GreenWest: inteligencia artificial para la predicción de créditos de carbono en proyectos de (re)forestación en España

Maider Araceli Urbón Jiménez[a,*], Jaime Gabriel Vegas[a], Ana de Luis Reboredo[a], Belén Pérez Lancho[a], Ana-Belén Gil-González[a]

[a] *Grupo B1, Equipo de investigación BISITE, Universidad de Salamanca, Facultad de Ciencias, Salamanca, España*

Abstract

Este trabajo presenta **GreenWest**, un modelo de inteligencia artificial diseñado para predecir la cantidad de carbono capturado en proyectos de forestación y reforestación en España. El modelo se entrena con datos multifuente: registros del **Inventario Forestal Nacional** (**IFN3–IFN4**, MITECO), variables climáticas derivadas de **Copernicus/ERA5-Land** e índices espectrales procedentes de **imágenes Landsat** (Collection 2, Level 2, USGS). Estos datos se integran en una base de datos relacional jerárquica que organiza la información por parcela, especie y clase diamétrica, manteniendo trazabilidad y coherencia estructural entre inventarios.

El modelo desarrollado responde a la pregunta: *Dado un cultivo forestal con características concretas de vegetación, clima y terreno, ¿cuánto CO_2 contendrá pasados unos años?* Esta capacidad predictiva permite su integración en marcos de optimización forestal, abordando cuestiones como la selección de especies o la asignación óptima de terrenos para maximizar la fijación de carbono.

Se evaluaron múltiples enfoques de aprendizaje supervisado, destacando **CatBoost** como el modelo con mejor rendimiento ($R^2 > 0{,}80$, RMSE<15), con alta capacidad de generalización temporal mediante validación cruzada

*Autora de correspondencia

Email addresses: murbon001@usal.es (Maider Araceli Urbón Jiménez), JaimeGabrielVegas@usal.es (Jaime Gabriel Vegas), adeluis@usal.es (Ana de Luis Reboredo), lancho@usal.es (Belén Pérez Lancho), abg@usal.es (Ana-Belén Gil-González)

por grupos. Los resultados demuestran el potencial del enfoque para estimar la absorción futura de CO_2 y optimizar decisiones de gestión forestal sostenible, contribuyendo a la transición hacia una economía baja en emisiones.

Keywords: créditos de carbono, inteligencia artificial, forestación, reforestación, modelado predictivo, cambio climático

---

Índice

1. Introducción

El cambio climático es uno de los mayores desafíos globales y su manifestación más directa es el aumento de las concentraciones atmosféricas de dióxido de carbono (CO_2), con impactos sobre criosfera, extremos climáticos y ecosistemas [1]. Los bosques actúan como sumideros naturales al fijar CO_2 en biomasa vía fotosíntesis, por lo que su gestión resulta clave para la mitigación.

A lo largo de las últimas décadas, instrumentos internacionales como la *Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC)* y el *Protocolo de Kioto* [2, 3] han establecido los marcos regulatorios para reducir las emisiones de gases de efecto invernadero mediante mecanismos basados en el mercado. En este contexto surgen los *créditos de carbono*, unidades que representan la cantidad de dióxido de carbono (CO_2), habitualmente una tonelada, que ha sido capturada o cuya emisión ha sido evitada a través de proyectos certificados de mitigación.

Entre las actividades elegibles, la forestación y reforestación destacan por su capacidad de actuar como sumideros naturales de carbono, fijando CO_2 en la biomasa y el suelo. No obstante, para que estas actuaciones puedan generar créditos de carbono válidos, deben cumplir una serie de criterios técnicos y legales establecidos en la normativa internacional sobre cambio climático y en su aplicación a nivel nacional. En particular, estos requisitos derivan de las reglas de contabilidad de sumideros forestales adoptadas en el marco de la Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC) y del Protocolo de Kioto, concretadas en los Acuerdos de Marrakech, así como de la definición nacional de bosque comunicada por España para estos fines [4, 5]. Dichos criterios incluyen:

- **Intervención humana directa:** Los árboles deben ser el resultado de actividades de intervención humana directa, tales como la plantación, la siembra o el fomento deliberado de la regeneración natural. Este requisito se deriva de la definición de *forestación* y *reforestación* establecida en el Protocolo de Kioto, que excluye expresamente la regeneración natural no inducida por la acción humana [4].
- **Período mínimo de permanencia:** El proyecto debe garantizar la permanencia del sumidero de carbono durante un período prolongado (habitualmente del orden de 20-30 años), con el fin de asegurar que el carbono capturado no sea liberado de forma prematura a la atmósfera. Este criterio responde al principio de permanencia exigido en la conta-

bilidad de sumideros forestales del régimen LULUCF y en los marcos de aplicación nacionales y europeos, lo que excluye cultivos de corta rotación cuyo carbono se libera tras la cosecha [4, 6].

- **Superficie mínima de 1 hectárea:** El área objeto del proyecto debe tener una extensión mínima de 1 hectárea. Este umbral procede de la definición nacional de bosque adoptada por España dentro de los rangos permitidos por los Acuerdos de Marrakech (0,05–1 ha), comunicada oficialmente a la CMNUCC [5].
- **Fracción mínima de cabida cubierta del 20 %:** Para que un terreno sea considerado bosque, la cobertura de copas de las especies arbóreas debe alcanzar al menos el 20 % de la superficie. Este valor corresponde a la elección nacional realizada por España para la definición de bosque a efectos de contabilidad climática [5].
- **Altura mínima de los árboles maduros de 3 metros:** Las especies arbóreas deben ser capaces de alcanzar una altura mínima de 3 metros en su madurez. No es necesario que dicha altura se alcance en el momento inicial del proyecto, pero sí que sea alcanzable bajo condiciones normales de crecimiento. Este criterio forma igualmente parte de la definición nacional de bosque comunicada por España conforme a las decisiones adoptadas bajo la CMNUCC [5].

Este trabajo presenta **GreenWest**, un modelo de inteligencia artificial para estimar la cantidad de carbono que capturará un cultivo forestal en España a partir de variables de vegetación, clima y terreno en un período de 20 a 30 años. Este enfoque innovador tiene el potencial de transformar la gestión de proyectos de forestación y reforestación, optimizando las prácticas de plantación y maximizando la cantidad de carbono que se puede capturar en estos ecosistemas.

La pregunta operativa es: *dadas las características iniciales de una plantación, ¿cuánto CO_2 contendrá tras t años? con t número natural.* Para responderla, se integran datos del **Inventario Forestal Nacional** (IFN2–IFN4, MITECO) [7], reanálisis **ERA5-Land** [8] e **índices espectrales Landsat** (Collection 2, L2) [9] en una base de datos relacional jerárquica descrita en un trabajo complementario [10].

Este modelo no solo mejorará la comprensión del comportamiento de los sumideros de carbono, sino que también proporcionará herramientas útiles para la toma de decisiones estratégicas tanto en el ámbito empresarial como en el ambiental. De esta forma, el proyecto *GreenWest* contribuye a la

transición hacia una economía baja en carbono, alineándose con los objetivos globales de sostenibilidad establecidos en el marco de la CMNUCC y el *Protocolo de Kioto*, y promoviendo la creación de un mercado de créditos de carbono más eficiente y accesible para los actores económicos involucrados en la gestión de los recursos naturales.

2. Objetivos y Justificación

El presente estudio tiene como objetivo principal desarrollar un modelo de inteligencia artificial capaz de predecir con precisión la capacidad de absorción de dióxido de carbono (CO_2) en cultivos forestales españoles. Este modelo se basa en variables que describen la especie arbórea, las características del terreno y las condiciones climáticas. A partir de este objetivo general se derivan varias metas específicas, que en conjunto justifican la relevancia y aplicabilidad del proyecto.

2.1. Objetivos específicos

- **Desarrollar un modelo predictivo robusto:** Construir un modelo de aprendizaje automático que estime la cantidad de CO_2 que será capturado a lo largo del tiempo por un cultivo forestal, a partir de datos como especie, tipo de suelo, clase diamétrica, clima y otras variables relevantes.
- **Optimizar la captura de carbono:** Utilizar el modelo para identificar combinaciones óptimas de especies y terrenos que maximicen la fijación de carbono, contribuyendo a la planificación eficiente de proyectos de (re)forestación.
- **Asegurar la compatibilidad con las normativas internacionales:** Garantizar que las predicciones y salidas del modelo sean compatibles con los marcos normativos definidos por la *Convención Marco de las Naciones Unidas sobre el Cambio Climático* (CMNUCC) y el *Protocolo de Kioto*, cumpliendo así los criterios necesarios para la validación de créditos de carbono.
- **Analizar los factores determinantes del desarrollo forestal:** Estudiar la influencia de variables climáticas (como la temperatura y la precipitación) y edáficas (como el tipo de suelo o la pendiente) sobre el crecimiento forestal y su capacidad de capturar carbono.
- **Apoyar la toma de decisiones ambientales y empresariales:** Proporcionar una herramienta práctica y validada que permita a técnicos, gestores y empresas seleccionar las especies más adecuadas y planificar actuaciones de forestación con la mayor eficiencia posible en términos de secuestro de carbono.

2.2. Justificación

La necesidad de contar con herramientas predictivas para estimar la captura de CO_2 se ha intensificado ante el crecimiento del mercado voluntario

de créditos de carbono, y las obligaciones adquiridas: cada país debe reportar sus emisiones y absorciones de gases de efecto invernadero, y puede utilizar actividades de (re)forestación como mecanismos de compensación.

Para que estos proyectos sean elegibles, deben cumplir criterios específicos, los cuales hacen imprescindible disponer de modelos que no solo estimen el carbono actual, sino que sean capaces de prever su evolución a futuro con base en condiciones iniciales y variables predictoras.

Este trabajo busca cubrir ese vacío mediante el uso de inteligencia artificial aplicada a datos reales y multifuente. Integrar su manejo dentro del sistema de créditos de carbono puede representar una importante oportunidad para la economía local y para la mitigación del cambio climático.

3. Revisión de la Literatura

El secuestro de carbono en ecosistemas forestales ha cobrado una importancia creciente en la literatura científica, impulsada tanto por los compromisos internacionales en materia de cambio climático como por el auge de los mercados de créditos de carbono. Esto ha motivado el desarrollo de modelos orientados a cuantificar la biomasa forestal y estimar el contenido de carbono, aprovechando avances recientes en sensores remotos y técnicas de inteligencia artificial (IA).

Una de las estrategias más consolidadas para la cuantificación del carbono forestal es la estimación del carbono almacenado en un momento dado a partir de datos de teledetección. Goetz et al. (2009) [11] revisan el uso de observaciones satelitales, incluyendo sensores ópticos como MODIS y Landsat, en modelos empíricos de biomasa aérea, destacando su aplicabilidad a escala regional, especialmente en ecosistemas boreales. Este tipo de estimaciones suele basarse en regresiones lineales o modelos de mínimos cuadrados generalizados, con coeficientes de determinación habitualmente entre 0.6 y 0.8, dependiendo de la resolución espacial y la heterogeneidad del ecosistema.

La aplicación de aprendizaje profundo ha permitido mejorar sustancialmente la precisión y resolución espacial de estas estimaciones. Por ejemplo, Zhang et al. (2022) [12] integran imágenes Sentinel-2 con redes neuronales convolucionales, alcanzando un R^2 de 0.84 para estimar el carbono en bosques subtropicales. Del mismo modo, Jiang et al. (2022) [13] desarrollan el modelo *ForestCarbonAI*, entrenado con datos multiespectrales y LIDAR, con el que generan mapas de carbono forestal de alta resolución (10 m), reportando errores medios absolutos (MAE) inferiores a 3.5 tC/ha en zonas templadas. Otros trabajos recientes, como Reiersen et al. (2022) [14] o Dong et al. (2023) [15], también demuestran la eficacia del *deep learning* para estimaciones estáticas, aunque se centran en contextos tropicales y no consideran el componente temporal.

