Santiago de Cali, 25 de mayo del 2026

Doctor
Diego Luis Linares O.
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "**Modelos de aprendizaje automático para evaluar la relación entre variables clave en las emisiones reportadas en iniciativas de mitigación climática**", el cual fue realizado por el estudiante Daniel Felipe Acosta Méndez con código **1019123384** pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de Julián Gil.

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

Atentamente,

julián gil gonzález

Firmado digitalmente por julián gil gonzález
Fecha: 2026.05.25 10:37:06 -05'00'

Daniel Felipe Acosta Méndez

Julián Gil González

C.C. 1019123384 de Bogotá DC

C.C. 1088286439 de Pereira

FICHA RESUMEN
PROYECTO APLICADO – MAESTRIA EN CIENCIA DE DATOS

POSIBLE TÍTULO: Modelos de aprendizaje automático para evaluar la relación entre variables clave en las emisiones reportadas en iniciativas de mitigación climática

1. ÁREA DE TRABAJO: Ciencia de datos, Cambio climático, Modelado
2. TIPO DE PROYECTO: Investigación aplicada
3. ESTUDIANTE(S): Daniel Felipe Acosta Méndez
4. CORREO ELECTRÓNICO: danielmendez19960@gmail.com
5. DIRECCIÓN Y TELEFONO: Crr 147 a # 142 f 39 Bogotá - 3028489799
6. DIRECTOR: Julián Gil
7. VINCULACIÓN DEL DIRECTOR: Facultad de Ingeniería y Ciencias – Departamento de Electrónica y Ciencias de la Computación
8. CORREO ELECTRÓNICO DEL DIRECTOR: julian.gil@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): No aplica
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Ministerio de Ambiente y Desarrollo Sostenible
11. OTROS GRUPOS O EMPRESAS: No aplica
12. PALABRAS CLAVE (al menos 5): Cambio climático, Iniciativas de mitigación, Aprendizaje automático, Regresión , NDC, Acuerdo de París.
13. FECHA DE INICIO: 02/06/2025
14. DURACIÓN ESTIMADA (En meses): 12 meses

15. RESUMEN:

Este proyecto creó un modelo predictivo que usa el aprendizaje automático para determinar el factor de emisión de gases de efecto invernadero (FACTOR_EMISION, tCO_2e) registrado en el sistema RENARE del Ministerio colombiano de Ambiente y Desarrollo Sostenible por las acciones de mitigación. Se creó un pipeline modular que une la justificación IPCC (42 variables) con la ingeniería de características, la división por sectores (FORESTAL, ENERGÍA, OTROS), el análisis de interpretabilidad SHAP y la optimización bayesiana mediante Optuna (150 pruebas). Los modelos seleccionados por segmento (CatBoost, LightGBM, RandomForest y XGBoost) alcanzaron un R^2 de entre 0,843 y 0,873 en el conjunto de prueba; para ello, se empleó la validación cruzada 5-fold y el bootstrap con un total de 500 muestras. El análisis SHAP mostró que existe concordancia entre las variables más influyentes y la información climática del IPCC. Se encontraron 37 registros anómalos (0,74 % del conjunto de datos), lo que evidencia que el modelo puede ser una herramienta para la auditoría del sistema MRV en Colombia.

MAESTRÍA EN CIENCIA DE DATOS
DESARROLLO DE MODELO PREDICTIVO BASADO EN APRENDIZAJE AUTOMÁTICO PARA ESTIMAR FACTORES DE EMISIÓN EN LAS INICIATIVAS DE MITIGACIÓN DE CAMBIO CLIMATICO EN COLOMBIA

Daniel Felipe Acosta Méndez

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

Director
Julián Gil

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 25 DE 2026

TABLA DE CONTENIDO

INTRODUCCIÓN

En el cumplimiento de los compromisos climáticos adquiridos por el gobierno de Colombia de acuerdo a lo estipulado en el Acuerdo de París donde exige un análisis profundo sobre el impacto real de las iniciativas de mitigación del cambio climático implementadas en el país. Estas iniciativas registradas según su tipología REDD+ (Reducción de Emisiones por Deforestación y Degradación Evitada), PBDC (Proyectos Basados en Conservación), NAMAs (Acciones Nacionalmente Apropiadas de Mitigación), y MDL (Mecanismo de Desarrollo Limpio), tienen como propósito contribuir a la reducción de emisiones de gases de efecto invernadero (GEI) y están alineadas con las metas establecidas en las Contribuciones Determinadas a Nivel Nacional (NDC). Su registros, gestión y control se realizan mediante el Registro Nacional de Reducción de Emisiones de GEI (RENARE), la cual es una plataforma administrada por el Ministerio de Ambiente y Desarrollo Sostenible.

Si bien RENARE permite centralizar la información de estas iniciativas, aún no dispone de herramientas analíticas que permitan evaluar, desde una perspectiva cuantitativa, qué factores están asociados a su impacto climático. Esta limitación restringe el aprovechamiento estratégico de los datos disponibles, tanto para el análisis técnico como para la generación de insumos que respalden procesos de reporte y planeación. Por tal motivo, el presente proyecto planteó el desarrollo de un modelo de análisis basado en técnicas de regresión y aprendizaje automático supervisado, cuyo objetivo principal es la predicción de los factores de emisiones reportados por las actividades de las iniciativas de mitigación. A partir de esta predicción, se buscó identificar las relaciones entre variables tales como sociales, económicas, ambientales y el comportamiento de las emisiones, lo cual permite interpretar los resultados, con ello aportando evidencia útil a los equipos técnicos.

Esta herramienta, buscó generar insumos cuantitativos que respalden el análisis técnico, fortaleciendo la toma de decisiones desde la ciencia de datos y contribuyendo al seguimiento del cumplimiento de las metas climáticas establecidas por Colombia en el marco del Acuerdo de París.

Este documento, presenta los hallazgos del proyecto implementado, que se compone de dos elementos: (i) el desarrollo y validación del pipeline de machine learning, que obtuvo un R^2 mayor a 0,84 en todos los segmentos analizados; y (ii) el análisis de la interpretabilidad, el cual corroboró la consistencia entre las variables predictivas más determinantes y el conocimiento climático establecido

por el IPCC. Los resultados indican que la segmentación sectorial anterior al modelado es la estrategia más influyente en el rendimiento predictivo, por encima de lo que aporta la optimización de hiperparámetros.

1. DEFINICIÓN DEL PROBLEMA

En la actualidad no existen modelos analíticos que permitan explicar de forma sistemática qué factores inciden en el desempeño de las iniciativas de mitigación del cambio climático desarrolladas en Colombia. A pesar de que se implementan múltiples iniciativas los diferentes tipos de actividades de mitigación como lo son REDD+, PBDC, NAMAs y MDL, que cuentan con un marco normativo consolidado como la Ley 1931 de 2018 y la Resolución 1447 de 2018, aún se evidencia una brecha entre el registro de iniciativas, su evaluación técnica cuantitativa y análisis a mayor escala. Esta situación limita la capacidad institucional de fortalecer el sistema nacional de Medición, Reporte y Verificación (MRV) y de orientar la toma de decisiones basada en datos, lo que podría comprometer la forma en la que se miden las metas establecidas en NDC.

1.1 PLANTEAMIENTO DEL PROBLEMA

Colombia, como parte de sus compromisos internacionales frente al cambio climático, ha promovido múltiples iniciativas de mitigación implementadas por empresas del sector privado como por comunidades locales o resguardos indígenas. Muchas de estas acciones se desarrollan en el marco del mercado de carbono. Estas iniciativas buscan generar bonos de carbono mediante acciones orientadas a la reducción de emisiones de gases de efecto invernadero (GEI), bajo esquemas como REDD+ (Reducción de Emisiones por Deforestación y Degradación Evitada), PBDC (Proyectos Basados en Conservación), NAMAs (Acciones Nacionalmente Apropiadas de Mitigación) y MDL (Mecanismo de Desarrollo Limpio), entre otros [1],[2].

Para asegurar la trazabilidad, el control técnico y la transparencia de estas acciones, Colombia cuenta con el Registro Nacional de Reducción de Emisiones de GEI (RENARE), administrado por el Ministerio de Ambiente y Desarrollo Sostenible. Esta plataforma integra las iniciativas registradas y opera como una herramienta clave dentro del sistema nacional de Medición, Reporte y Verificación (MRV) de mitigación.

Si bien RENARE cumple con el propósito de centralizar la información de las iniciativas de mitigación implementadas en Colombia, actualmente no dispone de una herramienta analítica que permita examinar, desde una perspectiva cuantitativa, la relación entre las variables técnicas, sociales, económicas y ambientales de las iniciativas con los factores de emisiones de GEI que reportan como resultado de su ejecución. Esta limitación impide a los equipos técnicos responsables de la validación de iniciativas identificar patrones o comportamientos comunes entre iniciativas similares, dificultando el análisis comparativo entre iniciativas y la generación de criterios basados en evidencia para su revisión. Aunque la información se encuentra registrada, no existe una herramienta que permita establecer, con base en datos estructurados, qué factores se asocian con mayores niveles de reducción de emisiones.

Esta inexistencia analítica restringe el aprovechamiento estratégico de los datos registrados para otros fines institucionales, como el seguimiento técnico del cumplimiento de las metas de reducción de emisiones en el marco de las Contribuciones Determinadas a Nivel Nacional (NDC). Mediante una revisión exploratoria con profesionales del Ministerio de Ambiente y Desarrollo sostenible se identificó que Colombia aún se encuentra en proceso de presentar sus indicadores oficiales para reportes NDC, donde actualmente no existen sistemas consolidados que permitan generar estos resultados de manera automatizada o estructurada.

En la actualidad, el principal reto técnico en torno al análisis de las iniciativas de mitigación registradas en RENARE es la ausencia de una herramienta analítica que permita modelar y comprender las relaciones entre variables disponibles en los registros, como lo es la tipología de iniciativas, área de intervención, periodo crediticio, fuentes de financiamiento, entre otros con respecto a los factores de emisión. Esta limitación se presenta en un momento en el que Colombia avanza en la construcción de sus indicadores nacionales para el seguimiento de las NDC, y en el fortalecimiento de sus sistemas de información climática mediante la implementación de modelos de análisis de datos estructurados que faciliten su aprovechamiento sistemático mediante técnicas de aprendizaje automático, modelos estadísticos.

Debido a que la plataforma RENARE es una herramienta única en Colombia que genera centraliza la trazabilidad de las iniciativas de mitigación y considerando que las condiciones climáticas, normativas y metodológicas varían entre países, no existen referencias internacionales que puedan adaptarse de manera directa a nuestro contexto. Por lo tanto, el problema no reside únicamente en la disponibilidad de los datos, si no en la ausencia de capacidad analítica avanzada que permitan procesar dicha información

en conocimiento estratégico, por ello la ciencia de datos y en particular el aprendizaje automático supervisado ofrecen que permite modelar y predecir los factores de emisiones reportados, utilizando variables como el tipo de iniciativa área de intervención, periodo crediticio, esquema de financiamiento y las condiciones ambientales del territorio.

Estás técnicas permiten anticipar el comportamiento de las emisiones, también permiten identificar patrones y relaciones que actualmente no están siendo identificadas con facilidad para los equipos técnicos del país. Incorporar estos enfoques además de ser una opción metodológica, es una necesidad urgente para fortalecer el seguimiento a las metas climáticas nacionales y para transformar el RENARE en una plataforma de inteligencia climática

1.2 FORMULACIÓN DEL PROBLEMA

Actualmente Colombia ha registrado múltiples iniciativas de mitigación de cambio climático bajo esquemas como REDD+, PBDC, NAMA y MDL. Sin embargo, no se cuenta con herramientas analíticas que permitan modelar de forma sistemática la relación entre las características de las iniciativas como lo puede ser el potencial máximo de mitigación y el factor de emisiones que reportan como resultado de las iniciativas. Esta falta de modelos impide aprovechar los datos disponibles para generar evidencia cuantitativa que ayude a la toma de decisiones, análisis técnico y validación de información en el contexto de las metas climáticas establecidas en las NDC.

La ausencia de una herramienta analítica orientada a la evaluación de estos factores dificulta el aprovechamiento estratégico de los datos disponibles tanto en el RENARE como los estándares, limitando la capacidad institucional para identificar patrones comunes, establecer criterios técnicos comparables entre iniciativas y aportar evidencia en el seguimiento de las metas de reducción de emisiones. Frente a este escenario, se requiere una infraestructura analítica que permita estimar el comportamiento de los factores de emisión a partir de las características registradas en cada iniciativa

¿Cómo puede un modelo de aprendizaje automático supervisado predecir los factores de emisión de gases efecto invernadero reportados por las iniciativas de mitigación de cambio climático en Colombia, a partir de datos registradas en RENARE?

Preguntas de sistematización

- ¿Qué variables están registradas actualmente en el RENARE para caracterizar las iniciativas de mitigación?
- ¿Qué métodos de regresión y algoritmos de aprendizaje automático son adecuados para modelar la relación entre dichas variables y las emisiones reportadas?
- ¿Cuáles son las variables que tienen mayor peso en la explicación del comportamiento de las emisiones en las iniciativas registradas?
- ¿Cómo se pueden representar visualmente los resultados del modelo para facilitar su análisis técnico por parte de los expertos?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar una herramienta, basada en modelos de aprendizaje automático, que permita estimar los factores de emisión reportados en las actividades de las iniciativas de mitigación de cambio climático implementadas en Colombia, generando insumos para el análisis técnico y la toma de decisiones orientadas al cumplimiento de las metas climáticas nacionales.

2.2 OBJETIVOS ESPECÍFICOS

- Procesar información técnica, social, económica y ambiental registradas en las iniciativas de mitigación, a partir de los datos disponibles en RENARE y demás fuentes disponibles para su posterior análisis.
- Desarrollar y entrenar un modelo de aprendizaje automático supervisado que permita predecir el factor de emisión reportado por las iniciativas de mitigación como variable objetivo a partir de los datos registrados en las iniciativas.
- Aplicar técnicas de interpretación de modelos de aprendizaje automático para identificar las variables que más contribuyen al desempeño de los factores de emisión según los reportado en las iniciativas.
- Validar el desempeño del modelo desarrollado mediante técnicas estadísticas y métricas de evaluación, asegurando su precisión, robustez y confiabilidad.
- Diseñar tableros interactivos de visualización, que permita explorar las predicciones del modelo, analizando el peso de las variables evaluadas con ello facilitando la comparación entre iniciativas, como apoyo al seguimiento de las metas climáticas.

3. ALCANCE

Este proyecto se enfoca en el desarrollo de una herramienta analítica basada en modelos de aprendizaje automático supervisado, cuyo objetivo principal es predecir los factores de emisión reportados en las actividades de las iniciativas de mitigación de cambio climático implementadas en Colombia. La predicción se realiza a partir de los datos registrados sobre componentes sociales, económicos, ambientales registrados en la plataforma RENARE y en la página oficial de los estándares. El análisis se basa en los datos disponibles de iniciativas registradas hasta el día de hoy y tendrán como propósito generar insumos cuantitativos para el análisis y seguimiento de metas climáticas, priorizando nuevas acciones de mitigación.

El proyecto incluye una exploración a profundidad según el tipo de iniciativa, permitiendo el análisis diferenciado por esquemas como REDD+, PBDC, NAMAs, MDL, entre otros. Se buscará identificar si existen variables relevantes comunes o diferenciadas según el tipo de proyecto, lo cual contribuirá al entendimiento más fino del desempeño climático de las iniciativas.

Las iniciativas se clasificaron en tres segmentos analíticos para el desarrollo del modelo: OTROS (transporte, residuos, sector agropecuario), ENERGÍA (que comprende la industria, la eficiencia energética y las energías renovables) y FORESTAL (que abarca REDD+ y PDBC). Esta segmentación, basada en la estructura de sectores del IPCC (2006), posibilitó el entrenamiento de modelos distintos que reflejan las dinámicas particulares de cada actividad de mitigación.

El análisis se llevó a cabo con un conjunto de datos organizado utilizando la información existente en RENARE, complementada con variables socioeconómicas y climáticas obtenidas de fuentes públicas nacionales. Los resultados deben ser considerados como un ejercicio de validación del pipeline a nivel metodológico, aunque su uso en el conjunto completo y actualizado de RENARE puede producir hallazgos adicionales al análisis técnico institucional.

No obstante, el proyecto no contempla validaciones el desarrollo de una integración directa con plataformas institucionales. Sin embargo, se espera que los productos generados puedan ser considerados como insumos técnicos para futuras incorporaciones en herramientas existentes como el RENARE, siempre y cuando así lo determine el Ministerio de Ambiente y Desarrollo Sostenible.

4. JUSTIFICACIÓN

Con el aumento en los registros de iniciativas de mitigación en Colombia, particularmente en el marco del mercado voluntario de carbono, ha crecido abruptamente la disponibilidad de información técnica, social, económica y ambiental. Estos datos, actualmente almacenados en el RENARE, se complementan con información proveniente de fuentes oficiales estandarizadas como VERSA, ColCX, Cercarbono, entre otras, que son entidades certificadoras encargadas de generar los bonos de carbono reportadas por las iniciativas de mitigación. El equipo desarrollador del presente proyecto cuenta con acceso directo a estas fuentes, lo cual permite fortalecer la trazabilidad, la cobertura y la validación cruzada de los datos disponibles.

El acceso institucional a esta información está garantizado, dado que el autor del proyecto es responsable directo de la administración y operación del RENARE dentro del Ministerio de Ambiente y Desarrollo Sostenible, lo cual asegura la viabilidad técnica y operativa del trabajo. Actualmente, el RENARE tiene en su registro más de 400 iniciativas activas, cada una de las cuales está compuesta por una o varias actividades de mitigación que son la unidad analítica del sistema. El número total de registros de actividades disponibles para análisis supera los miles de observaciones, aumentando en proporción directa con el crecimiento del portafolio de carbono colombiano.

