de cultivo.

A pesar de los esfuerzos por implementar alternativas más sostenibles, como la cosecha en verde, esta problemática enfrenta desafíos significativos que incluyen la falta de una legislación estricta que prohíba la quema de caña y la insuficiente información proporcionada por las empresas azucareras, lo que dificulta el control y la gestión efectiva de esta práctica. La combinación de estos factores ha generado insatisfacción entre la población, que exige soluciones más eficaces y sostenibles.

En este contexto, se propone desarrollar un modelo de aprendizaje ensamblado que incorpore técnicas avanzadas de aprendizaje automático, específicamente los árboles de decisión potenciados por gradiente y redes neuronales convolucionales para detectar y mapear las áreas con quema de caña de azúcar.

El objetivo de este modelo es proveer información precisa y actualizada sobre la extensión y ubicación de estas zonas, respaldando así una toma de decisiones más informada a fin de mejorar la fiscalización de esta práctica agrícola de las empresas azucareras por parte de los organismos competentes.

Se espera que el modelo facilite la identificación oportuna de dichas áreas, dirigiendo los esfuerzos de monitoreo hacia las unidades fiscalizables que lo requieran. De este modo, sería posible redefinir y modificar los instrumentos de gestión ambiental para prevenir y controlar la quema de caña de azúcar, incorporando métodos de cosecha más sostenibles. Además, su aplicación podría extenderse a otros países con desafíos similares, demostrando el potencial de

la inteligencia artificial para resolver problemáticas como la detección de áreas quemadas.

1.2. Planteamiento del problema

1.2.1. Descripción de la problemática

La caña de azúcar es el cultivo azucarero más importante a nivel mundial con presencia en ciento dieciséis países tropicales y subtropicales, se destaca por ser la principal fuente de azúcar y la segunda fuente de etanol después del maíz (OECD-FAO, 2023). De entre esos países, los mayores productores son Brasil, India, China, Tailandia y Estados Unidos, quienes representan juntos más del 70 % de la producción mundial (FAO, 2023).

Para su proceso de producción se utiliza entre el 75 y 80 % de la planta de caña de azúcar principalmente los tallos, mientras que el restante es el residuo compuesto de tallos secos, cogollos, raíces, hojas verdes y secas. Este residuo no es aprovechable en la industria y una práctica muy extendida a nivel mundial para su eliminación es la quema (Singh *et al.*, 2018).

França *et al.* (2012) menciona que la quema de los residuos de caña de azúcar genera un aumento significativo de contaminantes hacia la atmósfera como óxidos de carbono (CO_2 y CO), óxidos de nitrógeno (NO_x), hidrocarburos (HC), material particulado fino ($PM_{2.5}$) y grueso (PM_{10}). Estos contaminantes tienen efectos adversos en la calidad del aire y representan un riesgo significativo para la salud pública y el medio ambiente.

En la ciudad de Araquara en Brasil, los eventos de quema de caña de azúcar en los períodos de cosecha 2002-2004 generaron una concentración de PM_{10} tres veces mayor y una cantidad de hidrocarburos aromáticos policíclicos (PAH) cuatro veces mayor en comparación a los períodos sin quema (De Andrade *et al.*, 2010). De manera similar, Junpen *et al.* (2020) reportaron que en la región norte de Tailandia durante el período de cosecha 2018-2019, en los días que se llevó a cabo la quema de caña de azúcar, la concentración mensual promedio de $PM_{2.5}$ superó por más del doble los valores promedios anuales límites recomendados por la Organización Mundial de la Salud (2021) de $10 \mu g/m^3$.

En Piraicaba, Brasil, durante el período de cosecha de caña 1997-1998 se asoció el incremento de la concentración de $PM_{2.5}$ y PM_{10} , de hasta $10.2 \mu g/m^3$ y $42.9 \mu g/m^3$ respectivamente, con un aumento del 21.4 % en el número de niños y del 31.03 % en ancianos diagnosticados.

dos con enfermedades respiratorias (Cançado *et al.*, 2006). Posteriormente, Arbex *et al.* (2007) relacionó directamente el aumento de las emisiones de material particulado durante los períodos de quema de caña de azúcar con un incremento del 11.6 % en las admisiones hospitalarias por asma en un hospital de la ciudad de Araraquara, Brasil.

Silveira *et al.* (2013) observaron que los hidrocarburos aromáticos policíclicos (PAH), resultantes de la quema, tienen una asociación directa con efectos genotóxicos en trabajadores de caña en la ciudad brasileña de Barretos, lo que sugiere un incremento en los casos de riesgo de cáncer en las personas afectadas por las emisiones. Pestana *et al.* (2017) también asoció el aumento de los niveles de NO_2 con un incremento del 1.12 % en las admisiones hospitalarias por enfermedades cardiovasculares en la región del oeste de São Paulo.

Nowell *et al.* (2022) estimaron una mortalidad de 2.5 muertes por año en veinte condados del sur de Florida, en Estados Unidos, debido a la emisión de $PM_{2.5}$ producida por los incendios de caña de azúcar, las causas de las muertes estaban relacionadas con vías respiratorias, enfermedades pulmonares y cáncer de pulmón. Asimismo, se ha comprobado que la quema de caña genera síntomas inflamatorios oculares y afecta la calidad de la mucosidad nasal (Matsuda *et al.*, 2020).

En la tesis de Salas (2010, p. 42) se menciona que la quema de caña de azúcar realizada antes de la cosecha afecta a la biodiversidad del suelo, provocando la pérdida de nitrógeno, la población de microorganismos del suelo y el material orgánico disponible en el mismo.

En el Perú esta actividad, ha generado numerosas denuncias por parte de la población de las regiones de Piura, La Libertad, Lambayeque, Ancash y Lima debido a su impacto negativo en la salud y la calidad del aire. Estas denuncias reportadas por diversos medios de comunicación se encuentran documentadas en el proyecto de ley N° 06243/2020-CR, compilado por el Congreso de la República (2020, pp. 12-25).

El Organismo de Evaluación y Fiscalización Ambiental (OEFA), como institución clave en la supervisión de la quema controlada de caña de azúcar, desempeña un rol fundamental en la fiscalización ambiental. Su función incluye la vigilancia, monitoreo y control de esta práctica mediante la evaluación de variables críticas como la distancia a centros poblados, la velocidad y dirección del viento, el horario de la quema, la duración y la extensión del área quemada por día

(OEFA, 2023). Estas medidas se enmarcan en la Ley N° 31949, promulgada el 23 de noviembre del 2023, que establece la progresiva implementación de la cosecha en verde como alternativa al uso de la quema, promoviendo la salud pública y la sostenibilidad ambiental (AGROAURORA S.A.C., 2020; AGROLMOS S.A., 2021).

Sin embargo, a pesar de los esfuerzos normativos y de fiscalización, la efectividad de esta supervisión ha mostrado importantes deficiencias operativas. Hasta la fecha, solo siete de las veintiún empresas azucareras administradas por el OEFA han actualizado sus Instrumentos de Gestión Ambiental (IGA), como el Programa de Adecuación y Manejo Ambiental (PAMA) o los Estudios de Impacto Ambiental (EIA), para incluir compromisos de transición hacia la cosecha en verde. Esta situación refleja una falta de actualización y adecuación de los IGAs, lo que obstaculiza el control efectivo de la quema y la implementación de prácticas más sostenibles (Casa Grande S.A.A., 2018, pp. 1530 - 1536).

Dado que la quema de caña de azúcar no siempre es reportada por los responsables y que la vigilancia se basa principalmente en denuncias ciudadanas, se hace evidente la necesidad de un modelo de detección automatizado que permita identificar la quema de manera remota. Este modelo facilitaría la identificación y extensión de las áreas de quema, fortaleciendo las acciones de fiscalización del OEFA y la actualización de los instrumentos de gestión ambiental de las empresas, contribuyendo así a la protección de la salud pública y la calidad del aire.

1.2.2. Formulación del problema

1.2.2.1. Pregunta general. ¿Es posible desarrollar un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar?

1.2.2.2. Preguntas específicas.

- ¿Cómo generar el conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado para detectar la quema de caña de azúcar?
- ¿Cuáles son los algoritmos que mejor se ensamblan al modelo para detectar la quema de caña de azúcar?
- ¿El modelo de aprendizaje ensamblado presenta un desempeño superior en comparación

con los algoritmos individuales?

1.3. Objetivos de la investigación

1.3.1. Objetivo general

Desarrollar un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar a fin de mejorar su fiscalización.

1.3.2. Objetivos específicos

- Generar el conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado para detectar la quema de caña de azúcar.
- Determinar los algoritmos que mejor se ensamblan al modelo para detectar la quema de caña de azúcar.
- Determinar mediante métricas de evaluación si el modelo de aprendizaje ensamblado presenta un desempeño superior en comparación con los algoritmos individuales.

1.4. Importancia y alcance de la investigación

1.4.1. Importancia

El modelo desarrollado en este estudio será fundamental para la detección precisa de áreas con quema de caña de azúcar, permitiendo al Organismo de Evaluación y Fiscalización Ambiental (OEFA) identificar y validar los predios agrícolas y las empresas responsables de estas prácticas. Esta herramienta mejorará la capacidad de fiscalización ambiental, facilitando la toma de medidas correctivas contra los infractores y promoviendo el cumplimiento de las normativas ambientales.

Además, el modelo permite un seguimiento casi en tiempo real de las quemas, ya que puede procesar nuevas imágenes tan pronto como se actualicen los productos satelitales. También será útil para analizar eventos pasados donde se tiene conocimiento de quemas reportadas, pero cuya extensión no ha sido estimada. Esto permitirá un análisis retroactivo de los impactos, mejorando la planificación de estrategias de mitigación.

1.4.2. Alcance

El modelo desarrollado en este estudio se aplica específicamente a las regiones de la costa norte y centro del Perú, utilizando una base de datos y los modelos preentrenados para estas áreas, disponibles públicamente en [Hugging Face](#).

Este enfoque regional permite una adaptación precisa a las condiciones locales, pero el modelo también tiene un gran potencial de escalabilidad, es por ello que el código fuente está disponible en [GitHub](#) para su replicación y futura aplicación en otros contextos geográficos y tipos de cultivos, lo que permite su uso más allá de las zonas estudiadas.

Este estudio también sienta las bases para futuras investigaciones en la detección de quemas y en la aplicación de tecnologías de aprendizaje automático en la agricultura y el monitoreo ambiental, ofreciendo un marco flexible y adaptable para diversas necesidades de gestión y sostenibilidad.

1.5. Limitaciones de la investigación

En este estudio se identificaron varias limitaciones que pueden haber influido en los resultados y en la aplicación del modelo propuesto. Estas limitaciones se dividen en metodológicas, operacionales y tecnológicas, y abarcan desde aspectos relacionados con la calidad y disponibilidad de los datos hasta los recursos computacionales necesarios para ejecutar modelos de aprendizaje automático.

1.5.1. Limitaciones metodológicas

- **Disponibilidad y calidad de los datos:** La precisión del modelo puede estar limitada por la calidad y resolución de las imágenes satelitales disponibles, así como por la cantidad de datos etiquetados correctamente. Condiciones atmosféricas, como la nubosidad, han reducido el número de imágenes útiles para el entrenamiento y la validación del modelo.
- **Generalización del modelo:** El modelo se desarrolló utilizando datos de regiones específicas (costa norte y centro del Perú), por lo que **actualmente** su capacidad para generalizar a otras regiones geográficas o tipos de cultivos puede ser limitada. Es posible que el modelo no funcione tan bien en áreas con diferentes características climáticas o de suelo.

- **Simplificaciones y supuestos en la modelización:** La metodología podría haber implicado ciertas simplificaciones o suposiciones sobre las características de las quemadas, las condiciones de vegetación o el comportamiento de los incendios, que no siempre representan la complejidad real en el campo principalmente en los meses donde realizan la cosecha de caña de azúcar.

1.5.2. Limitaciones operacionales

- **Tiempo y recursos humanos para el etiquetado:** El proceso de etiquetado manual de las áreas quemadas es intensivo en tiempo y requiere personal capacitado. La cantidad de personal disponible es limitada debido a la necesidad de conocimientos especializados tanto en el uso del software IRIS como en temas de incendios, lo que dificultó el etiquetado rápido y preciso de los datos necesarios para el entrenamiento del modelo. Esta limitación también afecta la capacidad de generar conjuntos de datos más grandes y diversos, esenciales para mejorar la precisión y la robustez del modelo.
- **Replicabilidad y escalabilidad del modelo:** Aunque la investigación no se vio limitado tecnológicamente en su ejecución original, la replicabilidad y escalabilidad del modelo en otros contextos requieren acceso a recursos computacionales significativos, como GPUs de alto rendimiento, y un conocimiento técnico avanzado en aprendizaje automático y procesamiento de imágenes satelitales. Estas condiciones son necesarias para asegurar que el modelo funcione de manera eficiente en diferentes configuraciones y áreas de aplicación.

1.5.3. Limitaciones en la validación

- **Falta de datos en campo detallados:** La validación del modelo se ve limitada por la falta de datos de campo extensos y detallados, así como la estimación de la extensión de las áreas quemadas, que coincidan con las fechas de las imágenes satelitales y las coordenadas precisas del evento de quema. Esta ausencia de datos puede afectar la precisión y fiabilidad del modelo.

CAPÍTULO II

EL MARCO TEÓRICO

2.1. Antecedentes de la investigación

2.1.1. Antecedentes internacionales

La capacidad de detectar áreas quemadas en cultivos de caña de azúcar es esencial para el monitoreo y fiscalización de la industria azucarera. En los últimos años, la teledetección ha emergido como una técnica altamente eficiente en la identificación de áreas afectadas por incendios.

La convergencia de esta técnica con la abundante disponibilidad de imágenes satelitales y los avances recientes en el campo de la Observación de la Tierra (EO) han ampliado significativamente su aplicación y eficacia.

Según Knopp *et al.* (2020), los algoritmos empleados para la delimitación de áreas quemadas siguen dos enfoques: aquellos basados en reglas y los que se apoyan en técnicas de aprendizaje automático.

Los algoritmos basados en reglas están diseñados para identificar variaciones en la respuesta espectral de áreas quemadas en comparación con su entorno. Los más efectivos combinan los dominios del infrarrojo cercano (NIR) y el infrarrojo de onda corta (SWIR) como el Índice Normalizado de Área Quemada (NBR) descrito en Roy *et al.* (2006) y aquellos nativos al producto satelital que incorporan otras regiones del espectro como el Índice de Área Quemada para Sentinel-2 (BAIS2) introducido por Filipponi *et al.* (2018), el cual considera el rango del borde rojo (Red Edge).

Recientemente, Farhadi *et al.* (2023) presentó un índice espectral novedoso denominado Índice de Detección de Área Quemada (BADI) que aprovecha al máximo el espectro disponible en Sentinel-2 (rojo, borde rojo, NIR y SWIR) para la detección de áreas quemadas. Este índice fue evaluado utilizando imágenes post-incendio para los eventos ocurridos entre el 3 y 13 de mayo de 2020 en Irán, específicamente en la parte suroeste de la provincia de Fars. El índice BADI propuesto registró altos valores en el coeficiente Kappa y una precisión general de 0.92 en comparación con índices espectrales comunes como BAIS2 (0.87 y 0.85) y NBR (0.85 y

0.78).

Por otro lado, el avance de los algoritmos apoyados por técnicas de aprendizaje automático han llevado a la popularidad enfoques como los árboles de decisión y las redes neuronales convolucionales, estos últimos dentro del ámbito del aprendizaje profundo o deep learning.

En una investigación realizada por Lee *et al.* (2022), se compararon tres técnicas representativas de este enfoque para la detección de áreas quemadas: Random Forest, LightGBM y U-Net. Se utilizaron imágenes satelitales Sentinel-2 en relación con dos eventos de incendios forestales ocurridos en 2019 y 2020 en Corea del Sur y en el escenario que contempla el uso de imágenes post-incendio se compararon los modelos con el mapa de clasificación binaria NBR. Los resultados revelaron que el modelo U-Net presentó los mejores resultados en todas las métricas de evaluación, tales como Precisión General (0.97), Precisión (0.89), Recall (0.87), F1-score (0.88) y Kappa (0.88). Mientras tanto, los algoritmos basados en árboles de decisión, es decir LightGBM y Random Forest, obtuvieron resultados similares en todas las métricas.

La elección entre el modelo LightGBM y Random Forest puede preferirse en función de diferentes requisitos. Por ejemplo, Sun *et al.* (2022) recomendaron el uso de LightGBM para la predicción de la susceptibilidad a incendios forestales debido a su mayor eficiencia en el entrenamiento, menor consumo de memoria, capacidad para procesamiento en paralelo y compatibilidad con características categóricas.

La segmentación semántica, a través de las redes neuronales convolucionales, ha emergido como la técnica más prometedora para la detección de áreas quemadas en imágenes satelitales superando a los algoritmos basados en reglas (Cambrin *et al.*, 2023) y técnicas de aprendizaje automático (Tonbul *et al.*, 2023); sin embargo, la disponibilidad limitada de datos etiquetados representa un desafío significativo. En base a ello, Colomba *et al.* (2022) y posteriormente Arnaudo *et al.* (2023) proporcionaron un amplio conjunto de datos que abarca 433 eventos de incendios a nivel global concentrados en Europa, Australia y el continente americano desde el 2017 hasta los primeros meses del 2023. Cada evento contiene una imagen post-incendio Sentinel-2 de 10 m de resolución espacial, una máscara de área quemada derivada del Servicio de Gestión de Emergencias de Copernicus (CEMS), cobertura del suelo obtenido del producto ESA World Cover y una máscara de nubes obtenido del modelo CloudSEN12 para el entrena-

miento de un modelo de segmentación semántica.

La elección de la arquitectura U-Net como el más óptimo en la segmentación semántica de áreas quemadas se demuestra en Al-Dabbagh e Ilyas (2023) quienes entrenaron catorce modelos de aprendizaje profundo basados en combinaciones de cinco arquitecturas (U-Net, U-Net++, Attention ResU-Net, LinkNet y DeepLabV3+) y cuatro codificadores (ResNet101, ResNet50, ResNet152 y MobileNet) para U-Net y LinkNet, utilizando un conjunto de datos manual para cinco provincias de Turquía durante el año 2021. Los resultados obtenidos muestran que la arquitectura U-Net con un codificador ResNet50 (0.98, 0.99), Attention ResU-Net (0.98, 0.98) y ResNet101 (0.97, 0.98) lograron los mejores resultados en las métricas de puntuación IoU y F1-score respectivamente.

La integración de diferentes técnicas de aprendizaje automático, conocido como aprendizaje ensamblado, se logra mediante metaalgoritmos como el bagging, boosting y stacking que utilizan métodos para ensamblar los modelos con el promedio (averaging), voto mayoritario (voting) o un metamodelo. Este enfoque ha sido ampliamente reconocido por su capacidad para mejorar la eficacia de modelos predictivos reduciendo la varianza (bagging), sesgo (boosting) y mejorando la predicción mediante stacking (Saini, 2017). En el contexto de la detección de áreas quemadas, los más populares son el bagging y boosting presentes en algoritmos como Random Forest (bagging) y LightGBM (Mienye y Sun, 2022).

El uso de modelos ensamblados que combinan algoritmos de Machine Learning y Deep Learning tiene un gran potencial para mejorar la precisión y robustez de las predicciones. aunque hay poca información disponible específicamente sobre la técnica de stacking en el contexto de la detección de áreas quemadas, es importante considerar estudios como el de Zhang *et al.* (2022) quienes indican que el stacking proporciona un marco efectivo para aprovechar las ventajas de diferentes algoritmos, incluidos aquellos de aprendizaje profundo.

En este contexto, Hu *et al.* (2024) desarrollaron un modelo de aprendizaje ensamblado para el mapeo de la progresión de áreas quemadas en tiempo casi real, combinando las ventajas de algoritmos de Machine Learning y Deep Learning. El área de estudio se centró en varios eventos de incendios forestales en California, utilizando el incendio “Sand Fire” ocurrido cerca de la ciudad de Santa Clarita como caso de prueba para el período de julio del 2016. Los da-

tos utilizados abarcan imágenes multiespectrales tomadas durante y después de estos eventos, específicamente de los satélites Sentinel-2 y Landsat-8.

La metodología propuesta combina el algoritmo HRNet y Support Vector Machine (SVM), optimizándose iterativamente a medida que llegan nuevas imágenes mediante un metaalgoritmo similar al boosting y la combinación lógica entre los modelos base para mejorar la precisión global. Con HRNet, se alcanzó un índice Kappa del 96.77 % en la predicción de perímetros y al integrar el SVM, que se optimiza iterativamente, la precisión en la clasificación de píxeles individuales mejoró, aumentando el índice Kappa de 62.55 % a 70.75 %.

Al combinar ambos modelos en un marco de aprendizaje ensamblado, que aprovecha las fortalezas de HRNet y SVM, se lograron resultados aún más refinados, alcanzando un índice Kappa de 85.19 % durante el evento de “Sand Fire” en California.

2.1.2. Antecedentes nacionales

En el Perú, el mapeo de áreas quemadas ha sido poco abordado con técnicas de aprendizaje automático, reduciéndose a investigaciones utilizando algoritmos basados en reglas.

En Chulucanas, Piura, Jimenez Dios y García Núñez (2020) determinaron la resiliencia de la cobertura vegetal después de incendios forestales donde se utilizaron imágenes Sentinel - 2 capturadas desde agosto de 2017 hasta 2019 para generar una serie temporal del Índice Normalizado de Quema (NBR) como indicador de la severidad del incendio.

El análisis se centró en la dinámica de la cobertura vegetal posterior al incendio en el Km 34 de la Comunidad Campesina José Ignacio Távara Pasapera, Distrito de Chulucanas. Los resultados mostraron que, conforme se recuperaba la biomasa vegetal y aumentaban las precipitaciones, se observan valores positivos del NBR, lo que sugiere una recuperación de la cobertura vegetal en los dos años de estudio.

Añamuro-Luque *et al.* (2020) estimaron el área quemada y la severidad ocasionada por el incendio ocurrido en agosto del 2016 en el distrito de Macari, Puno. Se utilizaron imágenes Landsat-8 pre y post incendio para calcular la diferencia del Índice Normalizado de Quema (ΔNBR). Se encontró que el área quemada total fue de 2,458.673 hectáreas y que la mayor parte del incendio tuvo una severidad baja, abarcando 1,385.145 hectáreas, seguida de una severidad

media que abarcó 967.568 hectáreas.

En otro estudio similar, Gonzáles Vásquez (2022) utilizó imágenes de satélite tomadas antes y después de un gran incendio en agosto de 2017 en el distrito de Coporaque, provincia de Espinar, departamento de Cusco. Se empleó la diferencia en el Índice Normalizado de Quema (ΔNBR) para estimar la severidad del incendio forestal. Los resultados indicaron que el área total quemada fue de 1,836.89 hectáreas, siendo predominantemente de Severidad Baja con 790.83 hectáreas (equivalentes al 43 % del área afectada), seguida de una Severidad Moderada-Baja con 733.34 hectáreas (representando un 40 % del área quemada).

En el contexto de la quema de caña de azúcar, Luque y Echeverría (2010) evaluaron los impactos ambientales y socioeconómicos generados por la quema de caña de azúcar en el área de influencia de la Empresa Agroindustrial Laredo S.A.A. Los materiales utilizados incluyeron equipos de monitoreo ambiental para la medición de gases contaminantes como monóxido de carbono (CO), óxidos de nitrógeno (NO_x) y dióxido de azufre (SO₂), además de estaciones de monitoreo de la calidad del aire para la recolección de datos de material particulado (PM₁₀). La metodología empleada consistió en el uso de mediciones ambientales in situ, con estaciones de monitoreo instaladas a barlovento y sotavento de la planta azucarera para registrar la concentración de contaminantes en el aire.

Los resultados revelaron que, aunque las concentraciones de contaminantes no superaron los Estándares Nacionales de Calidad Ambiental, las emisiones de partículas y cenizas generaron malestar en la población de la zona de influencia, afectando la limpieza de los hogares y ocasionando problemas de salud respiratoria.

2.2. Bases teóricas

2.2.1. Detección de áreas quemadas mediante imágenes satelitales

Chuvieco *et al.* (2019) mencionan que los métodos utilizados para generar productos de áreas quemadas a partir de imágenes satelitales son numerosos y pueden clasificarse según varios aspectos, como enfoques basados en el uso exclusivo de una escena posterior al incendio (enfoque mono-temporal) o la integración adicional de datos previos al incendio o series temporales (enfoque multi-temporal). Por lo general, el mapeo de las áreas quemadas sigue dos

enfoques: aquellos basados en reglas y los que se apoyan en técnicas de aprendizaje automático.

2.2.1.1. Imágenes satelitales Sentinel-2. Dentro de la amplia disponibilidad de imágenes satelitales, el programa Sentinel-2 proporciona imágenes globales en doce bandas espectrales con una resolución de 10 a 60 m (Tabla 1) y un tiempo de revisita de aproximadamente cinco días con dos satélites (S2A y S2B) en condiciones libres de nubes. Este conjunto de datos representa el archivo global de Sentinel-2, desde 2017 hasta el presente, procesado en L2A y utilizando Sen2Cor como algoritmo para la detección de nubes (Planetary Computer, 2023).

Tabla 1

Bandas espectrales de Sentinel-2 disponibles en Planetary Computer.

Nombre	Nombre común	Longitud de onda central	Distancia de la muestra terrestre	Descripción
B01	Coastal	0.443 μm	60 metros	Aerosol costero
B02	Blue	0.49 μm	10 metros	Azul visible
B03	Green	0.56 μm	10 metros	Verde visible
B04	Red	0.665 μm	10 metros	Rojo visible
B05	Red Edge 1	0.704 μm	20 metros	Clasificación de bordes rojos de vegetación 1
B06	Red Edge 2	0.74 μm	20 metros	Clasificación de bordes rojos de vegetación 2
B07	Red Edge 3	0.783 μm	20 metros	Clasificación de bordes rojos de vegetación 3
B08	NIR	0.842 μm	10 metros	Infrarrojo cercano
B8A	Red Edge 4	0.865 μm	20 metros	Clasificación de bordes rojos de vegetación 4
B09	-	0.945 μm	60 metros	Vapor de agua

Nombre	Nombre común	Longitud de onda central	Distancia de la muestra terrestre	Descripción
B11	SWIR 1	1.61 μm	20 metros	Infrarrojo de onda corta, clasificación de nieve/hielo/nube
B12	SWIR 2	2.19 μm	20 metros	Infrarrojo de onda corta, clasificación de nieve/hielo/nube

Nota. Elaboración propia.

2.2.1.2. Enfoques basados en reglas. Estos enfoques detectan variaciones en la respuesta espectral de las áreas quemadas en comparación con su entorno, especialmente en los dominios NIR y SWIR, y luego establecen umbrales en índices espectrales para mapear las áreas afectadas por incendios (Knopp *et al.*, 2020).

2.2.1.2.1. Índice de Quema Normalizado (NBR). El NBR es un índice utilizado para evaluar el impacto de los incendios forestales y cambios en la vegetación, siendo utilizado como un clasificador de áreas quemadas (Ongeri y Kenduiywo, 2020).

La vegetación saludable y no afectada por incendios tiene un valor positivo de NBR, mientras que las áreas quemadas se expresan en valores negativos. Su cálculo, a partir del producto de reflectividad del Sentinel 2A, se basa en el cociente normalizado (Ecuación 1) de la banda 4 (NIR, 0.842 μm) y la banda 8 (SWIR 2, 2.19 μm).

$$NBR = \frac{NIR - SWIR2}{NIR + SWIR2} \quad (1)$$

Además, se utiliza el ΔNBR para la detección de cambios, definido como la diferencia entre los valores del índice NBR antes y después del incendio (Ecuación 2). Los valores positivos indican un aumento en la vegetación, mientras que los valores negativos suelen indicar un área afectada por el incendio.

$$\Delta NBR = NBR_{pre-fire} - NBR_{post-fire} \quad (2)$$

2.2.1.2.2. Índice de Detección de Área Quemada (BADI). Propuesto por Farhadi *et al.* (2023), espectralmente hace uso de las bandas: infrarrojo de onda corta (SWIR), infrarrojo cercano (NIR), borde del rojo (Red Edge) y rojo (Red) aprovechando información indirecta como la actividad fotosintética y la concentración de humedad que existe en la cobertura vegetal (Ecuación 3c).

Los valores positivos suelen indicar presencia de áreas quemadas. Su cálculo, a partir del producto de reflectividad del Sentinel 2A, se basa en la banda 12 (SWIR 2, 2.19 μm), banda 11 (SWIR 1, 1.61 μm), banda 8 (NIR, 0.842 μm), banda 5 al 8A (Red Edge 1-4, 0.7 μm -0.865 μm) y banda 4 (Red, 0.665).

$$F1 = \frac{(SWIR2 + SWIR1) - (NIR + Red\ Edge\ 4)}{(SWIR2 + SWIR1) + (NIR + Red\ Edge\ 4)} \quad (3a)$$

$$F2 = \left(2 - \sqrt{\frac{Red\ Edge\ 2 \times Red\ Edge\ 3 \times (NIR + Red\ Edge\ 4)}{Red + Red\ Edge\ 1}} \right) \quad (3b)$$

$$BADI = F1 \times F2 \quad (3c)$$

Tomando en cuenta el enfoque multi-temporal (antes y después de la quema), se define el $\Delta BADI$ (Ecuación 4).

$$\Delta BADI = BADI_{pre-fire} - BADI_{post-fire} \quad (4)$$

2.2.1.3. Enfoques basados en aprendizaje automático. A diferencia de los enfoques basados en reglas, los enfoques de aprendizaje automático de clasificación supervisada, aprenden las características de los píxeles a partir de un conjunto de muestras etiquetadas. En el caso de la detección de áreas quemadas, se consideran dos categorías para esta tarea: áreas quemadas y áreas no quemadas (Al-Dabbagh e Ilyas, 2023).

En este marco de técnicas de aprendizaje automático, el bagging, boosting y stacking, conocidos como metaalgoritmos de aprendizaje ensamblado, representan enfoques potentes para

combinar diversas técnicas de aprendizaje automático en un único modelo predictivo (Saini, 2017).

2.2.1.3.1. Regresión logística binaria. Según Militino *et al.* (2024) la regresión logística es un método estadístico utilizado para predecir la probabilidad de que un evento ocurra ajustando los datos a una curva logística, en función de uno o más predictores o variables independientes. Para ello el modelo requiere varios supuestos: (a) una variable de respuesta binaria, (b) observaciones independientes, (c) ausencia de multicolinealidad entre las variables explicativas, (d) ausencia de valores atípicos extremos y (e) las variables predictoras estén relacionadas linealmente con las probabilidades logarítmicas de la variable de respuesta.

Park (2013) señala que la regresión logística calcula la probabilidad de que ocurra un evento en relación con la probabilidad de que no ocurra, explicando el impacto de las variables predictoras en términos de “odd”. Los “odds” representan la razón entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra. Si la probabilidad de que ocurra un evento es p , entonces la probabilidad de que no ocurra es $1 - p$, por lo que el “odd” se define como:

$$\text{Odd del evento} = \frac{p}{1 - p} \quad (5)$$

La probabilidad de que ocurra el evento de interés no se puede modelar directamente mediante una combinación lineal de las variables predictoras usando la fórmula $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$. Esto se debe a que el resultado de esta combinación lineal y no está restringido al rango de probabilidades $[0, 1]$.