Frente a estos enfoques descriptivos, algunas iniciativas han intentado proyectar la evolución del carbono a futuro. En el ámbito nacional, el Ministerio para la Transición Ecológica (MITECO) ha implementado herramientas como la calculadora ex ante de absorciones [16], que permite obtener estimaciones simplificadas del carbono que puede fijarse en una plantación forestal en función de la especie y la zona agroclimática. No obstante, este instrumento se basa en coeficientes tabulados y no incorpora variables edafoclimáticas reales ni técnicas de modelización basadas en datos, lo que limita su precisión

y capacidad de adaptación a contextos específicos.

En este escenario, el presente trabajo propone una metodología innovadora centrada en la predicción dinámica de carbono a largo plazo. A diferencia de los modelos anteriores, que estiman el carbono ya almacenado, este estudio se enfoca en anticipar cuánto carbono capturará un cultivo forestal en un horizonte temporal concreto. Para ello, se estudian diversos modelos de aprendizaje supervisado entrenados con datos históricos del Inventario Forestal Nacional (IFN2, IFN3 e IFN4), variables climáticas de Copernicus, características edáficas y métricas espectrales derivadas de imágenes Landsat [17, 18, 19]. Los detalles sobre la arquitectura del modelo, las variables utilizadas, los algoritmos implementados y las métricas de evaluación se desarrollan en las siguientes secciones.

4. Estado del Arte

TODO: Todo ello

4.1. Contexto y formulación del problema

La estimación de *[nombre de la variable objetivo]* se aborda como un problema de regresión supervisada, donde el objetivo es aprender una función $f : \mathbb{R}^p \to \mathbb{R}$ que minimice el error de predicción bajo criterios como RMSE o MAE (**??**). Se requieren diseños de validación que eviten fuga de información (*leakage*) y respeten la estructura de los datos (por ejemplo, validación por grupos o espacio-temporal) (**?**).

4.2. Modelado predictivo para variables continuas

Los enfoques más empleados incluyen modelos lineales regularizados (Ridge, Lasso, Elastic Net) (**??**), métodos basados en árboles (Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost) (**?????**) y redes neuronales profundas para tabulares e imagen (**?**). La elección suele balancear interpretabilidad, robustez ante no linealidades e interacción entre variables, coste computacional y requisitos de datos.

4.3. Validación y evaluación

La literatura recomienda validación cruzada estratificada o por grupos para estimar el error fuera de muestra y evitar optimismo en la evaluación (**?**). Cuando existen dependencias (espaciales, temporales o por *grupo*), se emplean variantes como GroupKFold o bloqueos espacio-temporales (**?**). Las métricas habituales para regresión incluyen RMSE, MAE, R^2 y, cuando procede, métricas relativas (p. ej., MAPE). Es buena práctica reportar distribuciones (mediana, IQR) además de promedios y comparar contra *baselines* fuertes.

4.4. Selección de variables

Los métodos se agrupan en: (i) **filtro**, p. ej., correlación/ANOVA, información mutua y mRMR (**?**); (ii) **envoltura** (*wrapper*), como forward/backward selection o RFE (**?**); y (iii) **embebidos**, que integran la selección durante el ajuste del modelo (Lasso/Elastic Net, importancia en árboles/boosting) (**??**). Recientemente, se han popularizado enfoques de *stability selection* y métodos de importancia condicional para reducir sesgos por colinealidad (**??**).

4.5. Datos, preprocesado y fuga de información

La literatura subraya la importancia de: imputación apropiada, codificación de categóricas (one-hot, target encoding con CV anidada), tratamiento de outliers y escalado cuando el modelo lo requiere (**?**). Debe evitarse la fuga de información aplicando todo el preprocesado dentro del *pipeline* y re-ajustándolo por pliegue.

4.6. Explicabilidad e incertidumbre

Para interpretar predictores y robustez se usan curvas de dependencia parcial, perfiles acumulados y explicaciones SHAP (**??**). La estimación de la incertidumbre puede abordarse con ensambles, *quantile regression*, conformal prediction o bayesianos aproximados (**?**).

4.7. Trabajos relacionados y brechas

Estudios previos han aplicado *[modelos]* sobre *[dominio/datos]* con *[métricas]* y *[protocolos de CV]* (**??**). Persisten brechas en: (i) control explícito de fuga por grupos/espacio-tiempo; (ii) evaluación sistemática del impacto de la selección de variables; (iii) análisis de incertidumbre y generalización fuera de dominio.

4.8. Síntesis

En resumen, el estado del arte respalda: (1) protocolos de validación estrictos (p. ej., GroupKFold), (2) comparación de familias de modelos con *baselines* fuertes, (3) selección de variables combinando filtros (mRMR/MI) y técnicas embebidas, y (4) reporte de interpretabilidad e incertidumbre. Sobre esta base se diseña la metodología presentada en la Sección **??**.

5. Metodología

Esta sección describe el procedimiento seguido para el entrenamiento y validación de los modelos predictivos desarrollados. La metodología se fundamenta en la identificación de los factores que determinan el crecimiento forestal y, en consecuencia, la capacidad de los ecosistemas para capturar carbono a lo largo del tiempo. El enfoque integra información estructural, climática y espectral procedente del Inventario Forestal Nacional (IFN) y de otras fuentes ambientales, con el propósito de construir modelos robustos que permitan predecir el contenido de carbono acumulado en la biomasa viva.

El carbono fijado por los árboles se acumula progresivamente en su biomasa, en función del tamaño y vigor de los individuos, los cuales están condicionados por variables ambientales, topográficas y de competencia intraespecífica. Las condiciones meteorológicas, como la temperatura y la precipitación, inciden directamente en la fotosíntesis y en la disponibilidad hídrica; la orientación, la pendiente y la altitud modifican la radiación incidente y el microclima local; mientras que la densidad de árboles por unidad de superficie determina el nivel de competencia por los recursos, variando según la especie y su tolerancia ecológica [20].

A partir de estos fundamentos, se construyó una base de datos relacional que integra información forestal, climática y espectral a nivel de parcela, especie y clase diamétrica. Esta estructura permite caracterizar con precisión la dinámica del bosque entre inventarios sucesivos y alimentar modelos predictivos capaces de estimar el contenido futuro de carbono a partir de las condiciones observadas en el pasado.

5.1. Origen y estructura de los datos

La base de datos empleada en este trabajo integra información forestal, climática y espectral estructurada en torno a la parcela como unidad básica. Cada parcela se describe mediante sus coordenadas geográficas, características edáficas y su evolución a través de distintos inventarios (IFN2, IFN3, IFN4).

Los datos forestales incluyen información por especie y clase diamétrica, como número de pies o carbono aéreo, radical y total. Estos valores permiten caracterizar con precisión la estructura y crecimiento de la vegetación.

A cada parcela se asocian también estadísticas climáticas agregadas por estación e inventario: temperaturas (superficie, aire y subsuelo) y precipita-

ciones, resumidas mediante métricas como media, máxima, mínima y desviación típica.

Finalmente, se incorporan índices espectrales derivados de imágenes satelitales (NDVI, EVI, NDII, GNDVI), que permiten cuantificar propiedades biofísicas de la vegetación:

- **NDVI (Normalized Difference Vegetation Index):** estima la actividad fotosintética.
- **EVI (Enhanced Vegetation Index):** mejora la sensibilidad en zonas densamente vegetadas.
- **NDII (Normalized Difference Infrared Index):** refleja el contenido hídrico de la vegetación.
- **GNDVI (Green NDVI):** variante del NDVI basada en la banda verde, sensible a la cantidad de clorofila.

5.1.1. Estrutura de la base de datos

Estos datos se organizan en las siguientes entidades troncales (tablas):

- **parcelas**: contiene la información básica ligada a la localización de cada parcela (características edáficas y climáticas).
- **parcela_inventario**: describe el estado de cada parcela en un inventario determinado, incluyendo atributos edáficos y de contexto (p. ej., textura del suelo, fracción de cabida cubierta...).
- **parcela_inventario_especie**: detalla la presencia y condición de cada especie dentro de una parcela e inventario, incorporando descriptores de masa y tratamientos silvícolas.
- **parcela_inventario_especie_cd**: describe las poblaciones arbóreas por parcela, especie y *clase diamétrica*: n.º de pies, área basimétrica, volúmenes, altura de los especímenes, carbono captura···
- **parcela_especie_arbol**: caracteriza los pies mayores identificados por parcela y especie en el inventario cuarto. Recoge las caracteristicas particulares de cada pie como altura, diámetros, ubicación respecto del centro de la parcela, volumen y carbono capturado.
- **parcela_inventario_estacion**: almacena agregados climático-biofísicos por estación en la misma granularidad parcela–inventario, incluyendo variables como precipitación y temperatura, junto a índices de vegetación (NDVI, EVI, NDII, GNDVI).
- **especies** y **grupos**: recogen la información taxonómica y su clasificación jerárquica, estableciendo la relación entre especies individuales y

grupos funcionales.

Cada variable categórica posee una tabla de catálogo propia (`cat_`), donde se definen los valores posibles y sus descripciones. Por ejemplo, `cat_textura`, `cat_nivel1`, `cat_tratmasa` o `cat_origen`. Todas siguen un patrón uniforme: la clave primaria es el identificador de la variable (`<variable>_id`), y las tablas troncales referencian este mismo campo como clave foránea. Además la base de datos incluye una tabla llamada `meta_variables` que recoge los metadatos.

La Figura 5.1 muestra el esquema general de las tablas troncales y sus principales relaciones. Este diagrama resume la estructura interna de la base de datos y su jerarquía de dependencias.

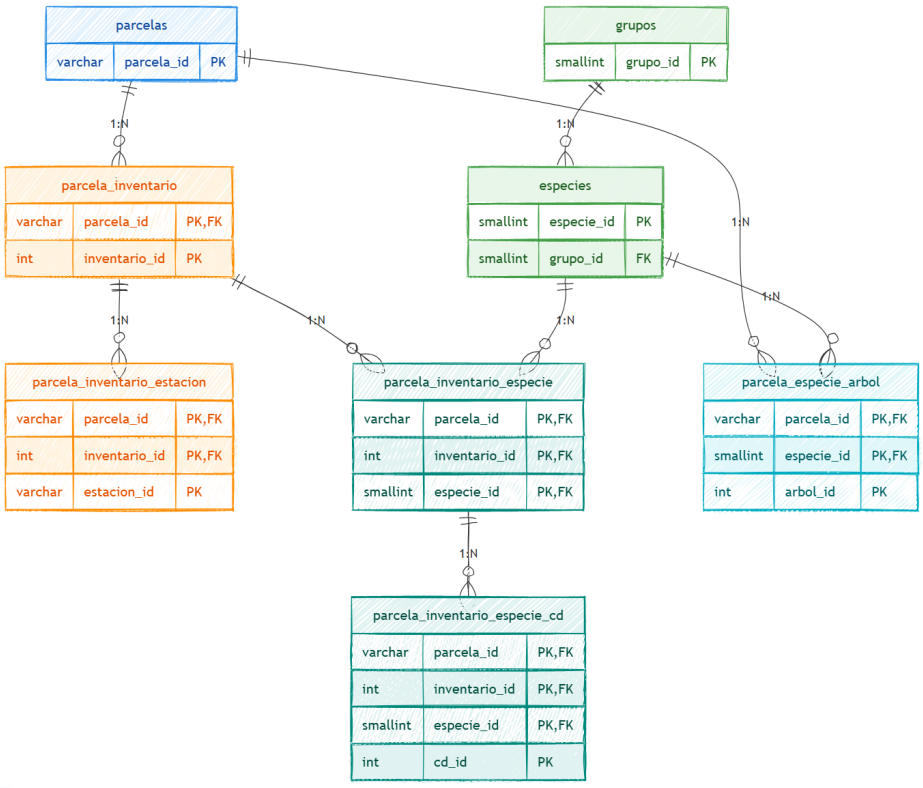

Figura 5.1. Esquema relacional de las tablas principales de la base de datos. Tabla extraida de [10], donde se pueden consultar más detalles sobre las variables.