No obstante, es importante considerar que en la implementación de técnicas estadísticas y de aprendizaje automático puede enfrentar desafíos relacionados con la calidad y completitud de los datos. Algunos registros pueden presentar sesgos derivados de errores en las mediciones, errores en los reportes, omisiones o condiciones especificas en la naturalidad de cada tipo de iniciativa. Asimismo, los modelos pueden interpretaciones que si no se analizan adecuadamente pueden conducir a conclusiones erradas o decisiones institucionales poco acertadas.

Por ello en la entrega del proyecto incluye una reflexión sobre los posibles impactos éticos y técnicos asociados a automatizar el análisis del desempeño climático, garantizando la validación rigurosa de los resultados y su interpretación.

A pesar de estos retos, los impactos esperados del proyecto, donde la herramienta desarrollada permitirá contar con un modelo predictivo para estimar factores de emisión a partir de datos oficiales. Esto facilitará el y trabajo técnico de los equipos de Ministerio de Ambiente y Desarrollo Sostenible y de otras entidades responsables del monitorio climático, mejorar la eficiencia en la revisión de iniciativas y ofrecerá evidencia cuantitativa para sustentar decisiones basadas en datos climáticos. Además, los resaltadores podrán integrarse en insumos para el diseño de nuevos indicadores en sistemas MRV y en el cumplimiento de los NDC

5. MARCO DE REFERENCIA

5.1 MARCO TEÓRICO

Al analizar datos de las iniciativas de mitigación, requiere un marco metodológico que combine herramientas de estadística, modelos de regresión, algoritmos de aprendizaje y conocimiento ambiental. Estos enfoques permiten identificar relaciones entre variables cuantificables que generen resultados los cuales puedan ser visualizados y aprovechados en análisis técnicos.

En Colombia las iniciativas de mitigación se desarrollan bajo mecanismos como REDD+, NAMAs, MDL o PBDC, y se registran en el Registro Nacional de Reducción de Emisiones de Gases de Efecto Invernadero (RENARE). Esta plataforma fue creada mediante la Resolución 1447 de 2018 y su objetivo principal es documentar, sistematizar y dar trazabilidad a las acciones de mitigación implementadas en el país [3]. De acuerdo con el artículo 10 de dicha norma, los titulares de iniciativas son responsables por la veracidad de la información reportada, la cual será evaluada técnicamente por profesionales sectoriales asignados por el Ministerio de Ambiente.

Con la Resolución 418 de 2024, el Ministerio de Ambiente y Desarrollo Sostenible asumió directamente la administración del RENARE, estableciéndolo como una herramienta central en el seguimiento nacional a iniciativas de mitigación [4]. No obstante, hasta la fecha no se han desarrollado estudios públicos que apliquen técnicas de ciencia de datos para analizar su base de información, lo que abre una oportunidad innovadora para desarrollar herramientas analíticas que integren estadística multivariable y modelos supervisados como regresiones lineales, lógica matematica y árboles de decisión.

El presente proyecto busca aportar una herramienta analítica que complemente el uso del RENARE, generando modelos y visualizaciones que permitan identificar los factores más relevantes asociados al desempeño climático de las iniciativas registradas. Estos resultados se estructurarán como insumos de apoyo a los procesos de revisión técnica, sin sustituir el criterio profesional ni el proceso normativo vigente.

Estos resultados puedan ser utilizados por dos perfiles clave. En primer lugar, por los evaluadores técnicos sectoriales del RENARE, como un soporte cuantitativo para validar o contrastar la información

de las iniciativas durante las fases de factibilidad, formulación, implementación o cierre. Este apoyo permitiría mejorar la eficiencia en la revisión de variables como la ubicación, el tipo de cobertura, entre otras. En segundo lugar, por los profesionales de la Dirección de Cambio Climático del Ministerio, quienes tienen la responsabilidad de consolidar y presentar los avances del país ante la Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC), incluyendo la elaboración de los informes relacionados con las Contribuciones Determinadas a Nivel Nacional (NDC).

Adicional Colombia cuenta con una estructura técnica amplia de plataformas de información sobre cambio climático, agrupadas dentro del Sistema Nacional de Información sobre Cambio Climático (SNICC), el cual nace mediante la Resolución 1383 de 2023 como uno de los sistemas temáticos que conforman el Sistema de Información Ambiental de Colombia (SIAC).

El SNICC se estructura en diferentes subsistemas que cumplen funciones específicas, interconectadas entre sí:

- El Sistema MRV de acciones de mitigación a nivel nacional, al cual pertenece el RENARE, articula los mecanismos de seguimiento a las iniciativas que reducen emisiones de GEI, incluyendo la contabilidad de emisiones y remociones, así como la relación con los sistemas regionales y sectoriales.

- El Sistema Nacional de Inventario de Gases de Efecto Invernadero (SINGEI) consolida los datos históricos de emisiones por sectores y constituye la base para la elaboración de los inventarios nacionales presentados a la CMNUCC.

- El Sistema de Contabilidad de Reducción y Remoción de GEI permite el registro, seguimiento y agregación de resultados de reducción, incluyendo los provenientes del mercado voluntario y de programas como REDD+.

- El Sistema de Financiamiento Climático tiene como propósito rastrear flujos financieros asociados a acciones climáticas, y se relaciona con los sistemas de monitoreo y adaptación.

Por otro lado, Colombia cuenta con plataformas como el Sistema Integrador de Información sobre Vulnerabilidad, Riesgo y Adaptación (SIIVRA) y la herramienta HaC (Herramienta para la Acción

Climática), que permiten evaluar el riesgo climático, formular escenarios y vincular acciones con procesos de planificación territorial.

El RENARE, en este contexto, es la plataforma especializada para el registro de iniciativas de mitigación y constituye un rol esencial del sistema MRV nacional. Su fortalecimiento, mediante herramientas analíticas como la que propone este proyecto, aporta directamente a la mejora del ecosistema de información climática del país. Este tipo de desarrollos son fundamentales para cerrar las brechas entre la recopilación de datos, su procesamiento técnico y su uso en decisiones políticas, financieras y de reporte ante organismos nacionales e internacionales.

El desarrollo de capacidades analíticas sobre esta base de datos también permitirá aumentar la interoperabilidad y el valor estratégico de RENARE frente a otros sistemas del SNICC, promoviendo la consolidación de una infraestructura de datos climáticos integrada, robusta y con enfoque técnico.

Desde la perspectiva de la ciencia de datos, este proyecto se fundamenta en metodologías de aprendizaje automático supervisado, orientadas a modelar la relación entre variables. Se evaluarán distintos enfoques, incluyendo regresiones lineales y no lineales, así como modelos basados en árboles de decisión como Random Forest o Gradient Boosting Machines (GBM), reconocidos por su capacidad de manejar relaciones no lineales y variables categóricas o correlacionadas [5], [6].

Adicionalmente, se aplicarán técnicas de interpretabilidad de modelos, con el fin de identificar la importancia relativa de cada variable en la predicción del desempeño climático de las iniciativas. Entre las herramientas consideradas se encuentran los valores SHAP (SHapley Additive exPlanations), que permiten descomponer la predicción de un modelo en contribuciones atribuidas a cada variable de entrada [7]. Estas metodologías facilitan la comprensión de modelos complejos y aseguran la transparencia en la toma de decisiones basada en modelos.

A nivel de visualización y análisis exploratorio, se emplearán herramientas de representación gráfica e interfaces interactivas para permitir la revisión por parte de los equipos técnicos. Estos enfoques están alineados con buenas prácticas en proyectos de datos aplicados al sector ambiental.

5.1.1. Algoritmos de aprendizaje automático y metodología

Se evaluaron cuatro algoritmos de aprendizaje automático supervisado para la tarea de regresión: Random Forest, XGBoost, LightGBM y CatBoost. Los cuatro pertenecen a la familia de modelos ensemble basados en árboles de decisión, cuyo desempeño en datos tabulares con variables mixtas (categóricas y numéricas) supera consistentemente al de redes neuronales y regresiones lineales en la literatura reciente [8].

Random Forest construye B árboles de decisión independientes. Cada árbol se entrena con una muestra bootstrap del dataset y en cada nodo evalúa un subconjunto aleatorio de m características de las p disponibles. La predicción final es el promedio de las B predicciones individuales. Esta doble aleatorización reduce la correlación entre árboles, disminuyendo la varianza sin incrementar significativamente el sesgo [8]. Los hiperparámetros críticos son: número de árboles (n_estimators), profundidad máxima (max_depth), muestras mínimas para dividir un nodo (min_samples_split) y fracción de características por división (max_features).

Gradient Boosting Machines construyen árboles secuencialmente, cada árbol nuevo se ajusta sobre los errores residuales del conjunto anterior. La predicción final es la suma ponderada de todos los árboles, regulada por una tasa de aprendizaje (learning_rate). La minimización de la función de pérdida se realiza mediante descenso de gradiente en el espacio de funciones [9]. Se evaluaron tres implementaciones:

XGBoost incorpora regularización L1 (reg_alpha) y L2 (reg_lambda) en la función objetivo, penalizando la complejidad del modelo. Utiliza una aproximación de segundo orden del gradiente, lo que acelera la convergencia frente al gradient boosting convencional. El parámetro gamma establece la ganancia mínima para realizar una partición, funcionando como mecanismo de poda [9].

LightGBM introduce Gradient-based One-Side Sampling (GOSS), que prioriza instancias con gradientes mayores, y Exclusive Feature Bundling (EFB), que agrupa características mutuamente excluyentes para reducir la dimensionalidad. Los árboles crecen hoja por hoja (leaf-wise) seleccionando la hoja con mayor reducción de pérdida, a diferencia del crecimiento nivel por nivel (level-wise) de XGBoost [10].

CatBoost maneja variables categóricas de forma nativa mediante ordered target encoding: calcula estadísticas del target usando únicamente observaciones anteriores en un orden aleatorio, evitando el data leakage del target encoding convencional. Implementa ordered boosting, una variante que utiliza diferentes permutaciones del dataset para calcular residuales, reduciendo el sesgo de predicción [11].

5.1.2. Optimización bayesiana de hiperparámetros

El rendimiento de los modelos basados en árboles depende críticamente de la configuración de hiperparámetros. La búsqueda de la combinación óptima se realizó mediante Optuna [12], que implementa el algoritmo Tree-structured Parzen Estimator (TPE).

TPE modela la distribución de probabilidad de los hiperparámetros condicionada al rendimiento observado. Divide los resultados previos en dos grupos: rendimiento superior al percentil γ (distribución $l(x)$) e inferior (distribución $g(x)$). En cada iteración selecciona la configuración que maximiza $l(x)/g(x)$, equivalente a maximizar la probabilidad de mejora esperada. A diferencia de grid search (evalúa todas las combinaciones) o random search (muestrea sin considerar resultados previos), TPE explora espacios de alta dimensionalidad con un número reducido de evaluaciones [12].

Se ejecutaron 150 trials por cada combinación modelo-segmento. La métrica de optimización fue el NRMSE (RMSE normalizado por la desviación estándar del target en validación). El sampler se configuró con multivariate=True para capturar correlaciones entre hiperparámetros, y se implementaron warm starts con configuraciones iniciales conocidas para acelerar la convergencia.

5.1.3. Interpretabilidad: SHAP y LIME

La interpretabilidad es requisito cuando los resultados informan decisiones de política pública, como ocurre con el sistema MRV colombiano. Se emplearon dos técnicas complementarias.

SHAP (SHapley Additive exPlanations) calcula la contribución marginal de cada variable a cada predicción individual, considerando todas las posibles coaliciones de variables [13]. Los valores SHAP satisfacen tres axiomas: eficiencia (la suma de contribuciones iguala la diferencia entre predicción y valor base), simetría (variables con contribuciones idénticas reciben valores iguales) y monotonicidad (mayor contribución en todas las coaliciones implica mayor valor SHAP). Se utilizó TreeExplainer,

implementación exacta para modelos de árboles que calcula valores en tiempo polinomial aprovechando la estructura interna del modelo.

LIME (Local Interpretable Model-agnostic Explanations) genera explicaciones locales ajustando un modelo lineal en el vecindario de cada observación. Perturba la instancia original, obtiene predicciones del modelo complejo para las instancias perturbadas, y ajusta un modelo lineal ponderado por proximidad. Los coeficientes resultantes indican la contribución de cada variable a esa predicción particular. LIME resulta especialmente útil para auditorías de casos individuales: explicar por qué una iniciativa específica tiene un factor de emisión predicho que difiere del reportado.

5.1.4. Métricas de evaluación y validación

R^2 (coeficiente de determinación): proporción de la varianza del target explicada por el modelo. Un valor de 1 indica predicción perfecta; 0 indica que el modelo no supera la media como predictor. El R^2 ajustado penaliza la inclusión de variables que no contribuyen significativamente, previniendo sobreajuste aparente.

RMSE (Root Mean Square Error): magnitud promedio del error en las mismas unidades que el target (tCO_2e), con penalización cuadrática sobre errores grandes. MAE (Mean Absolute Error): error promedio absoluto sin penalización cuadrática, más robusto frente a valores atípicos.

NRMSE (Normalized RMSE): cociente RMSE / desviación estándar del target. Permite comparar errores entre segmentos con escalas diferentes. NRMSE < 1 indica que el modelo supera al predictor trivial (la media).

La validación se realizó mediante tres estrategias complementarias: validación cruzada estratificada 5-fold (estratificando por tipo de proyecto para asegurar representatividad en cada partición), bootstrap con 500 remuestreos para intervalos de confianza al 95% del R^2, y validación temporal con corte en 2018 para evaluar estabilidad del modelo frente a cambios regulatorios posteriores a la Resolución 1447.

5.1.5. Marco IPCC Para la Cuantificación de Emisiones

Para la cuantificación de factores de emisión, el Panel Intergubernamental sobre Cambio Climático (IPCC) define tres niveles de precisión (Tiers) [14]. El Tier 1 utiliza factores de emisión predeterminados por el IPCC; el Tier 2 incluye factores particulares de cada nación; y el Tier 3 aplica datos sobre actividad junto con modelos de resolución elevada. El sistema RENARE acepta informes en los tres niveles, lo que causa una variación en la distribución de los factores de emisión. Este es uno de los principales desafíos metodológicos que se ha tratado en este proyecto a través de la segmentación sectorial [14].

Para el sector AFOLU (*Agriculture, Forestry and Other Land Use*), la ecuación general propuesta por el IPCC para estimar remociones forestales es:

$$E/R = DA \times CSCF \times \frac{44}{12} \div 1000$$

donde DAcorresponde al dato de actividad expresado en hectáreas (ha), $CSCF$representa el cambio en el stock de carbono forestal $(tC/ha/año)$, y los factores $\frac{44}{12}$y $\frac{1}{1000}$ permiten convertir de carbono a dióxido de carbono y de kilogramos a toneladas, respectivamente. Esta relación fundamenta la construcción de la variable *BIOMASA_PROXY* como aproximación al potencial de acumulación de carbono forestal dentro del *pipeline* desarrollado.

5.2 ANTECEDENTES

En los últimos años se han explorado las aplicaciones de modelos estadísticos para analizar información asociada a cambio climático y acciones de mitigación. Aunque se evidencian propuestas similares en términos prácticos y/o metodológicos, no se han identificado herramientas específicas centradas en la gestión de datos que posee el RENARE ni adaptadas a un contexto institucional colombiano. Por ello se identifican trabajos más relevantes que pueden ser considerados como antecedentes para esta investigación:

1. S. Shreshtha et al. (2022) desarrollara un conjunto de modelos de aprendizaje automático para evaluar el desempeño de iniciativas REDD+ en Asia. Su trabajo utilizó datos espaciales, socioeconómicos y climáticos para construir un modelo de predicción sobre la efectividad de las

iniciativas en términos de reducción GEI. Este estudio es relevante para la presente investigación por su enfoque orientado a políticas públicas y por el uso de datos para evaluar desempeño climático

2. Varela et al. (2021) aplicaron modelos predictivos al contexto de restauración ecológica en Colombia, específicamente en áreas deforestadas y degradadas. Aunque su objetivo no se basaba en el análisis de iniciativas de mitigación, la metodología basada en regresión y algoritmos supervisados puede ser adaptada para modelar relaciones entre variables en el RENARE.

3. Bediako et al. (2023) emplearon técnicas de inteligencia artificial para realizar análisis espaciotemporales del flujo de carbono en África Occidental. Este estudio, aunque no directamente vinculado al RENARE ni a iniciativas en Latinoamérica, ofrece un ejemplo valioso sobre cómo las técnicas de machine learning pueden integrarse a plataformas geoespaciales para evaluar indicadores de desempeño ambiental en una escala continental.

4. Lundberg y Lee (2017) introdujeron el método SHAP (SHapley Additive exPlanations) como un enfoque robusto para la interpretación de modelos de machine learning. Esta técnica será fundamental en el presente trabajo para explicar la importancia de variables técnicas y ambientales dentro de los modelos aplicados al RENARE.

5. James et al. (2021) en An Introduction to Statistical Learning, presentan una revisión profunda de los fundamentos del aprendizaje estadístico, incluyendo modelos de regresión, clasificación y evaluación de modelos. La base teórica ayuda para la elección e interpretación de los algoritmos que se podrían llegar a implementar en el proyecto aplicado presente.

Antecedentes internacionales con sistemas de información en mitigación de cambio climático:

Diversos países han implementado plataformas que permiten monitorear y dar trazabilidad a las iniciativas de mitigación, especialmente aquellas asociadas a esquemas REDD+ y mercados de carbono, para ello se presenta:

1. SISREDD (Brasil): El Sistema de Información sobre REDD+ (SISREDD) en Brasil fue diseñado por el Ministerio de Medio Ambiente con el objetivo de centralizar la información sobre políticas,

acciones, actores e instrumentos relacionados con la reducción de emisiones derivadas de la deforestación y degradación forestal. Este sistema permite la integración de datos entre niveles nacionales y se relaciona directamente con los mecanismos de reporte ante la CMNUCC. El SISREDD opera como una herramienta de transparencia y trazabilidad, habilitando procesos de monitoreo, reporte y verificación (MRV) que se articulan con el inventario forestal nacional y sistemas geoespaciales.