La regresión logística resuelve este problema al tratar la combinación lineal de las variables predictoras como un logit (el logaritmo de las odds). Al utilizar logits, el modelo permite convertir el resultado en una probabilidad utilizando la función logística, asegurando que la salida final esté siempre dentro del al rango de probabilidades $[0, 1]$.

$$\text{logit}(y) = \log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (6)$$

A partir de esta Ecuación 6, se puede definir la probabilidad **de que ocurra un evento**

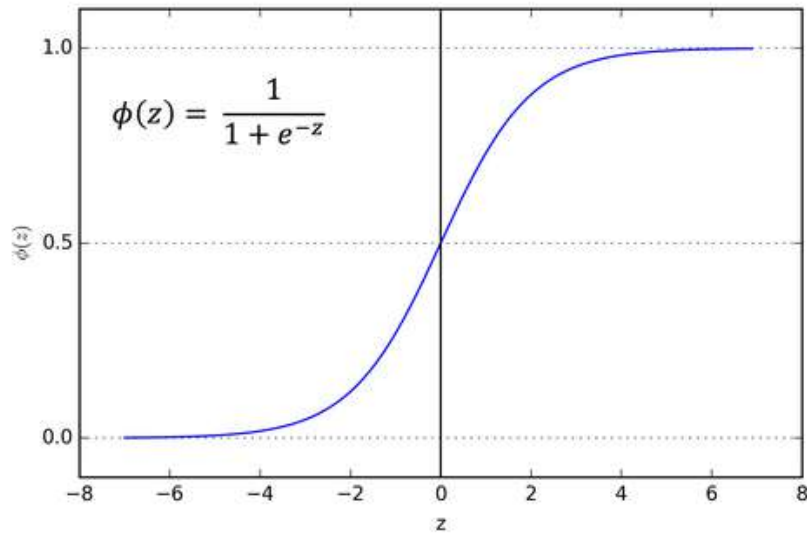
dado un conjunto de variables predictoras $X_1, X_2, \dots, X_n$ utilizando la función sigmoide.

$$P(Y = 1|X_1, X_2, \dots, X_n) = \frac{1}{1 + e^{-\text{logit}(y)}} \quad (7)$$

La función logística estándar o función sigmoide (Figura 1) transforma el logit (en este caso llamado z), en un rango entre 0 y 1, lo cual es interpretable como una probabilidad:

Figura 1

Función sigmoide utilizada en la regresión logística.



Nota. Extraído de Farrukh (2019).

Para Hosmer Jr *et al.* (2013) el método que busca estimar los valores óptimos de $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)$ es la función de máxima verosimilitud, el cual se define como el producto de las probabilidades en las que cada instancia pertenece a la clase correcta:

$$l(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (8)$$

Para facilitar los cálculos, se trabaja con el logaritmo de la función de verosimilitud, conocido como log-verosimilitud:

$$L(\beta) = \log l(\beta) = \sum_{i=1}^n y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (9)$$

Donde:

- $L(\beta)$ es la función de verosimilitud
- y_i es la etiqueta de clase de la observación i
- p_i es la probabilidad predicha de la observación i
- n es el número de observaciones

Los coeficientes β que maximizan la log-verosimilitud, se obtienen derivando $L(\beta)$ con respecto a cada coeficiente β_j e igualándolo a cero (Ecuación 10):

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^n (y_i - p_i) X_{ij} = 0 \quad \text{para } j = 0, 1, 2, \dots, m \quad (10)$$

Donde:

- X_{ij} es el valor de la variable j en la observación i
- m es el número de variables predictoras
- p_i es la probabilidad predicha de la observación i
- n es el número de observaciones
- y_i es la etiqueta de clase binaria de la observación i
- β_j es el coeficiente de la variable j
- $L(\beta)$ es la función de log-verosimilitud
- $\partial L(\beta)/\partial \beta_j$ es la derivada parcial de la función de verosimilitud con respecto al coeficiente β_j

2.2.1.3.2. Entropía Cruzada Binaria (BCE). Según Wali (2022) para abordar problemas de clasificación binaria se utiliza la entropía cruzada binaria (BCE) como función de pérdida, la cual mide la probabilidad de que un modelo asigne una clase correcta a la instancia, midiendo la disimilitud entre las dos distribuciones de probabilidad.

La probabilidad predicha está representada por un vector de probabilidades, donde la probabilidad predicha de la clase positiva se denota por p y la probabilidad predicha de la clase

negativa se denota por $1 - p$. La función de pérdida BCE se define la siguiente manera (Terven *et al.*, 2024):

$$L(y, p) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p) + (1 - y_i) \log(1 - p)] \quad (11)$$

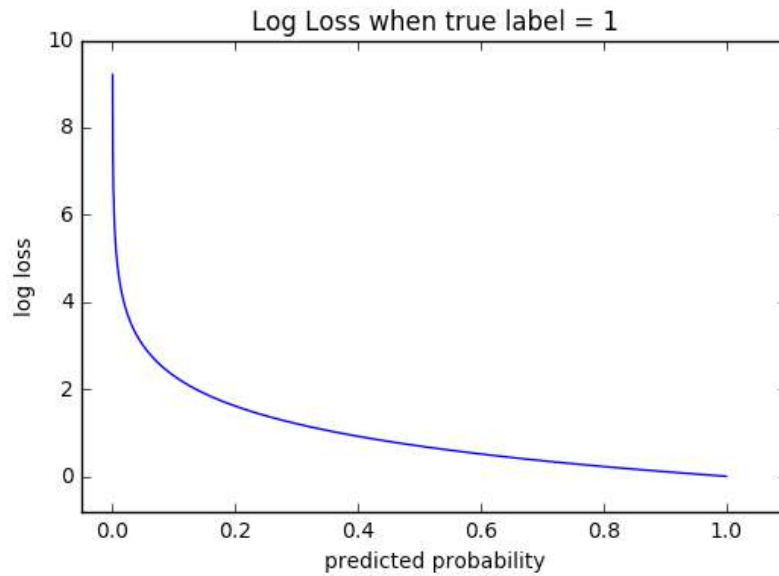
El cual puede ser intuitivamente dividido en:

$$\begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{if } y = 0 \end{cases}$$

Esta función de pérdida expresada como $-\log(x)$, como se visualiza en la Figura 2, convierte los valores en un rango de $[0, \infty)$.

Figura 2

Función de pérdida de entropía cruzada binaria cuando la clase es positiva ($y = 1$)



Nota. Extraído de Machine Learning Glossary (2017)

Este comportamiento es útil porque permite que la función de pérdida sea sensible a pequeñas diferencias en las probabilidades cercanas a 0 y 1, lo que es crucial en la clasificación. La entropía cruzada binaria penaliza fuertemente las predicciones incorrectas, acercando los valores de pérdida al infinito, mientras que las predicciones correctas reciben valores cercanos a 0.

2.2.1.3.3. Entropía Cruzada Binaria Ponderada (WBCE). La entropía cruzada binaria es sensible al problema de desequilibrio de clases que ocurre cuando el número de muestras de una clase es significativamente mayor que el de la otra. Para solucionar ello, se utiliza la entropía cruzada binaria ponderada para estos casos (Terven *et al.*, 2024).

En la entropía cruzada binaria ponderada (WBCE), se asigna un peso a cada muestra y la pérdida para cada muestra se calcula como:

$$L(y, p) = -\frac{1}{N} \sum_{i=1}^N w_i [y_i \log(p) + (1 - y_i) \log(1 - p)] \quad (12)$$

Donde w_i es el peso asignado a la muestra i , y es la etiqueta verdadera de clase binaria y p es la probabilidad predicha para la clase positiva. Terven *et al.* (2024) sostiene que al asignar un mayor peso a las muestras de clases subrepresentadas, se alienta al modelo a prestar más atención a estas muestras y se puede mejorar el rendimiento general del modelo.

2.2.1.3.4. Función de pérdida de Dice. Terven *et al.* (2024) define la función de pérdida de Dice como una medida de similitud entre la máscara de segmentación predicha y la máscara etiquetada para la clase positiva:

$$DSC = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (13)$$

Donde:

- X es la máscara de segmentación predicha
- Y es la máscara de segmentación verdadera
- $|X \cap Y|$ es el número de píxeles en los que las máscaras de segmentación predicha y verdadera coinciden
- $|X|$ es el número de píxeles en la máscara de segmentación predicha
- $|Y|$ es el número de píxeles en la máscara de segmentación verdadera

2.2.1.3.5. WCE - Dice Loss. Yali (2024) menciona que la combinación de múltiples funciones de pérdida es una estrategia viable para optimizar el rendimiento del modelo. En base a ello, formula la pérdida WCE-Dice que es una suma ponderada de las pérdidas de Entropía Cruzada Ponderada (WCE) y Dice (Ecuación 14).

$$L_{WCE-Dice} = (1 - \alpha) \times L_{WCE} + \alpha \times L_{Dice} \quad (14)$$

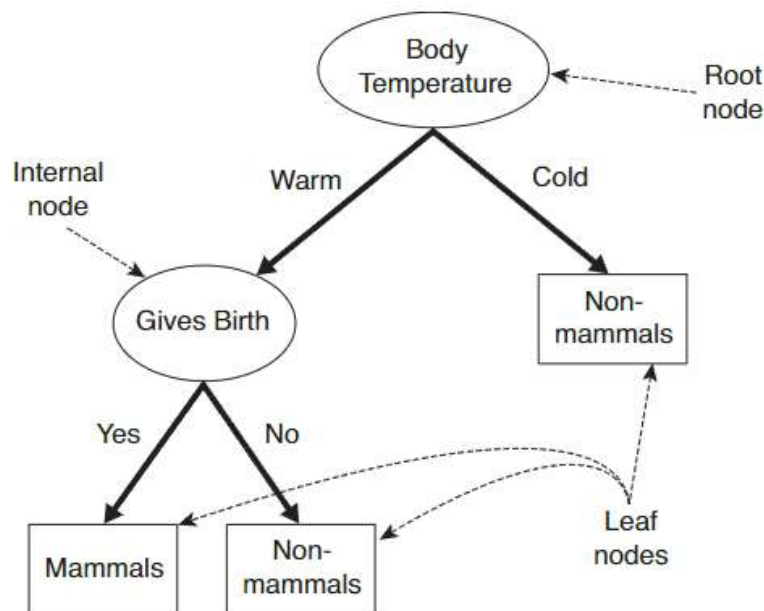
Donde α es el coeficiente de ponderación (entre 0 y 1) que controla la importancia relativa de cada función de pérdida.

2.2.1.3.6. Árbol de decisión. Steinbach *et al.* (2004) definen el árbol de decisión como una estructura de datos jerárquica que toma decisiones dividiendo repetitivamente un conjunto de datos en subconjuntos más pequeños en función de reglas o condiciones basadas en características o atributos específicos. Los atributos utilizados en un árbol de decisión pueden ser nominales (categorías sin orden), ordinales (categorías con un orden inherente) o continuos (valores numéricos dentro de un rango definido).

La estructura de un árbol de decisión se compone de nodos (Figura 3), donde cada nodo representa una condición de prueba de un atributo y cada rama representa el resultado de dicha condición. Existen tres tipos de nodos: los nodos raíz, que son el punto de inicio del árbol y contienen la primera condición de prueba; los nodos internos, que representan condiciones de prueba subsecuentes y dividen los datos en función de los atributos; y los nodos hoja, que no se dividen más y representan las clases o valores de salida finales.

Figura 3

Ejemplo de un árbol de decisión para un problema de clasificación de mamíferos.



Nota. Extraído de Steinbach *et al.* (2004).

Cada paso recursivo del proceso de construcción del árbol debe seleccionar una condición de prueba de atributos para dividir los registros en subconjuntos más pequeños hasta que todos los registros pertenezcan a la misma clase o ya no exista ninguna ganancia en seguir dividiéndose.

Para implementar este paso, el algoritmo debe proporcionar un método para especificar la condición de prueba para diferentes tipos de atributos, así como una medida objetiva para evaluar la bondad de cada condición de prueba.

En cuanto al tipo de atributos, Steinbach *et al.* (2004) señalan que las condiciones de prueba para los atributos binarios generan divisiones binarias, mientras que los atributos nominales, ordinales y continuos pueden generar tanto divisiones binarias como multidireccionales (multiway), respetando en el caso de los ordinales su orden inherente y utilizando pruebas de comparación o consultas de rango para los continuos.

Para el crecimiento un árbol de decisión, existen muchas medidas que se pueden utilizar para determinar la mejor manera de dividir los registros. Estas se definen en términos de la distribución de clases de los registros antes y después de la división, siendo $p(i|t)$ la proporción de registros que pertenecen a la clase i en el nodo t .

En el caso de clasificación binaria, la distribución de clases en cualquier nodo se puede escribir como (p_0, p_1) , donde $p_1 = 1 - p_0$.

Las medidas desarrolladas para seleccionar la mejor división suelen basarse en el grado de impureza de los nodos secundarios, cuanto menor sea el grado de impureza, más sesgada será la distribución de clases. Por ejemplo, un nodo con una distribución de clases $(0, 1)$ tiene cero impurezas, mientras que un nodo con una distribución de clases uniforme $(0.5, 0.5)$ tiene la mayor impureza porque las clases están igualmente repartidas. Algunos ejemplos de medidas de impurezas se describen en Steinbach *et al.* (2004):

$$\text{Impureza de Gini} = 1 - \sum_{i=1}^c p_i^2 \quad (15a)$$

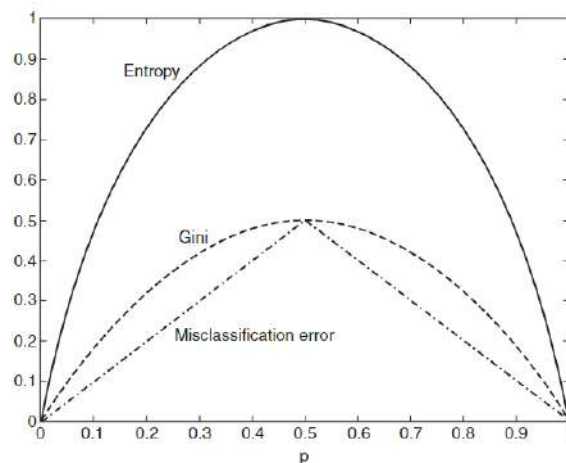
$$\text{Entropía} = - \sum_{i=1}^c p_i \log_2(p_i) \quad (15b)$$

$$\text{Error de clasificación} = 1 - \max(p_i) \quad (15c)$$

Donde c es el número de clases en el nodo y p_i es la proporción de registros que pertenecen a la clase i en el nodo específico. Los nodos con menor impureza presentan valores más bajos en estas medidas y, por lo tanto, se consideran mejores divisiones. La Figura 4 muestra cómo varían los valores de impureza según la proporción de clases para el índice de Gini, la entropía y el error de clasificación, donde las distribuciones más homogéneas ($p_i = 0.5$) resultan en los valores máximos (0.5, 1, y 0.5 respectivamente) para cada medida.

Figura 4

Comparación de las medidas de impureza de Gini, entropía y error de clasificación.



Nota. Extraído de Steinbach *et al.* (2004).

Para determinar el rendimiento de una condición de prueba, necesitamos comparar el grado de impurezas del nodo padre (antes de la división) con el grado de impurezas de los nodos hijos (después de la división). Cuanto mayor sea la diferencia, mejor será la condición de prueba.

La **ganancia** (Δ) es un criterio que se puede utilizar para determinar la bondad de una división:

$$\Delta = I(\text{parent}) - \sum_{i=1}^c \frac{N(v_i)}{N} \times I(v_i) \quad (16)$$

Donde $I(\text{parent})$ es la impureza del nodo padre, N es el número total de registros en el nodo padre, $N(v_i)$ es el número de registros en el nodo hijo v_i y $I(v_i)$ es la impureza del nodo hijo v_i . Los algoritmos de inducción de árboles de decisión suelen elegir una condición de prueba que maximiza la ganancia Δ . Dado que $I(\text{parent})$ es el mismo para todas las condiciones de prueba, maximizar la ganancia es equivalente a minimizar las medidas de impureza promedio ponderadas de los nodos hijo. Cuando la entropía es utilizado como medida de impureza, la ganancia (diferencia entre entropías) se le conoce también como **ganancia de información** (Steinbach *et al.*, 2004).

2.2.1.3.7. Gradient Boosting Decision Tree (GBDT). El GBDT, o árbol de decisión potenciado por gradiente, es un algoritmo de aprendizaje automático que utiliza boosting como técnica de ensamblado (Zhang *et al.*, 2022). Según Ke *et al.* (2017), este modelo entrena árboles de decisión en secuencia, ajustando en cada iteración los gradientes negativos, también conocidos como **errores residuales**.

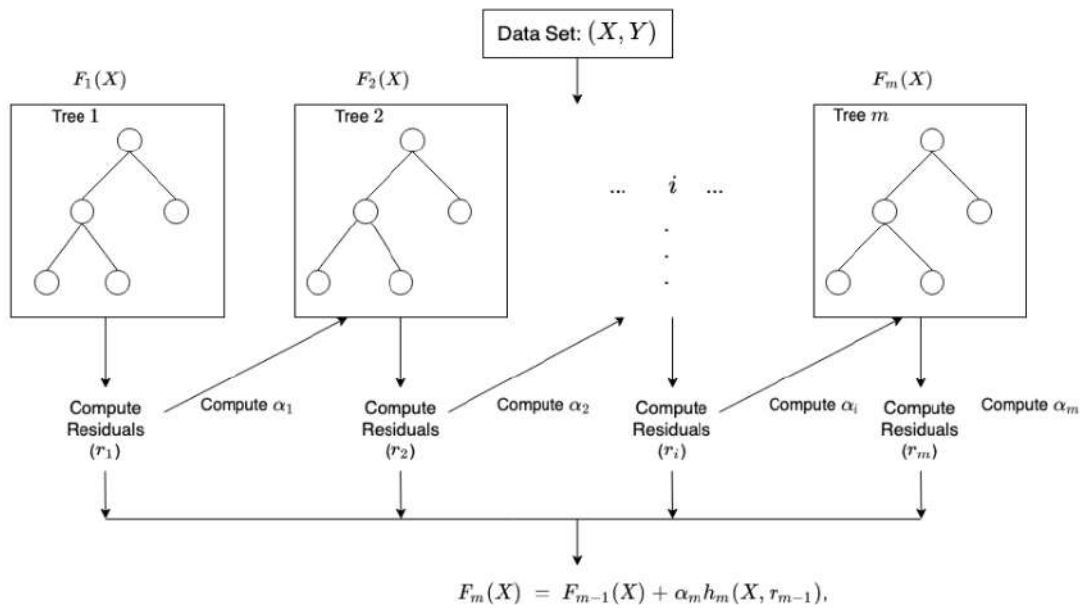
Dado el conjunto de entrenamiento $D = \{X, Y\} = \{x_j, y_j\}_{j=1}^N$, en donde $x_j \in R^c$ representa un vector predictivo de “c” características e $Y = \{y_j\}_{j=1}^N \in [0, 1]$ representa el conjunto de etiquetas de clases binarias para N instancias, la finalidad del GBDT es encontrar una función objetivo $F(x)$ que prediga la etiqueta Y para una nueva instancia de X .

El entrenamiento inicia con un modelo simple, considerada como la hoja inicial, que retorna un valor constante $F_0(X)$ que minimiza la función de pérdida en el conjunto de entrenamiento:

$$F_0(X) = \operatorname{argmin}_{\gamma} \sum_{j=1}^N L(y_j, \gamma) \quad (17)$$

Posteriormente, para cada árbol $i = 1, 2, \dots, m$ se calcula la función objetivo (donde m es el máximo número de árboles definido) como se observa en la Figura 5.

Figura 5
Ilustración de cómo trabaja el Gradient Boosting Tree.



Nota. Extraído de Amazon Web Services (2024).

En cada nuevo árbol i , se calcula el error residual r_i definido como la derivada de la función de pérdida con respecto a la función predictiva previa, esto se hace en el conjunto de entrenamiento o fracción dependiendo de la estrategia de muestreo utilizada (Ecuación 18):

$$r_i = - \left[\frac{\partial L(Y, F(X))}{\partial F(X)} \right]_{F(X)=F_{i-1}(X)} \quad (18)$$

En este contexto, $F(X) = F_{i-1}(X)$ es la predicción del anterior modelo y $L(Y, F(X))$ es la función de pérdida diferenciable.

Posteriormente, para predecir los **residuales** r_i se entrena un nuevo árbol de decisión (h_i), con un número máximo de hojas definido, que se agrega al modelo anterior $F_{i-1}(X)$. Luego, se realiza una optimización de búsqueda de línea para determinar la tasa de aprendizaje α_i (Ecuación 19):

$$\alpha_i h_i(X, r_{i-1}) = \operatorname{argmin}_{\alpha, h} \sum_{j=1}^N L(y_j, F_{i-1}(x_j) + \alpha h_i(x_j, r_{i-1})) \quad (19)$$

Finalmente, se actualiza el modelo $F(X)$ con el nuevo árbol $h_i(X, r_{i-1})$ y se repite el proceso hasta que se alcance el número máximo de árboles m .

Mienye y Sun (2022) definen a la función objetivo global, $F_m(x)$ como la suma ponderada de funciones:

$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}) \quad (20)$$

En donde α_m y r_{m-1} son la tasa de aprendizaje (entre 0.01 a 1) y los residuales calculados para el árbol $m-1$ respectivamente, siendo h_m el árbol de decisión que se entrena para predecir los residuales utilizando el conjunto de entrada X .

2.2.1.3.8. Algoritmo *LightGBM*. En el marco del Gradient Boosting Decision Tree, el algoritmo Light Gradient Boosting Machine (LightGBM) es una implementación eficiente del algoritmo de gradient boosting, desarrollada en 2017 por investigadores de Microsoft. Este algoritmo se utiliza en problemas de clasificación, ranking y otros problemas de aprendizaje automático (Sun *et al.*, 2022).

Ke *et al.* (2017) desarrollaron el algoritmo LightGBM incorporando dos métodos innovadores: Gradient-based One-Sided Sampling (GOSS) y Exclusive Feature Bundling (EFB), que reducen el número de instancias y de características respectivamente, mejorando la eficiencia y velocidad del algoritmo al manejar datos de alta dimensionalidad y gran volumen.

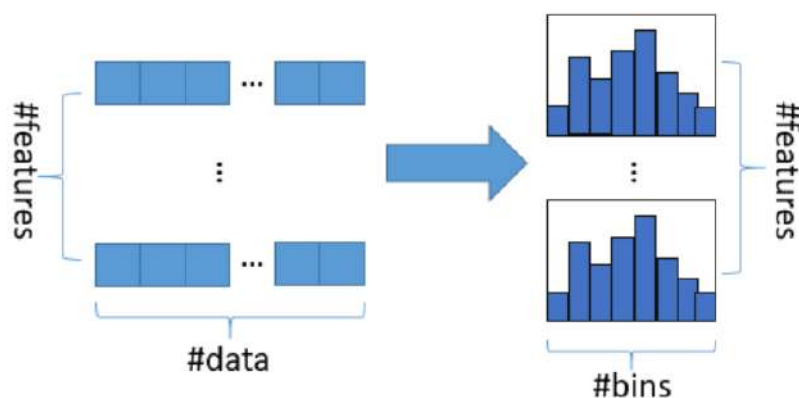
Asimismo, la construcción del árbol dentro del algoritmo se realiza mediante divisiones binarias asimétricas, con un número máximo de hojas predefinido, priorizando aquellas ramas que ofrecen la mayor ganancia de información (Ke *et al.*, 2017).

En el **Gradient-based One-Sided Sampling (GOSS)** se conservan todas las instancias con gradientes grandes y realiza muestreo aleatorio en los ejemplos con gradientes pequeños ya que las instancias con grandes gradientes contribuirán mejor a la **ganancia de información**. Al reducir el muestreo de las instancias de datos, es preferible mantener aquellas con gradientes grandes y eliminar aleatoriamente solo aquellas con gradientes pequeños.

Para ello, LightGBM parte del algoritmo basado en histogramas. En este enfoque, los valores de las características se discretizan en k intervalos y se construye un histograma con un ancho de k (Figura 6), durante el recorrido de los datos cada valor discretizado actúa como un índice en el histograma, lo que permite acumular estadísticas relevantes como la suma de gradientes. Una vez que se completa el primer recorrido de los intervalos, el histograma contiene las estadísticas necesarias, lo que permite un segundo recorrido para identificar el punto de división óptimo (Luo y Chen, 2020).

Figura 6

Algoritmo basado en histogramas.



Nota. Extraído de Luo y Chen (2020).

Esto se realiza para cada iteración de entrenamiento donde el modelo predice las etiquetas del conjunto de entrenamiento I y calcula el gradiente de la función de pérdida L para cada instancia j en I . Posteriormente, se identifican como muestras de gran gradiente los primeros $a \times 100\%$ de las muestras ordenadas por los valores absolutos de los gradientes de manera descendente, y las últimas $(1 - a) \times 100\%$ se denominan muestras de pequeño gradiente, en donde a y b son las tasas de muestreo definidas como parámetro (Figura 7).

De estas muestras de pequeño gradiente $(1 - a)$ se seleccionan de manera aleatoria una proporción de $b \times 100\%$ con un factor de peso se inicializa con 1 para todas las instancias y se actualiza en cada iteración para ajustar los pesos de las instancias con gradientes pequeños (Ke *et al.*, 2017).

Figura 7

Pseudocódigo del algoritmo Gradient-based One-Sided Sampling.

Algorithm 2: Gradient-based One-Side Sampling

Input: I : training data, d : iterations
Input: a : sampling ratio of large gradient data
Input: b : sampling ratio of small gradient data
Input: $loss$: loss function, L : weak learner
 $models \leftarrow \{\}$, $fact \leftarrow \frac{1-a}{b}$
 $topN \leftarrow a \times \text{len}(I)$, $randN \leftarrow b \times \text{len}(I)$
for $i = 1$ **to** d **do**
 $preds \leftarrow models.predict(I)$
 $g \leftarrow loss(I, preds)$, $w \leftarrow \{1, 1, \dots\}$
 $sorted \leftarrow \text{GetSortedIndices}(\text{abs}(g))$
 $topSet \leftarrow sorted[1:topN]$
 $randSet \leftarrow \text{RandomPick}(sorted[topN:\text{len}(I)], randN)$
 $usedSet \leftarrow topSet + randSet$
 $w[randSet] \times = fact$ $\triangleright$ Assign weight $fact$ to the small gradient data.
 $newModel \leftarrow L(I[usedSet], -g[usedSet], w[usedSet])$
 $models.append(newModel)$

Nota. Extraído de Ke *et al.* (2017).

Por otro lado, el algoritmo **Exclusive Feature Bundling (EFB)** tiene la capacidad de agrupar numerosas características exclusivas en un conjunto mucho más pequeño de características densas. Este proceso de agrupamiento reduce significativamente los cálculos innecesarios para los valores nulos.

De esta manera, la complejidad de construcción del histograma cambia de $O(\text{datos} \times \text{características})$ a $O(\text{datos} \times \text{grupo})$, siendo $\# \text{ grupo} \ll \# \text{ característica}$. Para ello, lo primero que

debe determinarse es que características deben agruparse juntas a través del método **Greedy Bundling** y lo segundo es unir las características con el método **Merge Exclusive Features** como se muestra en forma de pseudocódigo en la Figura 8 (Ke *et al.*, 2017).

Figura 8

Pseudocódigo del método Exclusive Feature Bundling.

Algorithm 3: Greedy Bundling Input: F: features, K: max conflict count Construct graph G $searchOrder \leftarrow G.sortByDegree()$ $bundles \leftarrow \{\}$, $bundlesConflict \leftarrow \{\}$ for i in $searchOrder$ do $needNew \leftarrow True$ for $j = 1$ to $len(bundles)$ do $cnt \leftarrow ConflictCnt(bundles[j], F[i])$ if $cnt + bundlesConflict[i] \leq K$ then $bundles[j].add(F[i])$, $needNew \leftarrow False$ break if $needNew$ then Add $F[i]$ as a new bundle to $bundles$ Output: $bundles$	Algorithm 4: Merge Exclusive Features Input: $numData$: number of data Input: F: One bundle of exclusive features $binRanges \leftarrow \{0\}$, $totalBin \leftarrow 0$ for f in F do $totalBin += f.numBin$ $binRanges.append(totalBin)$ $newBin \leftarrow new\ Bin(numData)$ for $i = 1$ to $numData$ do $newBin[i] \leftarrow 0$ for $j = 1$ to $len(F)$ do if $F[j].bin[i] \neq 0$ then $newBin[i] \leftarrow F[j].bin[i] + binRanges[j]$ Output: $newBin$, $binRanges$
--	---

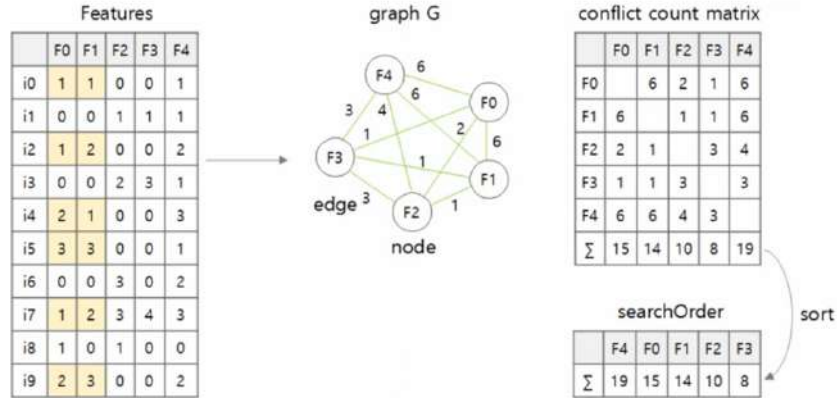
Nota. Extraído de Ke *et al.* (2017).

El método denominado **Greedy Bundling** tiene como objetivo agrupar las características en pares capturando la mayor cantidad de instancias exclusivas, es decir aquellas instancias donde uno de los valores es cero y el otro no (meanxai, 2023)

En la Figura 9 se observa cómo el método comienza con un conjunto de características $F_0, F_1, \dots, F_4$ y un conjunto de instancias $i_0, i_1, \dots, i_9$. Al realizar el grafo correspondiente G se capturan las instancias que generan conflicto, es decir que no son exclusivas, para cada par de características.

Luego, se suma el número de conflictos entre cada par de características y se ordena de manera descendente ($searchOrder$). Después, se inicia con la primera característica en “ $searchOrder$ ” y se procede a agrupar las características bajo la condición de que el número máximo de conflictos entre características sea menor o igual a K . En este caso, con $K = 1$ se obtienen tres grupos o bundles: F_4, F_{0-3}, F_{1-2} .