5.1.2. Diccionario resumido de variables

Tabla 5.1. Resumen de variables principales por entidad. Tabla extraida de [10].

Variable	Descripción	Unidad	Tipo de dato
parcelas			
`parcela_id`	Identificador único de parcela (IFN).	–	Identificador
`latitud`, `longitud`	Coordenadas geográficas (WGS84).	°	Geográfico
`coorx`, `coory`	Coordenadas UTM; `huso` especifica zona.	m (UTM)	Geográfico
`elevacion`	Cota sobre el nivel del mar (NASADEM).	m	Numérico
`pendiente`	Inclinación del terreno.	°	Numérico
`orientacion`	Orientación del terreno (0–360).	°	Numérico
`presencia_id`	Presencia en IFN → `cat_presencia`.	–	Categórico
`tipsuelo1_id`, `tipsuelo2_id`, `tipsuelo3_id`	Tipos de suelo → `cat_tipsuelo*`.	–	Categórico
`rocosidad_id`	Rocosidad → `cat_rocosidad`.	–	Categórico
`radio`, `superficie`	Radio de parcela y superficie derivada.	m; ha	Numérico
parcela_inventario			
`parcela_id`, `inventario_id`	Clave compuesta (parcela-inventario).	–	Identificador
`ano`	Año de apeo.	año	Numérico
`nivel1_id`, `nivel2_id`	Morfoestructura. → `cat_nivel*`.	–	Categórico
`textura_id`	Textura de suelo → `cat_textura`.	–	Categórico
`merosiva_id`	Manifestaciones erosivas → `cat_merosiva`.	–	Categórico
`matorg_id`	Materia orgánica → `cat_matorg`.	–	Categórico
`modcomb_id`	Modelo de combustible → `cat_modcomb`.	–	Categórico
`disesp_id`	Distribución espacial → `cat_disesp`.	–	Categórico
`comesp_id`	Composición específica → `cat_comesp`.	–	Categórico
`fccarb`, `fcctot`	Fracción de cabida cubierta (árboles).	%	Numérico
parcela_inventario_especie			
`parcela_id`, `inventario_id`, `especie_id`	Clave compuesta (parcela-inventario-especie).	–	Identificador
`ocupa`	Grado de ocupación de la especie.	(0–10)	Numérico
`estado_id`	Estado de desarrollo. → `cat_estado`.	–	Categórico

Continúa en la siguiente página

Variable	Descripción	Unidad	Tipo de dato
fpmasa_id	Forma principal de masa → cat_fpmasa.	–	Categórico
tratmasa_id	Tratamientos selvícolas → cat_tratmasa.	–	Categórico
orgmasa1_id	Origen de masa (IFN3/4)→ cat_orgmasa1.	–	Categórico
masa_id	Clasificación de masa → cat_masa.	–	Categórico
origen_id	Origen de la masa (IFN2) → cat_origen.	–	Categórico
parcela_inventario_especie_cd			
parcela_id, inventario_id, especie_id	Clave compuesta (parcela-inventario-especie-cd).	–	Identificador
cd_id	Clase diamétrica (CD) reglamento IFN.	cm	Numérico discreto
npies	Número de pies.	pies/ha	Numérico
abas	Área basimétrica.	m^2/ha	Numérico
vcc, vsc, vle	Volúmenes (con/sin corteza; leñas).	m^3/ha	Numérico
iavc	Incremento anual del volumen con corteza.	m^3/ha·año	Numérico
ca, cr	Carbono aéreo y radical.	t/ha	Numérico
ht	Altura media (modelo CatBoost).	m	Numérico
carbono_bruto	Carbono total estimado (alometrías).	t	Numérico
parcela_especie_arbol			
parcela_id, especie_id	Clave compuesta (parcela–especie–árbol).	–	Identificador
arbol_id	Identificador del árbol dentro de parcela y especie.	–	Entero
rumbo	Rumbo desde el centro de la parcela al árbol.	grados centesimales	Numérico
distancia	Distancia desde el centro de la parcela al árbol.	m	Numérico
cd	Clase diamétrica (reglamento IFN).	cm	Numérico discreto
ht	Altura total del árbol inventariado.	m	Numérico
dn1, dn2	Diámetros normales perpendiculares.	mm	Numérico
abas	Área basimétrica del pie medido.	m^2	Numérico
iavc	Incremento anual del volumen con corteza.	dm^3/año	Numérico
vcc, vsc, vle	Volúmenes (con corteza, sin corteza, leñas).	dm^3	Numérico
ca, cr	Carbono aéreo y radical del árbol.	t	Numérico

Continúa en la siguiente página

Variable	Descripción	Unidad	Tipo de dato
parcela_inventario_estacion			
parcela_id, inventario_id, estacion_id	Clave compuesta (agregado estacional).	–	Identificador
PR_*	Estadísticos de precipitación (mean, max, min, std, sum).	mm/(m^2·día), mm/m^2	Numérico
T2M_*, SKT_*	Aire 2 m y temperatura superficial (mean, max, min, std).	°C	Numérico
STL1_*–STL4_*	Temperatura del suelo por niveles (mean, max, min, std).	°C	Numérico
NDVI_*, EVI_*, NDII_*, GNDVI_*	Índices de vegetación (max, mean, median, min, std).	adimensional	Numérico
especies y **grupos**			
especie_id	Identificador de especie IFN.	–	Identificador
nombre, sinonimia	Denominación IFN y sinónimos.	–	Texto
tipo_especie	0 = conífera; 1 = frondosa.	–	Categórico
grupo_id	Grupo funcional → grupos.	–	Identificador
grupos.nombregrupo	Nombre del grupo.	–	Texto

5.1.3. Cardinalidad y completitud

El volumen de entradas por tabla es:

Tabla	Número de registros
parcelas	52,298
parcela_inventario	147,995
parcela_inventario_especie	417,119
parcela_inventario_especie_cd	1,191,070
parcela_especie_arbol	855,860
parcela_inventario_estacion	470,056
especies	195
grupos	33

5.2. Variables objetivo

El objetivo del modelo es estimar el **carbono total** que una parcela forestal capturará en un horizonte temporal de 20–30 años, a partir de las condiciones observadas en inventarios previos. Para ello se contemplan dos variables de respuesta complementarias, ambas derivadas de los datos del Inventario Forestal Nacional (IFN), que permiten analizar el contenido de carbono desde perspectivas distintas: una normalizada por superficie y otra en términos absolutos.

1. `c` (tC/ha): representa el **carbono total contenido en la biomasa viva aérea y subterránea** por unidad de superficie, expresado en *toneladas de carbono por hectárea*. Su cálculo se basa en la suma de las estimaciones de carbono aéreo (`ca`) y radical (`cr`) reportadas por el IFN. En los casos con valores faltantes, se completó la información mediante un modelo de *Random Forest Regressor* ajustado sobre variables dendrométricas observadas (Especie, CD, VSC, NPies, ABas, IAVC, VCC y VLE), alcanzando un rendimiento satisfactorio ($R^2_{test} > 0{,}90$). Esta variable es coherente con los formatos internacionales de reporte de inventarios forestales y permite comparar el contenido de carbono entre parcelas o especies.
2. `carbono_bruto` (tC): corresponde al **carbono total capturado por parcela y especie**, expresado en *toneladas de carbono totales*. Su estimación se realiza de forma trazable y físicamente interpretable a partir de variables medidas directamente en campo: número de pies (`npies`), altura media (`ht`), tipo de especie (`clase_especie`) y clase diamétrica (`cd_id`). El cálculo sigue un modelo alométrico adaptado de [21] y las directrices del IPCC [20], incorporando tanto la biomasa aérea como la biomasa radical mediante la relación Parte Radical:Parte Aérea (R). El resultado se expresa en toneladas de carbono totales por parcela, sin normalizar por superficie, lo que facilita la trazabilidad del proceso y la comparación entre inventarios sin depender de factores de expansión específicos del IFN. En coherencia con los criterios de proyectos de forestación y reforestación, las observaciones correspondientes a brinzales o plantones se consideran con valor de carbono nulo, dado que las fases tempranas de desarrollo no se contabilizan oficialmente como carbono capturado.

Estas dos variables resumen el contenido de carbono forestal desde enfoques complementarios: `c` (tC/ha) permite la comparación espacial y temporal

entre masas forestales, mientras que `carbono_bruto` (tC) ofrece una medida absoluta y directamente derivada de las observaciones de campo. Ambas constituyen los objetivos principales del modelado predictivo, orientado a estimar el carbono acumulado en el **IFN4** a partir de las condiciones registradas en los inventarios anteriores (**IFN2** e **IFN3**).

Para mayor detalle sobre el origen de estas variables consultar [10].

5.3. Supuestos de elegibilidad y verificación externa

Como ya se ha introducido, para que un proyecto forestal sea elegible en programas de *créditos de carbono* en España debe cumplir algunos requisitos técnicos [20, 19]. A continuación se resume cada criterio y la forma en que se aborda en este estudio:

- **Intervención humana directa.** El incremento de carbono debe proceder de actuaciones planificadas (reforestación, restauración o manejo sostenible). En nuestro caso, el modelo se entrena sobre datos observacionales (IFN2–IFN3–IFN4); por tanto, la *verificación de intervención* no se deduce del modelo, sino que se contempla como *condición externa* de elegibilidad del proyecto a evaluar.
- **Permanencia mínima.** Para caracterizar el crecimiento de las parcelas forestales en los datos que alimentan el modelo, es necesario disponer de dos mediciones sucesivas de cada parcela, separadas por un intervalo temporal conocido. Estas mediciones permiten cuantificar la evolución de las variables forestales y, por tanto, estimar el incremento de carbono asociado al crecimiento del arbolado durante dicho periodo.

  En este trabajo, el objetivo es predecir el contenido de carbono correspondiente al **IFN4**, utilizando como información explicativa las variables observadas en inventarios anteriores. Dado que los inventarios tercero y cuarto comparten una estructura homogénea y un conjunto de variables comparable la elección más directa para el entrenamiento del modelo sería emplear exclusivamente estos dos inventarios. Esta estrategia aprovecha la coherencia estructural de los inventarios más recientes, que incluyen un mayor número de variables y una caracterización más detallada del terreno.

  El intervalo de tiempo entre los inventarios **IFN3** e **IFN4** es relativamente corto: la Figura 5.2 muestra la distribución de la diferencia de años entre las mediciones del IFN3 y el IFN4. Como puede observarse, la mayoría de las parcelas presentan intervalos comprendidos entre 6 y

17 años, un rango demasiado estrecho para evaluar la estabilidad del modelo en horizontes más amplios.