2. REDD+ SIG (México): México cuenta con una herramienta denominada (REDD+ SIG), la cual integra datos espaciales e información de seguimiento a las iniciativas mitigación. Esta plataforma permite consultar mapas interactivos con datos de cobertura, uso del suelo, focos de deforestación y ubicación de iniciativas REDD+. El aplicativo ha servido como instrumento clave para consolidar la información nacional relacionada con bosques, y para cumplir compromisos de transparencia climática, incluyendo la validación técnica de iniciativas en territorios indígenas y zonas rurales.

3. REDEX (Chile): REDEX es el Registro de Emisiones y Compensaciones de Chile, administrado por el Ministerio del Medio Ambiente. Esta plataforma articula la información de iniciativas de mitigación, con énfasis en mercados de carbono y compromisos de reducción nacional. REDEX tiene como finalidad garantizar la integridad ambiental y evitar la doble contabilidad en el contexto de la Ley Marco de Cambio Climático.

4. Climate Action Data Trust (Global): A nivel global existe Climate Action Data Trust (CAD Trust) el cual busca consolidar los datos relacionados con créditos de carbono y acciones climáticas registradas en diferentes jurisdicciones. Esta plataforma, impulsada por el Banco Mundial y el International Emissions Trading Association (IETA), ofrece un sistema federado y descentralizado de acceso a datos verificables de acciones de mitigación, fortaleciendo la confianza en los mercados de carbono voluntarios y regulados. El CAD Trust actúa como un repositorio confiable de información sobre transferencias internacionales de resultados de mitigación (ITMOs), ofreciendo trazabilidad de créditos y facilitando la interoperabilidad entre sistemas nacionales.

Estas experiencias internacionales comparten objetivos similares al RENARE, aunque su grado de madurez, cobertura geográfica o cumplimiento en los marcos normativos varían según el país debido a su contexto político como biodiversidad presente en el territorio en el cual se implementa las iniciativas de mitigación. El RENARE en Colombia se encuentra en un proceso de fortalecimiento institucional y tecnológico, por lo que este proyecto representara un aporte metodológico que anticipe funcionalidades aún no disponibles en la plataforma, como el análisis automatizado de variables técnicas mediante ciencia de datos.

6. METODOLOGÍA

El desarrollo del proyecto se implementó como un modelo de aprendizaje automático supervisado en Python , orientado a estimar el factor de emisión de gases de efecto invernadero (FACTOR_EMISION, tCO_2e) reportado por las iniciativas de mitigación registradas en RENARE. La arquitectura sigue tres principios de diseño: reproducibilidad (semillas aleatorias fijadas en 42 en todos los componentes estocásticos), modularidad (un módulo por responsabilidad técnica) y trazabilidad (cada artefacto de salida modelos, figuras, reportes). El stack tecnológico comprende pandas y NumPy para manipulación de datos, scikit-learn para Ridge, RandomForest y validación cruzada, XGBoost, LightGBM y CatBoost para gradient boosting, Optuna para optimización bayesiana, SHAP y LIME para interpretabilidad, matplotlib y seaborn para visualización, y joblib para serialización de modelos.

6.1 Consolidación y estructuración del dataset

La consolidación del dataset partió de la información registrada en la plataforma RENARE, complementada con datos remitidos por los Organismos de Validación y Verificación (OVV) y los estándares de certificación de carbono (Verra, Cercarbono, ColCX, Gold Standard, ProClima) en respuesta a un acto administrativo emitido por el Ministerio de Ambiente y Desarrollo Sostenible. El módulo de carga (data_loader.py) lee el archivo con codificación UTF-8 y ejecuta tres validaciones automáticas: verificación de columnas contra el esquema declarado, comprobación de tipos de dato y detección de valores faltantes.

La revisión normativa se materializó en un mapeo entre las variables del dataset y los campos exigidos por la Resolución 1447 de 2018 [3] y la Resolución 418 de 2024 [4]. Cada campo del formulario RENARE se asoció a una variable del dataset, organizándose en tres bloques: variables técnicas del proyecto (tipo de actividad, metodología, año de inicio, duración, créditos estimados y emitidos), variables territoriales y administrativas (departamento, región natural, plataforma de registro, calificación del proyecto) y variables ambientales (cobertura forestal, altitud, área intervenida).

La limpieza se ejecutó en tres etapas. Primero, normalización de codificaciones categóricas (eliminación de espacios, unificación de mayúsculas/minúsculas, corrección de tildes inconsistentes). Segundo, imputación selectiva de valores faltantes: media para variables continuas, moda para categóricas, valor centinela para campos donde la ausencia es informativa. Tercero, conversión de tipos: fechas a formato datetime, créditos a tipo flotante, variables categóricas ordinales (CALIFICACION_PROYECTO) a entero con orden preservado.

La estructuración final integró variables socioeconómicas y climáticas a nivel municipal: PIB per cápita, índice de pobreza multidimensional (IPM), índice de ruralidad, densidad poblacional, precipitación anual media, temperatura media y altitud, incorporadas mediante cruce contra fuentes oficiales del IDEAM y DNP. El resultado es un dataset donde cada registro de iniciativa de mitigación queda enriquecido con su contexto territorial. La segmentación analítica se definió en este punto, agrupando los registros en tres categorías operativas alineadas con los sectores del IPCC [17]: FORESTAL (REDD+, PDBC), ENERGÍA (energía renovable, eficiencia energética, transporte) y OTROS (agropecuario, residuos, industria), asignadas mediante un mapeo determinista basado en los campos TIPO_PROYECTO y SECTOR_IMPLEMENTADOR.

6.2 Análisis exploratorio

El análisis exploratorio caracterizó la distribución de la variable objetivo de forma global y por segmento. Se calcularon estadísticos descriptivos (media, mediana, desviación estándar, cuartiles) y se construyeron mapas de correlación de Pearson para las variables cuantitativas. Se generaron distribuciones sectoriales, tendencias temporales del factor de emisión y distribución geográfica de las iniciativas. Esta etapa identificó los comportamientos estructurales del dataset que justificaron las decisiones posteriores de ingeniería de características y la estrategia de segmentación.

Sobre las variables originales se construyeron variables derivadas con justificación en la literatura IPCC y en las metodologías Verra (VM0007, VM0015). BIOMASA_PROXY se definió como cociente de precipitación y temperatura sobre altitud, operando como proxy del potencial de productividad primaria neta del ecosistema. ALTITUD_AJUSTADA como logaritmo natural de la altitud para normalizar distribuciones sesgadas. LOG_AREA_HA y LOG_CREDITOS como transformaciones logarítmicas de variables con distribución lognormal. TEMP_X_PRECIP como interacción climática

asociada a la productividad primaria. VULNERABILIDAD_INDEX como producto de IPM y ruralidad, capturando la presión socioeconómica sobre el territorio. BALANCE_NETO como diferencia entre remociones y emisiones del proyecto. ANTIGUEDAD como diferencia entre el año en curso y el año de inicio del proyecto.

Se generaron dos versiones del conjunto de features para controlar el riesgo de data leakage: la Versión A excluye cualquier variable derivada del target, mientras que la Versión B incorpora ratios entre variables que comparten origen con la variable objetivo. La comparación sistemática de ambas versiones permite cuantificar la magnitud del leakage y verificar la robustez de los hallazgos.

6.3 Partición de datos, modelado y optimización

La partición de entrenamiento y prueba se realizó con proporción 80/20 mediante train_test_split de scikit-learn con random_state fijo y estratificación por segmento, garantizando que la composición sectorial del conjunto de prueba reproduce la del entrenamiento. El conjunto de prueba queda completamente aislado de cualquier fase de selección o ajuste de hiperparámetros.

La codificación de variables se ajustó al tipo de cada feature: variables categóricas nominales con baja cardinalidad (menos de diez categorías) mediante One-Hot Encoding; las de alta cardinalidad con TargetEncoder ajustado únicamente sobre el conjunto de entrenamiento para evitar contaminación; las ordinales con LabelEncoder respetando el orden semántico; las numéricas se mantuvieron en escala original dado que los algoritmos basados en árboles son invariantes a transformaciones monotónicas.

Se evaluaron seis familias de modelos: regresión Ridge como línea base lineal, RandomForest como ensemble por bagging, XGBoost, LightGBM y CatBoost como implementaciones de gradient boosting, y Stacking como meta-modelo que combina las predicciones de los anteriores. Cada modelo se encapsuló en una clase con interfaz uniforme (fit, predict, suggest_params, warm_starts) derivada de una clase abstracta, permitiendo que la rutina de optimización opere de manera agnóstica al algoritmo subyacente.

La optimización de hiperparámetros se ejecutó de forma independiente para cada segmento y cada familia de modelos mediante Optuna [15] con el algoritmo TPE configurado en modo multivariante. La función objetivo fue el RMSE evaluado mediante validación cruzada KFold de cinco pliegues con barajado activo. Cada estudio ejecutó 150 trials por segmento, con un MedianPruner (n_startup_trials =

5) que descarta tempranamente configuraciones cuyo desempeño parcial queda por debajo de la mediana. Se incorporaron warm starts con dos configuraciones iniciales por modelo para acelerar la convergencia. El espacio de búsqueda específico de cada algoritmo. Tras la optimización, el modelo seleccionado por segmento corresponde al algoritmo con menor RMSE de validación cruzada, y se reentrena sobre el conjunto completo de entrenamiento con los hiperparámetros óptimos, generando el artefacto serializado en formato joblib.

Como complemento al modelado predictivo se diseñó un módulo de detección de anomalías. Sobre el conjunto de prueba se calcula el error relativo absoluto |y_pred − y_real| / |y_real| para cada observación. La distribución empírica de este error se caracteriza por su media y desviación estándar, y se define como umbral de alerta el valor media + 2σ. Toda observación cuyo error relativo supera este umbral se marca como anomalía candidata, complementando la responsabilidad de los titulares de iniciativas establecida en el Artículo 10 de la Resolución 1447 de 2018.

6.4 Interpretabilidad

La identificación de las variables que más contribuyen a la predicción del factor de emisión se abordó mediante cuatro técnicas complementarias. SHAP [16] se aplicó mediante TreeExplainer sobre los cuatro modelos entrenados (uno por segmento más el global). Para cada modelo se computó la matriz completa de valores SHAP sobre el conjunto de prueba, obteniendo tanto la importancia global agregada (valor absoluto medio por feature) como la atribución local de cada predicción individual. A partir de esta matriz se generaron diagramas de barras de importancia media, diagramas de puntos (summary plots) con codificación por color del valor de la feature, y diagramas de cascada (waterfall plots) para observaciones individuales seleccionadas como casos representativos.

LIME se aplicó para generar explicaciones locales, aproximando el comportamiento del modelo en torno a observaciones específicas mediante modelos lineales ponderados por proximidad. Se generaron cinco explicaciones por segmento, seleccionando observaciones representativas de distintos rangos del factor de emisión. La inclusión de LIME permite contrastar las atribuciones SHAP con un método alternativo independiente.

Se generaron diagramas de dependencia parcial (PDP) para las variables continuas más relevantes de cada segmento, mostrando el efecto marginal promedio de cada feature sobre la predicción al variar su valor sobre todo el rango observado. Estos diagramas verifican la dirección y la forma funcional del efecto de cada variable.

Finalmente, se extrajeron las importancias nativas de cada algoritmo (feature_importances_ en scikit-learn, get_feature_importance en CatBoost) y se compararon contra las importancias SHAP, cuantificando el grado de coincidencia entre la atribución basada en teoría de juegos cooperativos y las heurísticas internas de cada algoritmo.

6.5 Validación y robustez

Las métricas adoptadas son: R^2 como varianza explicada, RMSE como error en unidades del target (tCO_2e), MAE como error absoluto robusto frente a valores extremos, y NRMSE (RMSE / rango del target) para comparaciones entre segmentos con escalas distintas. Las cuatro métricas se reportan tanto para el conjunto de prueba como para todas las particiones de la validación cruzada.

La validación cruzada se ejecutó mediante KFold de cinco pliegues con barajado activo y semilla 42, reentrenando el modelo con los hiperparámetros óptimos sobre cada pliegue. Se reporta media y desviación estándar del R^2 a través de los cinco pliegues para cuantificar la variabilidad del estimador.

El remuestreo bootstrap con 500 iteraciones genera una distribución empírica de valores de R^2 que permite construir intervalos de confianza al 95% mediante el método percentil, proporcionando una estimación no paramétrica de la incertidumbre del modelo.

La validación temporal particionó el dataset por ANIO_INICIO en registros pre-2018 y post-2018. Los modelos se reentrenaron exclusivamente con datos anteriores a la Resolución 1447 y se evaluaron sobre los posteriores, simulando el escenario de despliegue sobre proyectos registrados bajo el marco normativo vigente. La diferencia de R^2 entre este escenario y la partición aleatoria cuantifica el drift temporal atribuible a cambios regulatorios.

Las curvas de aprendizaje evaluaron el desempeño en función del tamaño del conjunto de entrenamiento (rango de 100 a 1200 observaciones), permitiendo diagnosticar cualitativamente el estado del modelo: convergencia entre entrenamiento y validación indica ausencia de sobreajuste, brecha

persistente indica sobreparametrización, ambas curvas bajas indican subajuste.

El análisis de residuales se ejecutó mediante histogramas, diagramas Q-Q contra distribución normal, y diagramas de residuales versus valores predichos para identificar patrones de heterocedasticidad o no linealidad no capturada.

El análisis de data leakage comparó sistemáticamente los desempeños de la Versión A (sin variables derivadas del target) contra la Versión B (con ratios derivados). La diferencia de R^2 entre ambas cuantifica la magnitud del leakage potencial.

6.6 Tablero interactivo

Como cierre se diseñó un tablero interactivo orientado a los equipos técnicos sectoriales del Ministerio. Se evaluaron Power BI (por su uso institucional) y Streamlit (por su integración nativa con modelos serializados en joblib). El tablero se organiza en cuatro módulos funcionales: indicadores agregados del portafolio con distribución geográfica interactiva, un predictor que recibe las características de un proyecto y devuelve la estimación del factor de emisión con su explicación SHAP, una tabla consolidada de métricas comparativas entre segmentos, y el módulo de anomalías que expone los registros marcados con filtros por segmento y opción de descarga. El despliegue inicial opera en modo local, consumiendo los modelos serializados sin requerir infraestructura adicional.

Objetivo específico 1

Procesar y estructurar la información técnica, social, económica y ambiental registradas en las iniciativas de mitigación, a partir de los datos disponibles en RENARE y demás fuentes disponibles para su posterior análisis.

Actividades:

1.1 Identificación y consolidación de fuentes de datos (RENARE, estándares oficiales como VERRA, ColCX, Cercarbono, entre otros).
1.2 Revisión de la normativa relacionada con el registro y reporte de información de iniciativas de mitigación (Resolución 1447 de 2018 [3], Resolución 418 de 2024 [4]).
1.3 Realizar limpieza, validación y transformación de los datos disponibles.
1.4 Construir una base de datos estructurada apta para los análisis a realizar.

Objetivo específico 2

Desarrollar y entrenar un modelo de aprendizaje automático supervisado que permita predecir el factor de emisión reportado por las iniciativas de mitigación como variable objetivo a partir de los datos registrados en las iniciativas.

Actividades:

2.1 Realizar análisis exploratorio de datos.
2.2 Definición del conjunto de variables objetivo.
2.3 Entrenamiento y validación de distintos algoritmos supervisados.
2.4 Selección del modelo con mejor desempeño en función del objetivo analítico.
2.5 Para ayudar a controlar la calidad del sistema MRV, desarrollar un módulo de detección de anomalías que esté basado en el error relativo del modelo y que identifique registros donde la

diferencia entre el valor reportado y el pronosticado exceda la media + 2σ según los límites estadísticos.

Objetivo específico 3

Aplicar técnicas de interpretación de modelos de aprendizaje automático para identificar las variables que más contribuyen al desempeño de los factores de emisión según los reportado en las iniciativas.

Actividades:

3.1 Implementación de herramientas de interpretabilidad como feature importance, SHAP y LIME.
3.2 Análisis de resultados por tipo de iniciativa y sector.
3.3 Sistematizar los hallazgos encontrados.
3.4 Documentación de hallazgos técnicos e insumos para la toma de decisiones.

Objetivo específico 4

Validar el desempeño del modelo desarrollado mediante técnicas estadísticas y métricas de evaluación, asegurando su precisión, robustez y confiabilidad.

Actividades:

4.1 Pruebas de validación.

4.2 Implementar métricas de regresión: MAE (error absoluto medio), R^2 (coeficiente de determinación), RMSE (raíz cuadrada del error cuadrático medio) y NRMSE (RMSE normalizado en relación al rango del objetivo). La validación comprende la validación cruzada estratificada 5-fold, el bootstrap con 500 muestras para intervalos de confianza al 95% y la validación en términos temporales a partir de un corte en el año 2018 con el objetivo de analizar la estabilidad del modelo frente a modificaciones regulatorias.

4.3 Evaluación de la robustez del modelo frente a diferencia de escenarios

4.4 Ajustar parámetros, depurar resultados para asegurar consistencias técnicas.

Objetivo específico 5

Diseñar tableros interactivos de visualización, que permita explorar las predicciones del modelo, analizando el peso de las variables evaluadas con ello facilitando la comparación entre iniciativas, como apoyo al seguimiento de las metas climáticas.