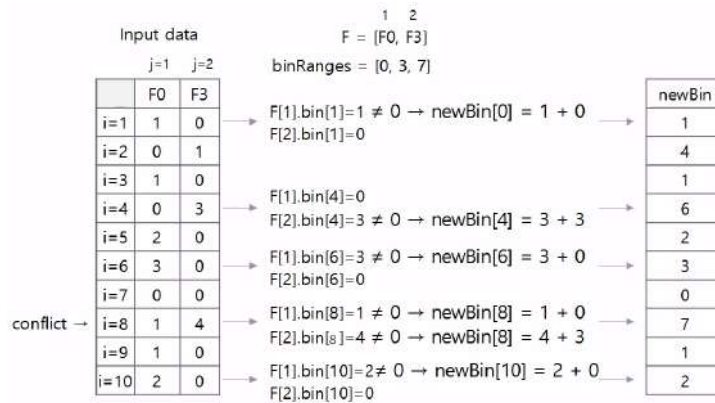
Figura 9
Ejemplo de cómo trabaja el Greedy Bundling - EFB.



Nota. Extraído de meanxai (2023).

En meanxai (2023), utilizando el método **Merge Exclusive Features** (Figura 10) se combinan los grupos en una única característica priorizando las instancias. Cuando la instancia de la característica no es igual a cero, $F \times bin \neq 0$, se suma el valor de la instancia en la característica con el correspondiente *binRanges* obtenido del acumulado del rango de los máximos valores de cada característica.

Figura 10
Ejemplo de cómo trabaja el Merge Exclusive Features.



Nota. Extraído de meanxai (2023).

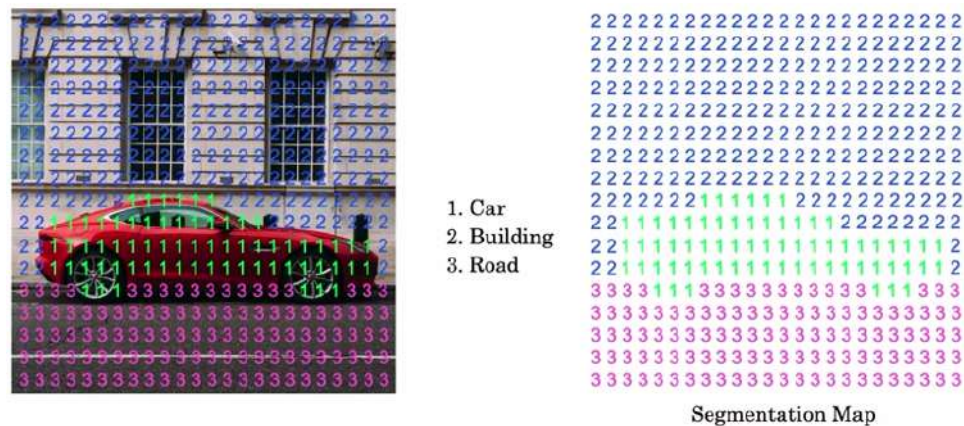
2.2.2. Segmentación Semántica

La segmentación semántica desempeña un papel fundamental en una amplia variedad de aplicaciones de visión por computadora, proporcionando información clave para la comprensión global de una imagen. En términos concretos, su objetivo es etiquetar los píxeles de la

imagen con las clases semánticas correspondientes y proporcionar límites de las clases de objetos, facilitando la comprensión de las apariencias de los objetos y las relaciones espaciales entre ellos. (Csurka *et al.*, 2023)

Figura 11

Ejemplo de segmentación semántica.



Nota. Extraído de Ng (2023).

2.2.2.1. Red Neuronal Convolutacional (CNN). Balas *et al.* (2020) lo definen como un tipo de red neuronal artificial (ANN) con una arquitectura de retroalimentación profunda que le permite aprender diversas características de los objetos, especialmente características espaciales. Las capas iniciales de un modelo CNN aprenden y extraen las características de alto nivel con una abstracción más baja y las capas más profundas aprenden y extraen las características de bajo nivel con mayor abstracción.

Katsaros *et al.* (2021) menciona que una CNN es apropiada para procesar entradas tipo cuadrícula como imágenes, videos y series de tiempo. Asimismo, su menor número de parámetros entrenables en comparación con una red neuronal artificial (ANN) convencional ayuda a reducir el riesgo de sobreajuste y mejora la capacidad de generalización del modelo (Balas *et al.*, 2020).

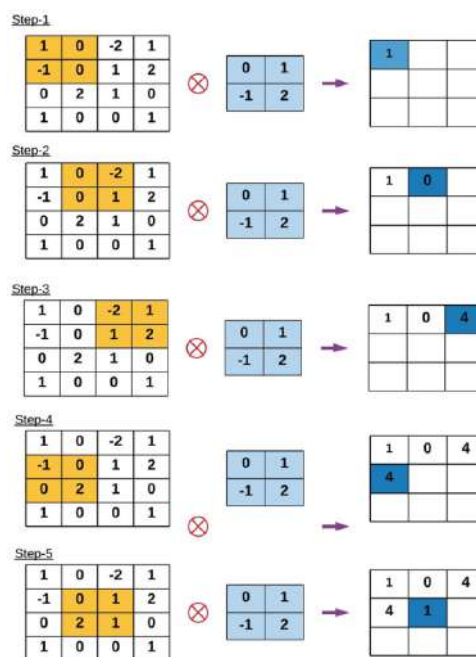
2.2.2.2. Arquitectura U-Net. Según Al-Dabbagh e Ilyas (2023), la arquitectura U-Net está diseñada específicamente para la tarea de segmentación semántica una forma de U y que consta de rutas de contracción (codificador) en el lado izquierdo y rutas expansivas, siendo un ejemplo típico de un modelo que consta de capas de convolución y agrupación.

2.2.2.2.1. Capa convolucional. La capa convolucional es considerada la parte principal de la CNN, contiene un conjunto de núcleos, también llamados filtros o kernel que convolucionan con la imagen de entrada para generar como resultado un mapa que abstrae las características relevantes de la imagen. Un núcleo puede describirse como una cuadrícula de valores o pesos que pueden ser aleatorios o inicializados.

Una operación convolucional se define como el producto punto entre el núcleo y la imagen de entrada, donde se multiplican los valores correspondientes de ambos y se suman todos los productos resultantes para generar un valor escalar en el mapa de características (Figura 12). Este proceso se realiza de manera sistemática a medida que el núcleo se desliza sobre la imagen de entrada en pasos definidos, conocidos como strides.

Figura 12

Ilustración de los primeros cinco pasos en una operación de convolución para una imagen de entrada 4x4x1 y un núcleo de 2x2.



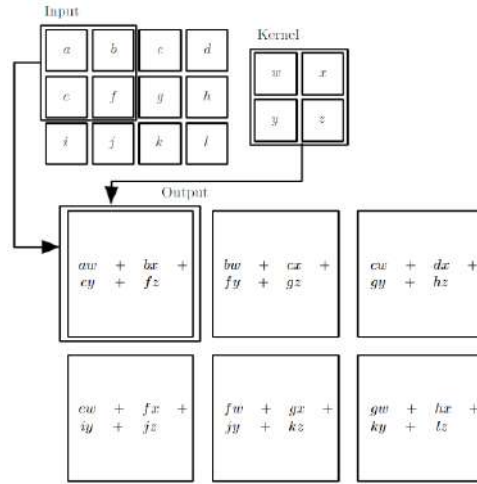
Nota. Extraído de Balas *et al.* (2020).

En la Figura 13, Goodfellow *et al.* (2016) muestran una versión generalizada de la operación de convolución en donde el núcleo (Kernel) se desliza sobre la imagen de entrada para generar un mapa de características. El resultado en cada pixel del mapa de características se obtiene mediante el producto punto, es decir la suma de los productos de los elementos correspondientes a la región de la imagen de entrada (Input) cubierta por el núcleo y el núcleo

mismo.

Figura 13

Mapa de característica resultante después de una operación de convolución 2×2 .



Nota. Extraído de Goodfellow *et al.* (2016).

2.2.2.2.2. Padding y Stride. El relleno o padding es crucial para preservar la información en los bordes de la imagen de entrada durante la convolución. Sin el uso de relleno, las características situadas en los bordes de la imagen se pueden perder rápidamente a medida que el núcleo de convolución no puede abarcar completamente estas regiones.

El relleno se utiliza para aumentar el tamaño de la imagen de entrada, lo que a su vez incrementa el tamaño del mapa de características de salida. Esto permite que todas las características, incluidas las de los bordes, se consideren en el proceso de convolución, manteniendo así más información relevante de la imagen original.

Por otro lado, las zancadas, pasos o stride son la cantidad de píxeles que se mueve el núcleo en cada paso de la convolución. La fórmula generalizada para calcular el tamaño del mapa de características de salida utilizando **padding y stride** se detalla en Balas *et al.* (2020):

$$h' = \left\lfloor \frac{h + 2p - f}{s} \right\rfloor + 1, \quad (21a)$$

$$w' = \left\lfloor \frac{w + 2p - f}{s} \right\rfloor + 1. \quad (21b)$$

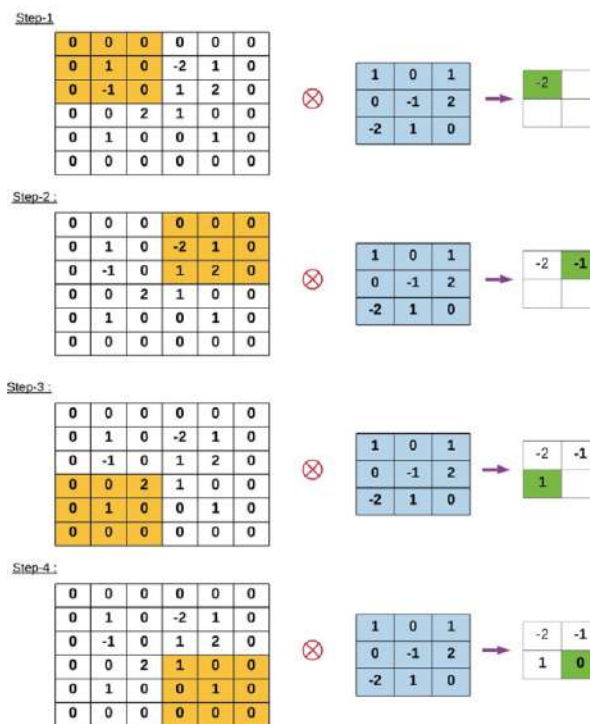
Donde h' y w' denotan la altura y ancho del mapa de características de salida, h y w

denotan la altura y ancho de la imagen de entrada, f es el tamaño del filtro, p es el relleno de la operación de convolución y s la zancada en la operación de convolución.

Se tiene el siguiente ejemplo para una operación de convolución utilizando un padding de 1 (relleno de ceros) y un stride de 3 (Figura 14); siguiendo con los cálculos de la Ecuación 21 se tiene un mapa de características de salida con dimensiones de 2×2 .

Figura 14

Ejemplo de una operación de convolución utilizando un padding de 1 y un stride de 3.



Nota. Extraído de Balas *et al.* (2020).

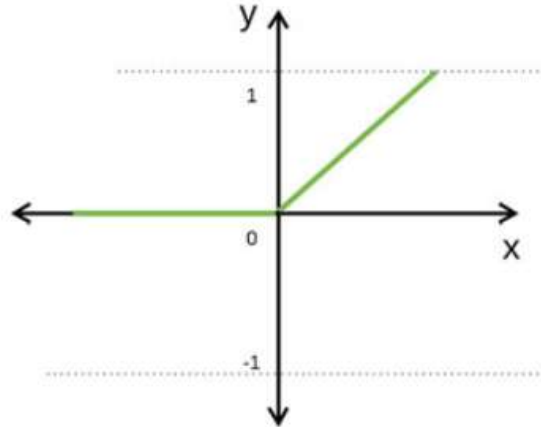
2.2.2.2.3. Funciones de activación. En la arquitectura CNN, después de cada capa con parámetros, se utilizan capas de activación no lineal. El comportamiento no lineal de esas capas permite que el modelo CNN aprenda cosas más complejas y logre asignar las entradas a las salidas de forma no lineal. Balas *et al.* (2020) mencionan que las funciones de activación más utilizada en redes neuronales profundas es la función ReLU (Rectified Linear Unit) (Ecuación 22).

$$f(x) = \max(0, x) \quad (22)$$

La función ReLU (Figura 15) activa solo los valores positivos y deja los negativos como cero, lo que introduce la no linealidad necesaria y acelera la convergencia del entrenamiento.

Figura 15

Función de activación ReLU.



Nota. Extraído de Balas *et al.* (2020).

2.2.2.2.4. Capa de agrupación. La capa de agrupación o pooling layer se utiliza para submuestrear los mapas de características producidos después de operaciones de convolución, es decir toma un mapa de características de entrada y reduce su tamaño conservando las características más dominantes en cada stride del grupo.

Existen diferentes tipos de técnicas de agrupación: agrupación máxima, mínima, promedio, cerrada, etc. De entre ellas, el max pooling es el método más popular y utilizado.

La fórmula para encontrar el tamaño del mapa de características de salida después de la operación de agrupación es:

$$h' = \left\lfloor \frac{h - f}{s} \right\rfloor + 1, \quad (23a)$$

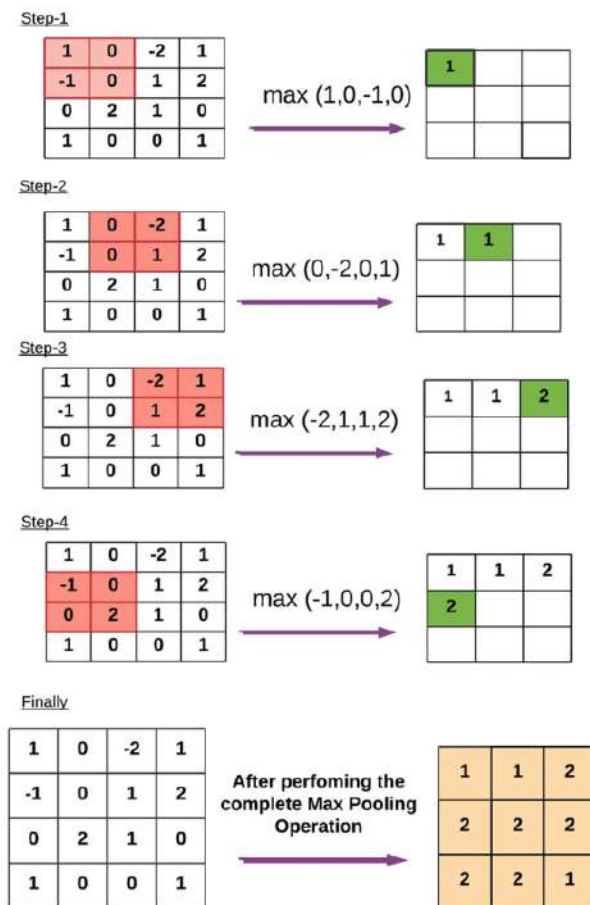
$$w' = \left\lfloor \frac{w - f}{s} \right\rfloor + 1. \quad (23b)$$

Donde h' y w' denotan la altura y ancho del mapa de características de salida, h y w denotan la altura y ancho de la imagen de entrada, f es el tamaño del filtro y s la zancada en la operación de convolución.

Balas *et al.* (2020) ejemplifican la operación de max pooling en una imagen de 4×4 con un filtro de 2×2 al cual da como resultado un mapa de características de dimensiones 3×3 (Figura 16).

Figura 16

Ejemplo de una operación de max pooling utilizando un stride de 1 y un tamaño de filtro de 2×2 .



Nota. Extraído de Balas *et al.* (2020).

2.2.2.2.5. Capa de Deconvolución. Somukesh (2023) define la capa de deconvolución como la capa que utiliza la operación de convolución transpuesta para aumentar la resolución espacial de una imagen.

A diferencia de la convolución estándar, la deconvolución expande el tamaño de la entrada al aplicar un filtro que redistribuye los valores del mapa de características de entrada a una salida más grande. Una operación de deconvolución se puede definir como un proceso donde un filtro o núcleo deconvolucional se aplica a cada valor de la entrada, esparciendo estos valores sobre la salida de acuerdo con los pesos del filtro. En cada paso, el núcleo multiplica los valores de entrada por sus pesos correspondientes y distribuye el resultado en la posición correspondiente de la salida.

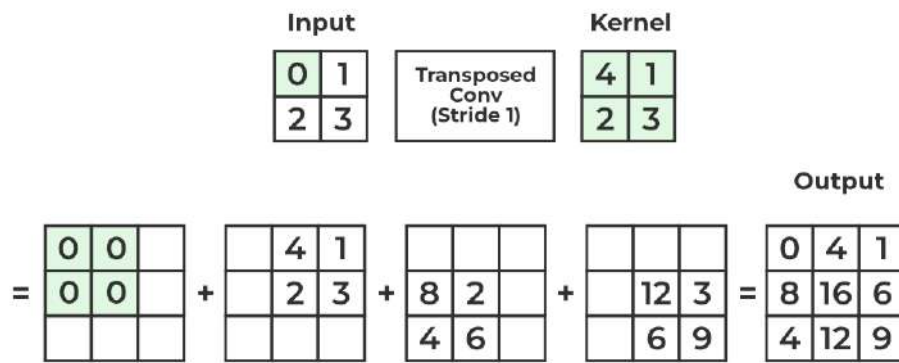
En la Figura 17, se ilustra cómo funciona la deconvolución con una entrada de 2×2 , un

filtro de 2×2 y un stride de 1. Para cada valor de la entrada, el filtro se aplica y se suma al mapa de salida en todas las posiciones afectadas.

Este proceso resulta en un mapa de características de salida de mayor tamaño, en este caso 3×3 , donde cada valor representa la acumulación de las multiplicaciones ponderadas de los valores de entrada y los pesos del filtro.

Figura 17

Ilustración de los pasos en una operación de deconvolución para una entrada de 2×2 y un núcleo de 2×2 con un stride de 1.



Nota. Extraído de Somukesh (2023).

Las dimensiones del mapa de características de salida se pueden calcular utilizando la siguiente fórmula descrita en Somukesh (2023):

$$h' = s(h - 1) + f - 2p + k \quad (24a)$$

$$w' = s(w - 1) + f - 2p + k \quad (24b)$$

Donde h' y w' denotan la altura y ancho del mapa de características de salida, h y w denotan la altura y ancho de la imagen de entrada, f es el tamaño del filtro, p es el relleno de la operación de deconvolución, s la zancada en la operación de convolución y k el tamaño del kernel.

2.2.2.2.6. Función de costo. Para Balas *et al.* (2020) la función de costo J , se define como el promedio de la función de pérdida de entropía cruzada sobre todas las muestras de

entrenamiento. Esta función de costo se expresa como:

$$J = \frac{1}{N} \sum_{i=1}^N L_i \quad (25)$$

Donde:

- J es la función de costo.
- L_i es la función de pérdida de la muestra i .
- N es el número total de muestras de entrenamiento.

2.2.2.2.7. Codificador y Decodificador. En la Figura 18 se ilustran los dos componentes principales de la arquitectura U-Net: el codificador (encoder) y el decodificador (decoder).

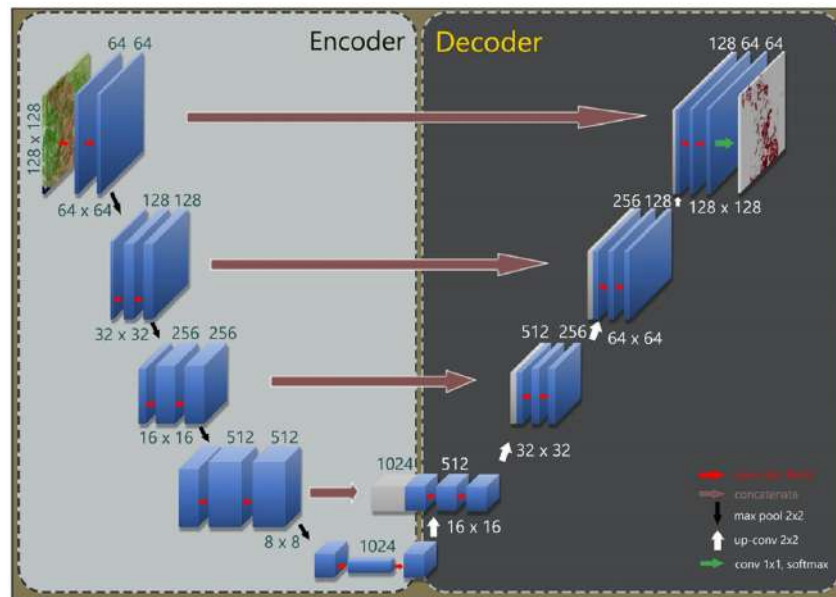
Según Ronneberger *et al.* (2015), la arquitectura de la U-Net está estructurada en dos trayectorias principales: una ruta de contracción, conocida como codificador, situada en el lado izquierdo y una ruta de expansión, llamada decodificador, ubicada en el lado derecho.

La ruta de contracción sigue la arquitectura clásica de una red convolucional, donde se aplican de forma repetida dos convoluciones de 3×3 (sin relleno), cada una seguida de una función de activación ReLU, y una operación de max-pooling de 2×2 con un paso de 2 para disminuir la resolución espacial.

En la ruta de expansión, cada paso comienza con un aumento de la resolución espacial del mapa de características mediante una convolución transpuesta de 2×2 (up-conv), que también reduce a la mitad el número de canales de características. Este mapa de características aumentado se concatena con el correspondiente mapa de características de la ruta de contracción, que ha sido recortado para alinearse espacialmente. Por último, se aplican dos convoluciones de 3×3 , cada una seguida de una ReLU, para obtener el mapa de características final de la segmentación.

Figura 18

Arquitectura U-Net de un modelo de red neuronal convolucional.



Nota. Extraído de Al-Dabbagh e Ilyas (2023).

2.2.2.3. Hiperparámetros del entrenamiento de una CNN. Wei (2024) define a los siguientes hiperparámetros de entrenamiento como los que controlan el proceso de entrenamiento de la red neuronal convolucional:

- Tasa de aprendizaje: Regula el tamaño del paso que el optimizador da al ajustar los pesos de la red en cada iteración. Una tasa de aprendizaje alta puede acelerar el entrenamiento, pero corre el riesgo de saltar sobre mínimos locales.
- Número de épocas: Indica cuántas veces el modelo verá el conjunto de entrenamiento completo. Más épocas pueden permitir un mejor ajuste, pero también pueden llevar a sobreajuste.
- Tamaño del lote: Es la cantidad de muestras del dataset en un lote (batch) que se utilizan en cada iteración del entrenamiento.
- Iteraciones: El número de actualizaciones de pesos que se realizan en una época. Se calcula como el número de muestras dividido por el tamaño del lote.
- Número de filtros: Cantidad de filtros que se aplican en cada capa convolucional.

- Número de capas: Cantidad de capas convolucionales y de agrupación que se utilizan en la red.

2.2.3. Métricas de evaluación

Los métodos de evaluación en la segmentación semántica tienen como objetivo calificar la similitud entre la segmentación predicha y la etiqueta real (ground truth). Aunque existen numerosas métricas de evaluación en la literatura, solo algunas se han demostrado apropiadas y se utilizan de manera estandarizada (Müller *et al.*, 2022).

Todas las métricas presentadas se basan en la computación de una matriz de confusión para una tarea de segmentación, la cual contiene el número de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN). En donde los rangos de valor de todas las métricas presentadas oscilan entre cero (peor) y uno (mejor).

2.2.3.1. Métricas basadas en F-measure. Las métricas basadas en F-measure, también conocidas como puntuaciones F, son una de las puntuaciones más extendidas para medir el rendimiento en visión por computadora. Se calculan a partir de la sensibilidad y precisión de una predicción, lo que evalúa la superposición entre la segmentación predicha y el ground truth. Aún así, al incluir la precisión, penalizan los falsos positivos, un factor común en conjuntos de datos altamente desequilibrados en clases.

Dos métricas populares utilizadas basadas en F-measure son el Índice de Intersección sobre Unión (IoU en Ecuación 26) y el coeficiente de similitud de Dice o F_1 en la Ecuación 27.

$$IoU = \frac{TP}{TP + FP + FN} \quad (26)$$

$$F_1 = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (27)$$

A pesar de que el IoU penaliza más la sub-segmentación y sobre-segmentación que el F_1 , ambos son métricas adecuadas y las más utilizadas en el campo de la observación de la Tierra (Al-Dabbagh e Ilyas, 2023; Arnaudo *et al.*, 2023; Cambrin *et al.*, 2023)

2.2.3.2. Sensibilidad y Especificidad. La sensibilidad y especificidad son métricas estándar establecidas para la evaluación del rendimiento.

La sensibilidad o Precision en la Ecuación 28 se centra en la capacidad de detección de verdaderos positivos, mientras que la especificidad o Recall en la Ecuación 29 evalúa la capacidad de identificar correctamente las clases verdaderas negativas.

$$Precision = \frac{TP}{TP + FP} \quad (28)$$

$$Recall = \frac{TP}{TP + FN} \quad (29)$$

2.2.3.3. Precisión o Índice Rand. La Precisión o Accuracy, también conocida como índice Rand, es una de las métricas más conocidas en estadística. Sin embargo, su uso está fuertemente desaconsejado en EO porque las imágenes de segmentación semántica suelen estar altamente desequilibradas en clases. Debido a la inclusión de verdaderos negativos (TN), esta métrica de precisión siempre dará puntuaciones muy altas de manera ilegítima (Müller *et al.*, 2022).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

2.2.3.4. Índice Kappa. El Índice Kappa de Cohen (κ) para clasificación binaria se define como:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Donde:

- P_o es la proporción de acuerdo observado entre los clasificadores.
- P_e es la proporción de acuerdo esperado por azar.

Footy (2020) menciona que para el caso simple de una matriz de confusión binaria, la proporción de acuerdo, P_o y P_e , se estiman de la siguiente forma:

- La proporción de acuerdo observado (P_o) se calcula como:

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

- La proporción de acuerdo esperado por azar (P_e) basado en los márgenes de la matriz de confusión se calcula como:

$$P_e = \frac{(TP + FP) \cdot (TP + FN) + (FN + TN) \cdot (FP + TN)}{(TP + FP + FN + TN)^2}$$

2.2.3.5. Curva ROC y Área Bajo la Curva (AUC). La curva ROC (Receiver Operating Characteristic) es una representación gráfica que muestra la tasa de verdaderos positivos (TVP) frente a la tasa de falsos positivos (TFP) para diferentes umbrales de discriminación.

El área bajo la curva ROC (AUC) es una métrica común para validar clasificadores de aprendizaje automático. Un AUC de 0.5 indica un clasificador aleatorio, mientras que valores cercanos a 1 indican un mejor rendimiento.

La fórmula del AUC se define como el área del trapecio según Powers (2020).

$$AUC = 1 - \left(\frac{fpr + tpr}{2} \right) \quad (31)$$

La tasa de verdaderos positivos (TVP) y la tasa de falsos positivos (TFP) se calculan como:

$$TVP = \frac{TP}{TP + FN} \quad (32)$$

$$TFP = \frac{FP}{FP + TN} \quad (33)$$

En donde:

- TP = Verdaderos positivos.
- TN = Verdaderos negativos.

- FP = Falsos positivos.
- FN = Falsos negativos

2.2.4. *Quema de caña de azúcar*

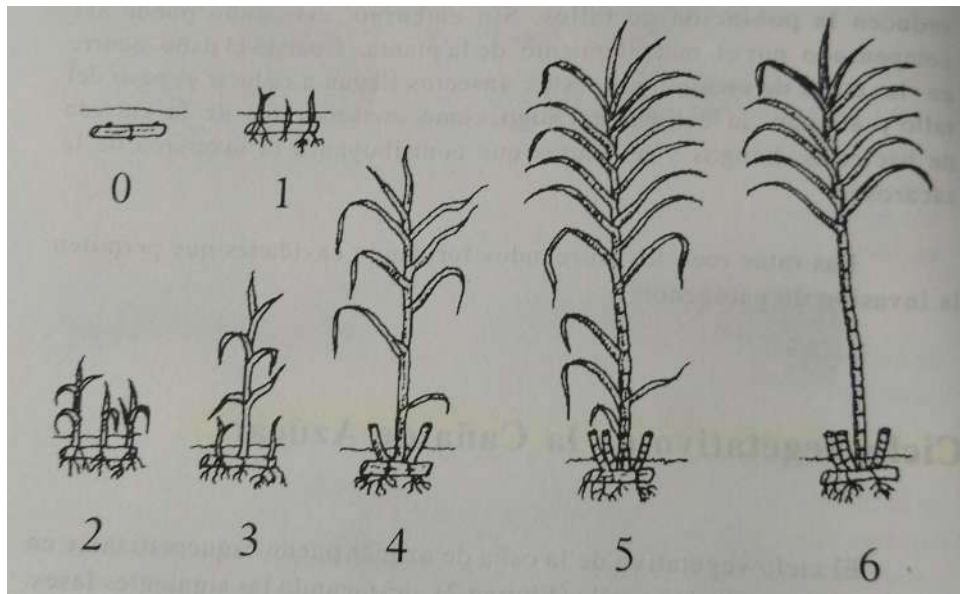
2.2.4.1. Fenología de la caña de azúcar. Según Helfgott (2016), la caña de azúcar, con nombre científico *Saccharum officinarum*, es una planta perenne de la familia de las gramíneas, originaria de Asia tropical y subtropical que se adapta bien a una variedad de suelos, siempre y cuando haya disponibilidad de agua, aunque prefiere las temperaturas tropicales y suelos de textura franca, con un pH cercano a 7.0, bien drenados y profundos.

El ciclo vegetativo de la caña de azúcar se desarrolla en las siguientes fases:

1. **Pre-brotamiento:** Desde el tapado de la semilla vegetativa (estaca) o desde la cosecha, hasta el inicio del brotamiento de las yemas.
2. **Brotamiento:** Desde el inicio del brotamiento de las yemas hasta el surgimiento de las raíces.
3. **Enraizamiento:** Durante esta fase, se desarrollan las raíces permanentes del brote.
4. **Macollaje:** Durante este período, se forman los vástagos o tallos de la caña. Esta etapa comienza entre 20 y 40 días después de la siembra o corte.
5. **Etapa inicial de crecimiento:** Durante esta fase, se forman los entrenudos y aumenta el sistema radicular permanente.
6. **Etapa de crecimiento:** Durante esta fase, el material foliar aumenta debido al desarrollo de los tallos, aproximadamente a razón de 1.25 cm por día. Además, comienza la concentración de azúcar en el tercio inferior del tallo, y las hojas inferiores comienzan a morir.
7. **Madurez:** En esta etapa, cesa el crecimiento y la sacarosa se concentra en el tallo. En esta etapa se realiza la cosecha aproximadamente entre 18 y 24 meses después de la siembra.

Figura 19

Etapas del desarrollo de la caña de azúcar.



Nota. Extraído de Helfgott (2016).