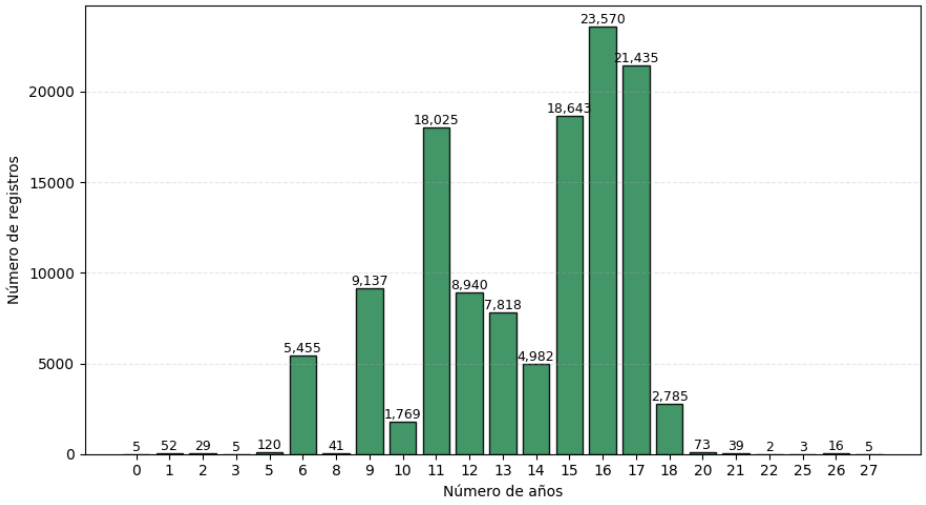

Figura 5.2. Distribución de la diferencia de años entre los inventarios tercero y cuarto.

Para ampliar la cobertura temporal y mejorar la capacidad de generalización del modelo, se optó por unificar la información de los inventarios **IFN2** e **IFN3** como base explicativa para la predicción del **IFN4**. Esta integración permite disponer de pares de mediciones de parcelas separadas por intervalos que oscilan entre 6 y 29 años, lo que constituye un rango mucho más representativo del horizonte de 20–30 años establecido como referencia.

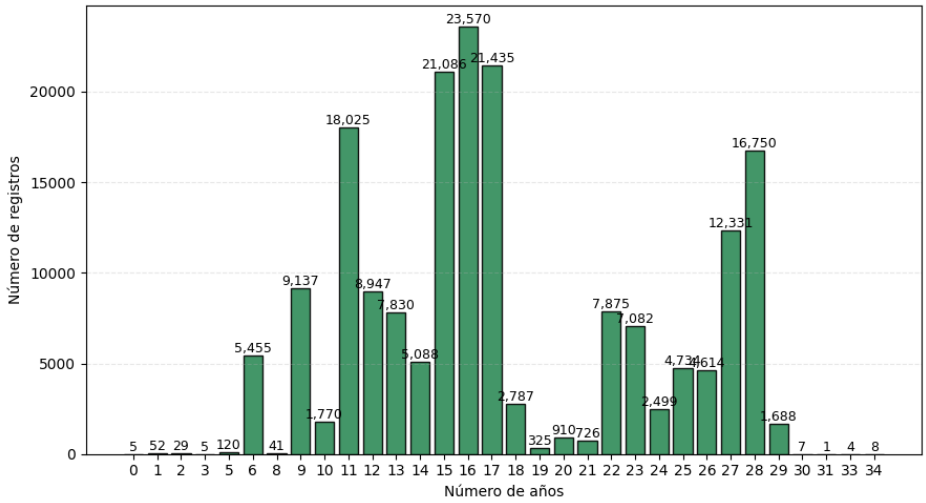

Figura 5.3. Distribución de la diferencia de años entre los inventarios IFN2–IFN3 e IFN3–IFN4.

De esta forma, el modelo se entrena y valida sobre un conjunto de datos más diverso y equilibrado, tanto en estructura como en amplitud temporal, manteniendo la coherencia metodológica y la trazabilidad de las estimaciones.

- **Superficie mínima de 1 ha.** Este criterio se considera externo al alcance del modelo predictivo, ya que el aprendizaje se realiza a nivel de parcela e inventario y no sobre polígonos de superficie total. En la práctica, la verificación de la superficie se realiza *ex ante*, sobre la geometría declarada del proyecto forestal. En los terrenos forestales generados a partir de intervención humana directa, como plantaciones o repoblaciones, la extensión suele presentar una estructura homogénea, con una especie dominante, edades coetáneas y densidades estandarizadas. Bajo estas condiciones, el carbono total es proporcional a la superficie: duplicar el área de una masa forestal homogénea implica aproximadamente duplicar su carbono almacenado. Por tanto, la variable de superficie no afecta al ajuste interno del modelo y su cumplimiento puede evaluarse fácilmente a nivel de proyecto, sin comprometer la validez de las predicciones.
- **Fracción mínima de cabida cubierta del 20 %.** La base de datos dispone de `fccarb` (arbórea) y `fcctot` (total). Este umbral se aplica como *filtro de elegibilidad* previo o posterior al modelado, sin modificar

la arquitectura del modelo (`fccarb`> 20).

- **Altura mínima de 3 m en la madurez.** Este requisito se refiere a la altura que alcanzan los árboles en su fase de pleno desarrollo, y no a la altura inicial de los plantones. Por tanto, las mediciones realizadas durante las etapas tempranas de crecimiento no determinan la elegibilidad del proyecto, siempre que las especies seleccionadas sean capaces de superar los 3 metros en la madurez. Este criterio se evalúa de forma externa al modelo, mediante la selección de especies forestales adecuadas y la verificación con fuentes auxiliares (catálogos silvícolas o tipologías de masa). En la práctica, el cumplimiento del requisito depende de una decisión de diseño del proyecto: no plantar especies cuyo tamaño adulto sea inferior a 3 metros. Por ello, la altura no interviene directamente en el entrenamiento, aunque sí condiciona la elegibilidad final del proyecto forestal.

5.4. Preparación y tratamiento de los datos

Como ya se ha introducido el entrenamiento se realiza en dos líneas según la variable objetivo: carbono en toneladas por hectárea (`c` de **IFN4**) o carbono en toneladas (`carbono_bruto` de **IFN4**); y según la información que se usa como explicativa: **IFN3** o **IFN3** e **IFN2**. Se plantea la preparación y filtrado de los datos en términos generales.

5.4.1. Filtrado de registros

Se descartan todas aquellas parcelas en las que el valor de carbono total (variable objetivo) en la segunda inventariación es inferior a la primera. Estos casos suelen deberse a episodios de deforestación, incendios u otras perturbaciones, y no representan un crecimiento forestal neto.

El conjunto de datos se restringe únicamente a las parcelas que presentan una `fccarb` (fracción de cabida cubierta arbórea) igual o superior al 20 % en el inventario explicativo. Este umbral define la proporción mínima de superficie ocupada por copas de árboles respecto al área total de la parcela, y constituye una de las condiciones esenciales para considerar una superficie como terreno forestal. La exclusión de parcelas con `fccarb` inferior al 20 % permite asegurar que las estimaciones de carbono se realicen sobre masas forestales consolidadas, evitando sesgos asociados a áreas agrícolas o matorrales. A los datos del **IFN2** no se les aplica dicho filtro porque no disponen de la variable `fccarb`.

5.4.2. Cálculo y agregación de variables

Cada registro de entrada se genera a nivel de combinación parcela–especie, incorporando las variables correspondientes de la primera medición y la variable objetivo (carbono) de la segunda medición (IFN4). Las variables de `parcela` y `parcela_inventario` se desdoblan para cada especie. Las entradas de la tabla `parcela_inventario_especie_cd` se agrupan por parcela y especie y se comprimen en una única entrada creando un conjunto de variables para cada clase diamétrica.

La Tabla 5.2 resume las variables empleadas como entrada al modelo, integradas desde las distintas tablas que conforman la base de datos relacional.

5.4.3. Reclasificación de las variables `pendiente` y `orientacion`

Las variables topográficas originales `pendiente` (en grados) y `orientacion` (acimut en grados) se registran de forma continua en las parcelas del IFN. Sin embargo, desde el punto de vista ecológico su efecto sobre la acumulación de carbono suele ser no lineal y está asociado a clases discretas (e.g. laderas suaves frente a escarpadas, exposición norte frente a sur), por lo que resulta más adecuado tratarlas como factores categóricos.

A partir de la distribución empírica y de criterios habituales en estudios de fisiografía forestal, se definió una variable categórica `pendiente_cat` mediante cortes en grados:

- $< 5°$: *muy suave*,
- $5-10°$: *suave*,
- $10-15°$: *moderada*,
- $15-20°$: *fuerte*,
- $20-30°$: *muy fuerte*,
- $30-50°$: *escarpada*,
- $> 50°$: *extrema*.

Esta reclasificación permite capturar diferencias funcionales relevantes (accesibilidad, estabilidad del suelo, escorrentía, profundidad efectiva del suelo) sin asumir una relación lineal entre la pendiente y el carbono almacenado.

De forma análoga, la variable `orientacion` se reclasificó en ocho sectores cardinales equiángulos: `N`, `NE`, `E`, `SE`, `S`, `SO`, `O` y `NO`. La nueva variable `orientacion_cat` agrupa orientaciones con condiciones de insolación y balance hídrico similares, lo que facilita la interpretación ecológica y reduce el ruido asociado a pequeñas variaciones angulares.

Resumen de Datos de Entrada del Modelo			
Variable	**Tipo**	**Descripción**	**Anexo**
parcela_id	varchar	Identificador único de parcela.	–
especie_id, tipo_especie, grupo_id	int (CF)	Especie, tipo y grupo taxonómico.	Apéndice A.17, Apéndice A.16
ocupa	int	Grado de ocupación (0–10).	–
estado_id, fpmasa_id, tratmasa_id, orgmasa_1_id	int (CF)	Estado, forma de masa, tratamiento, organización.	Apéndice A.2, Apéndice A.3, Apéndice A.4, Apéndice A.5
tipsuelo1-3_id	int (CF)	Tipos de suelo.	Anexo Apéndice A.6
rocosidad_id, textura_id, matorg_id, modcomb_id, disesp_id, comesp_id, merosiva_id	int (CF)	Variables edáficas y estructurales.	Apéndices varios
radio, orientacion, elevacion, pendiente	float	Topografía y geometría de parcela.	–
nivel1_id, nivel2_id, fccarb, fcctot	int/float	Niveles jerárquicos y cabida cubierta.	Apéndice A.14, Apéndice A.15
npies_{CD}	float	N.º de pies por clase diamétrica.	–
periodo	int	Años entre inventarios.	–
evi, gndvi, ndii, ndvi_{stat}_{est}	float	Índices de vegetación por estación.	–
pr, skt, stl1-4, t2m_{stat}_{est}	float	Variables climáticas por estación.	–
c4, carbono_bruto4	float	Carbono IFN4 (t/ha y t).	–

Tabla 5.2. Variables de entrada del modelo. Las variables en verde están disponibles en IFN2 e IFN3; el resto solo en IFN3.

5.4.4. Agrupación de la variable `periodo`

Como se puede observar en la Figura 5.3, la distribución de la variable `periodo`, definida como el número de años transcurridos entre la medición de las variables explicativas y la observación de la variable objetivo, presenta cierta heterogeneidad en su frecuencia. Aunque el rango total de valores se extiende aproximadamente entre 0 y 34 años, algunos intervalos temporales aparecen representados por un número muy reducido de observaciones.

Esta escasez de datos en determinados valores de `periodo` puede introducir inestabilidad en el entrenamiento de los modelos, al forzar al algoritmo a aprender patrones a partir de muestras poco representativas. Para mitigar este efecto y mejorar la robustez de las predicciones, se optó por agrupar ciertos valores de `periodo` en categorías temporales más amplias, dando lugar a una nueva variable denominada `periodo_agrupado`.

El procedimiento de agrupación se diseñó de forma conservadora, manteniendo sin modificar aquellos valores con suficiente soporte muestral y agrupando únicamente los intervalos más escasos. En concreto, los valores inferiores a 6 años se agruparon en la categoría 5; los periodos entre 7 y 10 años se agruparon en 10; los comprendidos entre 18 y 21 años se agruparon en 20; y los valores iguales o superiores a 29 años se agruparon en 30. El resto de valores intermedios se mantuvieron sin modificación.