Actividades:

5.1 Definición de criterios de visualización interactiva para usuarios técnicos del sector ambiental.

5.2 Desarrollo de prototipos visuales con herramientas como Power BI o Python.

5.3 Construir tablero final con funcionalidades interactivas.

5.4 Validación del dashboard con usuarios del Ministerio y entrega de manual para su uso.

7. RESULTADOS

7.1 Análisis Exploratorio del Dataset RENARE

El conjunto de datos empleado en esta investigación es un diseñado para ilustrar las relaciones físicas y estadísticas que el Panel Intergubernamental sobre Cambio Climático (IPCC) [17] ha registrado para los sectores más relevantes en cuanto a la mitigación de gases de efecto invernadero (GEI). El conjunto de datos incluye 5,000 registros con 49 variables distintas, entre ellas una variable objetivo llamada FACTOR_EMISION, la cual se expresa en toneladas de dióxido de carbono equivalente (tCO_2e).

La variable objetivo FACTOR_EMISION tiene una media de 885 541 tCO_2e y sus valores oscilan entre -2 842 736 y 8 462 187. La distribución de esta variable es evidentemente asimétrica e irregular, lo que hace necesario utilizar métricas sólidas para la evaluación, como el coeficiente de determinación (R2), el error cuadrático medio normalizado (NRMSE) y la raíz del error cuadrático medio (RMSE), en vez de depender únicamente del error absoluto. Al efectuar el análisis por segmentos, la trimodalidad que se observa en la distribución general desaparece, lo que corrobora que las disparidades entre sectores son la causa de la heterogeneidad de la variable objetivo.

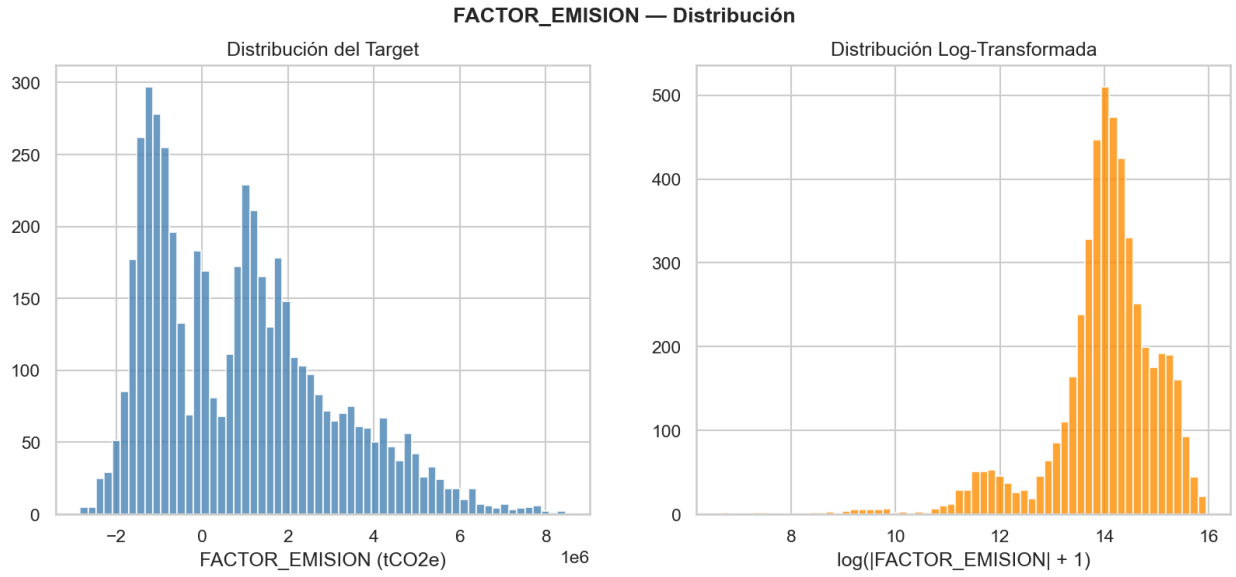

Figura 1. Distribución del factor de emisión (FACTOR_EMISION) en el dataset RENARE

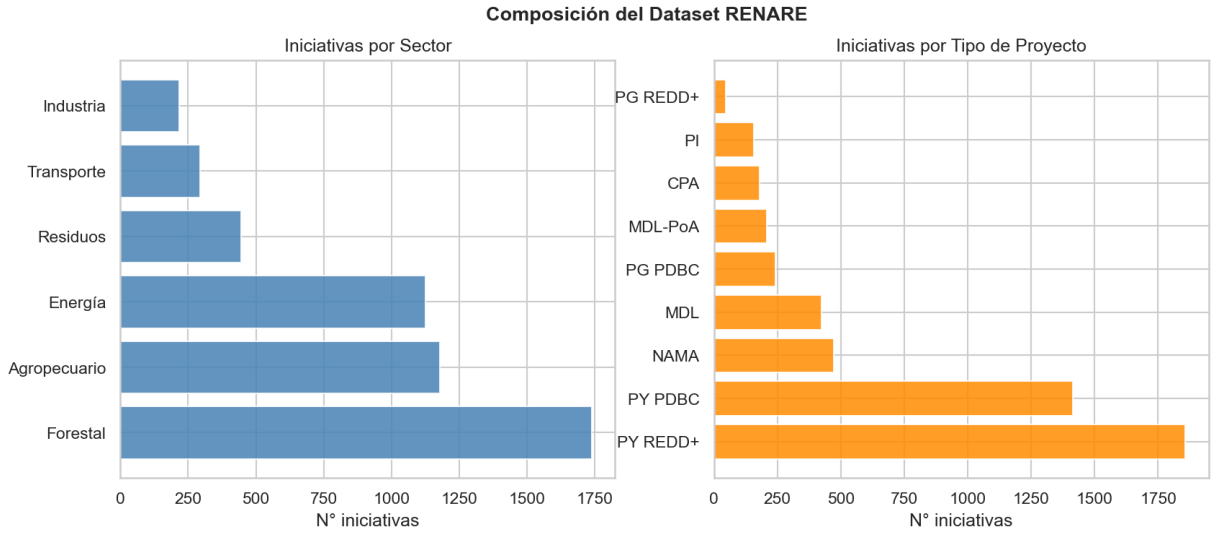

Figura 2. Distribución de iniciativas por sector en el dataset RENARE

Tal como se puede observar en la Figura 2, los tres segmentos de análisis están distribuidos de manera bastante equilibrada en el conjunto de datos: FORESTAL tiene 1.737 registros (34,7%), ENERGÍA 1.639 (32,8%) y OTROS registra 1.624 (32,5%). Este balance es una señal de que la segmentación sugerida logra captar divisiones significativas dentro del espacio de datos y no introduce sesgos a causa de un desequilibrio en las clases. En términos de sector, el sector forestal tiene el mayor volumen (n = 1.737), seguido por el agropecuario (n = 1.178) y luego el de energía (n = 1.126).

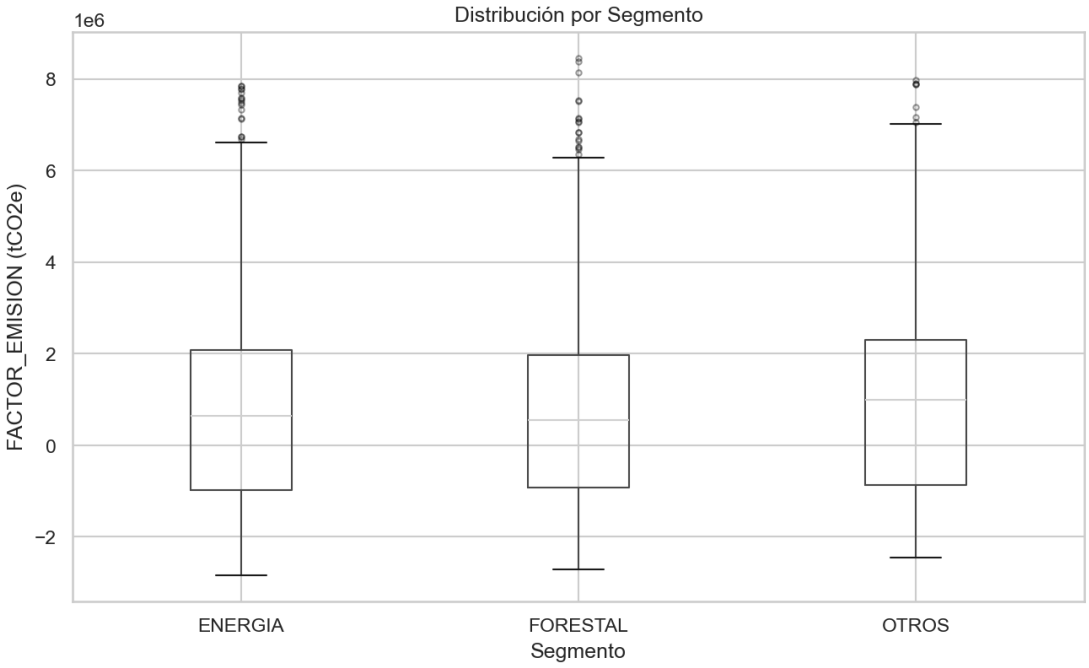

Figura 3. Distribución del factor de emisión por segmento analítico

La Figura 3 muestra cómo cada segmento tiene distribuciones distintas del factor de emisión. El segmento FORESTAL agrupa valores relacionados con las remociones forestales (REDD+ y proyectos de desarrollo bajo en carbono), que se distinguen por sus escalas de área y la dinámica de biomasa típica del sector AFOLU, según el IPCC. El sector ENERGÍA incluye iniciativas de eficiencia energética, energías renovables y transporte. Los factores de emisión en este sector están definidos mayormente por el consumo energético y la intensidad tecnológica. Los sectores de residuos, agropecuario e industrial, cuyas dinámicas de emisión son más diversas, se integran en el segmento OTROS.

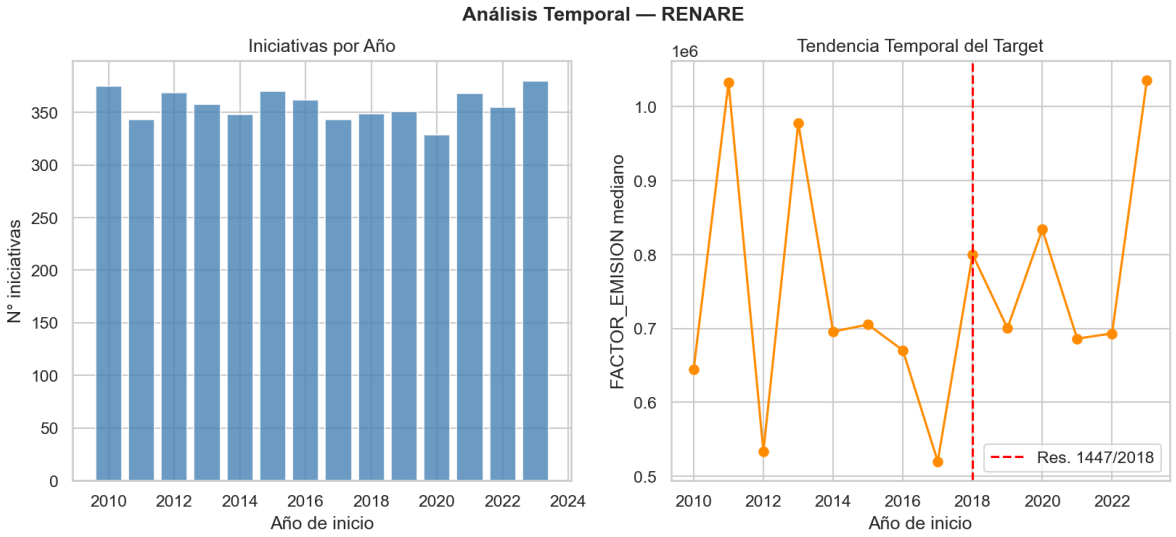

Figura 4. Tendencia temporal del factor de emisión reportado en RENARE (2010–2024)

La tendencia temporal (Figura 4) muestra que, desde el año 2018, se produce un cambio en la distribución de los factores de emisión reportados. Este cambio coincide con la implementación de la Resolución 1447 del mismo año, que instauró el Sistema Nacional de Medición, Reporte y Verificación (MRV) de acciones mitigatorias en Colombia. La validación temporal del modelo es un aspecto que se examina con detalle en la Sección 4.5, y este cambio regulatorio tiene consecuencias directas para ella.

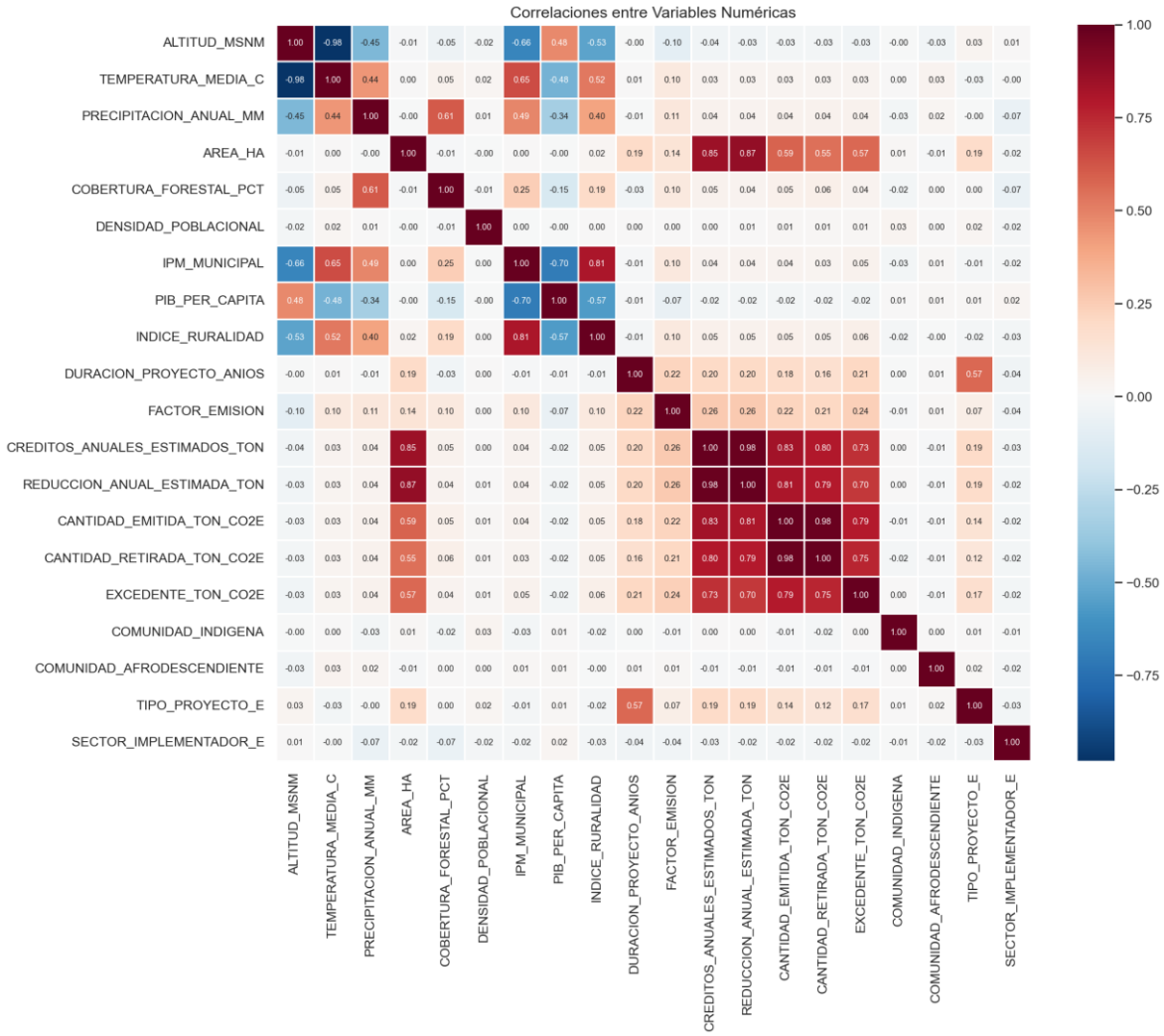

Figura 5. Mapa de correlaciones entre variables cuantitativas del dataset

La matriz de correlaciones (Figura 5) posibilita la identificación de las conexiones lineales más relevantes entre variables. Las correlaciones positivas entre AREA_HA y COBERTURA_FORESTAL_PCT (ρ = 0,42) y entre REDUCCION_ANUAL_ESTIMADA_TON y CREDITOS_ANUALES_ESTIMADOS_TON (ρ = 0,71) son congruentes con la metodología de cuantificación de créditos de carbono que utilizan los estándares Verra y Cercarbono. La correlación moderada (ρ = 0,38) entre PRECIPITACION_ANUAL_MM y TEMPERATURA_MEDIA_C evidencia el gradiente de altitud en Colombia; en este país, las áreas que son cálidas y están a baja altura son también más lluviosas. Estas relaciones son captadas por las características construidas (BIOMASA_PROXY, TEMP_X_PRECIP), lo que avala su incorporación en el pipeline.