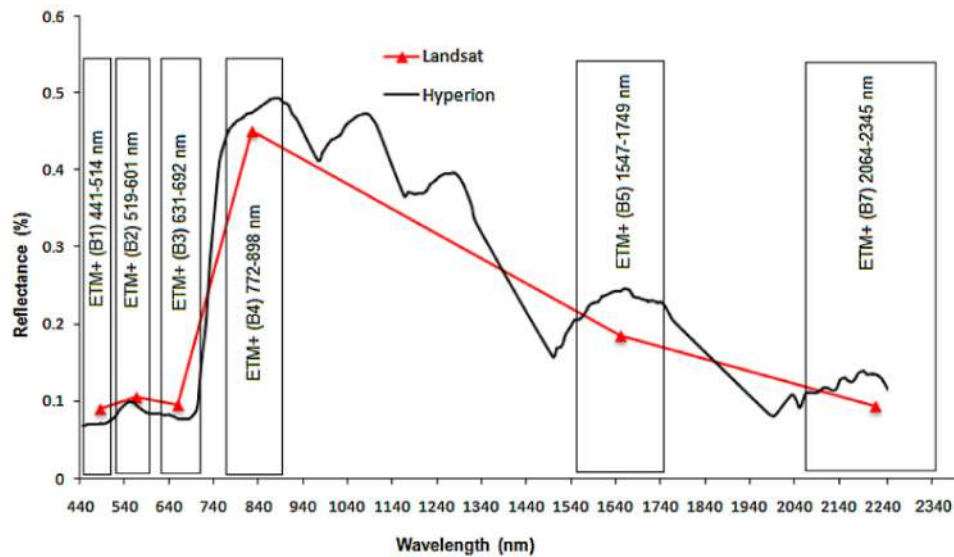
Cuando la caña planta alcanza su estado de maduración es cosechada aproximadamente a los veintiún meses después de la siembra, mientras que la caña-soca, la cual proviene de tallos de la caña planta cosechada, lo hace alrededor de los catorce meses después de su siembra.

2.2.4.2. Firma espectral de cultivos de caña de azúcar. Hamzeh *et al.* (2016) lograron ensamblar perfiles de firma espectral de campos de caña de azúcar al comparar datos de dos satélites EO diferentes: Landsat-7 ETM+ (multiespectral) e Hyperion (hiperespectral), ambos con una resolución espacial de 30 m. Los perfiles espectrales resultantes son típicos de la vegetación saludable ya que mostraron la mayor reflectancia en el infrarrojo cercano (NIR), 772 a 898 nm y valores más bajos en el infrarrojo de onda corta (SWIR), a 2064 a 2345 nm y en el visible ≤ 740 nm (Figura 20).

Este perfil espectral se obtuvo durante la etapa de madurez de la caña de azúcar en uno de los siete terrenos agrícolas e industriales de caña de azúcar ubicados en Juzestán, una provincia en el suroeste de Irán. Las imágenes, tomadas en septiembre de 2010, coinciden con la fase de madurez de la caña de azúcar en la región, cuyo período de crecimiento se extiende de abril a octubre (Shomeili, 2015).

Figura 20

Firma espectral representativa de la caña de azúcar en la etapa de madurez.

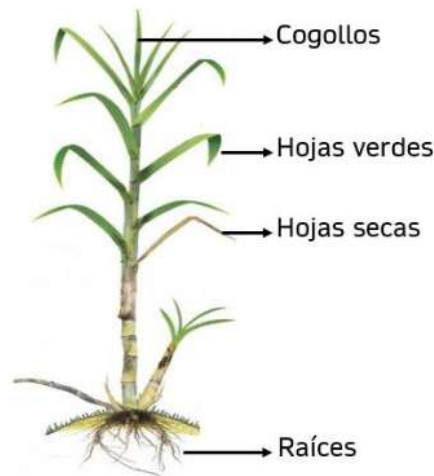


Nota. Extraído de Hamzeh *et al.* (2016).

2.2.4.3. Cosecha y quema de caña de azúcar. Pipicano y Restrepo (2022) mencionan que durante el proceso de cosecha del cultivo se tiene como residuos agrícolas a las hojas verdes y secas, raíces y el cogollo de la planta (Figura 21). En países como Brasil, Tailandia, Guatemala, Colombia, México y Costa Rica, la quema de estos residuos antes de la cosecha es una práctica común; en los Estados Unidos y Filipinas, los campos de caña de azúcar se queman antes o después de la cosecha. Sin embargo, en India y China la mayoría de los residuos de la caña de azúcar se queman en el campo después de la cosecha debido a la falta de técnicas de compostaje adecuadas (Junpen *et al.*, 2020; World Bank Group, 2018).

Figura 21

Partes de la caña de azúcar.



Nota. Extraído de Pipicano y Restrepo (2022).

En el caso del Perú, se tiene la particularidad de permitir la cosecha de caña durante todo el año, sin una temporada específica de zafra a excepción de los meses de otoño (abril/mayo) que es cuando la sacarosa en la caña disminuye y se aprovecha para realizar la Parada Anual, generalmente en esta época donde practica el agoste entre los meses de abril a setiembre que consiste en la suspensión del riego para reducir el crecimiento de las hojas verdes, contribuyendo así a su secado y permitiendo la acumulación máxima de sacarosa durante el período de madurez. (Helfgott, 2016).

En cuanto a la etapa de cosecha y procesamiento, ambas se llevan a cabo dentro de un lapso no mayor a 48 horas. En este lapso se realiza la quema de caña de azúcar que es una labor crucial, antes de la cosecha, con el propósito de eliminar las hojas verdes y secas facilitando de esta manera el corte y transporte de la caña. Para llevar a cabo este procedimiento consideran factores como la distancia y la accesibilidad a las áreas, las condiciones de los campos, la proximidad a otros cultivos, viviendas y lugares habitados (Casa Grande S.A.A., 2018, pp. 1530 - 1536).

El proceso de quema inicia en sentido contrario a la dirección del viento para prevenir la propagación del fuego hacia áreas no deseadas, ejecutándose de manera uniforme y completa en toda el área asignada. En situaciones donde la quema inicial no logra eliminar completamente todo el material no deseado, se realiza una segunda quema o requema empleando el mismo equipo, lanzallamas o palos de paja con kerosene, y personal involucrado en la tarea (Agroindustrias

San Jacinto S.A.A., 2014, pp. 434 - 436).

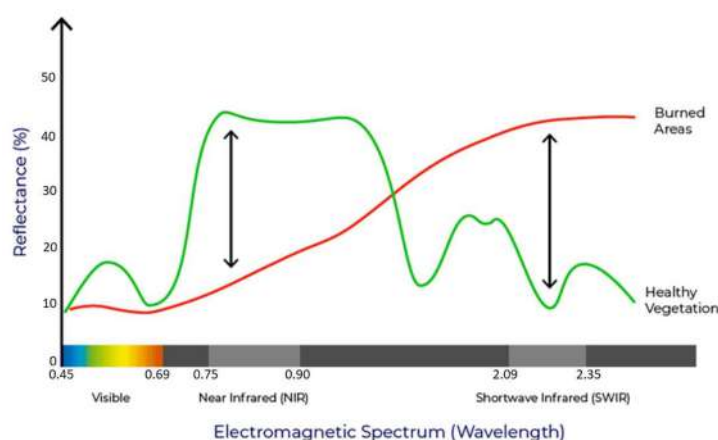
2.2.4.4. Firma espectral de áreas quemadas. Pereira *et al.* (1999) indican que el infrarrojo cercano (NIR), comprendido en los rangos de longitud de onda ($0.7 - 1.3 \mu\text{m}$), es la región espectral donde la señal de las cicatrices de incendios recientes es más fuerte y generalmente se considera la mejor región espectral para la detección y mapeo de áreas quemadas especialmente cuando las cargas de combustible previas al incendio son elevadas y la combustión produce grandes cantidades de carbón vegetal que se depositan en el suelo.

Dado que la vegetación verde es muy reflectante en el NIR, la quema suele causar disminuciones más o menos significativas de la reflectancia ya que tanto la vegetación verde y seca antes del incendio tienen reflectancias mayores que las quemaduras recientes, por lo que una caída en la reflectancia en el NIR es sistemático y sirve como un indicador confiable de áreas quemadas recientemente.

Alcaras *et al.* (2022) señalan que, tras un incendio, la reflectancia en el espectro del NIR disminuye considerablemente debido a la quema de la vegetación, mientras que la reflectancia en el espectro del SWIR aumenta como resultado de la pérdida de vegetación y, por ende, de agua (Figura 22). Asimismo, las regiones espectrales del Rojo y del Bordo Rojo son cruciales para la detección de áreas quemadas, ya que están relacionadas con una fuerte absorción de clorofila en las plantas.

Figura 22

Comparación de la respuesta espectral de la vegetación saludable y las áreas quemadas.



Nota. Extraído de Alcaras *et al.* (2022).

2.3. Definición de Términos

2.3.1. *Áreas con quema de caña de azúcar*

El término “área con quema de caña de azúcar” se refiere a la superficie de un campo de cultivo de caña de azúcar que ha sido afectada por la actividad de **quema de los residuos de caña de azúcar**. Esta expresión se abrevia a menudo como “área con quema” y puede usarse de manera intercambiable con “área quemada”, siempre que se trate de un campo de cultivo de caña de azúcar y que la quema se deba a esta práctica agrícola. El término abarca tanto la ubicación como la extensión total de la cicatriz de la quema, incluyendo todas las zonas donde el fuego ha dejado marcas visibles en el suelo, la vegetación o los residuos de caña de azúcar.

2.3.2. *Quema activa*

La quema activa o “quema en vivo” se refiere a la identificación de fuego en un campo de cultivo de caña de azúcar, mientras se eliminan sus residuos, capturado por el satélite en ese preciso momento. Esta situación, que se visualiza como un área que emite humo y calor, es la prueba fehaciente de una quema y puede ser detectada en vivo mediante imágenes satelitales térmicas o multiespectrales.

2.3.3. *Cosecha mecanizada en verde*

Es el proceso de recolección de caña de azúcar mediante el uso de maquinaria especializada sin la necesidad de quemar previamente los campos. Este método elimina la quema que generalmente se realiza para facilitar la recolección manual, presentando desafíos como la inversión en maquinaria adecuada, la gestión de los residuos agrícolas y la necesidad de ajustar los procesos de producción. Aunque la cosecha en verde ayuda a reducir el impacto ambiental, especialmente en áreas cercanas a poblaciones, puede resultar en menor productividad debido a la mayor presencia de materia extraña, lo que disminuye la eficiencia en la extracción de azúcar.

2.3.4. *Ingenio azucarero*

Un ingenio azucarero es una planta procesadora que convierte caña de azúcar en azúcar mediante procesos industriales como la extracción del jugo, purificación, evaporación, cristali-

zación y secado, y también produce subproductos como melaza, bagazo y bioetanol. Su enfoque principal es la transformación eficiente de la materia prima en productos derivados del azúcar, utilizando tecnología avanzada.

En el Perú, estos ingenios se encuentran principalmente en la costa norte y centro del país.

2.3.5. Fundo, Predio y Campo Azucarero

Un **fundo** azucarero es una propiedad rural de gran tamaño dedicada a la producción de caña de azúcar, que puede incluir varios predios, diversas operaciones agrícolas, infraestructuras, y diferentes tipos de cultivos de caña. Dentro de un fundo, un **predio** es una extensión de tierra que agrupa varios campos de cultivo bajo una gestión común, facilitando la planificación y el manejo integral de la producción. Un **campo**, en este contexto, es la unidad básica de producción agrícola: una parcela de tierra específica destinada al cultivo de caña de azúcar, donde se realizan actividades como la siembra, el riego y la cosecha.

2.3.6. Matriz de confusión en la segmentación binaria

En el contexto de la segmentación semántica binaria, considerando las dos clases “quema” y “no quema”, se tienen los siguientes términos que conforman la matriz de confusión:

- Verdadero positivo (TP): Píxeles identificados correctamente como “quema” y que corresponden a “quema”.
- Falso positivo (FP): Píxeles identificados incorrectamente como “quema” y que corresponden a “no quema”.
- Falso negativo (FN): Píxeles identificados incorrectamente como “no quema” y que corresponden a “quema”.
- Verdadero negativo (TN): Píxeles identificados correctamente como “no quema” y que corresponden a “no quema”.

2.3.7. Índice de Vegetación de Diferencia Normalizada (NDVI)

El NDVI (**Normalized Difference Vegetation Index**) es un índice utilizado para evaluar la salud de la vegetación. Se calcula utilizando las bandas de infrarrojo cercano (NIR) y rojo

(RED) de las imágenes satelitales y se define por la siguiente fórmula:

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

donde:

- *NIR* es la reflectancia en la banda del infrarrojo cercano.
- *RED* es la reflectancia en la banda roja.

El valor del NDVI varía entre -1 y 1. Los valores cercanos a 1 indican una vegetación densa y saludable, ya que la vegetación refleja más luz en el espectro del infrarrojo cercano y menos en el espectro rojo. Los valores cercanos a 0 indican poca o ninguna vegetación, mientras que los valores negativos (cercanos a -1) corresponden a superficies como agua, nieve, o nubes, que reflejan más luz en la banda roja que en el infrarrojo cercano.

2.3.8. Índice de Agua de Diferencia Normalizada (NDWI)

El NDWI (**Normalized Difference Water Index**) es un índice utilizado para identificar cuerpos de agua y contenido de humedad en las plantas. Se calcula utilizando las bandas de infrarrojo cercano (NIR) y del verde (GREEN) de las imágenes satelitales, y se define por la siguiente fórmula:

$$NDWI = \frac{(GREEN - NIR)}{(GREEN + NIR)}$$

donde:

- *GREEN* es la reflectancia en la banda verde.
- *NIR* es la reflectancia en la banda del infrarrojo cercano.

El valor del NDWI varía entre -1 y 1. Los valores positivos (cercanos a 1) indican la presencia de agua, ya que el agua absorbe mucho más la luz infrarroja cercana que la luz verde mientras que los valores negativos (cercanos a -1) sugieren áreas sin agua, como suelos desnudos o áreas urbanas, y valores cercanos a 0 pueden indicar áreas con vegetación moderada.

CAPÍTULO III

HIPÓTESIS Y VARIABLES

3.1. Hipótesis

3.1.1. Hipótesis general

El modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar permitirá resultados con un gran porcentaje de precisión.

3.1.2. Hipótesis específicas

- La correcta generación del conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado para detectar la quema de caña de azúcar optimizará la precisión del modelo.
- Los algoritmos de GBDT y U-Net serán los que mejor se ensamblan al modelo para detectar la quema de caña de azúcar.
- El modelo de aprendizaje ensamblado presentará un desempeño superior en comparación con los algoritmos individuales.

3.2. Variables

3.2.1. Variable independiente

Desarrollo de un modelo de aprendizaje ensamblado.

3.2.2. Variable dependiente

Detección de quema de caña de azúcar.

3.3. Operacionalización de variables

Tabla 2

Cuadro de Operacionalización de variables

Variables	Dimensión	Indicador	Unidad de medida	Instrumento
Variable Independiente				
Desarrollo de un modelo de aprendizaje ensamblado.	Modelo U-Net	Métricas de evaluación	Porcentaje	Librería Torch y PyTorchLightning
	Gradient Boosting Decision Tree	Métricas de evaluación	Porcentaje	Librería LightGBM en Python
	Regresión Logística	Métricas de evaluación	Porcentaje	Librería scikit-learn en Python
Variable Dependiente				
Detección de quema de caña de azúcar.	Ubicación de eventos de quema	Coordenadas geográficas	Grados decimales	Plataforma SIG (QGIS)
	Extensión de las áreas quemadas	Área quemada	Hectáreas	Plataforma de análisis de imágenes satelitales (R y Python)

Nota. Elaboración propia.

CAPÍTULO IV

MATERIALES Y MÉTODOS

4.1. Área de estudio (Unidad de análisis)

Se definió la unidad de análisis en la costa norte y centro del Perú, considerada como una zona altamente productiva para el cultivo de caña de azúcar debido a sus condiciones naturales que incluyen: temperatura con oscilaciones moderadas, disponibilidad constante de agua para riego y suelos fértiles de origen aluvial (Helfgott, 2016).

Figura 23

Ubicación de los principales ingenios azucareros del país que conforman el gremio PERUCANÑA.



Nota. El ingenio de Caña Brava es administrado por la empresa Agrícola del Chira S.A.

Elaboración propia con datos de **PERUCANÑA**.

La industria azucarera se concentra en cinco regiones: Piura, Lambayeque, La Libertad, Áncash y Lima (Figura 23). De estas regiones, La Libertad destaca en términos de superficie cultivada con un 47 % del total. Asimismo, las ocho empresas azucareras asociadas al gremio exportan más del 97 % del volumen comercializado a nivel nacional, según datos de PERUCAN (2022).

4.2. Diseño de investigación

4.2.1. Tipo de investigación

Es **básica**, ya que se busca producir nuevo conocimiento con respecto a la identificación y estimación precisa de las áreas de quema de caña de azúcar en la costa norte y centro del Perú mediante las imágenes satelitales.

4.2.2. Nivel de investigación

La investigación es de nivel **descriptiva** porque se busca especificar las características, propiedades y comportamiento de la quema de caña de azúcar mediante imágenes satelitales y en base a ello, determinar la ubicación y extensión de estos eventos en la costa norte y centro del Perú.

4.2.3. Enfoque de investigación

El enfoque es **cuantitativo**, ya que utiliza herramientas de análisis matemático y estadístico como parte de los algoritmos empleados para la detección de la quema de caña de azúcar en las imágenes satelitales.

4.2.4. Diseño de investigación

La investigación es de tipo **no experimental y longitudinal**.

El tipo de investigación adoptado es no experimental, dado que se emplearon imágenes satelitales para la detección de la quema de caña de azúcar. No se llevó a cabo ninguna manipulación de las variables, siendo preprocesadas las imágenes únicamente para generar las etiquetas correspondientes.

Además, se clasifica como un diseño longitudinal, puesto que se recolectaron imágenes

satelitales de la costa norte y centro del Perú a lo largo del periodo comprendido entre 2017 y 2022.

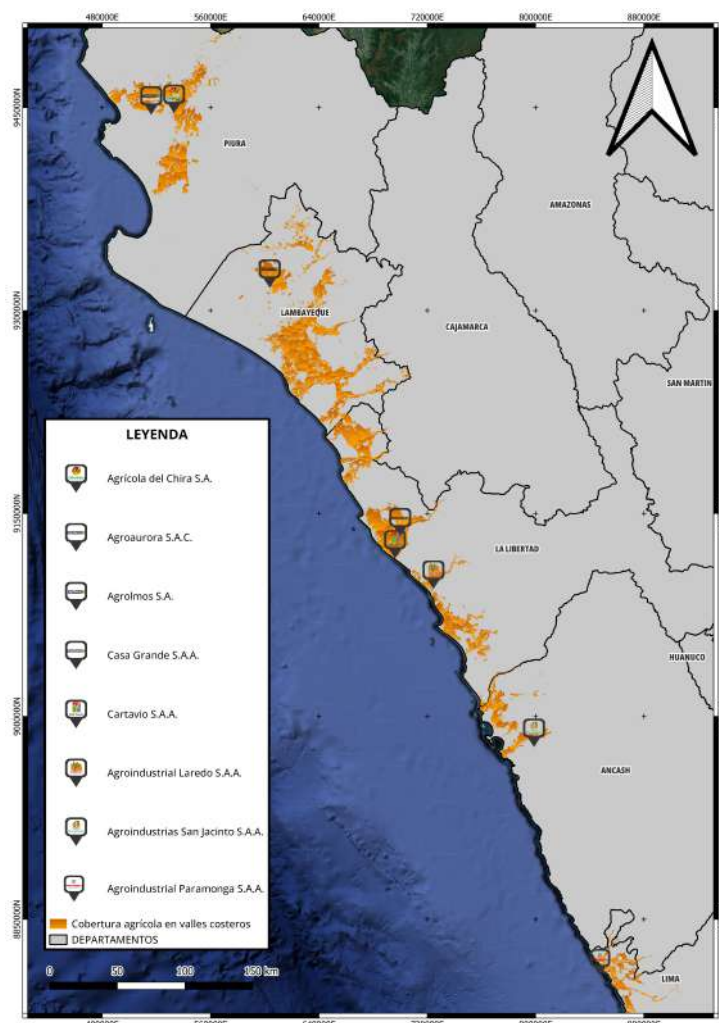
4.3. Población y muestra

4.3.1. Población

La población del presente estudio está conformada por **áreas de cultivo de caña de azúcar en la costa norte y centro del Perú**, desde el valle del Chira (Piura) hasta el valle Huaura (Lima). Se obtuvo a partir del área de superficie agrícola en los valles de la costa proporcionada por MIDAGRI y la ubicación referencial de los ingenios azucareros (Ver figura 24).

Figura 24

Áreas de cultivo de caña de azúcar.



Nota. Elaboración propia.

En la Tabla 3 se presentan los principales ingenios azucareros en la costa norte y centro del Perú, con su respectiva código administrado de la OEFA, ubicación y cuencas hidrográficas

donde se encuentran.

Tabla 3

Principales ingenios azucareros en la costa norte y centro del Perú.

Empresa azu- carrera	Código ad- ministrado OEFA	Departa- mento	Cuencas	Hidrográfi-	Coordenadas del ingenio (lat, lon)
Agroaurora S.A.C.	ADM17481	Piura	Cuenca Chira, Cuen- ca Piura e Intercuenca 1379.		(-4.929,-80.854)
Agrícola del Chira S.A.	ADM16435	Piura	Cuenca Chira, Cuen- ca Piura e Intercuenca 1379.		(-4.918,-80.702)
Agrolmos S.A.	ADM17036	Lambaye- que	Cuenca Olmos e Inter- cuenca 137779.		(-6.085,-80.063)
Casa Grande S.A.A.	ADM11828	La Libertad	Cuenca Chicama, Inter- cuenca 137719 e Inter- cuenca 13773.		(-7.745,-79.186)
Cartavio S.A.A.	ADM11829	La Libertad	Cuenca Chicama.		(-7.892,-79.220)
Agroindustrial Laredo S.A.A.	ADM11837	La Libertad	Cuenca Chicama, Cuenca Moche e Inter- cuenca 137715.		(-8.094,-78.962)
Agroindustrias San Jacinto S.A.A.	ADM11681	Áncash	Cuenca Nepeña.		(-9.146,-78.281)
Agroindustrial Paramonga S.A.A.	ADM11867	Lima	Intercuenca 137591.		(-10.674,-77.823)

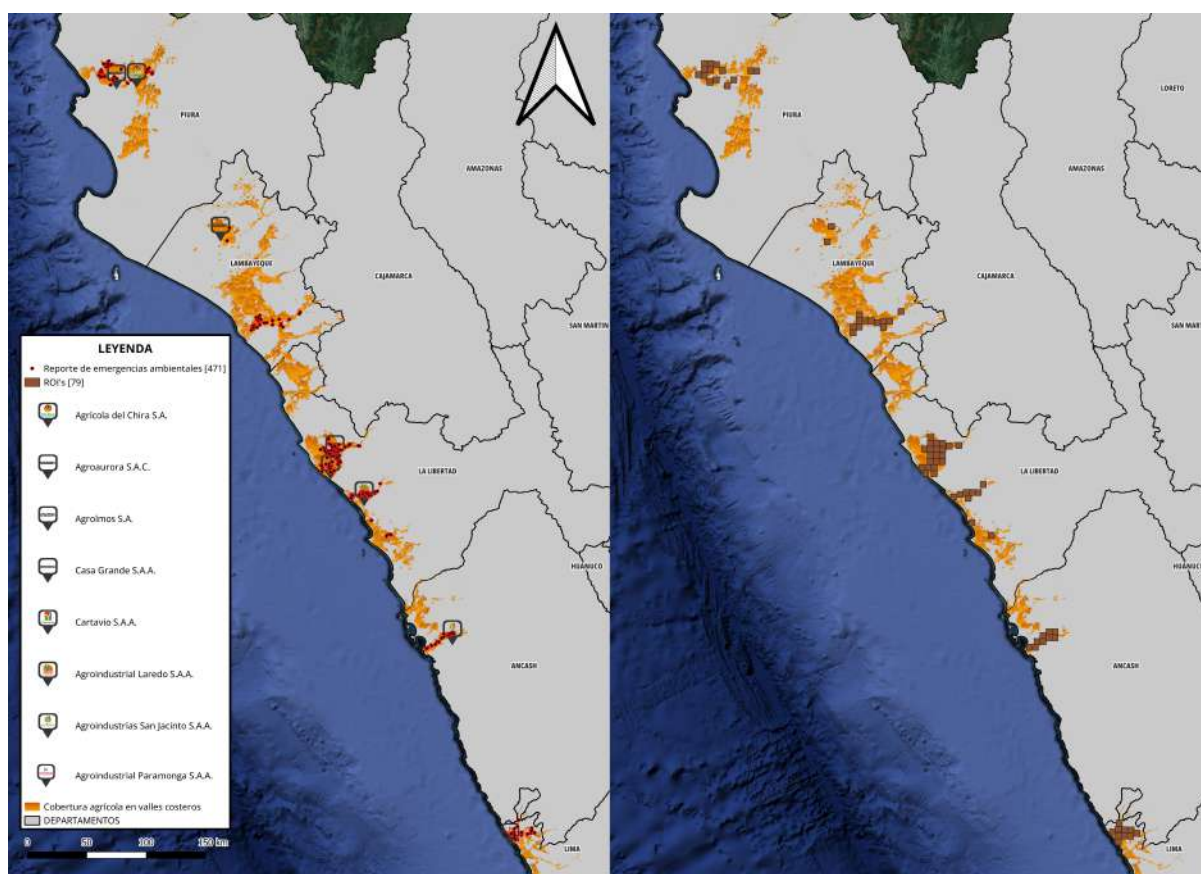
Nota. Extraído del [Portal Interactivo de Fiscalización Ambiental](#) de OEFA.

4.3.2. Muestra

La muestra de este estudio está constituida por las **áreas donde se ha reportado la quema de caña de azúcar**, ubicadas en la costa norte y centro del Perú, desde los valles del Chira (Piura) hasta el valle del Huaura (Lima). Esta delimitación se basa en los reportes de quema provenientes del registro histórico de 471 emergencias ambientales relacionadas con esta práctica, en el período comprendido entre enero de 2020 y agosto de 2023, proporcionado por el OEFA (Figura 25, parte izquierda).

Figura 25

Áreas con reporte de quema de caña de azúcar (ROI's).



Nota. Elaborado a partir del registro histórico de emergencias ambientales del OEFA.

Como resultado, se generan inicialmente 79 regiones de interés (ROI's), cada una correspondiente a una escena de imágenes Sentinel-2 L2A de 512×512 con una resolución de píxel de 10 m por píxel (Figura 25, parte derecha).

En la Figura 58 del Anexo, se muestra la tabla de los ROI's junto con sus coordenadas centrales y la cantidad de emergencias ambientales reportadas en cada uno de ellos.

4.4. Procedimiento, técnicas e instrumentos de recolección de información

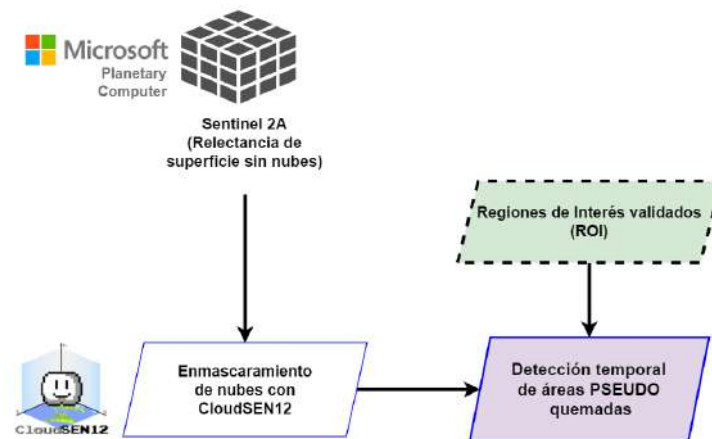
4.4.1. Pre-procesamiento de datos Sentinel-2 L2A

El producto Sentinel-2 L2A ofrece datos de reflectividad de superficie obtenidos a partir de un sensor óptico, pero las imágenes pueden verse afectadas por factores como nubes, sombras y aerosoles, lo que altera el espectro de los píxeles y complica la detección de áreas quemadas en series temporales de imágenes.

Para mitigar este problema, se emplea un modelo de detección de nubes y sombras entrenado con el conjunto de datos CloudSEN12 (Aybar *et al.*, 2022), con el objetivo de seleccionar únicamente aquellas imágenes de cada ROI cuya cobertura de nubes y sombra de nubes no supere el 10 %.

Figura 26

Esquema metodológico para el pre-procesamiento de datos Sentinel-2



Nota. Elaboración propia.

Se filtraron las imágenes históricas de las series temporales, obteniéndose un total de 10,149 imágenes en 74 ROI's de las 79 iniciales.

4.4.2. Detección temporal de áreas pseudoquemadas

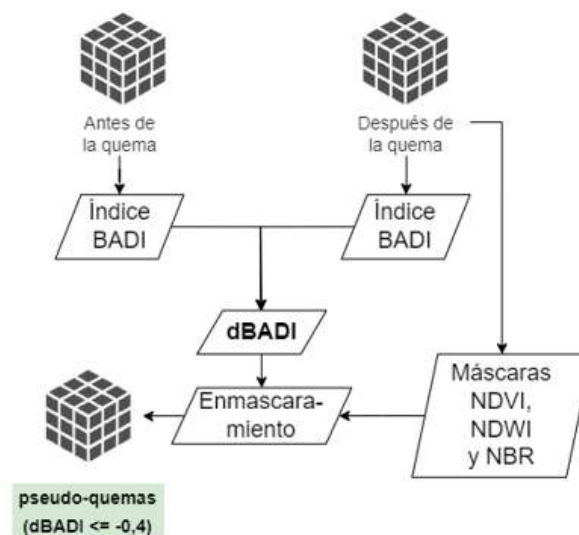
Los datos provenientes de Sentinel-2 L2A están disponibles desde el año 2017, lo que implica que cada región de interés (ROI) tiene un máximo aproximado de 400 imágenes disponibles. Esta cantidad considerable de datos dificulta el proceso de selección de imágenes que contengan áreas quemadas, ya que esto requiere una interpretación visual exhaustiva de cada

imagen de manera individual.

Para abordar esta dificultad, se emplean métodos basados en reglas para detectar posibles áreas quemadas o pseudoquemadas (ver Figura 27). Esta estrategia permite identificar automáticamente las posibles áreas con quema, facilitando el proceso de identificación y análisis posterior.

Figura 27

Esquema metodológico para la detección pseudoquemadas.