Cabe destacar que la variable `periodo_agrupado` no conserva la granularidad completa de la variable original, pero sí retiene su significado temporal esencial. Los experimentos realizados muestran que esta representación resulta más estable desde el punto de vista estadístico y conduce a modelos con un comportamiento predictivo más robusto, al reducir la sensibilidad a intervalos temporales con baja frecuencia de observaciones.

5.5. Partición y validación

Para obtener una estimación imparcial del rendimiento y evitar *fugas de información* derivadas de la estructura jerárquica de los datos, el proceso de entrenamiento se organiza en dos niveles: (i) una partición externa *hold-out* para la evaluación final y (ii) una validación cruzada interna para la selección de hiperparámetros.

Partición entrenamiento/test. Los datos, ya filtrados, se separan en un 80 % para entrenamiento y un 20 % para test. Dicha separación se hace de forma que todas las observaciones asociadas a una misma parcela queden asignadas íntegramente a uno de los subconjuntos.

Validación cruzada para selección de hiperparámetros. La selección de hiperparámetros se lleva a cabo exclusivamente sobre el conjunto de entrenamiento mediante `GridSearchCV` con métrica de optimización R^2. Se utiliza `GroupKFold` con $k = 5$ pliegues, imponiendo que no exista solape de parcelas entre pliegues (esto es, la agrupación se respeta tanto en el *hold-out* como en la validación interna).

Métricas de evaluación. El rendimiento se informa con un conjunto de medidas complementarias:

- **RMSE (Root Mean Squared Error):** raíz del error cuadrático medio entre valores observados y predichos; se expresa en las mismas unidades que la variable objetivo y penaliza con mayor peso los errores grandes. Valores más bajos indican mejor ajuste.
- $\boldsymbol{R^2}$ (coeficiente de determinación): proporción de la varianza observada explicada por el modelo (idealmente en $[0, 1]$). Valores cercanos a 1 denotan alta capacidad explicativa; puede ser negativo si el modelo es peor que la predicción constante.
- **MAE (Mean Absolute Error):** media aritmética del error absoluto, que cuantifica la desviación media entre las predicciones y los valores observados. Penaliza todos los errores de forma lineal y es más interpretable que el RMSE. Valores más bajos indican mejor ajuste.
- **SMAPE (Symmetric Mean Absolute Percentage Error):** error porcentual absoluto medio simétrico, que mide la discrepancia relativa entre valores observados y predichos normalizada por su magnitud media. Es especialmente útil para comparar el rendimiento del modelo entre distintos rangos de la variable objetivo y reduce la asimetría presente en métricas porcentuales tradicionales. Valores más bajos indican mejor ajuste relativo.

5.5.1. Codificación y normalización

Con el fin de garantizar coherencia metodológica y evitar *fugas de información* durante la validación cruzada, todas las etapas de preprocesado se integran explícitamente dentro de un `Pipeline` junto con el modelo de regresión. De este modo, los parámetros asociados al preprocesado se estiman *exclusivamente* a partir de los datos de entrenamiento de cada pliegue, y se aplican posteriormente a los datos de validación o test.

Las variables se tratan de acuerdo con su naturaleza:

- **Variables numéricas continuas**: se imputan mediante la mediana

para reducir la influencia de valores extremos y, posteriormente, se estandarizan mediante normalización *z-score* (media cero y desviación estándar unitaria). Este paso resulta especialmente relevante para modelos sensibles a la escala de las variables, como regresiones lineales, SVR o redes neuronales.

- **Variables estructurales de densidad** (`npies_*`): al representar recuentos por clase diamétrica, se imputan con valor cero cuando están ausentes y se estandarizan de forma análoga a las variables numéricas, preservando su contribución relativa en el modelo.
- **Variables categóricas**: se imputan mediante la moda, se convierten explícitamente a tipo cadena y se codifican mediante *one-hot encoding*. Se utiliza la opción `handle_unknown='ignore'` para garantizar la robustez del modelo frente a categorías no observadas durante el entrenamiento.

Esta estrategia asegura que la codificación, imputación y escalado de las variables se realicen de forma consistente en todos los modelos evaluados y que el rendimiento estimado refleje fielmente la capacidad de generalización del sistema, sin introducir sesgos derivados del acceso indebido a información del conjunto de evaluación.

5.6. Selección de variables explicativas

La selección de predictores se abordó mediante tres estrategias complementarias: (1) selección automática mediante *Featurewiz*, (2) selección basada en el criterio de mínima redundancia y máxima relevancia (*mRMR*) y (3) selección manual basada en criterios estadísticos y conceptuales.

5.6.1. Selección automática mediante Featurewiz

El algoritmo *Featurewiz* aplica un enfoque híbrido orientado a la relevancia predictiva. Primero ejecuta un filtrado por correlación, eliminando predictores altamente colineales (umbral $|r| > 0{,}70$), y posteriormente refina el conjunto mediante modelos de *Gradient Boosting* para estimar la importancia relativa de cada variable. El resultado es un subconjunto compacto de predictores con contribución significativa al rendimiento del modelo.

5.6.2. Selección mediante mRMR

El método *mRMR* (minimum Redundancy–maximum Relevance) selecciona las variables que mejor explican la variabilidad del objetivo a la vez

que minimizan la redundancia informativa entre ellas. Para ello emplea información mutua, permitiendo capturar relaciones potencialmente no lineales. Este enfoque prioriza predictores que aportan información complementaria sobre el proceso ecológico modelado, evitando duplicidades entre atributos altamente correlacionados.

5.6.3. Selección manual basada en criterios estadísticos y conceptuales

La selección manual integró criterios estadísticos (correlaciones, ANOVA y análisis de redundancia) con criterios ecológicos y de interpretabilidad. Se descartaron predictores sin asociación significativa con la variable objetivo y se redujo la colinealidad reteniendo un único representante por cada grupo altamente correlacionado. Asimismo, se garantizaron variables que describieran dimensiones esenciales del sistema (estructura del arbolado, topografía, suelo, clima e índices espectrales), asegurando un equilibrio entre precisión predictiva y coherencia biogeográfica.

5.6.4. Selección Secuencial Supervisada basada en Rendimiento Predictivo (SSSRP)

El método SSSRP complementó las estrategias anteriores mediante un enfoque explícitamente orientado al rendimiento predictivo. Se partió de un *bloque base* de variables estructurales y se evaluó el impacto marginal de cada candidato añadiéndolo individualmente y comparando el cambio en R^2 y RMSE mediante un modelo CatBoost con validación holdout estratificada por parcela. A continuación, se aplicó una estrategia de *forward selection* codiciosa, incorporando en cada iteración la variable que proporcionaba la mayor mejora y deteniendo el proceso cuando la ganancia resultaba inferior a un umbral predefinido ($\Delta R^2 > 10^{-5}$). Este procedimiento produjo un conjunto final de predictores reducido, no redundante y específicamente optimizado para maximizar el rendimiento del modelo.

5.7. Modelos evaluados

A continuación se describe el procedimiento seguido para la selección, optimización y combinación de modelos. El objetivo es construir un conjunto de predictores base sólidos y posteriormente integrarlos en un *stack-ensemble* capaz de mejorar la capacidad de generalización.

5.7.1. Modelos ensemble

Se utilizaron diversos métodos de *ensemble learning* con el fin de aumentar precisión y robustez del sistema predictivo. El principio fundamental consiste

en combinar predicciones de múltiples modelos, aprovechando su diversidad para reducir varianza, sesgo o ambos.

Técnicas empleadas:.

- **Bagging:** entrena modelos independientes sobre subconjuntos generados mediante muestreo bootstrap. Reduce varianza y mejora estabilidad.
- **Boosting:** construye modelos secuenciales donde cada uno corrige los errores del anterior. Tiende a reducir el sesgo y producir modelos altamente precisos.
- **Stacking:** integra múltiples modelos base mediante un meta-modelo entrenado sobre sus predicciones. Permite capturar relaciones no lineales entre las salidas de los modelos base.

5.7.2. Boosting y aprendizaje secuencial

El conjunto de modelos de boosting evaluados incluye:

- **XGBoost:** implementación avanzada del *gradient boosting*, que incorpora regularización L1/L2, optimización mediante segundo orden y manejo interno de valores faltantes.
- **LightGBM:** algoritmo especialmente eficiente, basado en crecimiento *leaf-wise*, capaz de manejar grandes volúmenes de datos y con soporte nativo para variables categóricas.
- **CatBoost:** optimizado para variables categóricas y robusto frente a ruido mediante técnicas como *ordered boosting*.
- **Gradient Boosting Decision Trees (GBDT):** implementación clásica del algoritmo basado en descenso por gradiente sobre residuos.
- **AdaBoost:** técnica que ajusta modelos simples (stumps) secuencialmente, asignando más peso a observaciones difíciles.

5.7.3. Bagging

Los modelos basados en bootstrap empleados fueron:

- **Random Forest:** conjunto de árboles de decisión que introduce aleatoriedad tanto en datos como en características. Suele ser robusto y relativamente estable.
- **Bagged Decision Trees (BaggedDT):** árboles no podados entrenados sobre muestras bootstrap, cuyas predicciones se promedian para reducir varianza.

5.7.4. Otros modelos evaluados

Además de los métodos ensemble, se evaluaron modelos representativos de paradigmas adicionales:

- **Support Vector Regression (SVR):** modelo de márgenes para regresión, evaluado con kernel lineal.
- **K-Nearest Neighbors (KNN):** modelo basado en vecinos más próximos; útil como referencia no paramétrica, aunque sensible a la escala.
- **Multi-Layer Perceptron (MLP):** red neuronal densa capaz de capturar relaciones no lineales.
- **Bayesian Neural Network (BayesianNN):** aproximación probabilística que permite cuantificar incertidumbre a través de regularización bayesiana.

5.7.5. Configuración del stacking

Tras evaluar todos los modelos anteriores, se construyeron diferentes configuraciones de modelos base (*base learners*) que se combinan mediante un meta-modelo. Estas combinaciones se diseñaron con dos criterios principales:

1. **Diversidad estructural:** mezclar métodos de boosting y bagging, así como variantes de boosting con distintas estrategias de crecimiento y regularización.
2. **Rendimiento individual:** incluir preferentemente los modelos con mayor R^2 y menor error (RMSE, MAE) en las pruebas individuales.

Los meta-modelos utilizados para integrar las predicciones fueron:

- **Modelos lineales:** Regresión Lineal, Ridge.
- **Modelos basados en árboles:** Random Forest, Gradient Boosting Regressor.
- **Modelos kernel:** SVR lineal.
- **Red neuronal:** MLP con una capa oculta.

Esta selección permite comparar desde combinadores lineales simples hasta integradores no lineales capaces de capturar interacciones complejas entre predicciones.

5.7.6. Comparación y justificación de modelos

La evaluación exhaustiva de múltiples algoritmos permite identificar no solo el modelo individual con mejor rendimiento, sino también combinacio-

nes sinérgicas para el *stacking*. La Tabla 5.3 resume los modelos finalmente entrenados y evaluados.

Modelo	**Tipo**	**Características**	**Observaciones**
Random Forest	Bagging	Bootstrap con selección aleatoria de atributos	Robusto y estable
BaggedDT	Bagging	Árboles sin poda sobre muestras bootstrap	Mejora por agregación
XGBoost	Boosting	Regularización L1/L2, segundo orden	Muy preciso; sensible a tuning
LightGBM	Boosting	Crecimiento leaf-wise, muy eficiente	Rápido; riesgo de sobreajuste
CatBoost	Boosting	Codificación ordenada; robusto al ruido	Excelente sin gran tuning
GBDT	Boosting	Árboles secuenciales ajustados a residuos	Buen rendimiento
AdaBoost	Boosting	Aumenta peso de obs. mal predichas	Menos robusto
KNN	Instancia	Predicción por proximidad	Sensible a escala y ruido
MLP	Red neuronal	Captura relaciones no lineales	Requiere normalización
SVR	Márgenes	Kernel lineal, gran margen	Robusto al sobreajuste
BayesianNN	Probabilístico	Cuantifica incertidumbre	Reduce sobreajuste

Tabla 5.3. Resumen de los modelos de aprendizaje supervisado evaluados.