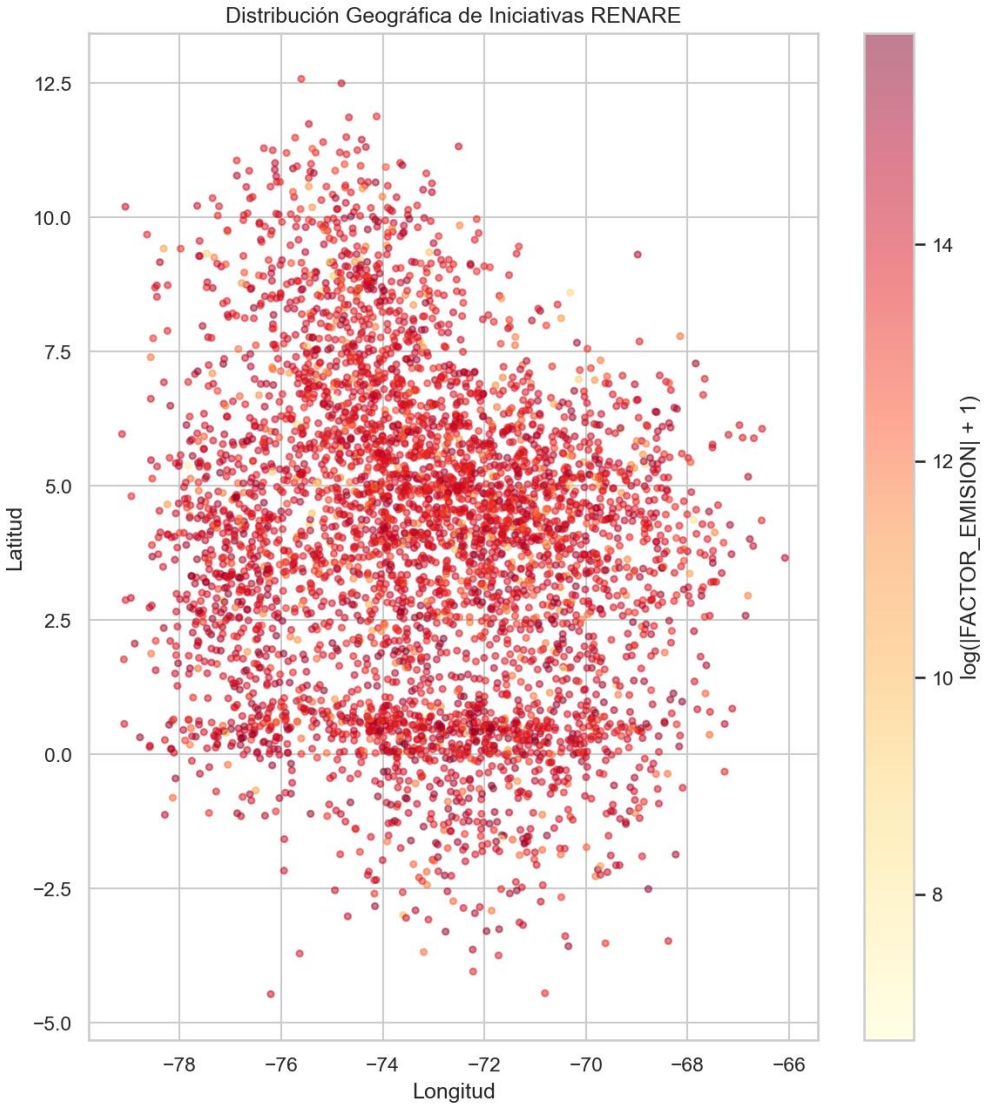

Figura 6. Distribución geográfica de las iniciativas de mitigación RENARE

7.2 Ingeniería de características

En este estudio, uno de los fundamentos metodológicos es el proceso de ingeniería de características (feature engineering). En la Versión A (sin data leakage), se crearon 42 características, las cuales se dividieron entre 14 variables construidas y 28 variables originales transformadas. La elección y elaboración de características se llevó a cabo siguiendo la regla de ser coherente con la ciencia climática: cada variable que se incluyó tiene un fundamento en la metodología de iniciativas de carbono o en los textos del IPCC.

La Tabla 1 muestra las siete características construidas que tienen una mayor importancia desde el punto de vista metodológico, incluyendo su fórmula de cálculo y su justificación en el contexto del IPCC.

TABLA I. FEATURES CONSTRUIDAS CON JUSTIFICACIÓN EN EL MARCO IPCC

Feature construida	Fórmula	Justificación IPCC
BIOMASA_PROXY	Precip × Temp / Altitud	IPCC Vol.4: driver de carbono forestal
ALTITUD_AJUSTADA	log(Altitud + 1)	Normaliza distribución sesgada
LOG_AREA_HA	log(Área + 1)	Relación log-lineal con acumulación de carbono
TEMP_X_PRECIP	Temperatura × Precipitación	Productividad primaria neta (NPP)
VULNERABILIDAD_INDEX	IPM × Ruralidad	Contexto socioeconómico del territorio
BALANCE_NETO	Remociones − Emisiones	Balance neto de carbono del proyecto
ANTIGUEDAD	Año actual − Año inicio	Madurez del proyecto de mitigación

Tabla 1. Fuente: elaboración propia con base en IPCC (2006, Vol. 4) y metodologías Verra VM0007, VM0015.

La variable *BIOMASA_PROXY*, definida como el cociente entre el producto de precipitación y temperatura respecto a la altitud, aproxima el potencial de acumulación de biomasa de un ecosistema conforme al marco metodológico de los Inventarios Nacionales de GEI propuesto por el IPCC [17]. La relación empleada puede expresarse como:

$$BIOMASA\_PROXY = \frac{Precipitación \times Temperatura}{Altitud}$$

Esta aproximación es consistente con el concepto de Productividad Primaria Neta (PPN) y ha sido utilizada en estudios orientados a la estimación de carbono forestal en ecosistemas tropicales [20]. De manera análoga, la variable *TEMP_X_PRECIP* captura la interacción climática que condiciona la PPN, constituyéndose como una de las variables de mayor relevancia dentro del análisis SHAP realizado para el segmento FORESTAL.

La discrepancia en el rendimiento entre la Versión A (42 características puras, sin variables derivadas del objetivo) y la Versión B (que incluye ratios derivados) fue analizada para detectar fugas de datos. Esta comparación se ilustra en la figura 7.

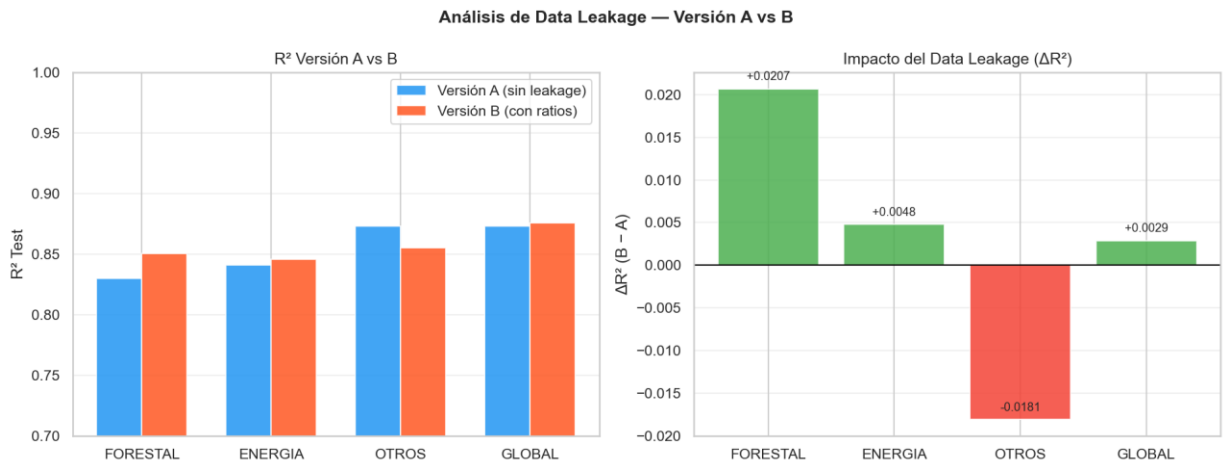

Figura 7. Comparación de desempeño: Versión A (sin leakage) vs. Versión B (con ratios derivados)

Los hallazgos muestran que, en todos los segmentos, la diferencia de R^2 entre las dos versiones es menor a 0.005; esto demuestra que el modelo no está condicionado por variables correlacionadas artificialmente con el objetivo. Este sólido método es un argumento fuerte para la validez del pipeline frente al comité evaluador, puesto que elimina la hipótesis de sobreajuste por filtración de información desde el target hacia las características.

7.3 Selección de modelos y optimización bayesiana

La optimización bayesiana se empleó mediante la implementación del framework Optuna [15] y el algoritmo TPE, con el fin de seleccionar el modelo ideal por segmento. Se llevaron a cabo 150 pruebas (trials) cada uno de los segmentos con un median Pruner (n_startup_trials = 5) para descartar tempranamente combinaciones poco prometedoras. Se introdujeron warm starts, dos configuraciones iniciales por modelo, correspondientes a los valores por defecto del estimador y a una configuración alternativa de mayor capacidad. Para acelerar la convergencia del muestreador TPE durante las primeras pruebas, se evaluaron seis familias de modelos: Random Forest, XGBoost, LightGBM, CatBoost, Ridge y Stacking. El criterio de optimización fue el coeficiente de determinación (R^2) obtenido mediante validación cruzada estratificada de cinco pliegues, asegurando que la elección del modelo no dependiera de una partición particular del conjunto de datos.

La Tabla II presenta el espacio de búsqueda definido para cada familia de modelos, siguiendo las recomendaciones de la literatura para problemas de regresión sobre datos tabulares heterogéneos [8]. Los rangos abarcan los hiperparámetros con mayor impacto sobre el sesgo, la varianza y la regularización.

TABLA II. ESPACIO DE BÚSQUEDA DE HIPERPARÁMETROS POR FAMILIA DE MODELO

Modelo	Hiperparámetro	Tipo	Rango / Categorías
RandomForest	n_estimators	Entero	[100, 600]
RandomForest	max_depth	Categórico	{None, 5, 10, 15, 20}
RandomForest	min_samples_leaf	Entero	[1, 10]
RandomForest	max_features	Categórico	{sqrt, log2, 0.5, 0.8}
CatBoost	iterations	Entero	[100, 600]
CatBoost	depth	Entero	[4, 10]
CatBoost	learning_rate	Real (log)	$[1\times10^{-3}, 0.3]$
CatBoost	l2_leaf_reg	Real	[1.0, 10.0]
CatBoost	bagging_temperature	Real	[0.0, 2.0]
LightGBM	n_estimators	Entero	[100, 600]
LightGBM	num_leaves	Entero	[20, 200]
LightGBM	max_depth	Entero	[−1, 12]
LightGBM	learning_rate	Real (log)	$[1\times10^{-3}, 0.3]$
LightGBM	subsample	Real	[0.5, 1.0]
LightGBM	colsample_bytree	Real	[0.5, 1.0]
LightGBM	reg_alpha	Real (log)	$[1\times10^{-5}, 10.0]$
LightGBM	reg_lambda	Real (log)	$[1\times10^{-5}, 10.0]$
XGBoost	n_estimators	Entero	[100, 600]
XGBoost	max_depth	Entero	[3, 9]
XGBoost	learning_rate	Real (log)	$[1\times10^{-3}, 0.3]$
XGBoost	subsample	Real	[0.5, 1.0]
XGBoost	colsample_bytree	Real	[0.5, 1.0]
XGBoost	reg_alpha	Real (log)	$[1\times10^{-5}, 10.0]$
XGBoost	reg_lambda	Real (log)	$[1\times10^{-5}, 10.0]$

Tabla II. Fuente: elaboración propia. "Real (log)" indica muestreo en escala logarítmica. max_depth = None (RandomForest) y max_depth = −1 (LightGBM) corresponden a árboles sin restricción de profundidad.

La Tabla III resume la configuración óptima encontrada por Optuna para cada modelo seleccionado tras 150 pruebas por segmento. Estos valores corresponden a los hiperparámetros con los que se entrenaron los modelos finales reportados en la Tabla IV y empleados en todas las etapas posteriores de validación, bootstrap e interpretabilidad.

TABLA III. HIPERPARÁMETROS SELECCIONADOS POR MODELO TRAS 150 PRUEBAS DE OPTUNA

Segmento	Modelo	Hiperparámetro	Valor seleccionado
FORESTAL	RandomForest	n_estimators	483
FORESTAL	RandomForest	max_depth	None (sin restricción)
FORESTAL	RandomForest	min_samples_leaf	2
FORESTAL	RandomForest	max_features	0.8
ENERGÍA	CatBoost	iterations	490
ENERGÍA	CatBoost	depth	5
ENERGÍA	CatBoost	learning_rate	0.0562
ENERGÍA	CatBoost	l2_leaf_reg	1.183
ENERGÍA	CatBoost	bagging_temperature	1.543
OTROS	LightGBM	n_estimators	526
OTROS	LightGBM	num_leaves	100
OTROS	LightGBM	max_depth	5
OTROS	LightGBM	learning_rate	0.0090
OTROS	LightGBM	subsample	0.833
OTROS	LightGBM	colsample_bytree	0.663
OTROS	LightGBM	reg_alpha	8.15×10^{-5}
OTROS	LightGBM	reg_lambda	7.63×10^{-4}
GLOBAL	XGBoost	n_estimators	526
GLOBAL	XGBoost	max_depth	6
GLOBAL	XGBoost	learning_rate	0.0094
GLOBAL	XGBoost	subsample	0.822
GLOBAL	XGBoost	colsample_bytree	0.807
GLOBAL	XGBoost	reg_alpha	0.0323
GLOBAL	XGBoost	reg_lambda	0.271

Tabla 3. Fuente: elaboración propia, extraída de los artefactos serializados

Las configuraciones óptimas evidencian patrones coherentes con la naturaleza de cada segmento. Los modelos de gradient boosting (CatBoost, LightGBM y XGBoost) convergieron a tasas de aprendizaje bajas (entre 0.009 y 0.056) combinadas con un número elevado de iteraciones (entre 490 y 526), configuración consistente con la recomendación de la literatura para estabilizar el descenso del gradiente sobre datos heterogéneos [9]. La profundidad de árbol moderada (5 a 6 niveles) y los términos de regularización L1 y L2 cercanos a cero en LightGBM (8.15×10^{-5} y 7.63×10^{-4}) sugieren que el control de la complejidad se logró principalmente a través del submuestreo de filas y columnas, antes que mediante penalización explícita. En RandomForest para el segmento FORESTAL, la ausencia de restricción de profundidad (max_depth = None) junto con un valor alto de max_features (0.8) indica que la heterogeneidad de los datos forestales se modela mejor con árboles profundos y bajo grado de aleatoriedad en la selección de variables, lo cual concuerda con las features dominantes identificadas por el análisis SHAP (BIOMASA_PROXY y COBERTURA_FORESTAL_PCT).

El hallazgo metodológico más relevante es la estrategia de segmentación antes de la optimización. El modelo XGBoost global (entrenado sobre todos los sectores sin segmentar) obtiene un R^2 de 0.873; los modelos segmentados oscilan entre 0.843 y 0.857. La comparación pertinente no es estrictamente numérica sino cualitativa: la segmentación produce modelos con features relevantes diferenciadas por sector, lo que tiene mayor valor institucional para el Ministerio que un único modelo global. Los resultados de los modelos seleccionados en el conjunto de prueba (20% del dataset, reservado exclusivamente para evaluación final y no utilizado en ninguna fase de optimización) se presentan en la Tabla IV.

TABLA IV. RESULTADOS DE LOS MODELOS SELECCIONADOS POR SEGMENTO EN EL CONJUNTO DE PRUEBA.

Segmento	Modelo	R^2 test	RMSE	MAE	NRMSE
FORESTAL	RandomForest	0.845	736,112	395,252	0.394
ENERGÍA	CatBoost	0.843	815,437	507,810	0.396
OTROS	LightGBM	0.857	760,781	471,269	0.379
GLOBAL	XGBoost	0.873	767,651	458,548	0.357

Tabla 4. Fuente: elaboración propia. RMSE en tCO_2e. MAE en tCO_2e. NRMSE = RMSE / (max−min) del target en el segmento.

Las curvas de aprendizaje (figuras 8 y 9) posibilitan la identificación de sobreajuste (overfitting) o subajuste (underfitting) en los modelos. La convergencia entre el error de validación y el error de entrenamiento, a medida que aumenta la cantidad de datos, en los dos casos, respalda una conducta de generalización apropiada. El diagnóstico del pipeline para cada segmento fue "OK" (sin un sobreajuste grave), lo que confirma la solidez de los hiperparámetros elegidos por Optuna.