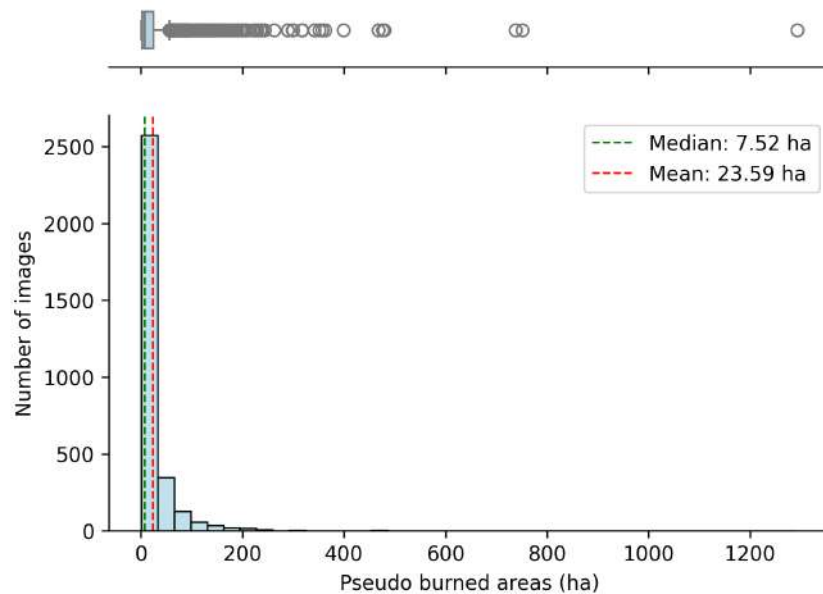
Nota. Elaboración propia.

El cálculo del índice $\Delta BADI$, basado en la propuesta de Farhadi *et al.* (2023), se lleva a cabo mediante el procesamiento del BADI utilizando imágenes antes y después del probable evento de quema de la serie temporal en cada ROI.

Posteriormente, el resultado del umbral $\Delta BADI \leq -0.4$ obtenido se enmascara mediante el uso de $NDVI \leq 0.2$, $NDWI \leq 0$ y $NBR \leq -0.15$ con el objetivo de eliminar áreas cubiertas por vegetación, cuerpos de agua y reducir los falsos positivos respectivamente. Esta estrategia detectó quema mayor a una hectárea (> 1 ha), equivalente a 100 píxeles de Sentinel-2, en un total de 3,195 imágenes distribuidas en 74 ROI's.

Figura 28

Histograma de áreas pseudoquemadas detectadas (en hectáreas) en las 74 ROI's.



Nota. Elaboración propia.

De la Figura 28 se observa que la mediana (7.52 ha) y la media (23.59 ha) se encuentran muy cerca del extremo izquierdo del histograma, lo que sugiere una distribución fuertemente sesgada hacia áreas quemadas de menor tamaño. Los outliers, que corresponden a áreas pseudoquemadas significativamente grandes, se localizan en el extremo derecho del histograma y suman un total de 337 imágenes.

De las 2,858 imágenes restantes, se seleccionan 2,000 mediante un muestreo estratificado, utilizando los percentiles 25, 50 y 75 como criterios para definir los estratos de la distribución de áreas quemadas.

4.4.3. Selección de imágenes con áreas quemadas

El primer paso para asegurar un alto grado de fiabilidad en la presencia de quema y no quema es a través de la interpretación visual de las imágenes a las cuales se le detectaron pseudoquemadas con extensión mayor a 1 hectárea. Para esta interpretación se utilizan cuatro combinaciones bandas de Sentinel-2 (Tabla 4).

Tabla 4*Descripción de las combinaciones RGB para diferentes tipos de áreas.*

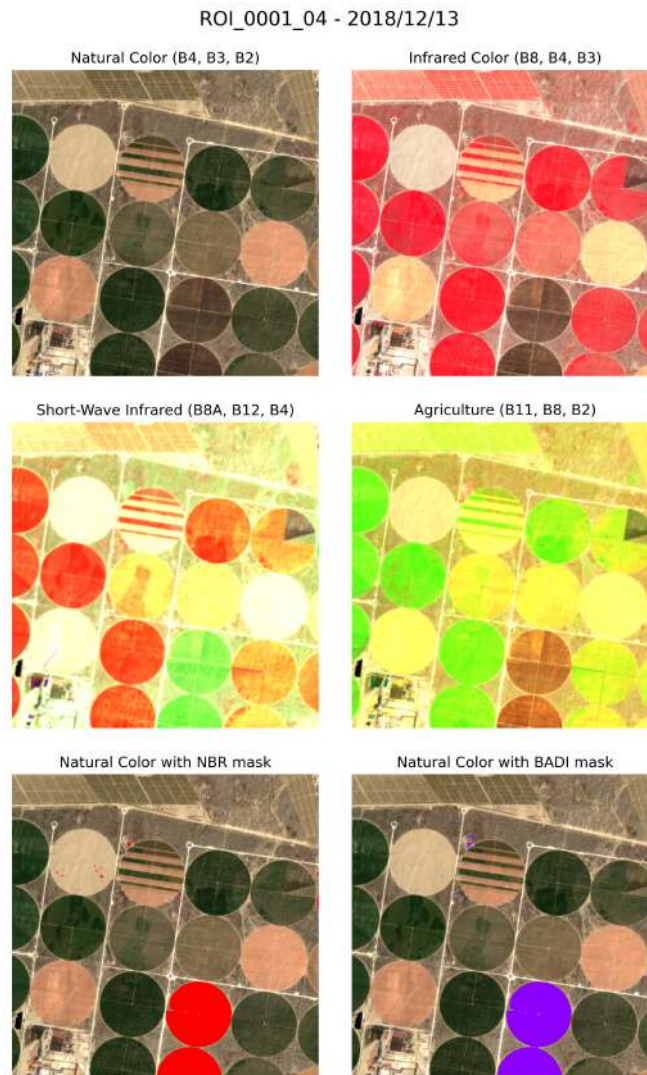
Nombre	Combinación RGB	Descripción
Color natural	R:B4 - G:B3 - B:B2	La vegetación sana es verde y las áreas de quema suelen ser de color gris o negro.
Falso Color convencional	R:B8 - G:B4 - B:B3	La vegetación densa es roja y las áreas de quema suelen ser de color gris o negro.
Infrarrojo de onda corta	R:B8A - G:B12 - B:B4	La vegetación densa es roja y las áreas de quema suelen ser de color verde.
Agricultura	R:B11 - G:B8 - B:B2	La vegetación tiene tonos de verde y las áreas de quema suelen ser de color rojizo.

Nota. Elaboración propia con información de Sentinel Hub (2023)

Asimismo, se emplean máscaras basadas en índices espectrales, a las cuales se les aplicó una reclasificación binaria utilizando los criterios $BADI \geq 0.4$ y $NBR \leq -0.15$. Con estos criterios, se generaron imágenes para identificar áreas quemadas, tal y como se muestra en la Figura 29 que muestra un área típica con la presencia de áreas quemadas.

Figura 29

Combinaciones RGB utilizadas para la selección de imágenes con áreas quemadas.



Nota. Elaboración propia.

En esta etapa se visualizaron las 2,000 imágenes para verificar la presencia de quema. En promedio, la **detección temporal de áreas pseudoquemadas** capturó el 52.7 % de las imágenes totales como áreas con quema. Es decir 1,054 imágenes en 66 ROI's contenían áreas quemadas y fueron incluidas en el **etiquetado de imágenes**.

4.4.4. Etiquetado de imágenes

Para generar el conjunto de datos se procede a etiquetar manualmente las imágenes utilizando el software IRIS (Figura 30) sin la asistencia predictiva por defecto. En este proceso se etiquetan las 1,054 imágenes seleccionadas en dos clases: “quema” y “no quema”. Además,

se incorpora un **composite de referencia** como visualización adicional, obtenido a partir de la composición espectral de 25 imágenes aleatorias de la serie temporal de cada ROI.

Figura 30

Interfaz de etiquetado en IRIS incorporando las combinaciones de banda y una imagen de referencia.

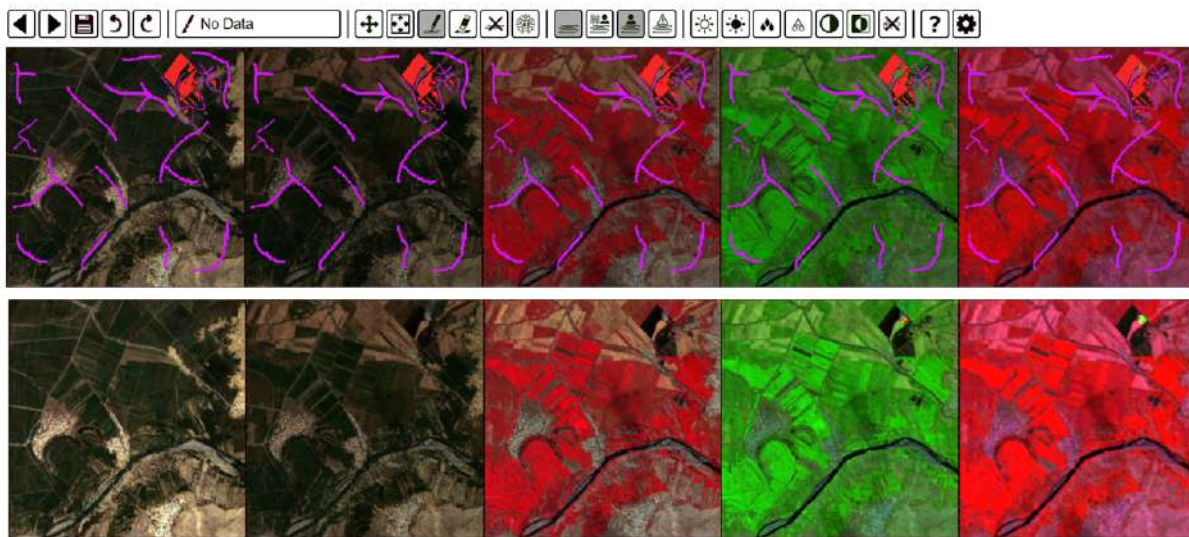


Nota. La visualización es de izquierda a derecha: Composite de referencia (en color natural), Color natural, Falso color convencional, Infrarrojo de onda corta y Agricultura.

Para realizar esta tarea, se utiliza el método “Scribble” (Figura 31), el cual implica etiquetar un pequeño conjunto de muestras en áreas de gran fiabilidad (Luo *et al.*, 2018). Se obtuvo un promedio por imagen de 1,317 píxeles para la clase de “quema” y 4,278 píxeles para la clase de “no quema”, manteniéndose en una proporción que no excedió la relación del 5 % y 95 % entre la clase “quema” y “no quema” respectivamente.

Figura 31

Ejemplo de etiquetado de imágenes utilizando el método “Scribble”.



Nota. En la parte inferior se muestra la imagen sin etiquetar perteneciente al ROI_0020_04, mientras que la parte superior se muestra el resultado del etiquetado con 3,715 píxeles clasificados como “quema” y 12,225 píxeles clasificados como “no quema”.

Se considera la diversidad espectral en las áreas etiquetadas, definiendo elementos como cuerpos de agua, suelo desnudo, diferentes tipos de vegetación, cultivos inundados, afloramientos rocosos, nubes y áreas artificializadas como ejemplos de “no quema”.

El protocolo para el etiquetado de la clase “quema” se basa en la evaluación visual de las imágenes, respaldada por las características espectrales de una quema en la región del espectro NIR y SWIR. Para determinar si en una imagen un área es específico corresponde a una quema o no, se establecen un protocolo con cuatro criterios, los cuales se muestran en la Tabla 5.

Si en una área de la imagen se cumple como mínimo uno de los cuatro criterios establecidos, se etiqueta como “quema”. Por el contrario, si ninguno de estos criterios se identifica claramente se etiqueta como “no quema”. Esto asegura un alto nivel de confiabilidad en la clasificación visual de las áreas quemadas.

Tabla 5*Criterios para el protocolo de generación de etiquetas de quema.*

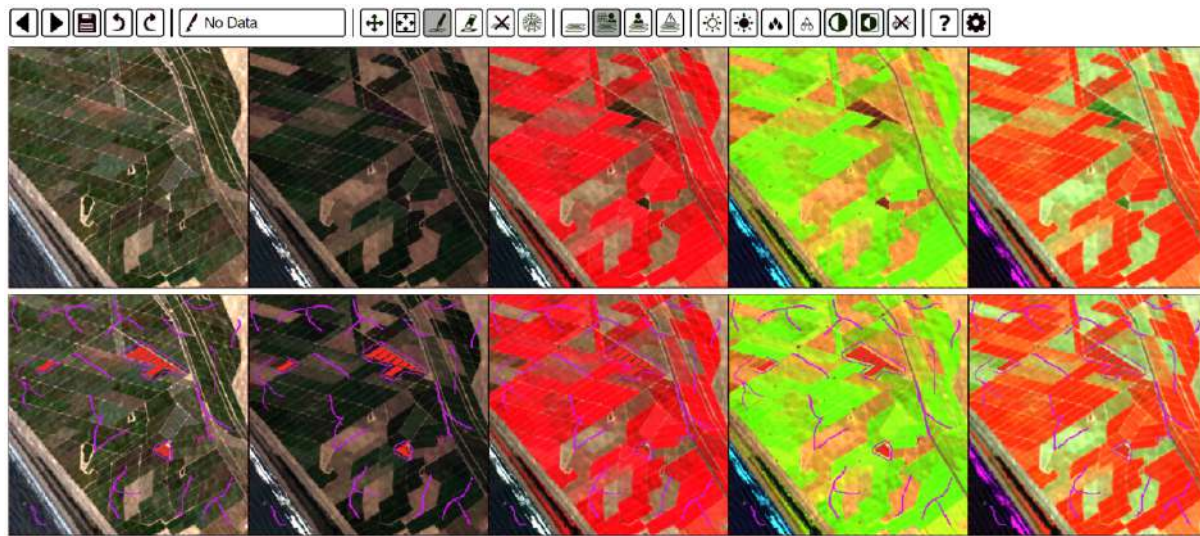
Criterio	No Quema (0)	Quema (1)
Visualización espectral	Los colores son típicos de otro uso de suelo.	La imagen cumple las tonalidades típicas en las cuatro visualizaciones.
Contraste	El área en cuestión, generalmente extensa, no se diferencia significativamente de sus alrededores.	Hay un contraste evidente con las áreas a su alrededor.
Foco de calor	No se detectan focos de calor en la imagen.	Se detectan focos de calor en la imagen y se aprecia el humo.
Ubicación y contexto espacial	Es difícil determinar si han ocurrido quemaduras en el área en cuestión.	Está en una ubicación conocida por tener quemaduras frecuentes.

Nota. Elaboración propia.

En las Figuras 32 - 35 se presentan ejemplos de imágenes etiquetadas en el software IRIS según los principales criterios establecidos, en donde las áreas clasificadas como “quema” se muestran en rojo y las áreas clasificadas como “no quema” se muestran en morado.

Figura 32

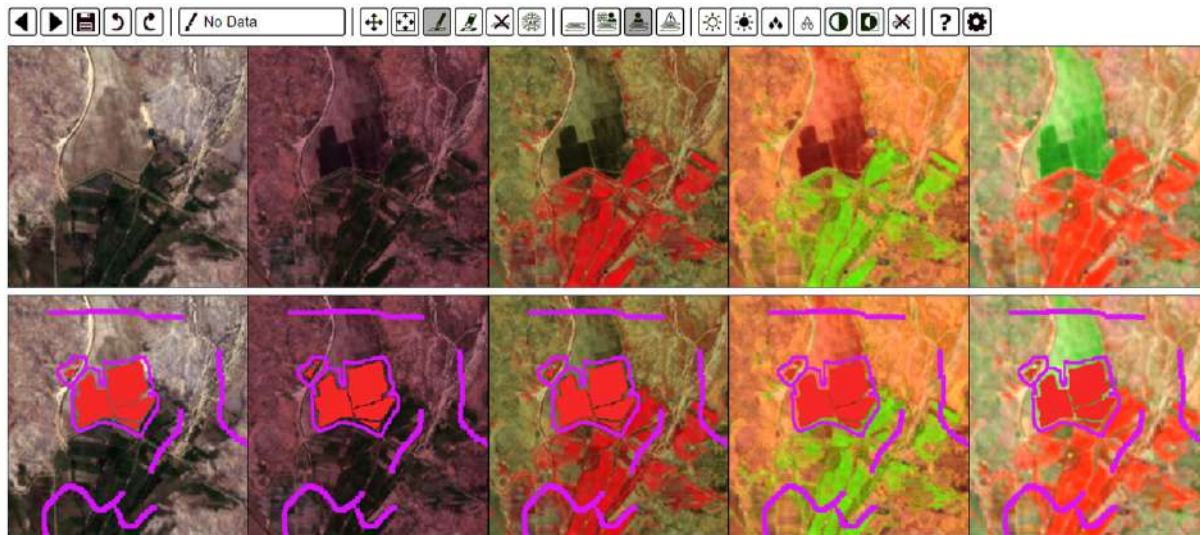
Visualización en IRIS según el criterio de visualización espectral para una imagen en la ROI_0097_04.



Nota. En la parte superior se muestra la imagen sin etiqueta, mientras que en la parte inferior se presenta la imagen con las etiquetas aplicadas. Elaboración propia.

Figura 33

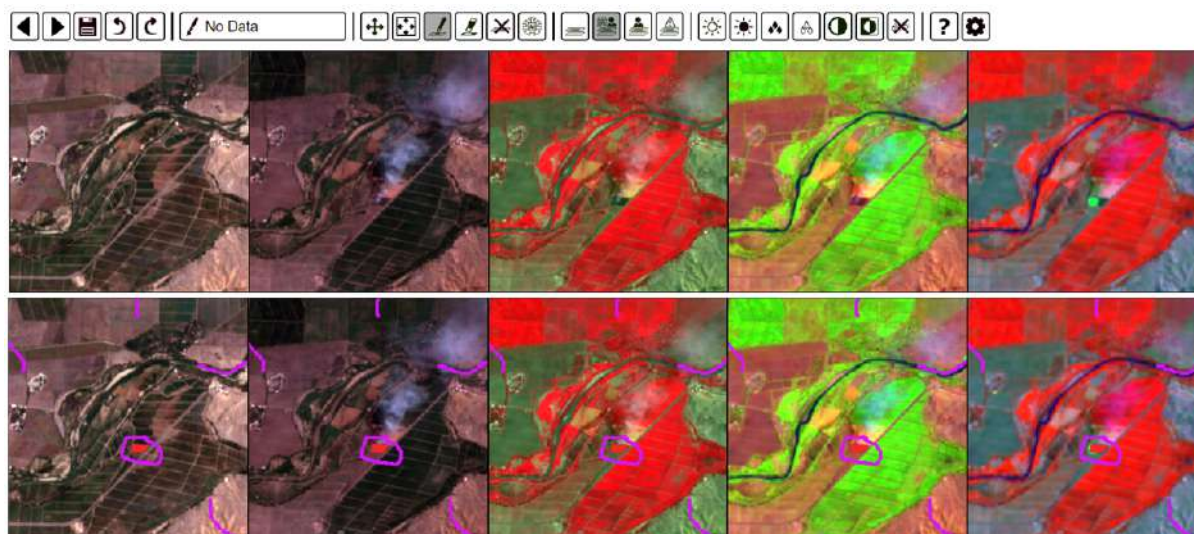
Visualización en IRIS según el criterio de contraste para una imagen en la ROI_0020_01.



Nota. En la parte superior se muestra la imagen sin etiqueta, mientras que en la parte inferior se presenta la imagen con las etiquetas aplicadas. Elaboración propia.

Figura 34

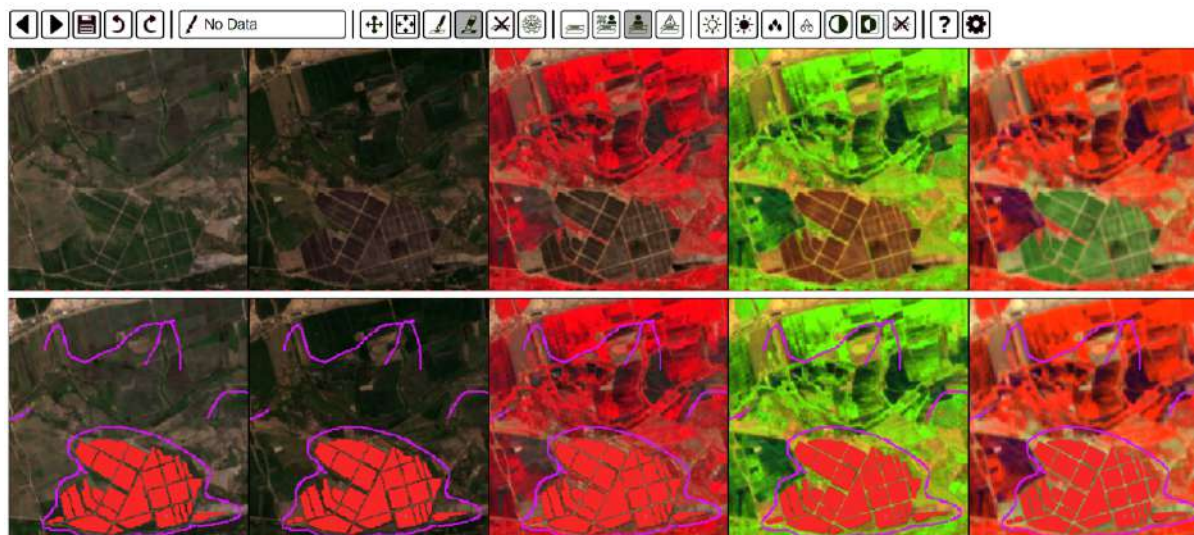
Visualización en IRIS según el criterio de foco de calor para una imagen en la ROI_0084_01.



Nota. En la parte superior se muestra la imagen sin etiqueta, mientras que en la parte inferior se presenta la imagen con las etiquetas aplicadas. Elaboración propia.

Figura 35

Visualización en IRIS según el criterio de ubicación y contexto espacial para una imagen en la ROI_0011_01.

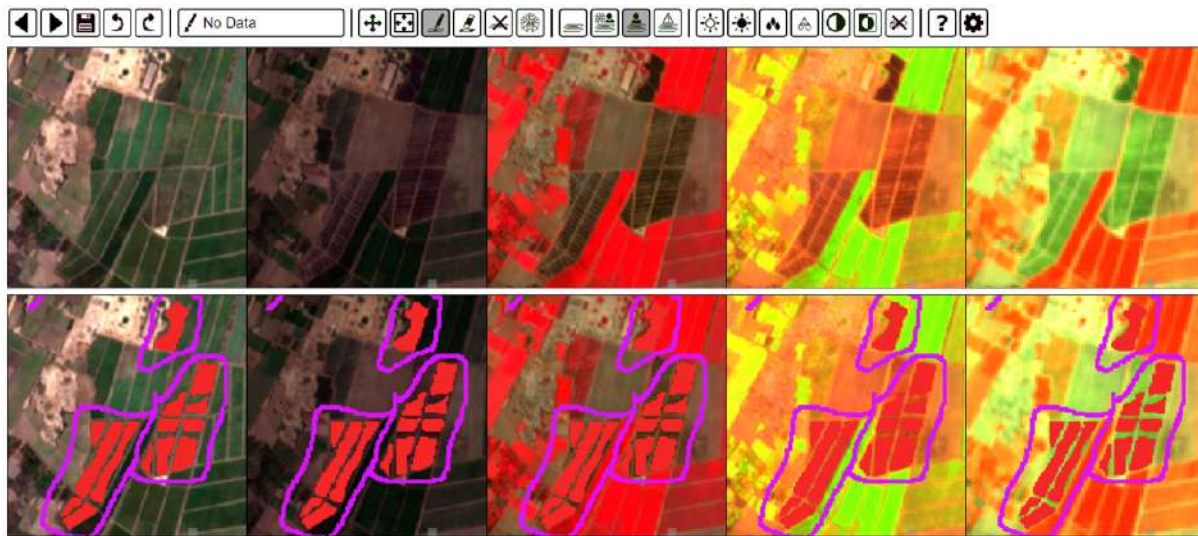


Nota. En la parte superior se muestra la imagen sin etiqueta, mientras que en la parte inferior se presenta la imagen con las etiquetas aplicadas. Elaboración propia.

En ciertos casos, se ha decidido establecer manualmente un umbral, especialmente cuando la quema exhibe tonalidades sutiles que no son consideradas significativas o relevantes.

Figura 36

Visualización en IRIS según el criterio de delimitación de umbral para una imagen en la ROI_0094_02.



Nota. En la parte superior se muestra la imagen sin etiqueta, mientras que en la parte inferior se presenta la imagen con las etiquetas aplicadas. Elaboración propia.

4.4.5. Generación del conjunto de datos para la detección de quemas

Una vez etiquetadas las imágenes, se procede a la generación de un conjunto de datos de entrenamiento para los modelos. Este conjunto de datos se compone de 1,054 imágenes de 512×512 píxeles con una profundidad de 16 bandas.

Estas 16 bandas comprenden el apilamiento de las 10 bandas de Sentinel-2, que abarcan las bandas del espectro visible (B02, B03, B04), el borde rojo (B05, B06, B07, B8A), el infrarrojo cercano (B8) y el infrarrojo de onda corta (B11, B12). Además, se incluyen índices específicos como los de quema (NBR y BADI), vegetación (NDVI), agua (NDWI), así como variables topográficas como la pendiente (en %) y la distancia a la cobertura agrícola (en metros). Siendo estas dos últimas variables definidas a continuación:

4.4.5.1. Pendiente. La pendiente es una variable topográfica que se calcula a partir de un modelo digital de elevación (DEM) y se utiliza para caracterizar la inclinación del terreno. En este estudio, se utilizó el producto **NASA DEM**, que ofrece datos topográficos globales con una resolución horizontal de 1 segundo de arco (30 m) y se derivan principalmente de los capturados durante la Shuttle Radar Topography Mission (SRTM).

La relevancia de la pendiente en la detección de la quema de caña de azúcar radica en que las áreas de menor pendiente son más aptas para el cultivo de este tipo de cosecha. Además, la pendiente permite diferenciar los suelos desnudos ubicados en los valles costeros con alta inclinación.

4.4.5.2. Distancia a la cobertura agrícola. La distancia a la cobertura agrícola es una variable utilizada para medir la proximidad de un píxel a áreas de cultivo. Esta variable se calculó utilizando el mapa de **ESRI Land Use**, que tiene una resolución de 10 metros.

Este conjunto de datos consiste en una serie temporal de mapas globales anuales de uso y cobertura del suelo (LULC) desde 2017 hasta 2023. Los mapas se derivan de imágenes Sentinel-2 de la ESA y proporcionan predicciones de LULC para 9 clases a lo largo del año, ofreciendo una representación precisa y representativa de cada año.

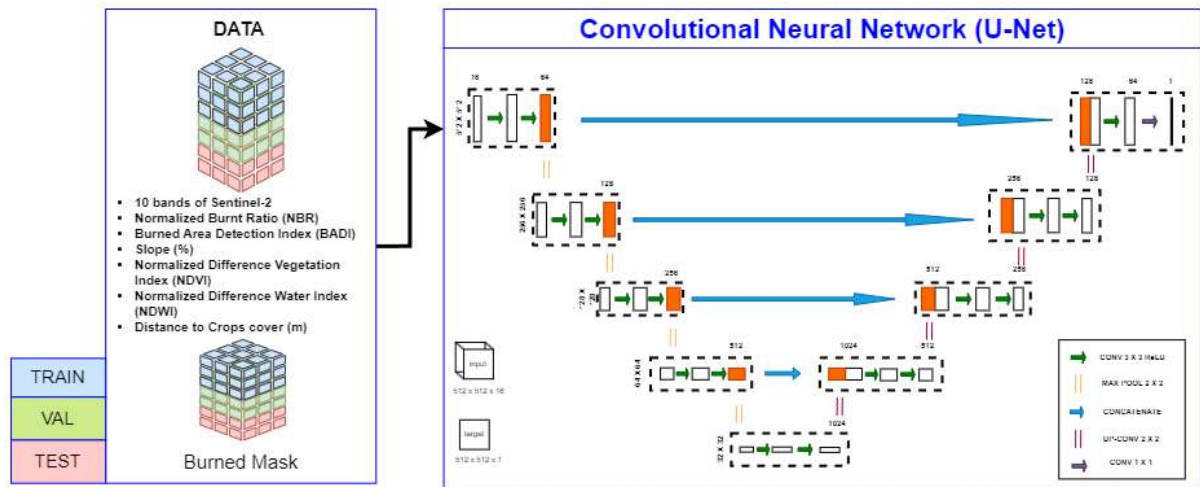
Su relación con la detección de quema de caña de azúcar se explica en que valores bajos de distancia a la cobertura agrícola indican áreas cercanas a cultivos. Esta variable es crucial para identificar las zonas que están directamente involucradas en actividades agrícolas y, por ende, potencialmente sujetas a quemas de caña de azúcar.

4.4.6. Implementación del modelo CNN

Se implementa una red neuronal convolucional con una arquitectura U-Net y un codificador **ResNet50** (Al-Dabbagh e Ilyas, 2023), preentrenado en el conjunto de datos ImageNet. En la Figura 37 muestra la arquitectura de la red U-Net, que toma como entrada una imagen con dimensiones de $512 \times 512 \times 16$ (profundidad de 16 bandas), y genera un mapa de segmentación de dimensiones $512 \times 512 \times 1$ que representa la probabilidad de quema en cada píxel.

Figura 37

Arquitectura U-Net para la segmentación semántica.



Nota. Elaboración propia.

4.4.7. Implementación del Gradient Boosting Decision Tree

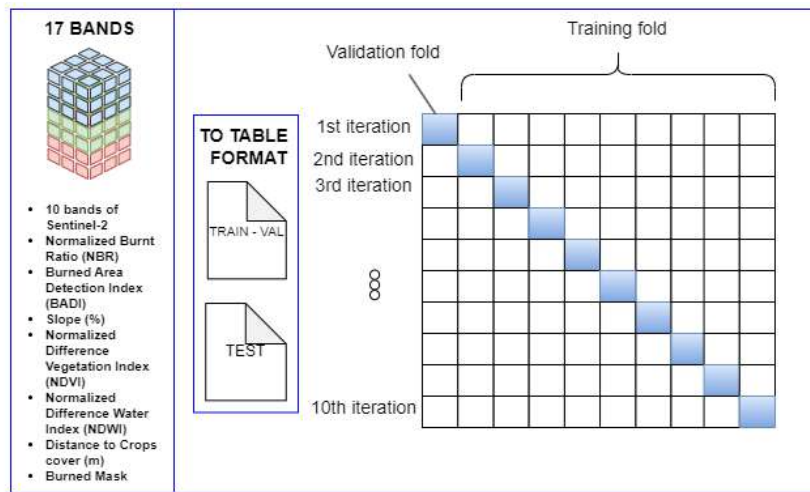
Se propone el modelo de Gradient Boost Decision Tree, implementado en el framework de LightGBM en Python, para generar predicciones. El modelo se entrena con un conjunto de datos estructurado en forma tabular, donde cada fila corresponde a un pixel y cada columna a una de las 17 bandas disponibles, incluyendo la etiqueta correspondiente.

Para prevenir el sobreajuste y asegurar una evaluación más precisa y confiable del rendimiento del modelo, se emplea la validación cruzada de 10 particiones o pliegues (ten-fold cross-validation). En este enfoque, el conjunto de datos se divide en 10 partes iguales; el modelo se entrena con 9 partes y se evalúa con la parte restante.

Este proceso se repite en 10 iteraciones, de modo que, en cada iteración, una parte diferente actúa como conjunto de validación mientras las restantes se utilizan para el entrenamiento. (Figura 38).

Figura 38

Esquema de la validación cruzada en diez pliegues.