6. Implementación del *pipeline*

El desarrollo y la evaluación de los modelos predictivos se realizaron íntegramente en **Python**, utilizando librerías como `scikit-learn`, `cuML` y `PyTorch`, junto con implementaciones específicas de *gradient boosting* como `XGBoost`, `LightGBM` y `CatBoost`. El proceso de entrenamiento se llevó a cabo en dos fases diferenciadas.

Para los modelos entrenados exclusivamente con datos del IFN3 se utilizó un equipo local equipado con un procesador Intel Core i7 y 32 GB de memoria RAM. En cambio, los modelos que empleaban conjuntamente datos del IFN2 e IFN3 se entrenaron en el sistema de computación de alto rendimiento (HPC) de la Universidad de Salamanca. Esta elección se debió a la disponibilidad de tarjetas gráficas Nvidia H100, que permiten acelerar de forma significativa el entrenamiento de aquellos modelos compatibles con ejecución en GPU gracias a su elevada capacidad de paralelización.

No obstante, cabe señalar que el entrenamiento también podría haberse realizado en un equipo de escritorio convencional equipado con una tarjeta gráfica comercial, ya que los requisitos computacionales del problema no son especialmente elevados.

6.1. Ingeniería práctica del entrenamiento y la validación

Desde el punto de vista de la implementación, el proceso de entrenamiento y validación se apoyó fundamentalmente en el ecosistema de `scikit-learn`, complementado con librerías especializadas para modelos de *gradient boosting*. La gestión de los datos se realizó mediante `pandas` y `numpy`, mientras que el cálculo de métricas y estadísticas adicionales del error se apoyó en `scipy` y los módulos de evaluación de `sklearn.metrics`.

A partir del conjunto de datos original, se aplicaron filtros de calidad sobre la variable objetivo utilizando operaciones vectorizadas de `pandas`, eliminando observaciones con valores nulos, inconsistentes o que no cumplían los criterios definidos en la Sección 5.4.1.

La partición de los datos en conjuntos de entrenamiento y prueba se realizó mediante `GroupShuffleSplit` del módulo `sklearn.model_selection`, con una proporción 80/20. Este esquema garantizó que todas las observaciones asociadas a una misma parcela se asignaran íntegramente a un único subconjunto, evitando fugas de información derivadas de la correlación espacial intra-parcela. Sobre el conjunto de entrenamiento se definió una validación cruzada de cinco pliegues utilizando `GroupKFold`.

El preprocesado de las variables y el ajuste de los modelos se integraron en un único objeto `Pipeline`, combinando `ColumnTransformer`, `SimpleImputer`, `StandardScaler` y `OneHotEncoder`. Esta integración aseguró que todas las transformaciones se estimaran exclusivamente con los datos de entrenamiento de cada pliegue durante la validación cruzada. El ajuste de hiperparámetros se llevó a cabo mediante `GridSearchCV`, definiendo rejillas específicas para cada algoritmo y utilizando el coeficiente de determinación (R^2) como métrica de selección.

Los modelos evaluados incluyen implementaciones de *gradient boosting* (`XGBoost`, `LightGBM`, `CatBoost` y `GradientBoostingRegressor`), métodos basados en *bagging* (`RandomForestRegressor`, `BaggingRegressor`), así como modelos de distinta naturaleza como `MLPRegressor`, `KNeighborsRegressor`, `LinearSVR`, `AdaBoostRegressor` y `BayesianRidge`. Para cada modelo se calcularon de forma sistemática las métricas de rendimiento sobre el conjunto de prueba: R^2, RMSE y MAE, junto con estadísticas adicionales del error absoluto (mediana y moda), almacenándose los resultados en estructuras tabulares para su análisis comparativo.

6.2. Implementación del stacking

La agregación de modelos mediante *stacking* se implementó de forma manual utilizando utilidades básicas de `scikit-learn`, con el objetivo de mantener un control estricto sobre el flujo de entrenamiento y validación. A partir de los mejores modelos individuales se generaron predicciones fuera de pliegue (*out-of-fold*, OOF) sobre el conjunto de entrenamiento, empleando el mismo esquema de validación cruzada (`GroupKFold`).

Estas predicciones OOF se organizaron en matrices de meta-variables mediante `numpy` y se utilizaron como entrada para el entrenamiento de los metamodelos. En paralelo, cada modelo base se reentrenó sobre la totalidad del conjunto de entrenamiento para generar las correspondientes predicciones sobre el conjunto de test, que se emplearon posteriormente para la evaluación final del *stack*.

Los metamodelos considerados incluyen `LinearRegression`, `Ridge`, `GradientBoostingRegressor`, `RandomForestRegressor`, `SVR` con kernel lineal y `MLPRegressor`. Antes de su ajuste, las meta-variables se estandarizaron mediante `StandardScaler`, integrando este paso en un `Pipeline` específico del segundo nivel. La evaluación del *stacking* se realizó exclusivamente sobre el conjunto de test independiente, calculando las métricas habituales (R^2, RMSE y MAE) para cada combinación de modelos base y metamodelo.

6.3. Datos finales de entrenamiento

Tras aplicar los criterios de elegibilidad y filtrado descritos en la Sección **??**, el conjunto de datos final utilizado para el ajuste de los modelos queda compuesto por:

- **IFN2:** Total de parcelas = **88.696**
  - Casos con $c4 > c$: **31.428**
  - Casos con $carbono\_bruto4 > carbono\_bruto$: **32.403**
- **IFN3:** Total de parcelas = **171.157**
  - Casos con $fccarb > 20$: **158.434**
  - Casos con $fccarb > 20$ y $c4 > c$: **57.401**
  - Casos con $fccarb > 20$ y $carbono\_bruto4 > carbono\_bruto$: **76.617**

La Tabla 6.1 resume las principales estadísticas descriptivas de las variables utilizadas en el modelado, adicionalmente en la Figura 6.1 se muestra la distribución de las mismas.

Tabla 6.1. Estadísticos descriptivos del conjunto de datos depurado.

Variable	N	Media	Desv. estándar	Mín.	Máx.
carbono_bruto4	136 325	24.6168	35.8198	0.000327	420.498829
carbono_bruto	114 485	15.9326	26.3052	0	359.805707
c4	105 714	38.3789	47.0348	0.484695	883.462735
c	92 372	23.4399	34.9622	0	842.739088
periodo	105 709	18.3167	6.4853	0	34

Se observa que la variable `carbono_bruto4` presenta una media de 24.62 y una desviación estándar de 35.82, mientras que la variable `c4` muestra valores notablemente superiores (media de 38.38 y desviación estándar de 47.03). La variable `c4` es más dispersa y heterogénea que `carbono_bruto4`. En general, una mayor variabilidad en la variable objetivo se traduce en un problema de predicción más complejo, ya que el modelo debe capturar relaciones más inestables y sujetas a mayor ruido.

Por tanto, incluso antes de evaluar los modelos, es razonable esperar que una misma familia de algoritmos obtenga valores de R^2 más elevados y errores más bajos (RMSE, MAE) al predecir `carbono_bruto4`, cuya estructura estadística es menos dispersa, que al predecir `c4`.

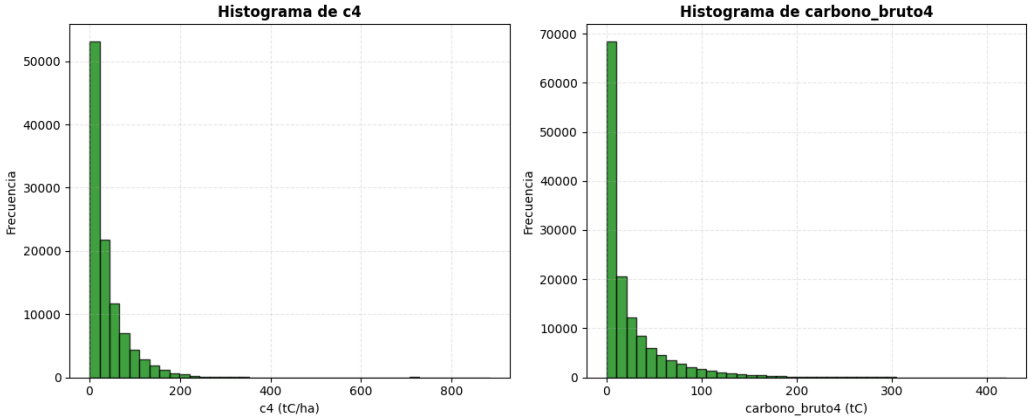

Figura 6.1. Distribución de las variables `c4` y `carbono_bruto4` en el conjunto depurado.

Observamos una clara distribución asimétrica a la izquierda con una larga cola en ambas variables. No existe un factor de escala único que lleve de una variable a la otra porque la generalización a hectárea tiene en cuenta la densidad forestal particular de cada parcela.

6.3.1. Efecto del periodo sobre el carbono

La influencia del *periodo* sobre las variables de carbono se evaluó mediante ANOVA de un factor. Los análisis realizados muestran que el *periodo* ejerce un efecto significativo sobre ambas variables. En `c4` se obtuvo un estadístico $F = 143{,}49$ ($p < 0{,}001$), mientras que en `carbono_bruto4` el valor fue $F = 161{,}08$ ($p < 0{,}001$). Estos resultados indican que las diferencias observadas entre periodos no son aleatorias, sino que reflejan variaciones sistemáticas asociadas al momento de muestreo, confirmando que el *periodo* constituye un factor explicativo relevante en la dinámica del carbono forestal.

7. Entrenamiento y validación

El proceso de entrenamiento se estructuró en varias fases orientadas a optimizar tanto la selección de variables predictoras como la robustez del modelo final. En primer lugar, se llevó a cabo una etapa de **selección de variables**, en la que se evaluaron distintos subconjuntos de características definidos por bloques temáticos con significado ecológico y funcional. Para esta tarea se adoptó un enfoque sistemático basado en la comparación del desempeño predictivo de las distintas combinaciones mediante el algoritmo `CatBoost`, seleccionado tras pruebas preliminares que mostraron su alta capacidad de ajuste y estabilidad frente a la heterogeneidad de los datos. En todas las configuraciones se mantuvo constante la variable objetivo (carbono capturado) y los parámetros del modelo, de modo que las variaciones en el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE) reflejaran exclusivamente la contribución informativa de cada bloque. Los resultados de esta fase son preliminares ya que se emplearon entrenamientos más sencillos (sin validación cruzada para la selección de hiperparámetros).

Las configuraciones analizadas incorporaron progresivamente variables relacionadas con las características de la especie, las propiedades edáficas, el terreno, las condiciones climáticas y los índices de vegetación. A partir de los resultados obtenidos, se identificaron los bloques con mayor aporte marginal al rendimiento del modelo, priorizando aquellos cuya inclusión mejoró consistentemente el R^2 sin aumentar de forma significativa la complejidad o redundancia del conjunto de predictores.

En una segunda fase, se procedió al **entrenamiento comparativo de modelos**, implementando un conjunto de algoritmos de aprendizaje supervisado con el fin de contrastar su capacidad predictiva. Cada modelo fue entrenado bajo las mismas condiciones experimentales, utilizando las configuraciones de variables seleccionadas en la fase anterior. Esta comparación permitió identificar los algoritmos con mejor ajuste global y menor error de predicción, destacando de nuevo el desempeño de `CatBoost`.

Posteriormente, se implementó una estrategia de **stacking**, combinando las predicciones de los modelos individuales mediante un metamodelo de segundo nivel, con el objetivo de aprovechar la complementariedad entre los distintos enfoques y mejorar la capacidad de generalización.