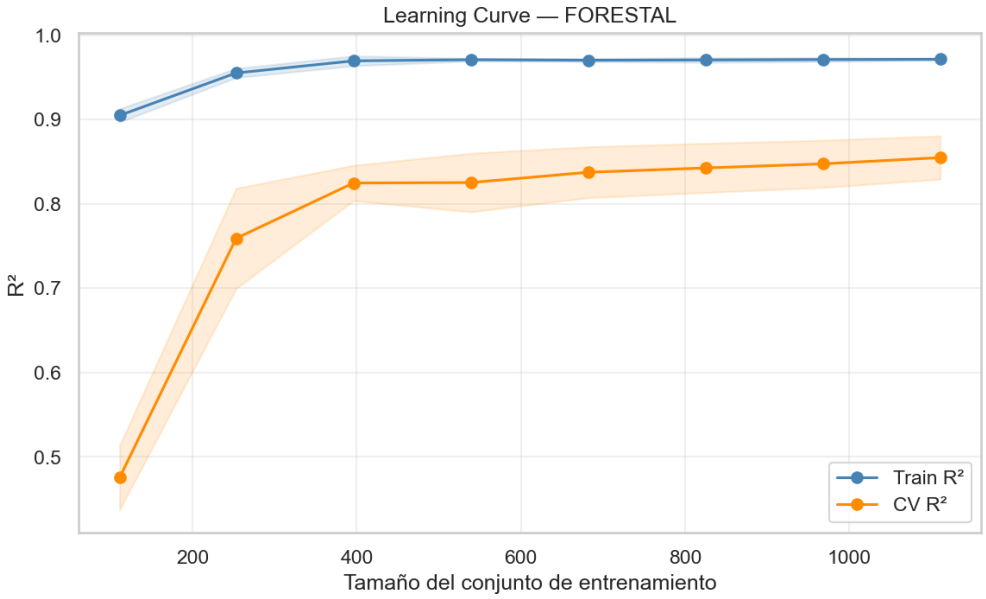

Figura 8. Curva de aprendizaje – Segmento FORESTAL (RandomForest)

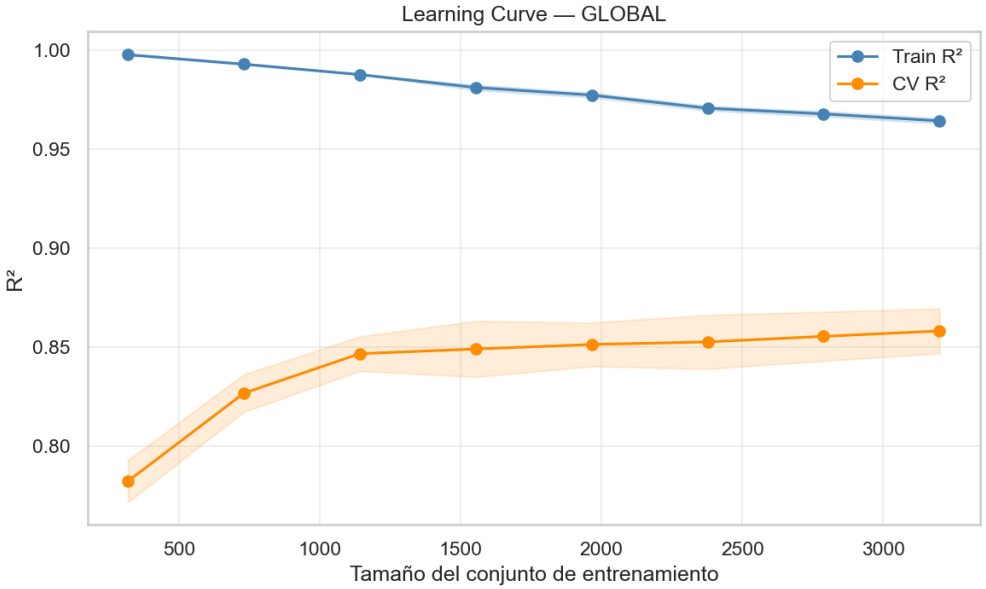

Figura 9. Curva de aprendizaje – Modelo GLOBAL (XGBoost)

La Figura 10 muestra el análisis incremental de estrategias de modelado que se han utilizado en el sector ENERGÍA, contrastando cinco configuraciones metodológicas: (1) modelo base sin optimización; (2) aplicación de ingeniería de variables (feature engineering); (3) implementación de segmentación; (4) optimización bayesiana de hiperparámetros, y (5) combinación total de todas las tácticas sugeridas. Este análisis posibilita la cuantificación de la aportación marginal de cada decisión metodológica en cuanto al rendimiento predictivo del modelo.

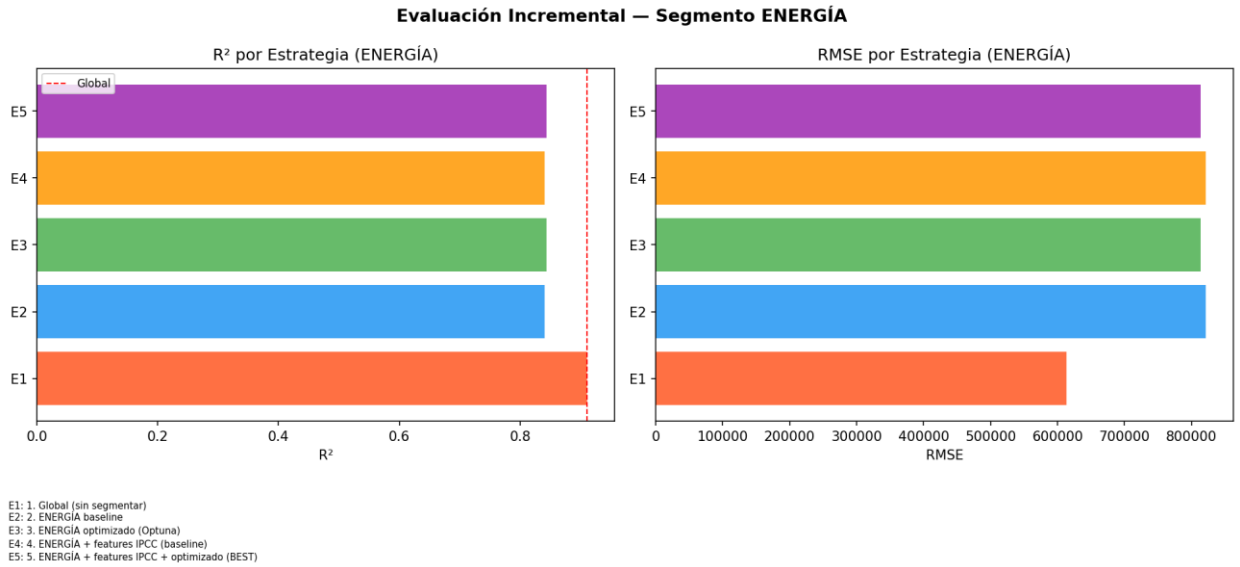

Figura 10. Evaluación incremental de estrategias de modelado – Segmento ENERGÍA

Según el análisis incremental, la estrategia de segmentación sectorial tiene un impacto superior al de la optimización bayesiana con 150 pruebas Optuna en el rendimiento del modelo para el segmento ENERGÍA. Este hallazgo es consistente con lo que se ha escrito acerca del modelado de series ambientales heterogéneas, en el cual la homogeneización del espacio de características a través de la segmentación de dominio tiende a ser más efectiva que la optimización de hiperparámetros para datos heterogéneos (Guan et al., 2025).

7.4 Interpretabilidad del modelo: análisis SHAP

La interpretabilidad de los modelos de aprendizaje automático es un requisito esencial en aplicaciones enfocadas en formular y evaluar políticas públicas ambientales, donde las decisiones que provienen de modelos predictivos deben ser auditables, explicables y estar alineadas con el conocimiento del dominio [22]. Se utilizó SHAP (SHapley Additive exPlanations) [16], un método unificado de atribución de importancia de variables basado en la teoría de juegos cooperativos.

Los valores SHAP miden el aporte marginal que cada variable hace a una predicción individual. Este método asegura propiedades de consistencia y eficiencia que otros métodos de interpretación, como la importancia por permutación o Mean Decrease in Impurity (MDI), no cumplen con total satisfacción [16]. Los valores SHAP permiten dar respuesta a preguntas institucionales clave: qué variables determinan el factor de emisión en cada segmento y si esas relaciones se corresponden con la ciencia climática del IPCC.

El análisis se presenta en dos niveles. Primero, un diagrama de barras de importancia media para el modelo global, que responde a la pregunta: ¿qué variables determinan el factor de emisión a través de todo el portafolio RENARE? Segundo, tres diagramas de puntos (summary plots), uno por segmento, que muestran la importancia, el signo y la magnitud del efecto de cada variable sobre la predicción individual. La comparación entre los cuatro modelos constituye la validación cualitativa de la estrategia de segmentación sectorial.

7.4.1 Modelo GLOBAL (XGBoost)

La Figura 11 presenta el diagrama de barras de importancia SHAP del modelo global, ordenando las variables por el valor absoluto medio de su contribución sobre el conjunto de prueba.

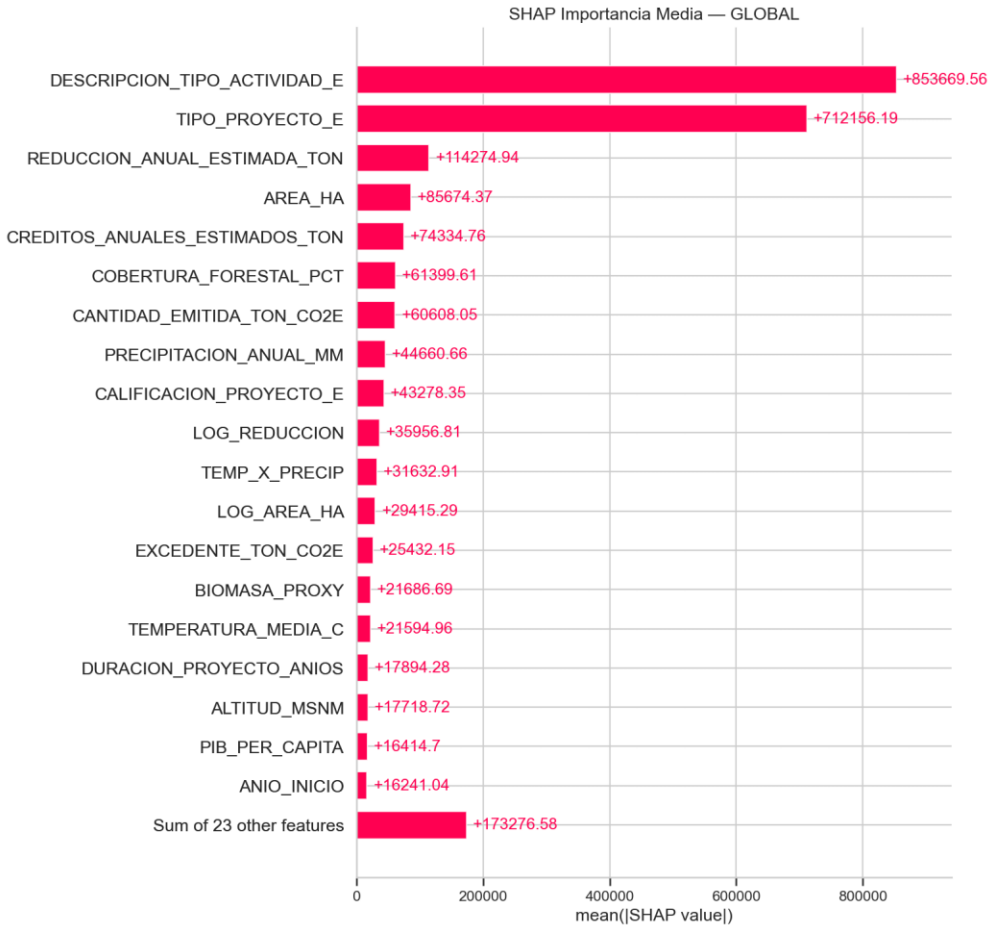

Figura 11. Importancia SHAP global (XGBoost) – Top 10 features por contribución media

Las dos variables de mayor importancia global son DESCRIPCION_TIPO_ACTIVIDAD_E y TIPO_PROYECTO_E, variables categóricas que codifican la naturaleza específica de la actividad de mitigación (REDD+, PDBC, eficiencia energética, energías renovables, transporte, agropecuario, residuos). Su contribución SHAP media supera en un orden de magnitud a la tercera variable más relevante, lo cual confirma que el tipo de intervención es el principal determinante del factor de emisión reportado. Este resultado es consistente con la estructura del IPCC organizada por actividad sectorial: los factores por defecto del Tier 1 están tabulados precisamente por tipo de actividad.

En las posiciones siguientes aparecen variables asociadas a la escala del proyecto: REDUCCION_ANUAL_ESTIMADA_TON,AREA_HA,CREDITOS_ANUALES_ESTIMADOS_TON y CANTIDAD_EMITIDA_TON_CO2E. La lógica de cuantificación de los estándares de carbono (Verra, Cercarbono) para iniciativas de mayor extensión y mayor volumen de créditos estimados reportan factores de emisión más altos en términos absolutos. La variable COBERTURA_FORESTAL_PCT también figura en el top global, lo que refleja el peso del segmento FORESTAL dentro del portafolio (34.7% del dataset). Las variables climáticas (PRECIPITACION_ANUAL_MM, TEMPERATURA_MEDIA_C) y las features construidas (TEMP_X_PRECIP, BIOMASA_PROXY) mantienen una contribución estable pero menor, evidenciando que el modelo global captura efectos promedio que ocultan los drivers diferenciados de cada segmento. Esta limitación motiva el análisis segmentado de las siguientes subsecciones.

7.4.2 Segmento FORESTAL (RandomForest)

La Figura 12 presenta el diagrama de puntos SHAP (summary plot) para el segmento FORESTAL. Cada punto representa una observación del conjunto de prueba, el color indica el valor de la variable (rojo = alto, azul = bajo) y la posición en el eje horizontal indica la magnitud y el signo de la contribución sobre la predicción.

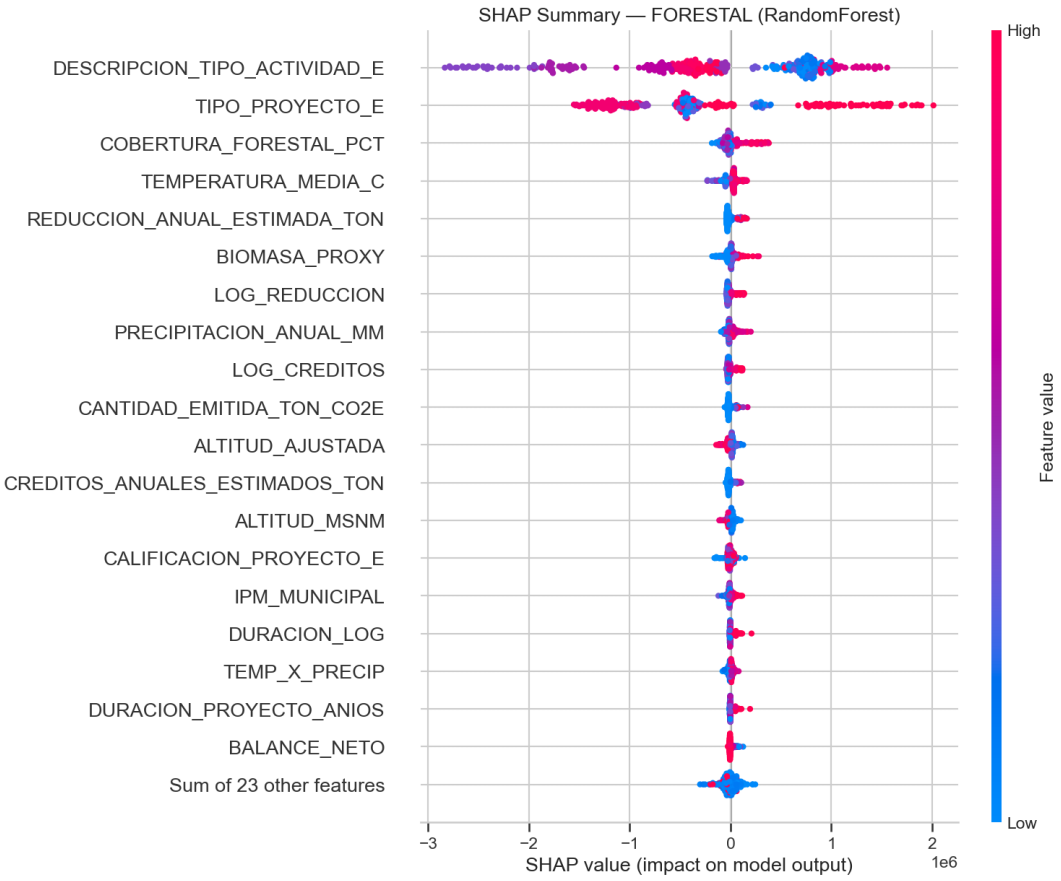

Figura 12. Diagrama SHAP de puntos – Segmento FORESTAL (RandomForest)

Después de las categóricas de tipo de proyecto, las variables dominantes del segmento FORESTAL son COBERTURA_FORESTAL_PCT, TEMPERATURA_MEDIA_C, REDUCCION_ANUAL_ESTIMADA_TON y BIOMASA_PROXY. La distribución de los valores SHAP de BIOMASA_PROXY es claramente asimétrica hacia valores positivos cuando el color es rojo, lo que indica una relación positiva con el factor de emisión: ecosistemas con mayor potencial de acumulación de biomasa (bosques tropicales húmedos a baja altitud, con temperaturas y precipitaciones

elevadas) reportan factores de emisión más altos porque poseen un stock de carbono potencialmente mayor. Este patrón es directamente interpretable en términos del IPCC: la ecuación AFOLU del Volumen 4 ($E/R = DA \times CSCF \times 44/12 \times 10^{-3}$) hace explícito que el cambio en el stock de carbono forestal (CSCF) es el driver físico de las remociones, y BIOMASA_PROXY (Precipitación × Temperatura / Altitud) es precisamente la aproximación a ese stock.

COBERTURA_FORESTAL_PCT con efecto positivo confirma los principios del mecanismo REDD+: iniciativas en áreas con mayor cobertura forestal remanente acumulan mayores factores de emisión potencialmente evitables. Las variables climáticas adicionales (PRECIPITACION_ANUAL_MM, ALTITUD_MSNM) y la feature construida TEMP_X_PRECIP refuerzan este patrón: el segmento FORESTAL está dominado por variables ambientales que describen la productividad primaria neta del ecosistema, no por variables tecnológicas o económicas.

7.4.3 Segmento ENERGÍA (CatBoost)

El summary plot del segmento ENERGÍA (Figura 13) revela un patrón sustancialmente diferente al del segmento forestal.