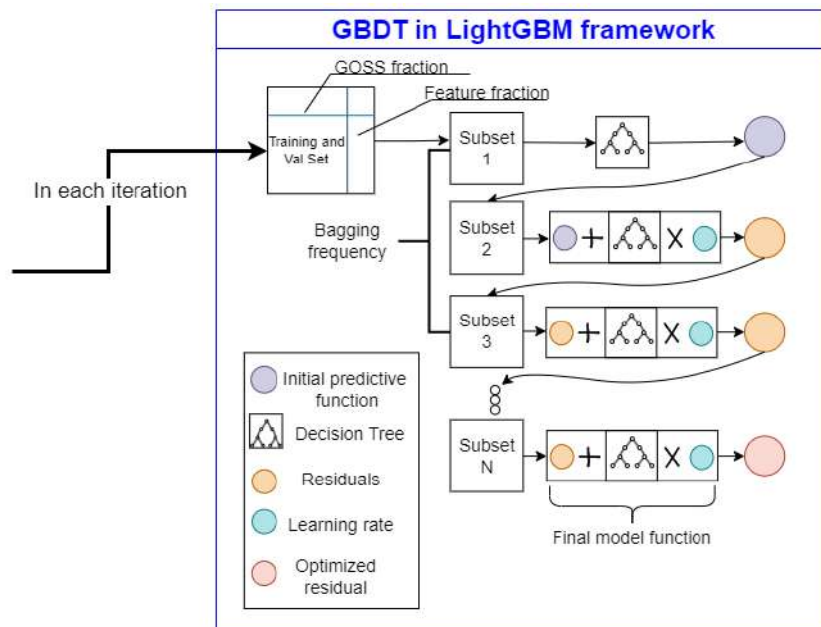


Nota. Elaboración propia.

La aplicación del validación cruzada de pliegues permitió obtener diez modelos utilizando LightGBM (Figura 39); se eligió el menor valor de probabilidad en cada píxel como medida de consenso entre los modelos debido que las etiquetas se crearon en áreas de alta fiabilidad, donde generalmente todos los modelos coinciden en clasificar el píxel como “quema”.

Figura 39

Modelo LightGBM para cada pliegue en la validación cruzada.



Nota. Elaboración propia.

4.4.8. Implementación del modelo Stacking

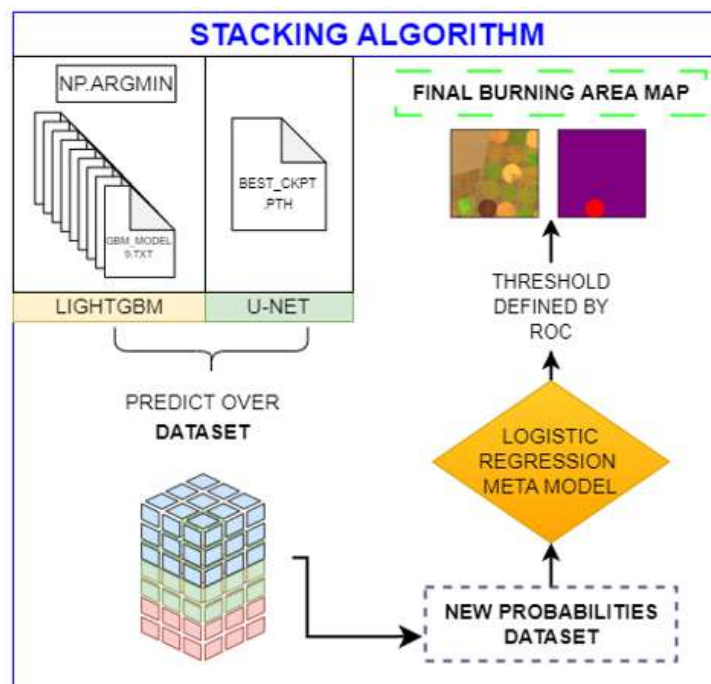
El modelo ensamblado de stacking utiliza como conjunto de datos los mapas de probabilidad generados por los dos modelos previos: U-Net y GBDT (Gradient Boosting Decision Tree).

A partir de estos mapas de probabilidad, se entrena un metamodelo de regresión logística cuyo propósito es combinar las predicciones de ambos modelos base para generar un mapa de probabilidades final que integra la información de ambos enfoques.

Una vez obtenido el mapa de probabilidades final del modelo Stacking, se procede a convertir en una máscara binaria. Para esto, se define un umbral óptimo basado en la curva ROC, seleccionando el punto que maximiza el equilibrio entre la sensibilidad (Precision) y la especificidad (Recall).

Figura 40

Esquema del modelo de ensamblado de stacking.



Nota. Elaboración propia.

4.4.9. Validación del umbral de clasificación para el modelo Stacking

La validación del umbral de clasificación definido anteriormente debe de verificarse visualmente en un conjunto de imágenes de validación. Estas imágenes de validación corresponden a áreas quemadas que son coherentes con los 471 reportes dentro del registro histórico del

OEFA.

Para este fin, de un total de 478 imágenes analizadas se obtuvieron 58 imágenes que cumplían con los siguientes criterios:

- La imagen muestra una extensión de quema dentro de un rango de ± 7 días respecto a la fecha reportada al OEFA.
- La imagen muestra una extensión de quema que es visualmente detectable.
- Las coordenadas del evento de quema son espacialmente coherentes con la ubicación mostrada en la imagen.

Este proceso asegura que el umbral de clasificación se corresponda con la realidad de las áreas quemadas, garantizando que el modelo Stacking sea capaz de detectar de manera precisa y confiable las áreas quemadas en las imágenes de Sentinel-2 con el umbral de clasificación definido.

4.4.10. Inferencia del modelo Stacking en un área piloto

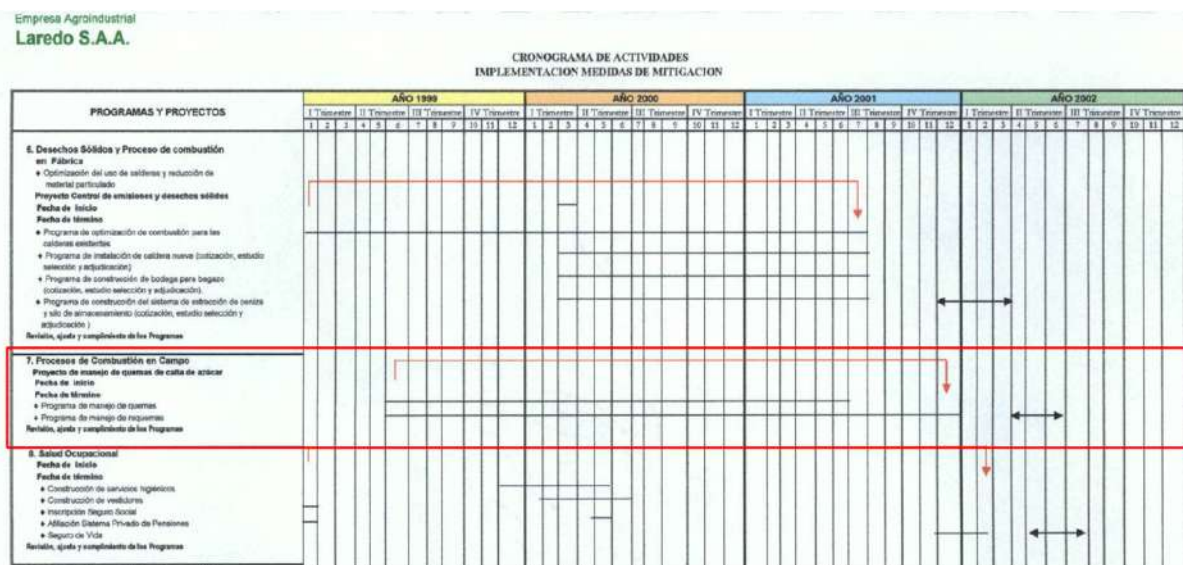
Tras validar el umbral de clasificación, se procede a realizar la inferencia en el distrito de Laredo (área piloto), ubicado en el departamento de La Libertad, uno de los doce distritos de la provincia de Trujillo, centrando el análisis en un período reciente de enero de 2023 a agosto de 2024.

Este distrito, con una población proyectada en 2022 de 44,440 habitantes según datos del Instituto Nacional de Estadística e Informática (2022), fue elegido debido a su proximidad a los fundos azucareros propiedad del ingenio azucarero Laredo S.A.A.

Esta empresa carece de un instrumento de gestión ambiental actualizado que incluya la cosecha verde y un cronograma más detallado para el manejo de quemas (ver Figura 41), ya que su último Programa de Adecuación y Manejo Ambiental (PAMA) data de 2004 (Rg.G. N° 002-2004-INRENA-OGATEIRN).

Figura 41

Cronograma de actividades para la implementa de medidas de mitigación en el PAMA de Laredo S.A.A.

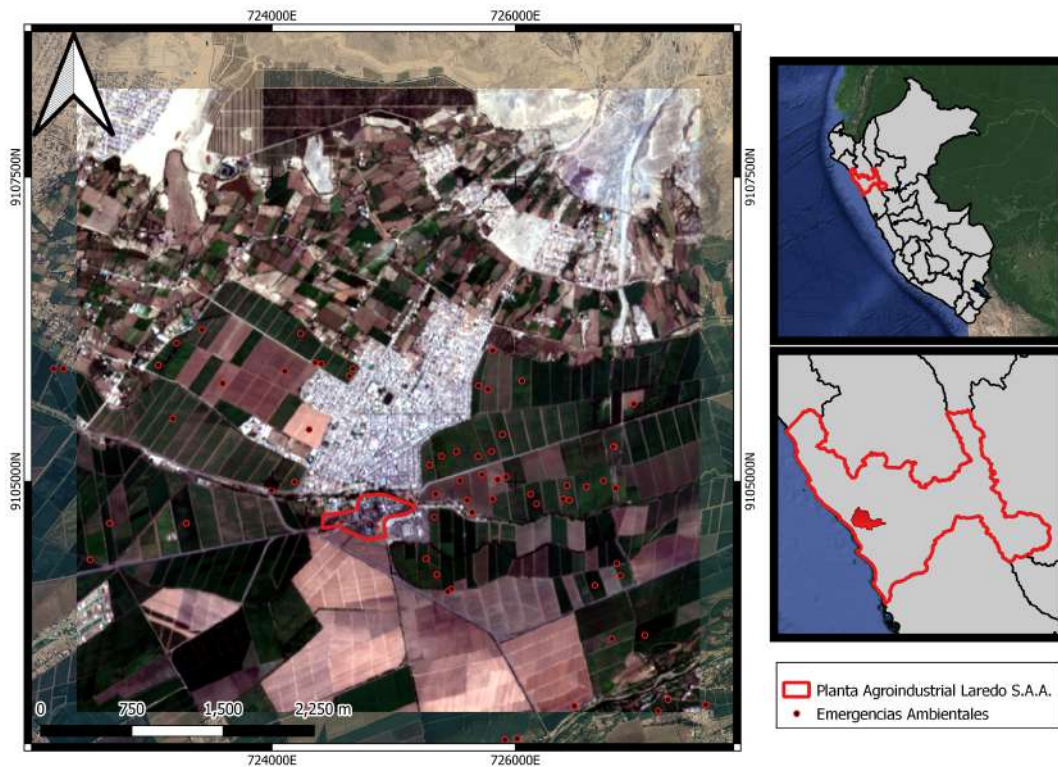


Nota. En la parte seleccionada de rojo, se observa que el PAMA de Laredo S.A.A. solo contempla este programa ante la quema de caña de azúcar. Extraído de Agroindustrial Laredo S.A.A. (2000, p. 24)

El criterio de selección es el mismo de la Sección 4.4.3, partiendo de la selección de la cobertura de nubosidad menor al 10 % y la detección de áreas pseudoquemadas. De esta selección se obtuvieron 30 imágenes de Sentinel-2 L2A para el área piloto de Laredo (Figura 42).

Figura 42

Área piloto seleccionada para la inferencia del modelo.



Nota. Elaboración propia.

4.5. Análisis estadístico

4.5.1. Evaluación del modelo

Para la evaluación del modelo, se utilizan cinco de las métricas definidas en la sección 2.2.3.

- **Precision:** Es útil para entender la exactitud del modelo al identificar áreas quemadas. Una alta precisión indica que el modelo tiene pocos falsos positivos, es decir, no clasifica incorrectamente áreas no quemadas como quemadas. Esto es importante para minimizar falsas alarmas y evitar la movilización innecesaria de recursos.
- **Recall:** Es útil porque garantiza que todas las áreas quemadas sean identificadas. Es crucial para asegurar que no se pasen por alto zonas afectadas, lo cual es fundamental en la planificación de la respuesta a emergencias.
- **F1-Score:** Es útil cuando es importante mantener un equilibrio entre evitar falsos positivos (precisión) y no perder áreas quemadas reales (recall). Es particularmente relevante

en situaciones donde tanto las omisiones como las falsas alarmas tienen consecuencias significativas y deben minimizarse.

- **IoU (Intersection over Union):** Es útil para medir la superposición entre la predicción del modelo y la verdad del terreno. Permite una evaluación precisa de cuán bien el modelo ha capturado los límites de la quema, lo cual es fundamental para aplicaciones que requieren una delimitación precisa de las áreas quemadas.
- **Índice Kappa:** Es útil como una medida más robusta en comparación con la simple precisión, ya que considera tanto los aciertos como los errores del modelo. Esto permite evaluar la concordancia del modelo más allá del azar, proporcionando una mejor comprensión de su fiabilidad y consistencia en la detección de áreas quemadas.

CAPÍTULO V

RESULTADOS, DISCUSIÓN, CONCLUSIÓN Y RECOMENDACIONES

5.1. Resultados

5.1.1. Implementación de los modelos de aprendizaje automático

Los experimentos realizados en la presente investigación se pueden visualizar en la plataforma de **Weights & Biases (W&B)**, en la cual se registraron los hiperparámetros, métricas y configuraciones de hardware durante el entrenamiento de modelos de aprendizaje automático.

5.1.1.1. Configuración de la Red Neuronal Convolutacional (CNN). Para el modelo, se utilizó una arquitectura de red neuronal convolutacional profunda, específicamente la arquitectura U-Net. En la siguiente configuración se define la estructura fundamental del modelo, incluyendo el tipo de red, el encoder preentrenado utilizado, y los detalles sobre cómo se manejan las entradas y salidas del modelo (Tabla 6).

Tabla 6
Configuración del modelo U-Net

Parámetro	Valor	Detalle
Modelo	<code>smp.Unet</code>	Es una clase que representa la arquitectura de red neuronal convolutacional profunda U-Net.
Nombre del encoder (<code>encoder_name</code>)	Resnet50	Es un modelo de red neuronal convolutacional pre-entrenado.
Pesos del encoder (<code>encoder_weights</code>)	ImageNet	Es una base de datos de imágenes etiquetadas.
Número de canales de entrada (<code>in_channels</code>)	16	12 bandas de Sentinel-2, BADI, NBR, Slope y Distance to Agriculture Cover
Número de canales de salida (<code>classes</code>)	1	Imagen con las probabilidades de quema.
Tasa de aprendizaje (<code>lr</code>)	0.0001	Tasa de aprendizaje.

Parámetro	Valor	Detalle
Peso de las clases (<code>class_weights</code>)	[1, 15]	Peso de las clases para la función de pérdida [“no quema”, “quema”].
Función de pérdida (<code>loss</code>)	WCE - Dice Loss	Suma ponderada entre la pérdida ponderada de entropía cruzada binaria y pérdida de DICE.
Métricas de evaluación	F1, Precisión, Recall, IoU, Kappa	Métricas de evaluación en el entrenamiento y validación del modelo

Nota. El experimento se visualiza en [Weights & Biases](#).

También se muestra la configuración del entrenamiento que define cómo se entrenará el modelo, especificando detalles sobre los datos utilizados, la estrategia de entrenamiento y los recursos computacionales. (Tabla 7):

Tabla 7
Configuración del entrenamiento del modelo UNet

Parámetro	Valor	Detalle
Dataset de entrenamiento	675	Conjunto de datos dividido para el entrenamiento.
Dataset de validación	168	Conjunto de datos dividido para la validación dentro del entrenamiento.
Tamaño de lote (<code>batch_size</code>)	16	Número de ejemplos de entrenamiento utilizados en una iteración.
Número de workers (<code>num_workers</code>)	16	Número de subprocesos (o procesos) que se crean para realizar tareas en paralelo.
Estrategia de entrenamiento	ddp	Distribuye los datos a través de múltiples GPU y nodos de manera paralela en sistemas distribuidos.

Parámetro	Valor	Detalle
Tipo de aceleración	GPU	Unidades de procesamiento gráfico para realizar cálculos intensivos en paralelo.
Dispositivo	0	GPU NVIDIA GeForce RTX 4070 con 12 GB de memoria.
Número de épocas (max_epochs)	50	Número máximo de veces que el modelo ve todo el conjunto de datos en el entrenamiento.
Callbacks	ModelCheck- point, EarlyStopping	ModelCheckpoint guarda el mejor modelo durante el entrenamiento, mientras que EarlyStopping detiene el entrenamiento si no hay mejoras, para prevenir el sobreajuste.
Precisión (precision)	16-mixed	Precisión mixta de 16 bits en cálculos de punto flotante.

Nota. El experimento se visualiza en [Weights & Biases](#).

5.1.1.2. Resultados del modelo U-Net en la detección de áreas quemadas. El resultado del entrenamiento de la red neuronal convolucional dio un valor de pérdida (WCE - Dice) de 0.289 para el conjunto de datos de prueba. Cuantitativamente, el modelo UNet obtuvo valores de F1 de 67.2 %, precisión de 60.2 %, recall de 77.2 %, IoU de 51.1 % y Kappa de 75.7 % para el conjunto de datos de prueba.

Los resultados cualitativos se pueden visualizar la Figura 43, donde cada columna representa un caso distinto, mostrando las predicciones del modelo U-Net en diversos contextos de áreas quemadas.

El modelo U-Net muestra un excelente desempeño al delimitar áreas quemadas extensas en todos los casos, ya que tiende a agrupar los píxeles dispersos y proporcionar una delimitación más coherente, especialmente en áreas de quema menores a 10 hectáreas (**quinto caso**).

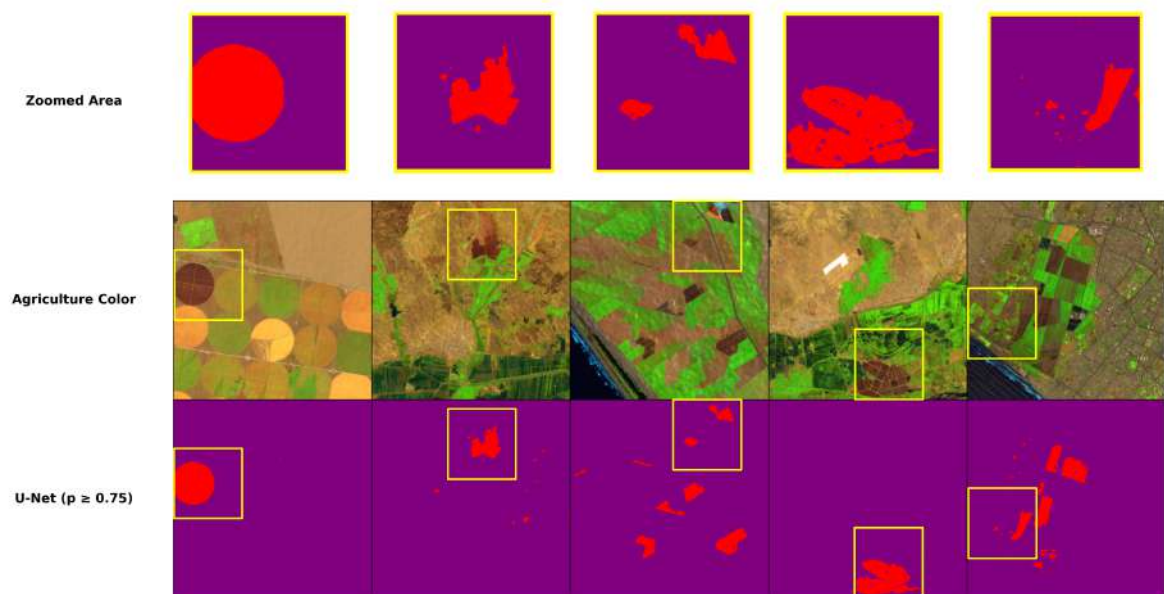
En situaciones de quemas activas (**tercer caso**) y en áreas con quema no reciente (**quinto caso**), el modelo logra una buena delimitación sin confundir las áreas de suelo desnudo que

rodean la área con quema.

Sin embargo, en casos donde se requiere una mayor precisión para diferenciar límites de los campos de cultivo, como en el **cuarto caso**, el modelo tiende a subestimar la extensión de las áreas quemadas, lo que resulta en delimitaciones incorrectas.

Figura 43

Resultados del modelo U-Net.



Nota. El color rojo representa las áreas quemadas y el morado las áreas no quemadas. Cada caso corresponde a una columna.

5.1.1.3. Configuración y resultados del Gradient Boosting Decision Tree (LightGBM).

Para la configuración del Gradient Boosting Decision Tree dentro del framework LightGBM se consideraron las principales parámetros (Tabla 8).

Tabla 8

Principales parámetros para el entrenamiento del modelo GBDT

Parámetro	Valor	Detalle
Tarea (task)	train	Tarea de entrenamiento del modelo.
Tipo de potenciación (boosting_type)	gbdt	Árboles de decisión potenciados por gradiente.
Objetivo de destino (objective)	binary	Los resultados son dos clases: “no quema” y “quema”.

Parámetro	Valor	Detalle
Métricas de entrenamiento (metric)	binary_logloss	Pérdida logarítmica binaria también conocida como entropía cruzada binaria.
Frecuencia de métricas (metric_freq)	1	Frecuencia de métricas por cada árbol construido.
Número de hojas (num_leaves)	31	Número de hojas máximo en un árbol.
Estrategia de muestreo (data_sample_strategy)	goss	Muestreo de datos utilizando el algoritmo GOSS.
Tasa de aprendizaje (learning_rate)	0.5	Tasa de aprendizaje.
Fracción de características (feature_fraction)	0.9	Fracción de características seleccionadas para la construcción de árboles.
Tasa de retención de altos gradientes (top_rate)	0.2	Se utilizarán el 20 % de los datos con los gradientes más altos para la construcción de árboles.
Tasa de retención de pequeños gradientes (other_rate)	0.1	Se utilizarán el 10 % de los datos con los gradientes más bajos para la construcción de árboles.
Dispositivo (device)	gpu	Unidades de procesamiento gráfico para realizar cálculos intensivos en paralelo.
ID de la plataforma (gpu_platform_id)	0	GPU NVIDIA GeForce RTX 4070 con 12 GB de memoria.

Nota. El experimento se visualiza en [Weights & Biases](#).

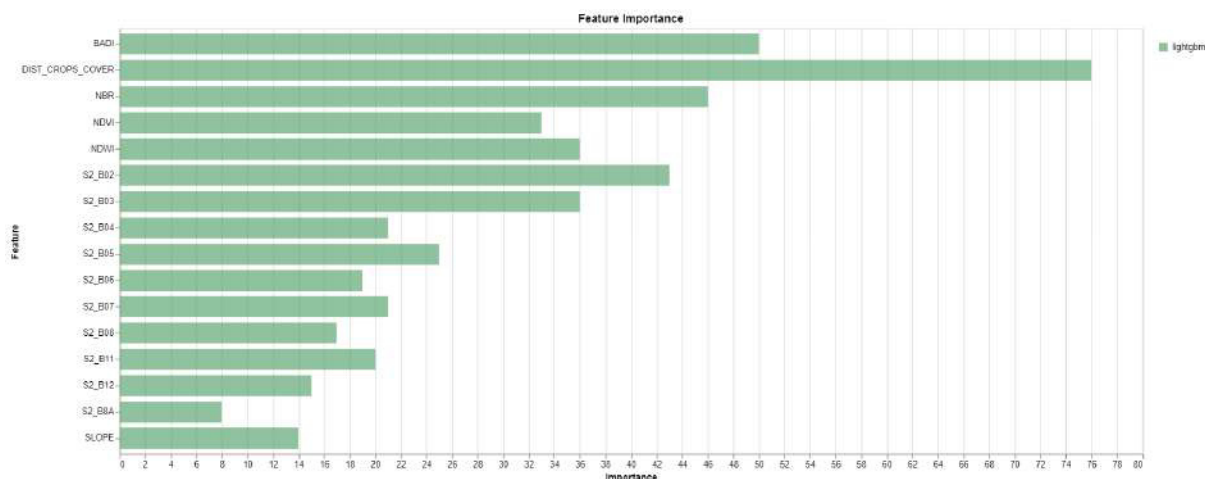
Los resultados del modelo LightGBM arrojaron una pérdida logarítmica binaria de 0.155 para el conjunto de datos de prueba. Cuantitativamente, los resultados de las métricas de evaluación fueron F1 de 86.0 %, precisión de 79.3 %, recall de 93.8 %, IoU de 75.4 % y Kappa de 81.2 %.

Asimismo, en la Figura 44, se ilustra la importancia de las características empleadas en

el modelo LightGBM. Se observa que la distancia a la cobertura agrícola (DIST_CROPS_COVER) con un 76 % y los índices de quema (BADI y NBR) con un 50 % y 46 % respectivamente son las más relevantes para la clasificación de áreas quemadas.

Figura 44

Importancia de las características en el modelo LightGBM.



Nota. Elaboración propia con datos de Weights & Biases.

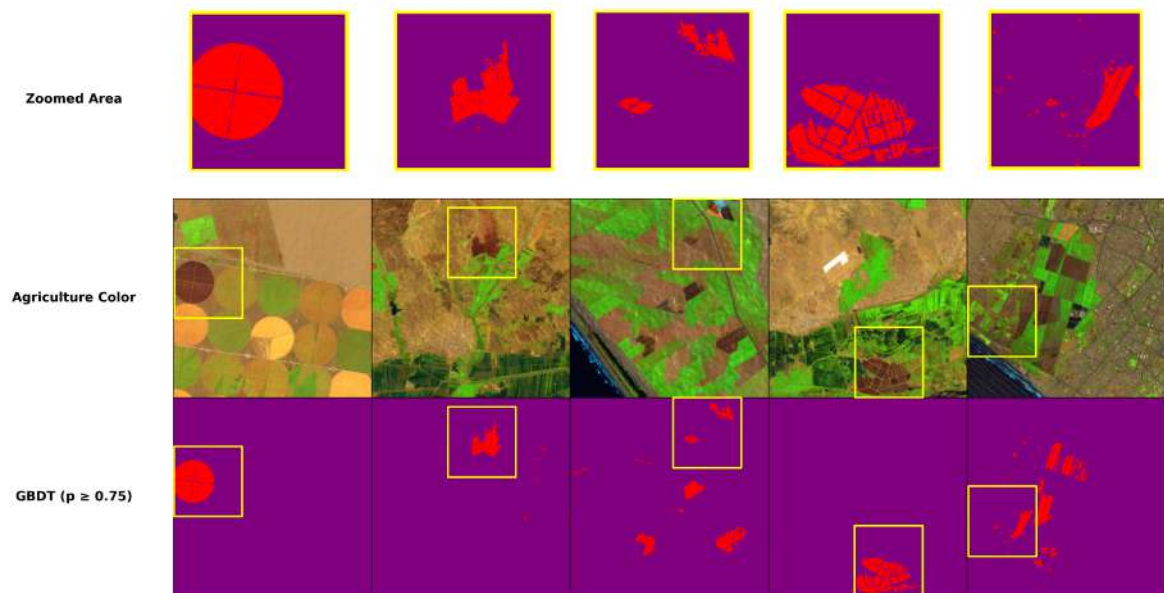
Para resultados cualitativos se puede visualizar la Figura 45, donde cada columna representa un caso distinto, mostrando las predicciones del modelo LightGBM en diversos contextos de áreas quemadas.

El modelo LightGBM muestra mejoras significativas en la delimitación de áreas quemadas para grandes extensiones, especialmente en el **primer caso** y el **cuarto caso**, donde logra capturar los contornos de manera más precisa en comparación con el modelo U-Net.

Sin embargo, en el **segundo caso** y en el **tercer caso**, se observa una dispersión considerable de los píxeles, lo que dificulta una delimitación clara de las áreas quemadas, especialmente dentro de las áreas quemada con características complejas, como zonas mixtas de vegetación y suelo.

En el **quinto caso**, esta dispersión se vuelve aún más pronunciada, resultando en un efecto de ruido o “salt n’ peper” que afecta significativamente la delimitación en áreas donde las características espectrales del terreno circundante, como áreas de quema no reciente y suelo desnudo, son similares a las áreas quemadas.

Figura 45
Resultados del modelo *LightGBM*.



Nota. El color rojo representa las áreas quemadas y el morado las áreas no quemadas. Cada caso corresponde a una columna.

5.1.1.4. Configuración y resultados del modelo Stacking. Para el modelo Stacking, se utilizó como entrada los mapa de probabilidades del modelo UNet y LightGBM, y se entrenó un metamodelo de regresión logística con la siguiente configuración (Tabla 9):

Tabla 9
Configuración del modelo de regresión logística

Parámetro	Valor	Detalle
Parámetro de regularización (C)	1.0	Inversa de la fuerza de regularización.
Penalidad (penalty)	l2	Regularización L2 para evitar el sobreajuste.
Solucionador (solver)	lbfgs	Algoritmo de optimización para minimizar la función de pérdida (binary cross-entropy).

Nota. Elaboración propia.

Los resultados del modelo ensamblado de Stacking arrojaron una pérdida logarítmica binaria de 0.157 para el conjunto de datos de prueba. Los resultados cuantitativos de las métricas

de evaluación fueron F1 de 90.6 %, precisión de 86.4 %, recall de 95.3 %, IoU de 82.8 % y Kappa de 87.5 %.

En la Figura 46, se presentan los resultados cualitativos donde cada columna representa un caso distinto, mostrando las predicciones del modelo Stacking en diversos contextos de áreas quemadas.