7.1. Elección de variables

7.1.1. Resultados de la selección de variables manual

La selección manual de variables partió de una organización temática del conjunto de predictores, agrupando las variables según el tipo de información ecológica, estructural o climática que representan. Esta clasificación permitió estructurar el proceso de reducción dimensional en torno a los siguientes bloques conceptuales:

- **Bloque de variables fijas**: describe la estructura básica de la masa forestal y los atributos esenciales de identificación y caracterización general de cada parcela.
- **Bloque de variables de especie**: recoge información relativa a la composición, estado y características específicas de las formaciones forestales.
- **Bloque sustrato**: integra variables edáficas y de manejo susceptibles de variar en el tiempo.
- **Bloque de terreno**: agrupa propiedades físicas del medio que permanecen estables a escala temporal de inventarios (pendiente, orientación, tipo de suelo, etc.).
- **Bloque climático resumido**: representado por el índice de aridez de Martonne, que sintetiza la interacción entre temperatura y precipitación.
- **Bloque climático detallado**: incluye métricas estacionales explícitas de temperatura y precipitación.
- **Bloque de índices de vegetación**: recoge información espectral relacionada con el estado hídrico, vigor y actividad fotosintética de la vegetación.

En total, la base de datos contenía inicialmente 445 variables candidatas distribuidas entre estos bloques temáticos. Tras aplicar el procedimiento de selección manual, apoyado en criterios estadísticos, ecológicos y en la comparación del rendimiento del modelo; el conjunto se redujo a 44 variables representativas. Las variables finalmente seleccionadas dentro de cada bloque fueron las siguientes:

- **Bloque de variables fijas**: `especie_id`, `tipo_especie`, `grupo_id`, `periodo`, `radio`, `ocupa`, `npies_1`, `npies_2`, `npies_5`, `npies_10`, `npies_15`, `npies_20`, `npies_25`, `npies_30`, `npies_35`, `npies_40`, `npies_45`, `npies_50`, `npies_55`, `npies_60`, `npies_65`, `npies_70`.

- **Bloque de variables de especie**: `estado_id`, `fccarb`, `disesp_id`.
- **Bloque sustrato (dinámico)**: `modcomb_id`, `nivel2_id`, `tratmasa_id`.
- **Bloque de terreno**: `rocosidad_id`, `orientacion_cat`, `elevacion`, `pendiente_cat`.
- **Bloque climático resumido (Martonne)**: `martonneidx_id`.
- **Bloque climático detallado (temperatura y precipitación)**: `skt_mean_primavera`, `skt_mean_verano`, `skt_std_primavera`, `skt_std_verano`, `pr_sum_invierno`, `pr_sum_otoño`, `pr_sum_primavera`, `pr_sum_verano`.
- **Bloque de índices de vegetación**: `gndvi_mean_verano`, `ndii_mean_primavera`, `gndvi_std_primavera`, `evi_mean_primavera`.

Este proceso permitió sintetizar la información original manteniendo una representación equilibrada de todos los ámbitos ecológicos implicados en la estimación del carbono.

La comparación de modelos entrenados con combinaciones incrementales de bloques mostró que todos ellos aportan información relevante, siguiendo el orden de contribución aproximado: *variables fijas* > *variables de especie* > *sustrato* > *terreno* > *índices de vegetación* > *Martonne* > *temperatura y precipitación*. Es decir, la mayor parte de la capacidad predictiva se explica por la estructura y composición de la masa forestal, mientras que las condiciones edáficas, topográficas y climáticas actúan como moduladores adicionales de la acumulación de carbono.

7.1.2. Selección de variables mediante Featurewiz

Aplicado al conjunto completo de predictores, *Featurewiz* seleccionó **67 variables**. El patrón resultante muestra una clara preferencia por dos grandes grupos: (i) **índices de vegetación** derivados de Sentinel-2 y (ii) **variables térmicas estacionales**. El algoritmo retuvo numerosas estadísticas de NDII, EVI, GNDVI y NDVI (medias, máximos, mínimos, medianas y desviaciones estándar), especialmente durante primavera y verano, reflejando la relevancia del estado hídrico y el vigor fotosintético en la estimación del carbono.

Asimismo, se seleccionaron múltiples métricas de temperatura del aire y del suelo (`t2m_*`, `skt_*`, `stl_*`) y diversas variables de precipitación (`pr_sum_*`, `pr_max_*`, `pr_min_*`), lo que muestra sensibilidad del método a las condiciones climáticas estacionales. El índice de aridez de Martonne tam-

bién fue seleccionado, aportando una medida sintetizada del balance térmico-hídrico.

Finalmente, el algoritmo incluyó un conjunto contenido pero representativo de variables estructurales (número de pies por clase diamétrica), de especie y de terreno, indicando que dichas variables aportan información complementaria necesaria para la predicción.

7.1.3. Selección de variables mediante mRMR

El método *mRMR* seleccionó un total de **50 variables**, priorizando aquellas con alta información mutua respecto al carbono y baja redundancia entre sí. El conjunto final integra predictores estructurales (identificación de especie, radio, clases diamétricas, orientación y pendiente), variables topográficas y edáficas (rocosidad, tipos de suelo), métricas climáticas estacionales (temperatura del aire y del suelo, índice de Martonne) e índices de vegetación representativos del estado estacional de la copa.

La presencia sistemática de valores medios, máximos y medianos de NDII, GNDVI y EVI en verano y primavera confirma que la actividad fotosintética y el estado hídrico son predictores directos del carbono almacenado. De igual modo, la selección de múltiples métricas térmicas refleja la relevancia de los pulsos climáticos sobre la productividad forestal.

En conjunto, mRMR produjo un conjunto compacto y equilibrado, asegurando diversidad informativa y evitando redundancias, lo que lo convierte en un complemento eficaz a los métodos anteriores.

7.1.4. Discusión de la selección de variables

De los tres conjuntos de variables seleccionados se mantuvo la selección manual al demostrar un mejor rendimiento con mayor simplicidad como se aprecia en la tabla 7.1.

TODO: Esto igual debería ir en resultados?

Tabla 7.1. Comparación de configuraciones de selección de variables y rendimiento del modelo CatBoost sobre los datos del IFN 2-3 y 4 para predecir `c4`.

Configuración	Modelo	n_{vars}	R^2	RMSE	MAE	Moda error (aprox.)
Manual	CatBoost	44	0.80	21.77	11.48	1
mRMR	CatBoost	67	0.79	21.91	11.69	1
FeatureWiz	CatBoost	50	0.72	25.65	13.08	2

7.2. Ensamblado tipo stacking de modelos de regresión

Con el objetivo de estudiar el compromiso entre diversidad del ensamble, coste computacional y rendimiento, se definieron cinco configuraciones de modelos base (Tabla 7.2). Los modelos AdaBoost, BayesianNN, SVR, MLP y KNN se descartaron como candidatos.

Tabla 7.2. Configuraciones de modelos base para *stacking*.

Config.	Modelos base
1	CatBoost, LightGBM, XGBoost, Random Forest, GBDT, BaggedDT
2	CatBoost, LightGBM, Random Forest, GBDT
3	LightGBM, XGBoost, GBDT
4	CatBoost, Random Forest, GBDT
5	LightGBM, Random Forest

- **Configuración 1:** incluye todos los modelos con rendimiento competitivo. Esta configuración es la más rica en términos de variedad de arquitecturas, aunque también la más costosa computacionalmente y potencialmente más propensa al sobreajuste si no se controla adecuadamente.
- **Configuración 2:** reduce el número de modelos, eliminando XGBoost y BaggedDT, que aportan menos mejora marginal respecto a sus alternativas (LightGBM y Random Forest). Esta combinación mantiene una buena diversidad con menor complejidad y coste computacional.
- **Configuración 3:** agrupa únicamente modelos de la familia de *gradient boosting*. El objetivo es analizar el efecto de combinar variantes de un mismo paradigma y evaluar hasta qué punto diferentes implementaciones de boosting proporcionan suficiente diversidad como para ser beneficiosa en un ensamble.
- **Configuración 4:** combina un modelo de boosting basado en manejo robusto de variables categóricas (CatBoost) con Random Forest (bagging de árboles) y GBDT (boosting clásico). La idea es mezclar enfoques de bagging y boosting, manteniendo un número moderado de modelos y una buena diversidad estructural.
- **Configuración 5:** es la configuración más simple. LightGBM compite con CatBoost en rendimiento, mientras que Random Forest aporta un

sesgo diferente al basarse en bagging en lugar de boosting. Esta configuración sirve como referencia de un ensamble muy ligero, con bajo coste computacional y, al mismo tiempo, razonablemente diverso.

El objetivo es que el meta-modelo reciba como entradas predicciones de alta calidad y suficientemente diversas, en lugar de introducir ruido procedente de modelos débiles.

Sobre las predicciones apiladas de cada configuración se entrenan distintos meta-modelos $g(\cdot)$, definidos en la Tabla 7.3.

Tabla 7.3. Meta-modelos utilizados en el *stacking* junto con sus parámetros.

Meta-modelo	**Parámetros**
Gradient Boosting	Configuración por defecto
Regresión Lineal	Sin regularización
Ridge	Regularización L2 con validación cruzada ($\alpha \in \{0{,}01, 0{,}1, 1, 10, 100\}$)
Random Forest	50 árboles
SVR	Kernel lineal
MLP	Una capa oculta con 50 neuronas, 500 iteraciones máximas

Estos meta-modelos representan diferentes formas de combinar las predicciones de los modelos base:

- **Modelos lineales** (Regresión Lineal y Ridge): permiten comprobar si una combinación lineal de las predicciones base es suficiente para mejorar el rendimiento. Ridge añade regularización L2 para controlar el sobreajuste.
- **Modelos no lineales basados en árboles** (GradientBoostingRegressor, RandomForestRegressor): pueden capturar interacciones complejas entre las predicciones de los modelos base, a costa de una mayor complejidad.
- **Modelos de *kernel*** (SVR con kernel lineal): permiten una combinación robusta y, en algunos casos, menos sensible a valores extremos en las predicciones.
- **Red neuronal (MLPRegressor)**: introduce una capa adicional de flexibilidad, capaz de aproximar combinaciones no lineales complejas entre las salidas de los modelos base.

Al evaluar todas las combinaciones de `stack_configs` con los diferentes `meta_models`, se obtiene un conjunto de ensambles apilados que permiten estudiar de forma sistemática: (i) qué subconjuntos de modelos base son más complementarios, y (ii) qué tipo de meta-modelo aprovecha mejor la información contenida en sus predicciones.

8. Resultados

Para estructurar de forma clara el análisis, dividiremos este apartado en dos secciones, sesgún se hayan entrenado los modelos solo con el IFN3 o bien con el IFN2 e IFN3 como explicativos. A su vez, en sección contendrá una explicación distinta para cada una de las cantidades predichas (tC o tC/ha).

8.1. Resultados IFN3

8.2. Resultados IFN2 e IFN3

8.2.1. Toneladas de carbono por hectárea

Modelos base

Una vez entrenados los modelos, algunos parámetros globales como el R^2, RMSE y MAE se presentan en la Tabla **??**

Tabla 8.1. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en tC/ha con el conjunto de datos que emplea IFN2 e IFN3 como explicativos.

Modelo	R^2_{test}	$\text{RMSE}_{\text{test}}$	MAE_{test}
LightGBM	**0.787**	**22.767**	**11.650**
XGBoost	0.784	22.952	11.590
CatBoost	0.783	22.990	11.607
GBDT	0.783	23.014	11.658
MLP	0.771	23.607	12.287
BaggedDT	0.740	25.142	13.021
Random Forest	0.732	25.547	12.908
BayesianNN	0.678	28.021	14.689
SVR	0.551	33.065	13.708

En la Figura 8.1 se muestra la distribución de predicciones frente a valores reales para el modelo LightGBM (el que se ha escogido como el mejor) con IFN2 e IFN3 como explicativos para el rango de predicciones entre 0 y 300 tC/ha (para facilitar la visualización). También se incluye un código de colores para facilitar la visualización de la densidad de puntos y un histograma de la distribución tanto de los valores reales como de las predicciones.