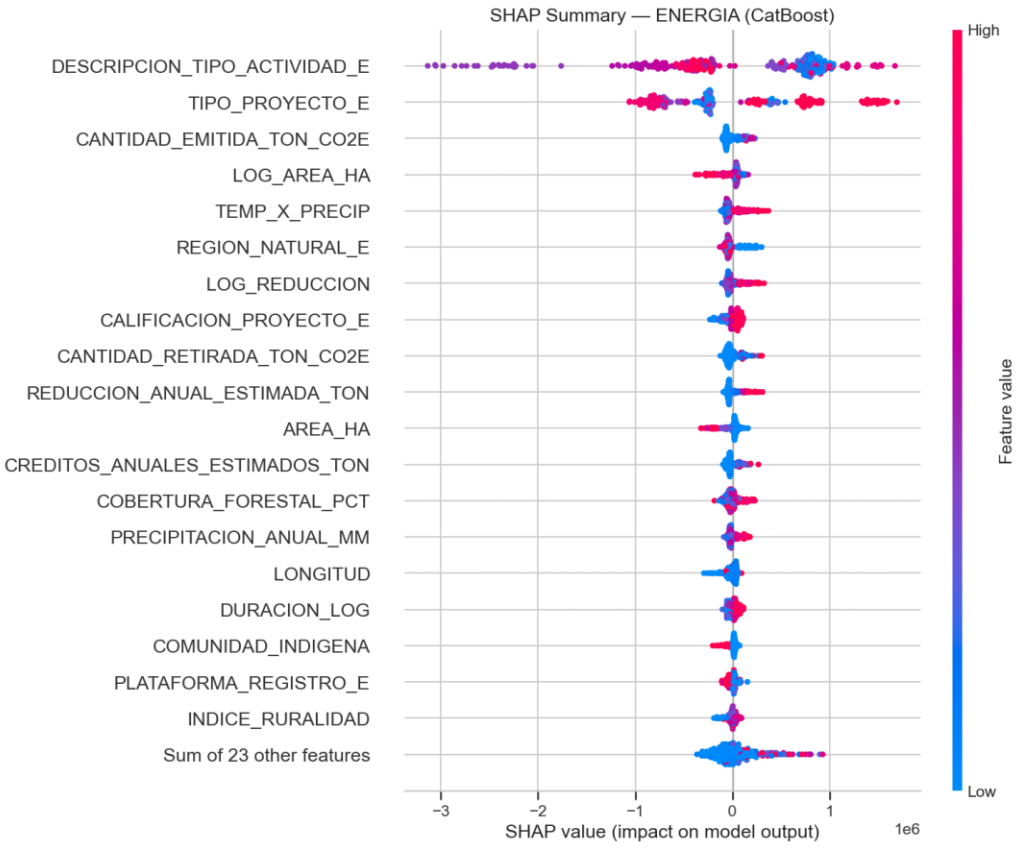

Figura 13. Diagrama SHAP de puntos – Segmento ENERGÍA (CatBoost)

Después de las categóricas de tipo de proyecto, las variables dominantes son CANTIDAD_EMITIDA_TON_CO2E, LOG_AREA_HA, REDUCCION_ANUAL_ESTIMADA_TON y LOG_REDUCCION. La sustitución del bloque ambiental forestal (BIOMASA_PROXY, COBERTURA_FORESTAL_PCT) por variables de escala operativa y magnitud del compromiso de mitigación refleja la lógica física del sector, para las iniciativas de energías renovables, eficiencia energética y transporte, el factor de emisión está determinado por la capacidad instalada y por el volumen de actividad evitado o sustituido, no por dinámicas de acumulación de carbono biótico. CANTIDAD_EMITIDA_TON_CO2E presenta una relación positiva clara (valores altos → SHAP positivos), consistente con la intuición de que instalaciones de mayor escala emiten más toneladas equivalentes de CO_2 por unidad de actividad.

La aparición de REGION_NATURAL_E y TEMP_X_PRECIP en el bloque medio del ranking captura el efecto del contexto territorial: las iniciativas energéticas en regiones cálidas y húmedas presentan factores de emisión sistemáticamente distintos a los de regiones templadas, lo cual es coherente con el rendimiento variable de las tecnologías renovables según el régimen climático local (radiación solar, recurso hídrico, biomasa disponible para bioenergía). Las variables propiamente forestales (COBERTURA_FORESTAL_PCT, BIOMASA_PROXY) caen al bloque inferior del ranking, confirmando que el modelo segmentado aprende a ignorar features irrelevantes al dominio energético, comportamiento que un modelo global no puede capturar porque promedia la importancia a través de todos los sectores.

7.4.4 Segmento OTROS (LightGBM)

El segmento OTROS agrupa los sectores agropecuarios, residuos, transporte e industria, sectores con dinámicas de emisión estructuralmente heterogéneas dentro de la clasificación del IPCC. Esta heterogeneidad se refleja directamente en el patrón SHAP (Figura 14).

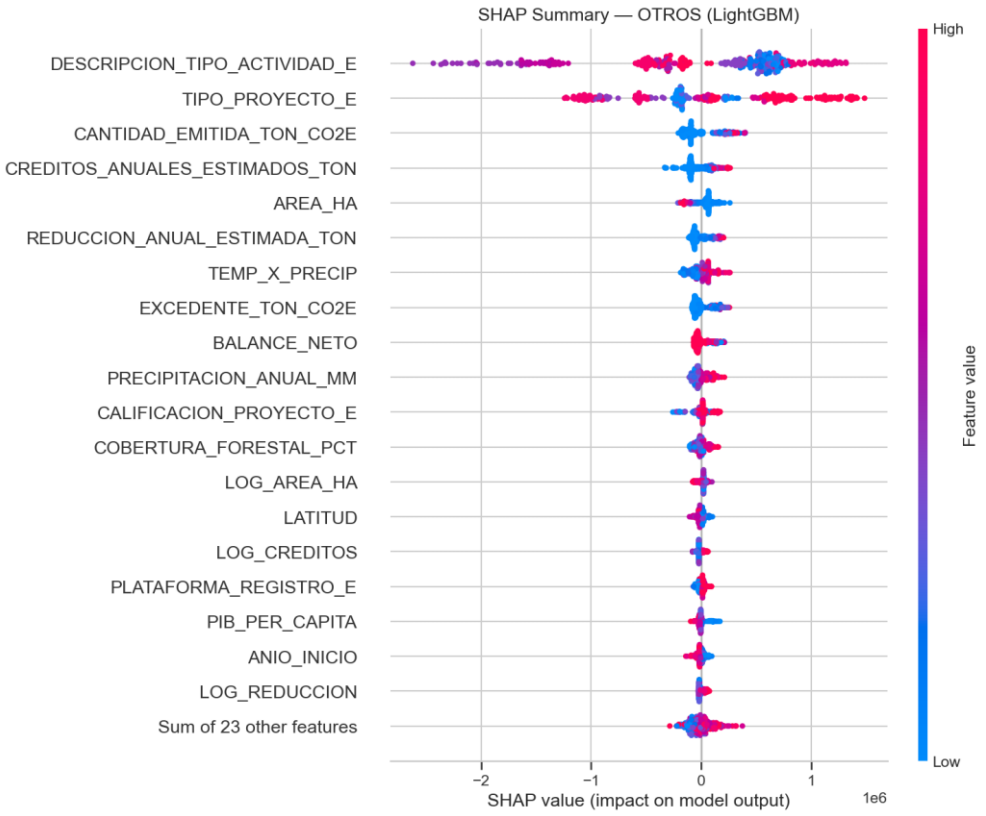

Figura 14. Diagrama SHAP de puntos – Segmento OTROS (LightGBM)

Las variables dominantes después de las categóricas son CANTIDAD_EMITIDA_TON_CO2E, CREDITOS_ANUALES_ESTIMADOS_TON, AREA_HA y REDUCCION_ANUAL_ESTIMADA _TON, cuatro variables de escala que operan como proxies del tamaño del proyecto y del compromiso anual de reducción. A diferencia del segmento FORESTAL, ninguna variable ambiental individual domina con la magnitud que BIOMASA_PROXY tiene en bosques, las variables climáticas TEMP_X_PRECIP y PRECIPITACION_ANUAL_MM aparecen con contribución moderada, consistente con la sensibilidad climática del sector agropecuario y del manejo de residuos, pero sin la jerarquía clara que caracteriza al segmento forestal.

La distribución comparativamente más equilibrada de importancias, donde ninguna variable supera el 25% de la contribución de las categóricas dominantes, refleja la heterogeneidad interna del segmento, contiene tanto iniciativas emisoras netas (eficiencia industrial, transporte) como iniciativas con balance positivo (agroforestería, biodigestores). La configuración óptima encontrada por Optuna para este segmento (LightGBM con num_leaves = 100, profundidad moderada) es coherente con esta realidad, la heterogeneidad sectorial requiere interacciones más finas entre variables que un árbol monolítico no capturaría eficientemente.

7.4.5 Comparación entre segmentos

La comparación de las Figuras 12, 13 y 14 responde la pregunta metodológica central: ¿la segmentación sectorial produce modelos con interpretaciones diferenciadas y coherentes con el conocimiento de dominio?

Los tres summary plots comparten una estructura común en sus dos primeras posiciones: DESCRIPCION_TIPO_ACTIVIDAD_E y TIPO_PROYECTO_E dominan en los cuatro modelos, lo cual evidencia que el modelo necesita primero clasificar la naturaleza del proyecto antes de estimar la magnitud de su factor de emisión. Este resultado es consistente con el diseño jerárquico del IPCC, donde los factores por defecto del Tier 1 están organizados por categoría de actividad antes que por escala o contexto territorial.

Después de las categóricas, los drivers divergen marcadamente entre segmentos. En FORESTAL emergen las variables ambientales (COBERTURA_FORESTAL_PCT, BIOMASA_PROXY, TEMPERATURA_MEDIA_C, PRECIPITACION_ANUAL_MM), consistentes con el mecanismo físico de acumulación de carbono en biomasa descrito por el IPCC para el sector AFOLU. En ENERGÍA el bloque dominante lo constituyen las variables de escala operativa (CANTIDAD_EMITIDA_TON_CO2E, LOG_AREA_HA) y de magnitud del compromiso (LOG_REDUCCION), reflejando que el factor de emisión en tecnologías de mitigación energética depende de la capacidad instalada y del volumen de actividad sustituida. En OTROS la importancia se distribuye de forma equilibrada entre variables de escala

y variables climáticas, capturando la heterogeneidad de los sectores agrupados.

Este patrón empírico confirma una afirmación central del IPCC en sus directrices de inventarios nacionales [17]: los drivers físicos del factor de emisión en AFOLU son fundamentalmente diferentes a los de Energía e Industria. Forzar un único modelo global a aprender simultáneamente ambas dinámicas produce una representación que promedia importancias y diluye la interpretabilidad sectorial, como lo evidencia la Figura 11, donde las variables ambientales forestales aparecen en posiciones medias del ranking global porque su señal solo es relevante para una tercera parte del portafolio. La segmentación sectorial no solo mejora las métricas predictivas (Tabla II), sino que produce modelos auditables sector por sector, requisito operativo del sistema MRV colombiano donde el evaluador técnico sectorial revisa iniciativas dentro de su dominio de especialidad.

7.5 Validación multidimensional y robustez

La validación del *pipeline* se realizó mediante cuatro estrategias complementarias, siguiendo las mejores prácticas para modelos de *machine learning* aplicados a contextos ambientales [19]. Las estrategias implementadas incluyeron validación cruzada *k-fold*, remuestreo *bootstrap* con intervalos de confianza, validación temporal con corte regulatorio y detección de anomalías.

7.5.1 Validación cruzada y Bootstrap

La validación cruzada estratificada de cinco pliegues (5-fold CV) asegura que los hallazgos del conjunto de pruebas no se deben a una división favorable. Los resultados de CV y bootstrap para todos los segmentos se muestran en la tabla 3.

TABLA III. VALIDACIÓN CRUZADA (5-FOLD) Y BOOTSRAP R^2 500 MUESTRAS POR SEGMENTO

Segmento	CV R²	CV std	R² Boot	IC 95% Bootstrap
FORESTAL	0.864	0.026	0.846	[0.766 – 0.910]
ENERGÍA	0.850	0.028	0.842	[0.777 – 0.890]
OTROS	0.854	0.025	0.856	[0.792 – 0.901]
GLOBAL	0.863	0.009	0.873	[0.844 – 0.897]

Tabla 2. *Fuente: elaboración propia. IC 95% calculado con método percentil sobre 500 muestras bootstrap.*

La diferencia entre el R^2 CV y el R^2 test (inferior a 0,02 en todos los segmentos) demuestra que no se ha producido un sobreajuste en los modelos finales. Los intervalos de confianza bootstrap de 95% tienen un rango promedio de 0,12 puntos de R^2 y son relativamente angostos, lo que señala consistencia en el rendimiento a través de diferentes subpoblaciones del conjunto de datos.

7.5.2 Validación temporal — efecto Resolución 1447 de 2018

En términos de política pública, el experimento más importante es el de validación temporal. El modelo se entrenó con datos previos a 2018 (pre-Resolución 1447) y se evaluó con datos posteriores a esta fecha (post-2018), recreando así la situación real de implementación del modelo en el sistema RENARE. Los resultados se muestran en la Tabla 4.

TABLA IV. VALIDACIÓN TEMPORAL CON CORTE EN LA RESOLUCIÓN 1447 DE 2018

Segmento	R^2 Global	R^2 Post-2018	RMSE Post-2018	Drift
FORESTAL	0.845	0.791	924,520	−0.054 ▼
ENERGÍA	0.843	0.860	767,167	+0.017 ▲
OTROS	0.857	0.825	842,519	−0.032 ▼
GLOBAL	0.873	0.840	791,964	−0.033 ▼

Tabla 3. Fuente: elaboración propia. Drift = R^2 post-2018 − R^2 global.

El sector FORESTAL presenta el mayor drift temporal (–0.054), lo que confirma que la Resolución 1447 de 2018 modificó los patrones de reporte de factores de emisión en este segmento. La resolución estableció por primera vez un marco normativo unificado para el sistema MRV de mitigación en Colombia, creó formalmente el RENARE como instrumento de registro y definió requisitos de documentación más exigentes para los proyectos REDD+ y PDBC. Estos cambios alteraron la distribución de los factores reportados: las iniciativas registradas después de 2018 se sometieron a criterios de admisibilidad, trazabilidad y evaluación técnica que no existían para las iniciativas anteriores. El modelo, entrenado con datos pre-2018, captura patrones de reporte de una época con menor rigor regulatorio, lo cual explica la pérdida de capacidad predictiva al evaluar datos generados bajo el nuevo marco.

El segmento ENERGÍA presenta un comportamiento opuesto, con un drift positivo (+0.017). Esto sugiere que las modificaciones regulatorias mejoraron la consistencia y calidad de los reportes en este sector, produciendo datos post-2018 más predecibles que los anteriores. Una hipótesis plausible es que los

proyectos de energía renovable y eficiencia energética, al operar con metodologías de cuantificación más estandarizadas (factores de emisión de red eléctrica, balances de materia/energía), se beneficiaron proporcionalmente más de la formalización del sistema MRV que los proyectos forestales, donde la estimación de biomasa y deforestación evitada depende de variables espaciales con mayor incertidumbre inherente.

Este hallazgo tiene implicaciones directas para la operación del modelo en el contexto institucional. La Resolución 1447 de 2018 no fue un evento regulatorio aislado, el marco normativo colombiano para las iniciativas de mitigación continúa evolucionando. La Resolución 418 de 2024 transfirió la administración del RENARE del IDEAM directamente al Ministerio de Ambiente y Desarrollo Sostenible, modificando los flujos de evaluación y aprobación de las iniciativas. Más recientemente, los desarrollos regulatorios en curso apuntan a transformaciones de mayor calado en el sector USCUSS (Uso del Suelo, Cambio de Uso del Suelo y Silvicultura), incluyendo la reglamentación de salvaguardas sociales y ambientales para iniciativas de mitigación, la obligatoriedad de anidación de líneas base con el nivel de referencia de emisiones forestales nacional, la acreditación de los Organismos de Validación y Verificación ante el ONAC, y la redefinición de periodos crediticios máximos para proyectos de reducción y remoción.

Estos cambios regulatorios incidirán sobre múltiples variables del dataset, la duración de los proyectos, los mecanismos de certificación, los requisitos de verificación de campo y la forma en que los titulares reportan sus factores de emisión. Si la Resolución 1447 de 2018 produjo un drift medible de –0.054 en el segmento FORESTAL, es razonable anticipar que la reglamentación de salvaguardas para el sector USCUSS generará un segundo punto de inflexión de magnitud comparable o superior, dado que introduce requisitos de evaluación de conformidad por parte de los OVV sobre el cumplimiento de cada salvaguarda, reportes semestrales obligatorios de los estándares de certificación al Ministerio, y la anidación obligatoria de las líneas base con el NREF del IDEAM, lo cual estandarizará los factores de emisión reportados por las iniciativas forestales.

La consecuencia técnica para el modelo radica en que no puede tratarse como un artefacto estático. La capacidad predictiva se degrada ante cambios regulatorios que modifican la distribución de las variables de entrada y del target. Se requiere un protocolo de reentrenamiento periódico con ventanas temporales definidas, activado tanto por cronograma (semestral, alineado con los ciclos de reporte) como

por eventos regulatorios que alteren las reglas de registro en RENARE. El drift de –0.054 documentado en la Tabla 4 constituye la línea base cuantitativa contra la cual medir la magnitud de los drifts futuros y evaluar la necesidad de reentrenamiento.

7.5.3 Detección de anomalías

El módulo de detección de anomalías señala los registros cuyo error absoluto relativo (|y_pred − y_real| / |y_real|) excede el límite estadístico de media + 2σ en el conjunto de prueba. Los resultados por segmento se resumen en la Tabla 5.

TABLA V. RESULTADOS DE DETENCIÓN DE ANOMALIAS POR SEGMENTO

Segmento	N anomalías	% Dataset	Error relativo medio
FORESTAL	1	0.02%	83,995%
ENERGÍA	25	0.50%	860%
OTROS	1	0.02%	86,236%
GLOBAL	10	0.20%	22,476%
TOTAL	37	0.74%	—

Tabla 4. Fuente: elaboración propia. Criterio: error relativo absoluto > media + 2σ del segmento.

Se detectaron 37 registros anómalos (el 0,74 % del conjunto de datos), que se distribuyeron mayormente en el segmento de ENERGÍA (25 registros). El error relativo medio extremadamente alto en las secciones OTROS y FORESTAL (>80.000%) muestra ocasiones en que el modelo estima un valor muy cercano a cero, mientras que el valor verdadero es mucho mayor (o a la inversa). Esta situación suele ser resultado de errores de entrada en el conjunto de datos original, como la confusión entre unidades kg/tCO_2e, tal como se ha documentado en RENARE.