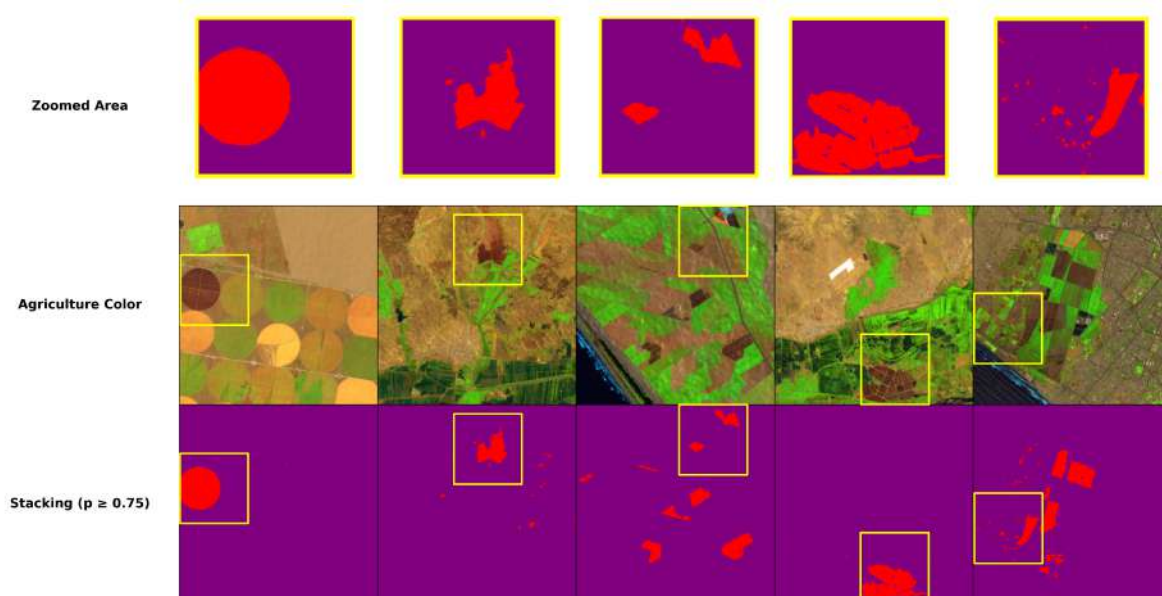
El modelo de stacking muestra resultados similares al U-Net en los tres primeros casos, logrando una buena delimitación de las áreas quemadas extensas y agrupando los píxeles dispersos de manera efectiva.

En el **cuarto caso**, el modelo de stacking se desempeña un poco mejor que U-Net, proporcionando una delimitación más precisa de las áreas quemadas. Sin embargo, no es tan claro ni detallado como el modelo LightGBM en esta situación.

En el **quinto caso**, aunque el modelo de stacking hereda los problemas del “salt n’ pepper”, logra agrupar la mayoría de este ruido, proporcionando una forma más clara y reduciendo en cierta medida la dispersión, lo que resulta en una delimitación más definida en las pequeñas áreas, en comparación a los otros modelos.

Figura 46

Resultados del modelo ensamblado.



Nota. El color rojo representa las áreas quemadas y el morado las áreas no quemadas. Cada caso corresponde a una columna.

5.1.2. Evaluación de los modelos

Para la evaluación de los modelos, se utilizó la métrica de evaluación de F1, precisión, recall, IoU y Kappa sobre el conjunto de datos de prueba, las cuales se presentan en la Tabla 10. Se observa que el modelo de Stacking supera consistentemente a U-Net y LightGBM en cada una de estas métricas.

Tabla 10
Evaluación de los modelos en porcentaje (%).

Métrica	U-Net	LightGBM	Stacking Model
F1 (↑)	67.2	86.0	90.6
Recall (↑)	77.2	93.8	95.3
Precision (↑)	60.2	79.3	86.4
IoU (↑)	51.1	75.4	82.8
Kappa (↑)	67.5	81.2	87.5

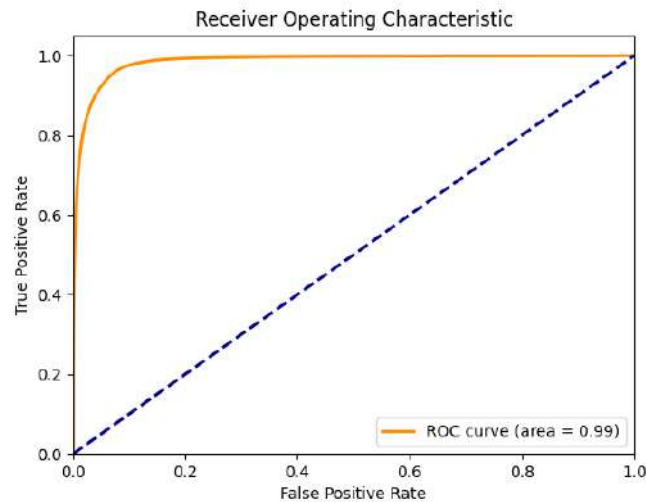
Nota. Elaboración propia.

5.1.3. Validación del umbral de probabilidad del modelo Stacking.

En la Figura 47, se presenta la curva ROC para el modelo ensamblado, con un valor $AUC = 0.99$ que indica una capacidad excelente para identificar áreas quemadas frente a áreas no quemadas. En otras palabras, el modelo es muy eficaz para distinguir entre las dos clases en casi todas las circunstancias.

Figura 47

Curva ROC para el modelo ensamblado.



Nota. Elaboración propia.

El valor de umbral que optimiza la sensibilidad y especificidad del modelo según la curva ROC fue de 0.20.

Posteriormente, se validó visualmente los resultados del umbral definido en el ROC para las siguientes 58 imágenes que cumplían los criterios mencionados en la Sección 4.4.9.

Con el umbral 0.20, se observó que el modelo Stacking logró una correcta identificación de áreas quemadas en el 100 % (58 de las 58 imágenes) de casos.

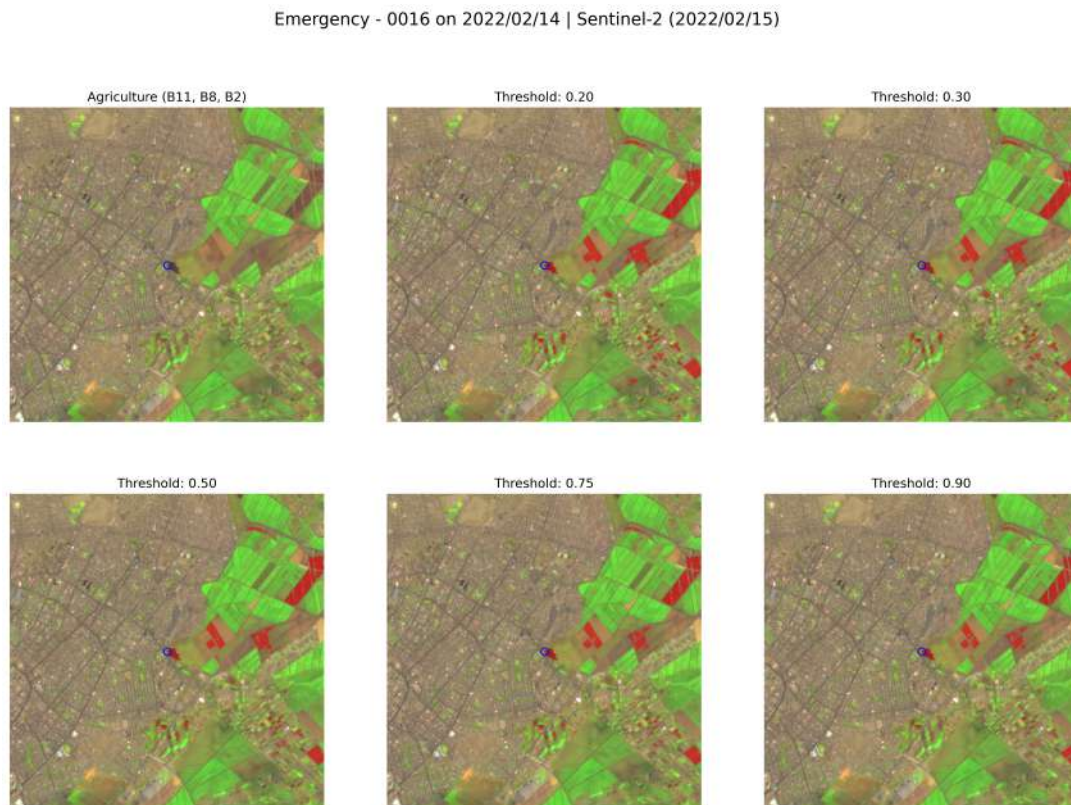
Se probó que con otros valores de umbral, como 0.30, 0.50, 0.75 y 0.90, también lograron identificar correctamente las áreas quemadas en todos los casos.

A continuación, se presentan los resultados cualitativos de la identificación de estas áreas quemadas en diferentes emergencias, referenciadas en las imágenes dentro de una circunferencia azul.

Para la emergencia **0016** reportada cerca de la ciudad de La Rinconada en Trujillo (Figura 48), se observa que todos los umbrales lograron identificar correctamente la ubicación y delimitación de la quema reciente. A medida que se incrementa el umbral, la delimitación no se ve tan afectada, siendo que el umbral de 0.90 estimó la quema reciente en 2.54 hectáreas.

Figura 48

Comparación de umbrales de probabilidad para la emergencia 0016.



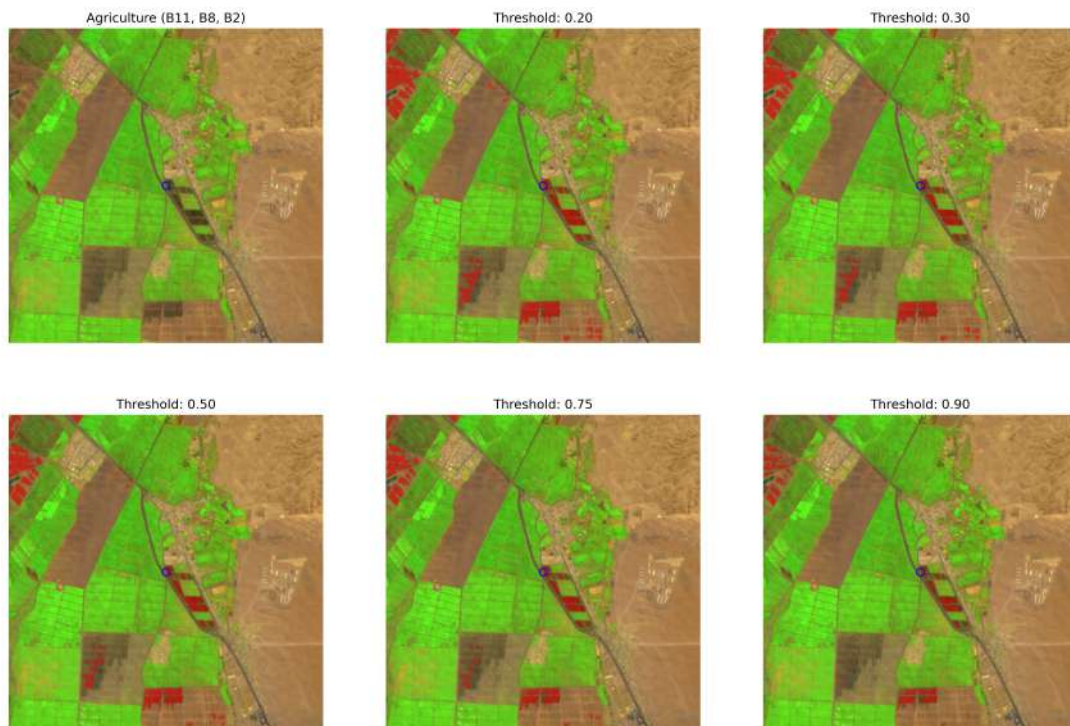
Nota. Elaboración propia.

En la ciudad de Chicama, se reportó una emergencia **0376** en el campo San Jorge, propiedad de Agroindustrial Casa Grande S.A.A., debido a una quema accidental. En la imagen, se delimita un área de quema distribuida en nueve campos de cultivo. En este caso, como en la mayoría de los casos, con el umbral en 0.20 se estimó la ubicación y delimitación de la quema en un total de 13.94 ha (Figura 49).

Figura 49

Comparación de umbrales de probabilidad para la emergencia 0376.

Emergency - 0376 on 2020/05/07 | Sentinel-2 (2020/05/11)



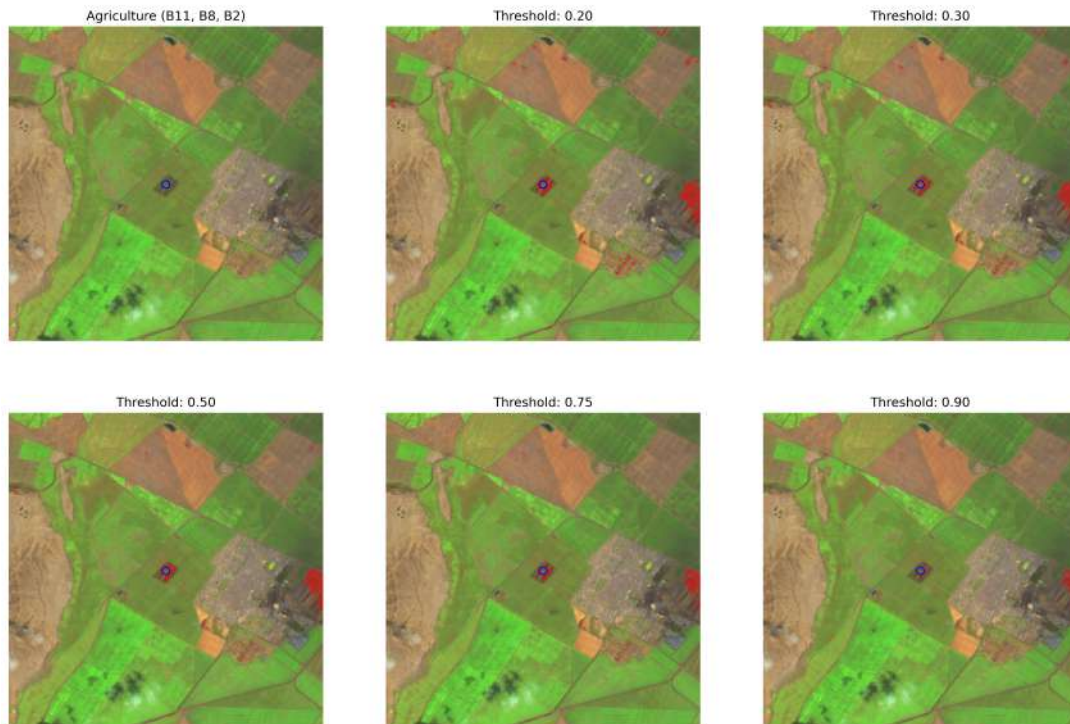
Nota. Elaboración propia.

En la ciudad de Casa Grande, se reportó una emergencia **0068** en un área de 9 hectáreas perteneciente a Agroindustrial Casa Grande S.A.A., debido a una quema accidental. En este caso, la quema no fue intensa y se vio afectada por la presencia de vegetación que aún no había sido cosechada. Los umbrales más bajos (0.20 - 0.50) lograron delimitar claramente el área afectada con un total de 4.5 ha, mientras que los umbrales más altos (0.70 y 0.90) subestimaban estas áreas (Figura 50).

Figura 50

Comparación de umbrales de probabilidad para la emergencia 0068.

Emergency - 0068 on 2022/11/16 | Sentinel-2 (2022/11/17)



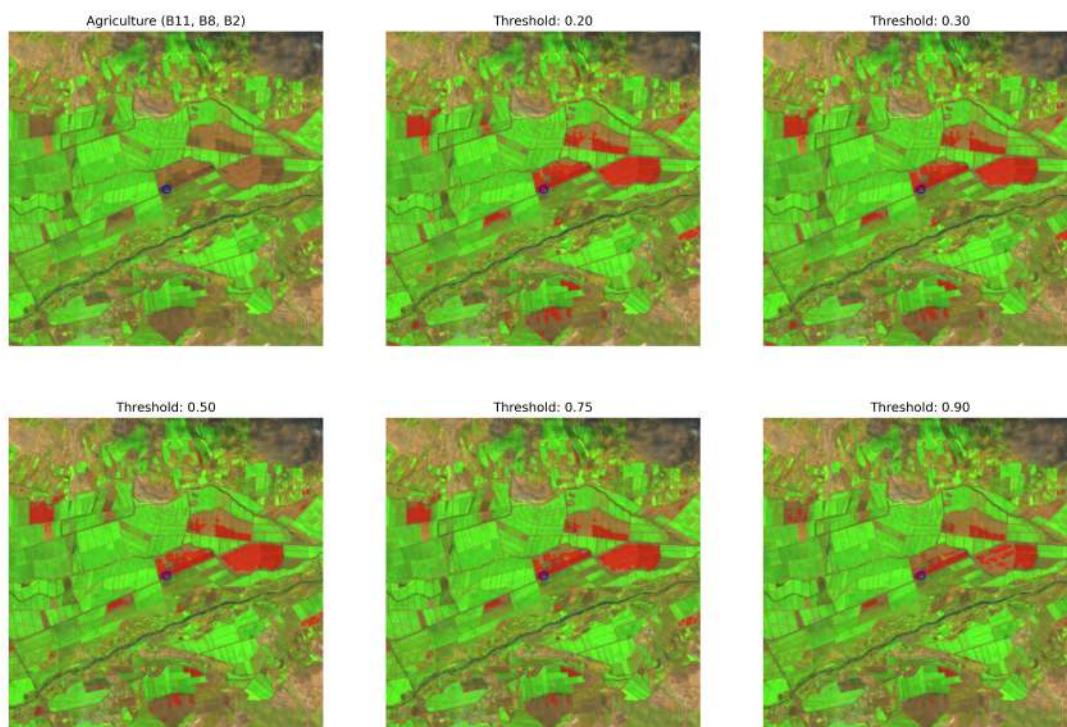
Nota. Elaboración propia.

La emergencia **0456** en los campo de cultivo Bambas 2, perteneciente a Agroindustrial Laredo S.A.A., reportó una quema accidental que afectó 10 hectáreas de campo. En este caso, los umbrales más bajos (0.20 - 0.30) tendieron a sobrestimar el área de quema debido a la reciente actividad y la presencia de suelo desnudo. Por lo tanto, un umbral alto de 0.90 se considera el más adecuado para este escenario logrando delimitar un área de 10.29 ha (Figura 51).

Figura 51

Comparación de umbrales de probabilidad para la emergencia 0456.

Emergency - 0456 on 2021/05/02 | Sentinel-2 (2021/05/06)



Nota. Elaboración propia.

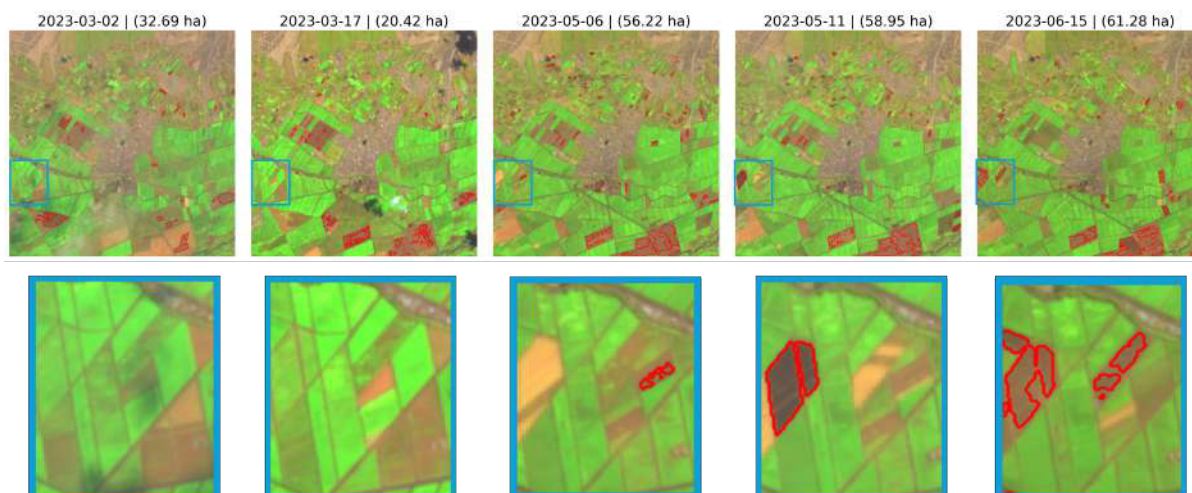
5.1.4. Inferencia en el área piloto.

El área analizada corresponde a los fundos azucareros cercanos a la ciudad de Laredo, durante un período reciente que abarca desde enero de 2023 hasta agosto de 2024. A partir de un total de 30 imágenes disponibles para esta área, se han generado tres series temporales.

5.1.4.1. Marzo a junio de 2023. Durante los meses de marzo y junio de 2023 (Figura 52), el análisis del 11 de mayo identificó 5.65 ha de áreas quemadas pertenecientes a dos campos de cultivo contiguos (recuadro celeste). Estas áreas, que cinco días antes eran suelos desnudos (zonas amarillentas), fueron requemadas debido a la presencia de paja o residuos de caña de azúcar. Para el 15 de junio, las áreas quemadas se vuelven más tenues, indicando que el suelo ha sido reacondicionado y preparado para la siembra.

Figura 52

Resultados de la inferencia en el área piloto (Marzo a Junio de 2023).



Nota. Los valores de umbral utilizados fueron 0.85, 0.98, 0.75, 0.75, 0.75 respectivamente.

Elaboración propia.

5.1.4.2. Octubre a noviembre de 2023. Durante los meses de octubre y noviembre de 2023 (Figura 53), el análisis del 18 de octubre reveló 4.88 ha de áreas quemadas recientemente en cuatro campo de cultivos cercanos (recuadro celeste), que hace cinco días eran suelo desnudo, similar al caso anterior fue producto de la requema. En las tres fechas posteriores, estas áreas, aunque con menor intensidad, continúan siendo detectadas por el modelo.

Para el 22 de noviembre, se identificó un área con quema activa de 2.94 ha, la cual cinco días después se observa con un tono más tenue y una extensión de 4.61 ha.

Figura 53

Resultados de la inferencia en el área piloto (Octubre a Noviembre de 2023).



Nota. Los valores de umbral utilizados fueron 0.85, 0.80, 0.80, 0.80, 0.80 respectivamente.

Elaboración propia.

5.1.4.3. Marzo a abril de 2024. El análisis temporal durante los meses de marzo y abril de 2024 (Figura 54), muestra diferentes escenarios de quema: el recuadro amarillo identifica una pequeña quema activa de 0.64 hectáreas el 15 de abril, que se atenúa cinco días después, pero sigue siendo detectada.

El recuadro azul muestra tres áreas quemadas con un total de 9.75 hectáreas, que también se vuelven más tenues cinco días después.

En el recuadro celeste se destaca áreas quemadas en dos campos contiguos el 26 de marzo, donde cinco días antes se observaba caña adulta, y un mes después de este evento, se encuentran recuperadas para la siembra y no son delimitadas por el modelo. Asimismo, otra área quemada se detecta el 15 de abril en un campo de cultivo donde se observa suelo desnudo producto de la cosecha, que cinco días después se muestra más tenue y se está reacondicionando para la siembra.

Figura 54

Resultados de la inferencia en el área piloto (Marzo a Abril de 2024).



Nota. Los valores de umbral utilizados fueron 0.85, 0.75, 0.75, 0.80, 0.80 respectivamente.

Elaboración propia.

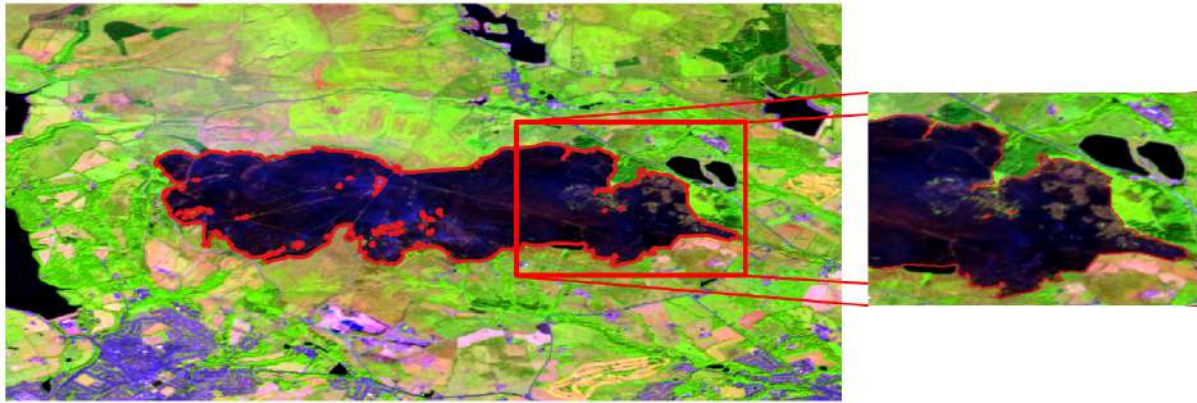
5.2. Discusión

5.2.0.1. Discusión de la generación y etiquetado del conjunto de datos. La detección de áreas quemadas en cultivos de caña de azúcar presenta un desafío significativo, principalmente debido a la necesidad de disponer de una base de datos confiable y de alta calidad. En este estudio, se propuso una metodología específica para la selección de imágenes con el objetivo de etiquetar correctamente las áreas quemadas.

A pesar de los esfuerzos por utilizar grandes conjuntos de datos globales de áreas quemadas, como los presentados en el trabajo de Arnaudo *et al.* (2023), la aplicación de estos datos a los modelos individuales enfrenta limitaciones significativas. Los estudios globales a menudo se centran en grandes incendios forestales que cubren extensiones superiores a las 5,000 hectáreas Seydi *et al.* (2022). Este enfoque no es el ideal para el contexto de la detección de quema en cultivos de caña de azúcar, donde las áreas quemadas son generalmente mucho más pequeñas, menores a 100 hectáreas y con condiciones locales propias. Como resultado, las etiquetas utilizadas en estudios globales no logran capturar con precisión los detalles de áreas quemadas de menor escala (ver Figura 55), lo cual es crítico en este contexto.

Figura 55

Ejemplo de etiquetado de áreas quemadas a nivel mundial.



Nota. Se visualiza la etiqueta EMSR291_AOI02_01 de 714 ha de área quemada. Extraído de Arnaudo *et al.* (2023).

El etiquetado de nuestra base de datos se realizó mediante el uso de “scribbles”, lo cual en áreas no etiquetadas puede llevar a definiciones poco claras sobre qué debe considerarse como quema. Este problema es general en la segmentación de áreas quemadas como se discute en Knopp *et al.* (2020), donde la ambigüedad no solo se limita a las áreas parcialmente afectadas por el fuego, sino que también se extiende a las áreas quemadas con cierta antigüedad, donde es crucial definir en qué momento una zona previamente quemada se ha recuperado lo suficiente para dejar de ser clasificada como área quemada.

En el contexto nacional, la caña de azúcar no tiene un período de zafra único, lo que significa que la quema de caña y su cosecha pueden ocurrir en cualquier época del año (Helfgott, 2016); esto complica aún más la interpretación de las áreas quemadas, especialmente en la diferenciación con suelos desnudos que se están recuperando y reacondicionando para la siembra de nueva caña soca que no es visible hasta el primer mes en las etapas de macollamiento y crecimiento.

Durante el período de cosecha, las quemas y requemas son comunes, lo que implica que el suelo desnudo con residuos de caña (straw) suele preceder a la quema, al igual que ocurre con la caña madura como se ha visto en los resultados de la Sección 5.1.3. Por lo tanto, que la imagen anterior a la quema no muestre vegetación no excluye la posibilidad de que se trate de un área quemada (ver Figura 56).

Figura 56

Ejemplo de distintos estados del área de cultivo de caña de azúcar durante la época de cosecha.



Nota. La imagen muestra tres estados: caña de azúcar madura (adult sugarcane), suelo desnudo (bare soil) y residuos de caña (straw). Extraído de Campos *et al.* (2022).

5.2.0.2. Discusión de los resultados de los modelos base. En cuanto a los resultados del modelo U-Net, se observa que produce una segmentación de áreas quemadas más gruesa en comparación con una clasificación basada en píxeles, lo cual coincide con los hallazgos de Knopp *et al.* (2020). Esto se evidencia aún más con el modelo LightGBM delimitando de manera fina las áreas de cultivo extensas diferenciándolos de los píxeles que pertenecen a los caminos que cruzan los campos de cultivo.

Este estudio representa un primer intento en el campo de la teledetección de combinar LightGBM y U-Net mediante un enfoque de stacking. Según Zhang *et al.* (2022), el stacking permite combinar las fortalezas de diferentes modelos base, mejorando significativamente la precisión de las predicciones en comparación con el uso de modelos individuales.

Al utilizar U-Net, que captura bien las relaciones no lineales en las imágenes, junto con LightGBM, un potente algoritmo de aprendizaje automático para la clasificación por píxel, se pueden aprovechar las capacidades complementarias de ambos modelos, logrando un equilibrio más robusto y flexible en la detección de áreas quemadas en cultivos de caña de azúcar.

Sin embargo, el fenómeno de “salt n’ pepper”, como señala Sdraka *et al.* (2024), es un inconveniente significativo que introduce ruido en la clasificación de píxeles y dificulta una delimitación precisa de las áreas quemadas. Este error se propaga a partir de los índices espectrales y afecta especialmente a los modelos de aprendizaje automático, como LightGBM y la

regresión logística, provocando confusión entre áreas quemadas y superficies similares, como cuerpos de agua, sombras y estructuras urbanas.

Aunque el modelo ensamblado de stacking, que combina U-Net como algoritmo de deep learning, puede mitigar este problema en cierta medida, sigue siendo susceptible a este tipo de ruido, especialmente en situaciones con alta confusión espectral o condiciones de imagen complejas.

Las variables del modelo LightGBM se alinean con la literatura existente sobre la detección de áreas quemadas, en la que la mayoría de los estudios se enfocan en utilizar bandas visibles, así como bandas en el rango del infrarrojo cercano, borde rojo, y de onda corta debido a su mayor sensibilidad para detectar quemas, como se ha demostrado en investigaciones previas (Quintano *et al.*, 2011; Van Dijk *et al.*, 2021). Además, es habitual incluir índices como el NDVI y el NBR para mejorar la precisión en la identificación de áreas quemadas (Lee *et al.*, 2022).

La mayor relevancia de la distancia a las coberturas agrícolas en el modelo se debe a su enfoque en priorizar las quemas dentro de las áreas de cultivos y diferenciar los terrenos eriazos con suelos desnudos ajenos a las áreas de cultivo.

5.2.0.3. Discusión de los resultados del modelo Stacking y su validación. El modelo ensamblado de Stacking supera cuantitativamente a cada uno de estos modelos individuales (Tabla 10). En particular, posee un **F1-Score** de 90.6 %, lo cual es adecuado para garantizar valores consistentes y reducir el impacto de falsas alarmas al combinar múltiples modelos de detección. Como se menciona en estudios anteriores, el uso de técnicas de ensemble learning puede mejorar significativamente la precisión y robustez en la detección de áreas quemadas (Lee *et al.*, 2022)

Mediante los reportes del OEFA, se validó que el modelo muestra un buen rendimiento con la mayoría de los umbrales. Un mayor umbral ayuda a eliminar falsos positivos, lo cual es útil en áreas donde la cobertura es similar, como suelo desnudo o sombras de nubes, y se requiere discriminar la quema de manera consistente, como en el caso de la emergencia **0546**.

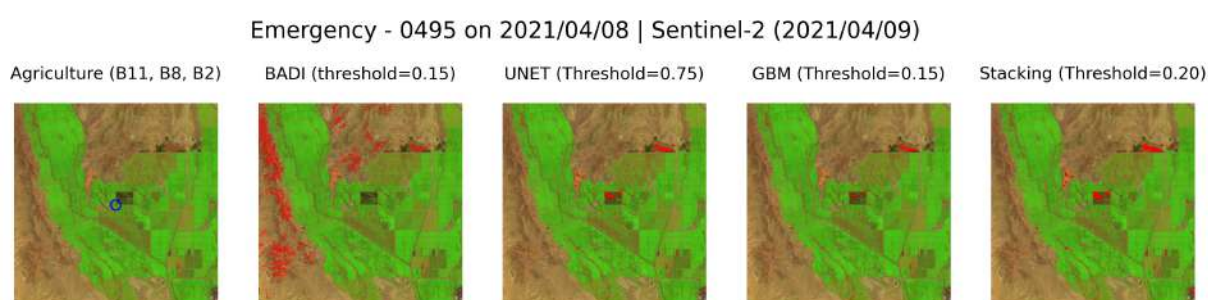
Por otro lado, umbrales más bajos son más efectivos para delimitar correctamente áreas quemadas en zonas con mezcla de vegetación, como se observa con el umbral de la emergencia

0068.

En la Figura 57, se evidencia que el modelo ensamblado está alineado con la información proporcionada por la OEFA y supera a los modelos utilizados basados en reglas más comunes, como el BADI, mostrando su eficacia en escenarios complejos y su capacidad para integrar diferentes fuentes de información para una mejor detección y delimitación de áreas quemadas.

Figura 57

Ejemplo de la detección de áreas quemadas en la costa norte y centro del Perú.



Nota. Se visualiza la detección de áreas quemadas en tres áreas de validación con diferentes umbrales de probabilidad. Elaboración propia.

5.2.0.4. Discusión de la inferencia en el área piloto. Inicialmente, al aplicar un filtro de nubosidad inferior al 10 % para seleccionar imágenes satelitales de Sentinel-2 en el período de enero de 2023 a agosto de 2024, se lograron obtener solo 96 imágenes de las 222 disponibles, que se redujeron a 30 de acuerdo a los criterios de la Sección 4.4.10. Esta limitación se debe a la alta presencia de nubosidad en la región de estudio, lo que complica la obtención de imágenes satelitales claras y de alta calidad. A pesar de estas restricciones, el modelo ensamblado fue capaz de detectar áreas quemadas en la región piloto utilizando las imágenes seleccionadas.

En la inferencia del área piloto, se observó que el modelo ensamblado es especialmente eficaz para detectar la ubicación de eventos de quema, resultando útil para identificar y delimitar tanto quemas activas como recientes, principalmente en áreas mayores a 1 ha. El modelo se distingue claramente al diferenciar áreas quemadas de otras superficies, como zonas urbanas, cuerpos de agua y suelo desnudo fuera de las áreas de cultivo, lo que proporciona una alta fiabilidad en la toma de decisiones en estos escenarios.

El uso de umbrales altos de probabilidad en la inferencia permitió minimizar los falsos positivos y mejorar la precisión en la detección de áreas quemadas, lo cual es crucial para la

toma de decisiones. Sin embargo, dependiendo de la situación, se puede ajustar el umbral para priorizar una detección más amplia o una más precisa. Esta elección depende de si se prefiere obtener más puntos de datos para un análisis exhaustivo o si se requiere mayor exactitud en la identificación de áreas quemadas.

La decisión final debe alinearse con las necesidades de fiscalización y monitoreo de las empresas azucareras administradas por el OEFA.

5.3. Conclusiones y Recomendaciones

5.3.1. Conclusiones

Este estudio demuestra que es posible desarrollar un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar. Al combinar el aprendizaje profundo con métodos de clasificación por árboles de decisión, el modelo ensamblado fue capaz de detectar y delimitar eficazmente las áreas quemadas alcanzando una efectividad del 100 % en la identificación de las áreas quemadas correspondiente a los casos de validación reportados por el OEFA, utilizando un umbral óptimo de 0.20.

El conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado destinado a detectar la quema de caña de azúcar, está consituido por 1,054 imágenes satelitales Sentinel-2, junto con índices de vegetación como el NDVI y NDWI, y variables topográficas como la pendiente y la distancia a coberturas agrícolas. Además, se incluyeron etiquetas de áreas quemadas, generadas mediante una metodología específica detallada que combina la detección temporal de quema basada en reglas y la validación visual de las imágenes detectadas.

Los algoritmos que mejor se ensamblan al modelo para detectar la quema de caña de azúcar son los de aprendizaje profundo y de clasificación por píxeles. En particular, U-Net, un modelo de deep learning que captura bien las relaciones no lineales en las imágenes, y LightGBM, un potente algoritmo de aprendizaje automático basado en árboles de decisión, han demostrado ser complementarios en la detección de áreas quemadas.

Como resultado, el modelo de aprendizaje ensamblado presenta un desempeño superior en comparación con los algoritmos individuales según las métricas de evaluación: 82.8 % en IoU, 86.4 % en precisión, 95.3 % en recall, 90.6 % en F1-Score y 87.5 % en coeficiente Kappa.

El modelo de Stacking supera consistentemente a U-Net y LightGBM en cada una de estas métricas. Además, los resultados visuales muestran una delimitación más equilibrada y coherente de las áreas quemadas debido a capacidad de combinar las fortalezas de ambos enfoques, heredando la coherencia espacial de U-Net y la potente capacidad de clasificación a nivel de píxel de LightGBM.

5.3.2. Recomendaciones

1. **Realizar estudios específicos sobre el ciclo de cultivo de la caña de azúcar:** Implementar investigaciones detalladas utilizando imágenes satelitales para monitorear todas las etapas del proceso de cultivo de caña de azúcar, desde la preparación del terreno hasta la cosecha, para identificar y gestionar mejor las quemadas.
2. **Incluir detalles en los reportes de emergencia y de recuperación de campos:** Incorporar en los reportes de emergencia datos específicos sobre el área quemada estimada, junto con información detallada proporcionada por los administrados, como cronogramas de fechas planificadas para las quemadas, la extensión estimada de las áreas afectadas por las quemadas una vez realizadas, y el tiempo estimado de recuperación de las tierras hasta que estén listas para una nueva siembra. Esta información adicional permitirá una validación más precisa y temporalmente coherente de los modelos mediante el uso de imágenes satelitales, mejorando así la capacidad de monitoreo y gestión de las áreas quemadas.
3. **Aumentar la base datos con nuevas imágenes en diferentes condiciones:** Ampliar la base de datos con imágenes de satélite en diferentes condiciones climáticas, de iluminación y diversas con presencia de superficies similares a la quema en términos de reflectancia, para mejorar la generalización del modelo y su capacidad para detectar áreas quemadas en diversas situaciones. Asimismo, se recomienda explorar otros métodos de etiquetado, incorporando la información de los reportes de emergencia y de recuperación de campos mencionados en el punto anterior, con el fin de incrementar la cantidad y la calidad de los datos disponibles.

4. **Actualización de los Instrumentos de Gestión Ambiental:** Revisar y actualizar los Instrumentos de Gestión Ambiental (IGA) de las empresas azucareras, incluyendo los Planes de Adecuación y Manejo Ambiental (PAMA) para empresas en operación y los Estudios de Impacto Ambiental (EIA) para la ampliación, modificación o formalización de sus operaciones. Esta actualización debe incorporar la implementación de la cosecha mecanizada en verde como una alternativa viable a la quema de caña de azúcar.

CAPÍTULO VI

REFERENCIAS Y ANEXOS

6.1. Referencias bibliográficas

- AGROAURORA S.A.C. (2020). Actualización del Plan de Manejo Ambiental (PMA) del Estudio de Impacto Ambiental Semidetallado (EIASd) del proyecto "CULTIVOS DE CAÑA DE AZÚCAR Y ELABORACIÓN DE AZÚCAR AGRIAURORA S.A.C. Technical report.
- Agroindustrial Laredo S.A.A. (2000). Programa de Adecuación y Manejo Ambiental (PAMA).
- Agroindustrias San Jacinto S.A.A. (2014). Programa de Adecuación y Manejo Ambiental (PAMA) - Capítulo VII - Descripción de la actividad agrícola. Technical report.
- AGROLMOS S.A. (2021). Modificación del Estudio de Impacto Ambiental Semidetallado del Proyecto Agroindustrial AGROLMOS S.A. Technical report.
- Al-Dabbagh, A. M. e Ilyas, M. (2023). Uni-temporal Sentinel-2 imagery for wildfire detection using deep learning semantic segmentation models. *Geomatics, Natural Hazards and Risk*, 14(1):2196370. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/19475705.2023.2196370>.
- Alcaras, E., Costantino, D., Guastaferro, F., Parente, C., y Pepe, M. (2022). Normalized burn ratio plus (nbr+): A new index for sentinel-2 imagery. *Remote Sensing*, 14(7).
- Amazon Web Services (2024). How XGBoost Works - Amazon SageMaker.
- Arbex, M. A., Martins, L. C., De Oliveira, R. C., Pereira, L. A. A., Arbex, F. F., Cancado, J. E. D., Saldiva, P. H. N., y Braga, A. L. F. (2007). Air pollution from biomass burning and asthma hospital admissions in a sugar cane plantation area in Brazil. *Journal of Epidemiology & Community Health*, 61(5):395–400.
- Arnaudo, E., Barco, L., Merlo, M., y Rossi, C. (2023). Robust Burned Area Delineation through Multitask Learning.
- Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Bautista, L., Yali, R., Flores, A.,

- Díaz, L., Cuenca, N., Espinoza, W., Prudencio, F., Llactayo, V., Montero, D., Sudmanns, M., Tiede, D., Mateo-García, G., y Gómez-Chova, L. (2022). CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Scientific Data*, 9(1):782.
- Añamuro-Luque, H. H., Larico-Mamani, C. J., Ruiz-Vásquez, C. E., Monteza-Rosales, O. M., y Quiñonez-Choquecota, J. (2020). ANÁLISIS DE INCENDIOS FORESTALES EN PAJONALES ANDINOS UTILIZANDO SISTEMAS DE INFORMACIÓN GEOGRÁFICA Y TELEDETECCIÓN, MACARI – PUNO. *REVISTA EL CEPROSIMAD*, 8(2):30–37. Number: 2.
- Balas, V. E., Kumar, R., y Srivastava, R., editores (2020). *Recent Trends and Advances in Artificial Intelligence and Internet of Things*, volumen 172 de *Intelligent Systems Reference Library*. Springer International Publishing, Cham.
- Cambrin, D. R., Colomba, L., y Garza, P. (2023). CaBuAr: California burned areas dataset for delineation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):106–113.
- Campos, L. R., Demattê, J. A., Bellinaso, H., Poppiel, R. R., Greschuk, L. T., Rizzo, R., Rosin, N. A., y Rosas, J. T. F. (2022). Detection of bare soils in sugarcane areas by temporal satellite images: A monitoring technique for soil security. *Soil Security*, 7:100057.
- Cançado, J. E., Saldiva, P. H., Pereira, L. A., Lara, L. B., Artaxo, P., Martinelli, L. A., Arbex, M. A., Zanobetti, A., y Braga, A. L. (2006). The Impact of Sugar Cane-Burning Emissions on the Respiratory System of Children and the Elderly. *Environmental Health Perspectives*, 114(5):725–729.
- Casa Grande S.A.A. (2018). *Actualización del Plan de Manejo Ambiental del Programa de Adecuación y Manejo Ambiental (PAMA) de la empresa CASA GRANDE S.A.A.*
- Chuvieco, E., Mouillot, F., Van Der Werf, G. R., San Miguel, J., Tanase, M., Koutsias, N., García, M., Yebra, M., Padilla, M., Gitas, I., Heil, A., Hawbaker, T. J., y Giglio, L. (2019). Historical background and current developments for mapping burned area from satellite Earth observation. *Remote Sensing of Environment*, 225:45–64.

- Colomba, L., Farasin, A., Monaco, S., Greco, S., Garza, P., Apiletti, D., Baralis, E., y Cerquitelli, T. (2022). A Dataset for Burned Area Delineation and Severity Estimation from Satellite Imagery. En *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 3893–3897, New York, NY, USA. Association for Computing Machinery.
- Congreso de la República (2020). LEY QUE REGULA LA GESTIÓN Y MANEJO DEL CULTIVO DE LA CAÑA DE AZÚCAR, EN FORMA SANITARIA Y AMBIENTALMENTE ADECUADA, PROHIBIENDO SU QUEMA Y PROMOVRIENDO LA MINIMIZACIÓN DE SUS RESIDUOS SÓLIDOS.
- Csurka, G., Volpi, R., y Chidlovskii, B. (2023). Semantic image segmentation: Two decades of research.
- De Andrade, S. J., Cristale, J., Silva, F. S., Julião Zocolo, G., y Marchi, M. R. (2010). Contribution of sugar-cane harvesting season to atmospheric contamination by polycyclic aromatic hydrocarbons (PAHs) in Araraquara city, Southeast Brazil. *Atmospheric Environment*, 44(24):2913–2919.
- FAO (2023). *World Food and Agriculture – Statistical Yearbook 2023*.
- Farhadi, H., Ebadi, H., y Kiani, A. (2023). BADI: A NOVEL BURNED AREA DETECTION INDEX FOR SENTINEL-2 IMAGERY USING GOOGLE EARTH ENGINE PLATFORM. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4-W1-2022:179–186. Conference Name: ISPRS WG IV/3
ISPRS GeoSpatial Conference 2022, Joint 6th Sensors and Models in Photogrammetry and Remote Sensing (SMPR) and 4th Geospatial Information Research (GIResearch) Conferences - 19–22 February 2023, Tehran, Iran (virtual) Publisher: Copernicus GmbH.
- Farrukh, M. (2019). Modeling on Feature Vectors in Compressed Spaces by the use of Neural network techniques. Technical report.
- Foody, G. M. (2020). Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment*.

- França, D. D. A., Longo, K. M., Neto, T. G. S., Santos, J. C., Freitas, S. R., Rudorff, B. F. T., Cortez, E. V., Anselmo, E., y Carvalho, J. A. (2012). Pre-Harvest Sugarcane Burning: Determination of Emission Factors through Laboratory Measurements. *Atmosphere*, 3(1):164–180.
- González Vásquez, A. M. (2022). *APLICACIÓN DEL ÍNDICE NORMALIZADO DE ÁREA QUEMADA (NBR) Y SISTEMAS DE INFORMACIÓN GEOGRÁFICA PARA ESTIMAR LA SEVERIDAD DE LOS INCENDIOS FORESTALES EN EL DEPARTAMENTO DE CUSCO*. Tesis doctoral, Universidad Nacional Federico Villareal, Lima.
- Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep Learning*. MIT Press.
- Hamzeh, S., Naseri, A. A., AlaviPanah, S. K., Bartholomeus, H., y Herold, M. (2016). Assessing the accuracy of hyperspectral and multispectral satellite imagery for categorical and quantitative mapping of salinity stress in sugarcane fields. *International journal of applied earth observation and geoinformation*, 52:412–421.
- Helfgott, S. E. (2016). *El cultivo de la caña de azúcar en la costa peruana*. Universidad Nacional Agraria La Molina. Google-Books-ID: UnbttAEACAAJ.
- Hosmer Jr, D. W., Lemeshow, S., y Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Hu, X., Wen, H., Zhang, P., Yuen, K.-V., y Zhong, P. (2024). Near Real-Time Burned Area Progression Mapping With Multispectral Data Using Ensemble Learning. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5.
- Instituto Nacional de Estadística e Informática (2022). PERÚ: Proyecciones de Población Total según Departamento, Provincia y Distrito, 2018 – 2022 Boletín Especial N°27.
- Jimenez Dios, A. D. y García Núñez, A. (2020). *Resiliencia de la cobertura vegetal post incendio forestal aplicando índices espectrales de vegetación – Distrito de Chulucanas – Departamento de Piura, 2019*. Tesis doctoral, UNIVERSIDAD NACIONAL DE TUMBES, Tumbes.
- Junpen, A., Pansuk, J., y Garivait, S. (2020). Estimation of Reduced Air Emissions as a Result of

- the Implementation of the Measure to Reduce Burned Sugarcane in Thailand. *Atmosphere*, 11(4):366.
- Katsaros, D., Fragkou, E., y Papakostas, D. (2021). Backpropagation for Convolutional Neural Networks. *Department of Electrical & Computer Engineering, University of Thessaly*.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., y Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree.
- Knopp, L., Wieland, M., Rättich, M., y Martinis, S. (2020). A Deep Learning Approach for Burned Area Segmentation with Sentinel-2 Data. *Remote Sensing*, 12(15):2422. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.
- Lee, C., Park, S., Kim, T., Liu, S., Md Reba, M. N., Oh, J., y Han, Y. (2022). Machine Learning-Based Forest Burned Area Detection with Various Input Variables: A Case Study of South Korea. *Applied Sciences*, 12(19):10077.
- Luo, S. y Chen, T. (2020). Two Derivative Algorithms of Gradient Boosting Decision Tree for Silicon Content in Blast Furnace System Prediction. *IEEE Access*, 8:196112–196122.
- Luo, Y., Wang, Z., Huang, Z., Yang, Y., y Zhao, C. (2018). Coarse-to-Fine Annotation Enrichment for Semantic Segmentation Learning. En *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 237–246. arXiv:1808.07209 [cs].
- Luque, J. P. C. y Echeverria, E. L. (2010). Impacto ambiental generado por la quema de la caña de azúcar en laredo – trujillo.
- Machine Learning Glossary (2017). Loss Functions — ML Glossary documentation.
- Matsuda, M., Braga, A. L., Marquezini, M. V., Monteiro, M. L., Saldiva, P. H., y De Santos, U. (2020). Occupational effect of sugarcane biomass burning on the conjunctival mucin profile of harvest workers and residents of an adjacent town - A Brazilian panel study. *Experimental Eye Research*, 190:107889.
- meanxai (2023). [MXML-12-03] Light GBM [3/4] - Exclusive Feature Bundling (EFB) - Greedy Bundling.

- Mienye, I. D. y Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, 10:99129–99149.
- Militino, A. F., Goyena, H., Pérez-Goya, U., y Ugarte, M. D. (2024). Logistic regression versus XGBoost for detecting burned areas using satellite images. *Environmental and Ecological Statistics*, 31(1):57–77.
- Müller, D., Soto-Rey, I., y Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210.
- Ng, A. (2023). Semantic Segmentation with U-Net.
- Nowell, H. K., Wirks, C., Val Martin, M., Van Donkelaar, A., Martin, R. V., Uejio, C. K., y Holmes, C. D. (2022). Impacts of Sugarcane Fires on Air Quality and Public Health in South Florida. *Environmental Health Perspectives*, 130(8):087004.
- OECD-FAO (2023). OECD-FAO Agricultural Outlook 2023-2032.
- OEFA, editor (2023). *Caña de azúcar: ruta hacia la cosecha en verde. Determinación de distancias mínimas para la quema de caña de azúcar en las localidades de La huaca y La Rinconada (provincia de Paita, Piura) mediante un modelamiento de dispersión de material particulado en el año 2018*. Lima.
- Ongeri, D. y Kenduiywo, B. K. (2020). BURNT AREA DETECTION USING MEDIUM RESOLUTION SENTINEL 2 AND LANDSAT 8 SATELLITES. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B5-2020:131–137.
- Organización Mundial de la Salud (2021). Directrices mundiales de la OMS sobre la calidad del aire. Technical report.
- Park, H.-A. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal of Korean Academy of Nursing*, 43(2):154.
- Pereira, J. M. C., Sá, A. C. L., Sousa, A. M. O., Silva, J. M. N., Santos, T. N., y Carreiras, J.

- M. B. (1999). *Spectral characterisation and discrimination of burnt areas*, pp. 123–138. Springer Berlin Heidelberg, Berlin, Heidelberg.
- PERUCAÑA (2022). Industria de la caña del Perú Informe Anual 2022. Technical report.
- Pestana, P. R. D. S., Braga, A. L. F., Ramos, E. M. C., Oliveira, A. F. D., Osadnik, C. R., Ferreira, A. D., y Ramos, D. (2017). Effects of air pollution caused by sugarcane burning in Western São Paulo on the cardiovascular system. *Revista de Saúde Pública*, 51(0).
- Pipicano, A. F. V. y Restrepo, J. A. V. (2022). ANÁLISIS ECONÓMICO DE ALTERNATIVAS DE DENSIFICACIÓN DE RESIDUO AGRÍCOLA DE COSECHA (RAC) DE CAÑA DE AZÚCAR PARA PROCESOS DE COMBUSTIÓN.
- Planetary Computer (2023). Sentinel-2 Level-2A.
- Powers, D. M. W. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061.
- Quintano, C., Fernández-Manso, A., Stein, A., y Bijker, W. (2011). Estimation of area burned by forest fires in mediterranean countries: A remote sensing data mining perspective. *Forest Ecology and Management*, 262(8):1597–1607.
- Ronneberger, O., Fischer, P., y Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.
- Roy, D., Boschetti, L., y Trigg, S. (2006). Remote Sensing of Fire Severity: Assessing the Performance of the Normalized Burn Ratio. *IEEE Geoscience and Remote Sensing Letters*, 3(1):112–116.
- Saini, Rashmi y Ghosh, S. (2017). Ensemble classifiers in remote sensing: A review. En *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1148–1152, Greater Noida. IEEE.
- Salas, G. L. (2010). La Quema de la Caña de Azúcar en Guanacaste Impacto Ambiental.
- Sdraka, M., Dimakos, A., Malounis, A., Ntasiou, Z., Karantzalos, K., Michail, D., y Papoutsis, I. (2024). FLOGA: A Machine-Learning-Ready Dataset, a Benchmark, and a Novel Deep Learning Model for Burnt Area Mapping With Sentinel-2. *IEEE Journal of Selected*

- Topics in Applied Earth Observations and Remote Sensing*, 17:7801–7824. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
- Sentinel Hub (2023). Simple rgb composites (sentinel-2).
- Seydi, S. T., Hasanlou, M., y Chanussot, J. (2022). *Burnt-Net*: Wildfire burned area mapping with single post-fire Sentinel-2 data and deep learning morphological neural network. *Ecological Indicators*, 140:108999.
- Shomeili, M. (2015). Sugarcane cultivation in iran.
- Silveira, H. C. S., Schmidt-Carrijo, M., Seidel, E. H., Scapulatempo-Neto, C., Longatto-Filho, A., Carvalho, A. L., Reis, R. M. V., y Saldiva, P. H. N. (2013). Emissions generated by sugarcane burning promote genotoxicity in rural workers: a case study in Barretos, Brazil. *Environmental Health*, 12(1):87.
- Singh, N., Thumar, V., Leua, A., y Chaudhari, D. (2018). Impact of Straw Burning Practices on Economics of Sugarcane in South Gujarat Region, India. *International Journal of Current Microbiology and Applied Sciences*.
- Somukesh, M. (2023). Apply a 2D Transposed Convolution Operation in PyTorch. Section: AI-ML-DS.
- Steinbach, T., Tan, K., Steinbach, y Kumar (2004). Data mining classification : Basic concepts , decision trees , and model evaluation.
- Sun, Y., Zhang, F., Lin, H., y xu, S. (2022). A Forest Fire Susceptibility Modeling Approach Based on Light Gradient Boosting Machine Algorithm. *Remote Sensing*, 14:4362.
- Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., Chavez-Urbiola, E. A., y Romero-Gonzalez, J. A. (2024). Loss Functions and Metrics in Deep Learning. arXiv:2307.02694 [cs].
- Tonbul, H., Yilmaz, E., y Kavzoglu, T. (2023). Comparative Analysis of Deep Learning and Machine Learning Models for Burned Area Estimation Using Sentinel-2 Image: A Case Study in Muğla-Bodrum, Turkey. pp. 1–5.

- Van Dijk, D., Shoaie, S., Van Leeuwen, T., y Veraverbeke, S. (2021). Spectral signature analysis of false positive burned area detection from agricultural harvests using Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 97:102296.
- Wali, R. (2022). Xtreme Margin: A Tunable Loss Function for Binary Classification. *Machine Learning*.
- Wei, D. (2024). Understanding Parameters in ML Training — Batch Size, Iteration, Epoch, Learning Rate.
- World Bank Group (2018). Agricultural Pollution Field Burning. *World Bank Group*.
- Yali, R. (2024). *Evaluación de modelos de segmentación semántica para el monitoreo de deslizamiento de tierra utilizando imágenes satelitales*. Tesis doctoral, PUCP, Lima.
- Zhang, Y., Liu, J., y Shen, W. (2022). A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences*, 12(17):8654.

6.2. Anexos

Tabla 11
Matriz de consistencia.

PREGUNTA GENERAL	OBJETIVO GENERAL	HIPÓTESIS GENERAL
¿Es posible desarrollar un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar?	Desarrollar un modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar a fin de mejorar su fiscalización.	El modelo de aprendizaje ensamblado para la detección de quema de caña de azúcar permitirá resultados con un gran porcentaje de efectividad.
Preguntas Específicas	Objetivos Específicos	Hipótesis Específicas
¿Cómo generar el conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado para detectar la quema de caña de azúcar?	Generar el conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado para detectar la quema de caña de azúcar.	La correcta generación del conjunto de datos necesario para el funcionamiento del modelo de aprendizaje ensamblado para detectar la quema de caña de azúcar optimizará la precisión del modelo.
¿Cuáles son los algoritmos que mejor se ensamblan al modelo para detectar la quema de caña de azúcar?	Determinar los algoritmos que mejor se ensamblan al modelo para detectar la quema de caña de azúcar.	Los algoritmos LightGBM y U-Net serán los que mejor se ensamblan al modelo para detectar la quema de caña de azúcar.

Continúa en la siguiente página ...

Tabla 11
(continuación)

Preguntas Específicas	Objetivos Específicos	Hipótesis Específicas
¿El modelo de aprendizaje ensamblado presenta un desempeño superior en comparación con los algoritmos individuales?	Determinar mediante métricas de evaluación si el modelo de aprendizaje ensamblado presenta un desempeño superior en comparación con los algoritmos individuales.	El modelo de aprendizaje ensamblado presentará un desempeño superior en comparación con los algoritmos individuales.

Figura 58*Resumen de las ROI's.*

ROI	Nro de Denuncias	Longitud central	Latitud central
ROI_0001_04	1	-79.99562711	-6.064082555
ROI_0003_03	7	-81.00710867	-4.911311097
ROI_0005_04	1	-79.99652254	-6.156361876
ROI_0011_01	6	-80.96099109	-4.872692172
ROI_0011_02	12	-80.9609884	-4.919011018
ROI_0011_03	1	-80.91481319	-4.872687908
ROI_0012_04	4	-80.91570412	-5.011146631
ROI_0013_02	4	-80.96018527	-4.826881078
ROI_0013_04	21	-80.91401049	-4.826876799
ROI_0020_01	1	-80.86879286	-4.834755346
ROI_0020_04	1	-80.82260558	-4.881063518
ROI_0021_01	5	-80.86979688	-4.949505501
ROI_0024_02	1	-80.77798448	-4.996637233
ROI_0024_03	1	-80.73181716	-4.950301736
ROI_0030_04	20	-80.60949864	-4.882875788
ROI_0040_02	9	-80.56377293	-4.883663964
ROI_0056_04	8	-79.82778042	-6.827387201
ROI_0057_03	1	-79.82878899	-6.873380612
ROI_0061_02	6	-79.7808008	-6.736110607
ROI_0062_01	17	-79.78182118	-6.782116101
ROI_0062_02	7	-79.78170444	-6.828421277
ROI_0062_03	4	-79.73549791	-6.781997609
ROI_0065_01	1	-79.68952481	-6.783031358
ROI_0065_03	3	-79.64320314	-6.782904041
ROI_0065_04	1	-79.64307311	-6.829206681
ROI_0069_01	2	-79.59712839	-6.783916682
ROI_0069_02	2	-79.59699392	-6.830218418
ROI_0069_03	1	-79.55080845	-6.783780528
ROI_0083_04	4	-79.28220085	-7.756273431
ROI_0085_01	5	-79.29941053	-7.895218639
ROI_0085_02	3	-79.29922061	-7.941511523
ROI_0085_03	2	-79.25298216	-7.895027271

ROI_0086_02	2	-79.23510451	-7.665011181
ROI_0087_01	2	-79.23620077	-7.711090455
ROI_0087_02	2	-79.23600843	-7.757382187
ROI_0087_03	5	-79.18979429	-7.71089667
ROI_0087_04	21	-79.18959689	-7.757187225
ROI_0088_01	18	-79.23710697	-7.803471637
ROI_0088_02	2	-79.23691241	-7.849763191
ROI_0088_03	4	-79.19069034	-7.803275645
ROI_0088_04	1	-79.19049066	-7.849566022
ROI_0089_01	20	-79.2068146	-7.896164954
ROI_0089_02	6	-79.20661432	-7.942455539
ROI_0090_01	1	-79.14360468	-7.712006692
ROI_0090_02	4	-79.14340222	-7.758296043
ROI_0090_03	1	-79.09720051	-7.711802845
ROI_0090_04	4	-79.096993	-7.758090958
ROI_0091_01	1	-79.14451094	-7.804397934
ROI_0091_02	11	-79.14430613	-7.850687107
ROI_0093_01	1	-79.05100852	-7.71289299
ROI_0093_03	2	-79.0046068	-7.712679081
ROI_0094_01	6	-79.0546342	-8.082488192
ROI_0094_02	13	-79.05441175	-8.128774321
ROI_0094_03	49	-79.00819111	-8.082264512
ROI_0094_04	2	-79.00796336	-8.128549344
ROI_0096_01	59	-78.96183824	-8.083404803
ROI_0096_02	7	-78.96160516	-8.129688306
ROI_0096_03	5	-78.91539771	-8.083170558
ROI_0096_04	7	-78.91515932	-8.129452702
ROI_0097_04	1	-78.91704593	-8.31425995
ROI_0098_04	1	-78.82156203	-8.037917446
ROI_0099_01	4	-78.86914222	-8.084291469
ROI_0103_04	5	-78.76783244	-8.40811741
ROI_0118_01	5	-78.47145185	-9.256982837
ROI_0118_03	3	-78.42488315	-9.256651161

ROI_0119_02	3	-78.41321109	-9.213768822
ROI_0119_03	6	-78.3669922	-9.167169982
ROI_0119_04	2	-78.36665022	-9.213431151
ROI_0120_01	3	-78.3217488	-9.138372981
ROI_0120_02	1	-78.32140206	-9.184632519
ROI_0120_03	13	-78.27520123	-9.138026322
ROI_0120_04	1	-78.27484848	-9.184284075
ROI_0123_03	6	-77.81976814	-10.64794542
ROI_0123_04	4	-77.81928717	-10.69417994
ROI_0124_01	1	-77.77550875	-10.59132411
ROI_0124_02	3	-77.77502374	-10.63755679
ROI_0124_04	3	-77.72828809	-10.63707243
ROI_0125_01	14	-77.77631793	-10.68382624
ROI_0125_02	5	-77.77582871	-10.73005868
ROI_0125_03	6	-77.72957516	-10.68333999
ROI_0127_02	7	-77.68151455	-10.63836159
ROI_0127_04	3	-77.6347832	-10.63786327
ROI_0128_01	2	-77.68282289	-10.68463543
ROI_0128_03	1	-77.63608442	-10.68413517
ROI_0138_04	3	-77.35782093	-11.10272343
ROI_0141_01	1	-77.31198537	-11.10138372
ROI_0144_03	2	-79.45771678	-6.704937831
ROI_0144_04	1	-79.45757068	-6.751236801
ROI_0145_02	1	-77.77421883	-10.54505489

Nota. Elaboración propia.