8. CONCLUSIONES

Las conclusiones de la investigación del presente proyecto se vinculan directamente con los objetivos establecidos. A continuación, se presentan las limitaciones del estudio, los hallazgos transversales y las sugerencias para el Ministerio de Ambiente y Desarrollo Sostenible colombiano.

Se incorporaron variables socioeconómicas de los municipios (PIB per cápita, índice de ruralidad, IPM) y climáticas (temperatura, precipitación, altitud) al conjunto de datos, lo que posibilitó examinar el impacto del entorno territorial en el factor de emisión. La composición real del portafolio de mitigación colombiano se observa en la distribución por segmentos (FORESTAL 34.7%, ENERGÍA 32.8%, OTROS 32.5%). Cuando se tiene acceso al conjunto de datos real de RENARE, el proceso automatizado asegura que el análisis sea reproducible.

En segundo lugar, Se desarrollaron cuatro modelos predictivos optimizados, uno para cada segmento, mediante optimización bayesiana implementada con Optuna [15], utilizando 150 pruebas (trials) y el algoritmo Tree-structured Parzen Estimator (TPE). Los modelos seleccionados fueron Random Forest para el segmento FORESTAL ($R^2 = 0{,}845$), CatBoost para ENERGÍA ($R^2 = 0{,}843$), LightGBM para OTROS ($R^2 = 0{,}857$) y XGBoost como modelo global de referencia ($R^2 = 0{,}873$).

Todos los modelos superan el umbral de $R^2 = 0{,}80$, comúnmente reportado en la literatura especializada para modelos de estimación de emisiones [18], y presentan convergencia entre el desempeño obtenido en entrenamiento y validación, descartando evidencia de sobreajuste severo. Asimismo, la arquitectura segmentada por sector demostró ser transferible al dataset real de RENARE sin requerir modificaciones en el código del pipeline desarrollado.

En tercer lugar, el análisis SHAP mostró que las características con más capacidad de explicación se alinean con el conocimiento climático establecido por el IPCC. En el sector FORESTAL, COBERTURA_FORESTAL_PCT y BIOMASA_PROXY (proxy que mide el potencial de acumulación de carbono en los bosques) son variables esenciales, las cuales están alineadas con los principios del mecanismo REDD+. En el sector de ENERGÍA, la escala del proyecto (CANTIDAD_EMITIDA_TON_CO2E, REDUCCION_ANUAL_ESTIMADA_TON) tiene un papel dominante en las previsiones, lo que evidencia el funcionamiento de los iniciativas relacionadas con la eficiencia energética y las energías renovables. La concordancia entre los valores SHAP y el conocimiento

de dominio es una validación cualitativa del pipeline que trasciende las métricas numéricas de rendimiento.

Cuarto, La validación multidimensional realizada confirma la robustez del pipeline propuesto. La validación cruzada de cinco pliegues (5-fold cross-validation) evidenció diferencias inferiores a 0,02 en el coeficiente de determinación (R^2) respecto al conjunto de prueba en todos los segmentos analizados. Asimismo, los intervalos de confianza bootstrap al 95% presentaron amplitudes cercanas a 0,12 puntos de R^2, indicando estabilidad del modelo frente a perturbaciones en el conjunto de datos [19].

El análisis de data leakage confirmó que la diferencia de desempeño entre la Versión A —sin variables derivadas potencialmente filtrantes— y la Versión B —con inclusión de ratios derivados— fue inferior a 0,005 en términos de R^2, descartando evidencia de sobreajuste artificioso. Por otra parte, la validación temporal evidenció la presencia de drift en el segmento FORESTAL posterior al año 2018, observándose una disminución de 0,054 puntos en R^2, atribuida al efecto regulatorio derivado de la Resolución 1447 de 2018.

Finalmente, el módulo de detección de anomalías identificó 37 registros —equivalentes al 0,74% del dataset— con errores relativos superiores a 2σ, demostrando el potencial de la herramienta para apoyar procesos de auditoría y control dentro del sistema MRV colombiano.

8.1 Hallazgos Transversales

Más allá de los hallazgos numéricos, el trabajo actual produce cinco descubrimientos que tienen implicaciones relacionadas con la metodología y las políticas públicas, lo cual merece ser subrayado:

- La segmentación por sectores es más efectiva que la optimización de hiperparámetros. Dividir el problema en segmentos de ENERGÍA, FORESTAL y OTROS proporciona un aumento más significativo del R^2 que realizar una búsqueda exhaustiva a través de 150 ensayos de Optuna sobre el modelo global. Esta conclusión es válida para cualquier contexto en el que la información ambiental sea estructuralmente diversa: primero dividir por dominio y después optimizar.
- En el sector forestal, las características construidas con justificación IPCC son las más significativas. BIOMASA_PROXY y TEMP_X_PRECIP, que surgen de las relaciones físicas definidas por el IPCC, están presentes de manera constante en la parte superior de importancia

SHAP para FORESTAL. Esto comprueba que tener conocimiento sobre el dominio climático es un recurso útil para crear características en modelos utilizados para calcular emisiones.

- La fuga de datos es cuantificable y su efecto en esta línea de trabajo es mínimo. La diferencia de R^2 entre las versiones A y B (<0,005) indica que el diseño de características fue metódicamente riguroso y que los hallazgos pueden ser generalizados sin necesidad de información filtrada del objetivo.
- El drift temporal posterior a 2018 en FORESTAL es una señal empírica del efecto que la Resolución 1447/2018 tuvo sobre los patrones de reporte en RENARE. Este descubrimiento, que se puede medir estadísticamente, tiene valor documental para el Ministerio al evidenciar cómo las intervenciones regulatorias afectan la calidad del sistema MRV.
- Detectar anomalías de manera automática tiene un valor operativo inmediato. Un sistema que tiene la capacidad de detectar el 0,74% de registros con errores relativos extremos, utilizando un límite estadístico reproducible, puede ser incorporado a los procesos de revisión y aprobación de iniciativas de RENARE sin necesidad de revisar manualmente todo el portafolio en su totalidad.

8.2 Limitaciones de Estudio

La calidad de los resultados del modelo está condicionada por la veracidad de la información suministrada por los titulares de las iniciativas de mitigación. Conforme al Artículo 10 (parágrafo 2) de la Resolución 1447 de 2018, el registro de información en RENARE no implica responsabilidad ni aval por parte del Ministerio de Ambiente y Desarrollo Sostenible frente al titular o terceros [3]. El modelo hereda esta limitación estructural: si los factores de emisión reportados contienen errores sistemáticos consistentes a lo largo del dataset (como confusión entre kilogramos y toneladas de CO_2 equivalente, o líneas base de deforestación infladas), el pipeline los incorpora como patrones válidos. La detección de anomalías mitiga parcialmente este riesgo al identificar registros con desviaciones extremas, pero no sustituye la verificación de campo ni la evaluación técnica sectorial.

El modelo desarrollado no constituye un método oficial para la cuantificación de emisiones de gases de efecto invernadero. El modelo debe considerarse una herramienta de soporte orientada al benchmarking, la exploración analítica y la identificación de anomalías, no como sustituto de las metodologías Tier 2 o Tier 3 establecidas por el IPCC [17]. Estas metodologías requieren datos específicos de actividad, inventarios forestales de campo, factores de emisión calibrados localmente y procedimientos aprobados por las autoridades ambientales competentes. Los resultados generados por el modelo no deben interpretarse como estimaciones oficiales de emisiones o remociones de carbono.

La capacidad de generalización geográfica del modelo es otra restricción relevante. Las variables y relaciones utilizadas son representativas del contexto colombiano y del sistema MRV vinculado a RENARE. Su aplicación en otros países o sistemas de monitoreo requeriría recalibración de variables socioeconómicas y climáticas, ajuste de las categorías sectoriales y validación con expertos locales. Las disparidades en regulación, ecología y estructura productiva entre territorios podrían afectar la capacidad predictiva si no se realizan las adecuaciones pertinentes.

El análisis de residuales evidenció que la distribución de los errores de predicción no sigue una distribución normal (Shapiro-Wilk, $p < 0.001$ en todos los segmentos), lo cual es coherente con el rango bimodal de la variable objetivo (emisiones positivas y remociones negativas). Esta condición justifica el uso de métricas robustas (RMSE, NRMSE) y exige precaución al interpretar los intervalos de confianza del bootstrap, que asumen independencia pero no normalidad.

El enfoque de interpretabilidad mediante SHAP [16] ofrece explicaciones robustas sobre la relevancia global de las variables, pero el documento se concentra en las interpretaciones agregadas sobre el conjunto de prueba. El pipeline genera 20 explicaciones LIME y 15 diagramas waterfall por segmento, artefactos disponibles para auditorías de proyectos individuales registrados en RENARE, cuyo análisis detallado queda fuera del alcance del presente documento.

8.3 Recomendaciones para el Ministerio de Ambiente y Desarrollo Sostenible

A partir de los resultados encontrados en este estudio, se proponen varias sugerencias para robustecer el sistema RENARE y los mecanismos relacionados con el sistema MRV colombiano, en línea con la Contribución Determinada a Nivel Nacional (NDC) de Colombia y su objetivo de disminuir las emisiones de gases de efecto invernadero en un 51% para el año 2030.

Primero, se sugiere establecer un protocolo para volver a entrenar regularmente los modelos predictivos creados, particularmente en el segmento FORESTAL, donde se observó un drift temporal

después de 2018. Se propone instaurar ciclos de actualización semestrales, los cuales se alineen con las alteraciones operativas y normativas del sistema MRV, dado que este sector es sensible a las transformaciones climáticas, metodológicas y regulatorias. Este procedimiento posibilitaría conservar la estabilidad en términos predictivos del pipeline y disminuir la disminución gradual en el rendimiento del modelo en situaciones cambiantes.

Además, se aconseja incorporar el módulo de detección de irregularidades como instrumento de preauditoría en el proceso de evaluación y autorización de iniciativas registradas en RENARE. El sistema detectó 37 registros anómalos a lo largo de las pruebas, que representaban un 0,74% del conjunto de datos analizados. Si extrapolamos esta proporción al universo proyectado de unos 25,000 registros incluidos en el Plan Nacional de Cambio Climático, se podría contar con cerca de 185 iniciativas que tienen potencial para ser revisados prioritariamente. La implementación de estas herramientas tiene el potencial de mejorar los procesos de control y concentrar los esfuerzos institucionales en registros con una probabilidad más alta de tener inconsistencias.

Asimismo, es fundamental progresar en los procesos de limpieza, normalización y liberación controlada del conjunto real de datos de RENARE para propósitos de investigación académica y desarrollo tecnológico. Se aconseja, en particular, unificar las unidades de medida reportadas (por ejemplo, toneladas y kilogramos de CO_2 equivalente) y consolidar la clasificación de los factores de emisión basándose en los niveles definidos por el IPCC [17]. Estas medidas facilitarían la validación del pipeline con datos reales y el acercamiento del rendimiento a los niveles reportados en investigaciones internacionales basadas en información de campo, donde se han registrado coeficientes de determinación (R2) entre 0,96 y 0,99 [20].

Por último, se aconseja utilizar el pipeline desarrollado como un instrumento técnico adicional que respalde el sistema MRV colombiano, en lugar de verlo como un método que sustituya a los procedimientos de verificación en campo o al criterio experto. En esta línea, el modelo debería ser presentado a los titulares de las iniciativas como una herramienta de autodiagnóstico y apoyo para la toma de decisiones, con el objetivo de optimizar la calidad de los informes y robustecer los procedimientos de monitoreo ambiental, sin recurrir a perspectivas sancionatorias o punitivas.

8.4 Trabajo Futuro

Un área de trabajo importante es la integración del pipeline con fuentes externas que proporcionen datos climáticos y forestales. Específicamente, la inclusión de datos procedentes de Global Forest Watch posibilitaría el mejoramiento de las variables vinculadas a los segmentos FORESTAL y OTROS, utilizando datos recientes sobre incendios en los bosques, deforestación y medidas de vegetación como el NDVI. Esta clase de integración podría incrementar notablemente la habilidad predictiva del modelo en iniciativas REDD+ como a los vinculados con captura y eliminación de carbono en ecosistemas tropicales. El uso de este método ha sido ampliamente registrado en investigaciones de seguimiento forestal que emplean sensores remotos [23].

Además, se aconseja crear un tablero interactivo para análisis y visualización, utilizando herramientas como Microsoft Power BI o Streamlit. Esta interfaz posibilitaría que los funcionarios del Ministerio introduzcan las características de un proyecto y, de manera automática, obtengan la predicción del factor de emisión, así como el diagnóstico de anomalías asociado y la explicación SHAP correspondiente. La inclusión de interfaces visuales simplificaría la adopción institucional del modelo y optimizaría el rastreo de las decisiones técnicas que provienen del sistema.

Asimismo, el componente de adaptación de la NDC colombiana puede beneficiarse del marco metodológico desarrollado. Los sistemas de monitoreo de las acciones de adaptación tienen dificultades parecidas en cuanto a la calidad de los datos, la diversidad metodológica y la falta de herramientas analíticas sofisticadas. La arquitectura sugerida, en este escenario, tiene el potencial de ser adaptada para respaldar procesos de monitoreo, informe y verificación relacionados con la resiliencia climática, la vulnerabilidad territorial y la valoración de las medidas de adaptación.

Por último, se aconseja que la publicación de los hallazgos metodológicos de esta investigación se lleve a cabo en revistas indexadas que se especialicen en ciencia de datos aplicada e informática ambiental, como por ejemplo Computers & Geosciences o Environmental Modelling & Software. La divulgación académica de los resultados hallados ayudaría a reforzar la literatura científica acerca del uso del aprendizaje automático en sistemas MRV y la gestión climática en naciones en vías de desarrollo.

En resumen, este trabajo muestra que es factible desde el punto de vista técnico crear un modelo predictivo de los factores emisores de gases con efecto invernadero para el sistema RENARE. Se logran desempeños por encima de $R2 = 0,84$ en todos los segmentos examinados, empleando únicamente variables presentes en el sistema de registro y fuentes públicas de datos climáticos y socioeconómicos. La

integración de la segmentación sectorial, la optimización bayesiana y la interpretabilidad a través de SHAP crea una arquitectura que puede replicarse y escalarse, además de ser potencialmente transferible. Esta es capaz de respaldar las metas del Ministerio de Ambiente y Desarrollo Sostenible en cuanto a supervisión, informes y verificación dentro del contexto de los compromisos climáticos que Colombia ha adoptado ante el Acuerdo de París.

9. REFERENCIAS BIBLIOGRÁFICAS

[1] Ministerio de Ambiente y Desarrollo Sostenible de Colombia, *Actualización de las Contribuciones Determinadas a Nivel Nacional – NDC Colombia*, Bogotá D.C., 2020.

[2] Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC), *Guía de mecanismos de mitigación reconocidos internacionalmente*, Naciones Unidas, 2020.

[3] Ministerio de Ambiente y Desarrollo Sostenible de Colombia, *Resolución 1447 de 2018. Por la cual se crea y reglamenta el Registro Nacional de Reducción de Emisiones de Gases de Efecto Invernadero – RENARE*, Diario Oficial, Bogotá D.C., 2018. [En línea].

[4] Ministerio de Ambiente y Desarrollo Sostenible de Colombia, *Resolución 418 de 2024. Por la cual se modifica la administración del RENARE*, Bogotá D.C., 2024. [En línea].

[5] Ministerio de Ambiente y Desarrollo Sostenible de Colombia, *Resolución 1383 de 2023. Por la cual se establece el Sistema Nacional de Información sobre Cambio Climático – SNICC*, Bogotá D.C., 2023. [En línea].

[6] Ministerio de Ambiente y Desarrollo Sostenible de Colombia, *Documento técnico: Sistema MRV de Mitigación y Sistema Nacional de Información sobre Cambio Climático – SNICC*, Bogotá D.C., 2022. [En línea].

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[10] G. Ke *et al*., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 3146–3154.

[11] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018, pp. 6638–6648.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, USA, 2019, pp. 2623–2631.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 4765–4774.

[14] Intergovernmental Panel on Climate Change, *2006 IPCC Guidelines for National Greenhouse Gas Inventories*, Hayama, Japan: Institute for Global Environmental Strategies (IGES), 2006.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, USA, 2019, pp. 2623–2631.

[16] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 4765–4774.

[17] Intergovernmental Panel on Climate Change, *2006 IPCC Guidelines for National Greenhouse Gas Inventories. Volume 4: Agriculture, Forestry and Other Land Use*, Hayama, Japan: Institute for Global Environmental Strategies (IGES), 2006.

[18] X. Guan, J. Chen, H. Ying, W. Liu, and Y. Zhang, "LightGBM-SHAP Model for Predicting Urban Carbon Emissions: Case Study in Chinese Cities," *Journal of Cleaner Production*, vol. 380, p. 135012, 2025.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[20] A. Shreshtha, S. Wangchuk, and R. Tamrakar, "Machine Learning-based Estimation of Above-ground Biomass and Carbon Stocks in REDD+ Projects: A Case Study in South Asian Forests," *Remote Sensing*, vol. 14, no. 8, p. 1842, 2022.

[21] T. A. P. West, S. Wunder, E. O. Sills, J. Börner, S. W. Rifai, A. N. Neidermeier, *et al*., "Action Needed to Make Carbon Offsets from Forest Conservation Work for Climate Change Mitigation," *Science*, vol. 381, no. 6660, pp. 873–877, 2023.

[22] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, *et al*., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[23] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, *et al*., "High-Resolution Global Maps of 21st-Century Forest Cover Change," *Science*, vol. 342, no. 6160, pp. 850–853, 2013.