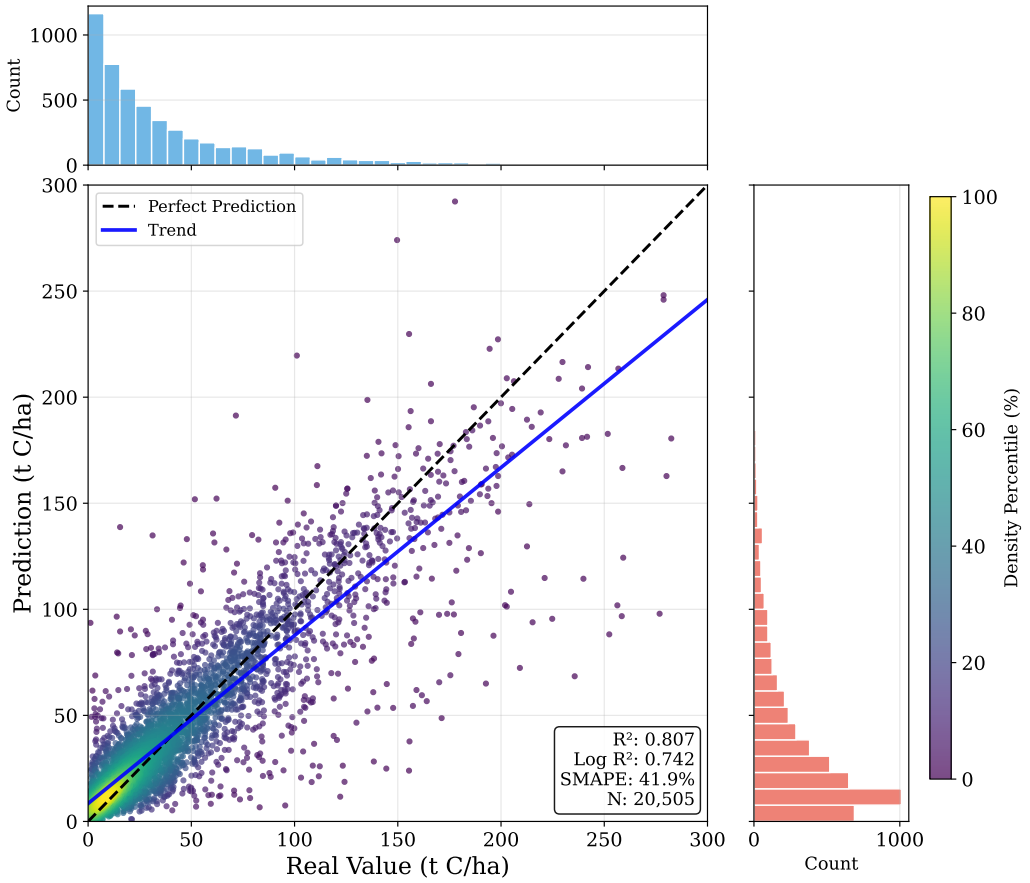

Figura 8.1. Distribución de predicciones frente a valores reales para el modelo LightGBM con IFN2 e IFN3 como explicativos. Solo se muestra el intervalo de predicciones entre 0 y 300 tC/ha.

Modelos con stacking

8.2.2. Toneladas de carbono

Modelos base

En la Tabla 8.2 se muestra el resumen del rendimiento de los modelos para la predicción de la variable de carbono en toneladas con el conjunto de datos que emplea el IFN2 e IFN3 como explicativos.

Tabla 8.2. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en toneladas (carbono_bruto4) con el conjunto de datos que emplea IFN2 e IFN3 como explicativos.

Modelo	R^2_{test}	$\text{RMSE}_{\text{test}}$	MAE_{test}
CatBoost	**0.845**	**13.846**	**6.615**
LightGBM	0.841	14.006	6.654
XGBoost	0.840	14.054	6.655
GBDT	0.838	14.159	6.722
MLP	0.832	14.410	6.931
BaggedDT	0.821	14.858	7.282
Random Forest	0.819	14.950	7.135
BayesianNN	0.775	16.674	8.906
SVR	0.679	19.897	8.137

Al igual que en apartado anterior, en la Figura 8.2 se muestra la distribución de predicciones frente a valores reales para el modelo CatBoost (el que se ha escogido como el mejor) con IFN2 e IFN3 como explicativos para el rango de predicciones entre 0 y 150 tC/ha (para facilitar la visualización). También se incluye un código de colores para facilitar la visualización de la densidad de puntos y un histograma de la distribución tanto de los valores reales como de las predicciones.

Por otro lado, en las Figuras 8.3 y 8.4 se muestra la evolución del RMSE en cuantiles y el RMSE porcentual en cuantiles en función de la variable `periodo` para el modelo CatBoost con IFN2 e IFN3 como explicativos, junto con el histograma de distribución de los datos de entrenamiento en función de la variable periodo.

8.3. Resultados

En esta sección se presentan los resultados obtenidos por los modelos descritos en la Sección 5.7. Las tablas de resultados completas se pueden consultar en el Apéndice A.18.

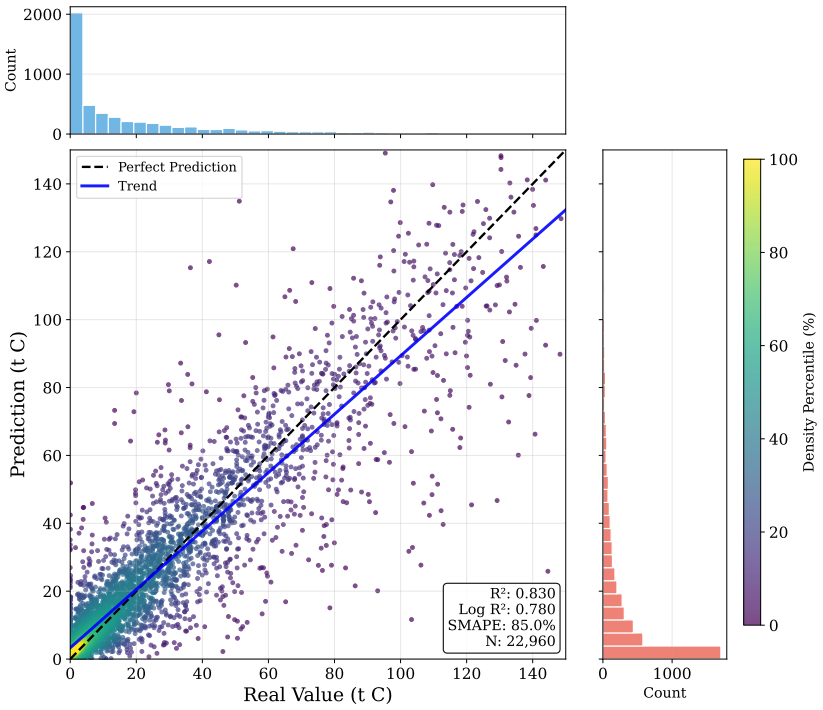

Figura 8.2. Distribución de predicciones frente a valores reales para el modelo CatBoost con IFN2 e IFN3 como explicativos para el rango de predicciones entre 0 y 150 tC/ha (para facilitar la visualización).

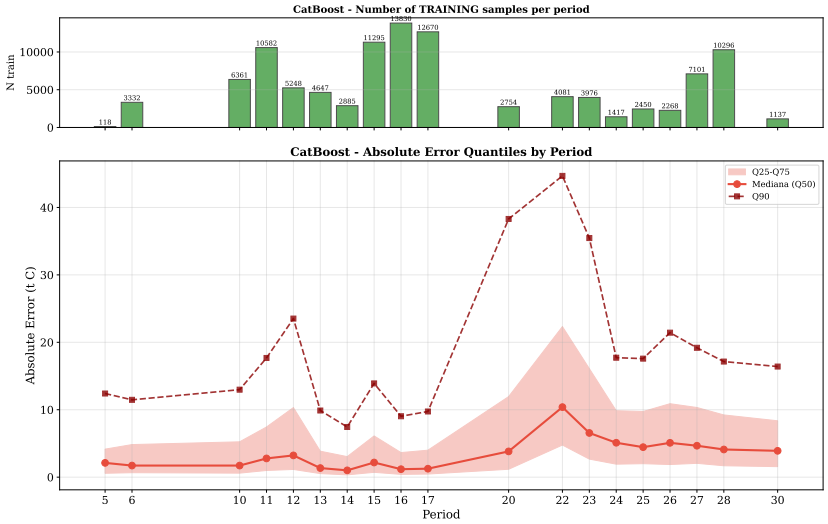

Figura 8.3. Evolución del RMSE en cuantiles en función de la variable `periodo` para el modelo CatBoost con IFN2 e IFN3 como explicativos.

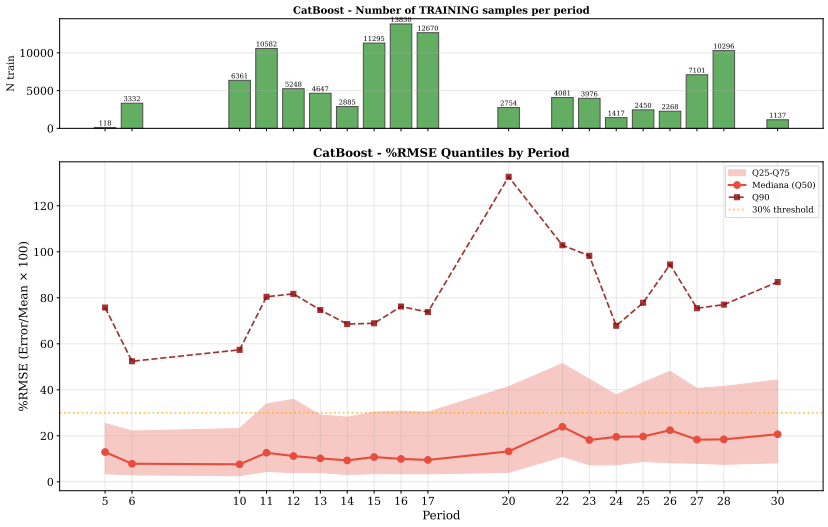

Figura 8.4. Evolución del RMSE porcentual en cuantiles en función de la variable `periodo` para el modelo CatBoost con IFN2 e IFN3 como explicativos.

En conjunto, los resultados obtenidos a lo largo de las cuatro configuraciones de entrenamiento analizadas (IFN3 o IFN2 y 3 como explicativos / variable objetivo en tC o tC/ha) muestran un comportamiento notablemente estable y coherente entre versiones, tanto en términos de capacidad predictiva como de generalización. De forma sistemática, los modelos basados en árboles de decisión y *gradient boosting* son los que alcanzan los mejores niveles de rendimiento, destacando de manera consistente CatBoost y LightGBM como las alternativas más competitivas entre los modelos individuales, independientemente del inventario empleado o de la forma en que se expresa la variable objetivo.

Un aspecto especialmente relevante es la alta similitud entre los valores de R^2 obtenidos en validación cruzada y en el conjunto de test, lo que indica que los modelos presentan una buena capacidad de generalización y no muestran síntomas apreciables de sobreajuste. Esta estabilidad se observa tanto en los escenarios con mayor volumen de información (IFN2+IFN3) como en aquellos más simples (IFN3), reforzando la robustez de los enfoques basados en árboles frente a variaciones en la disponibilidad de datos.

La incorporación de esquemas de *stacking* no produce incrementos sustanciales en el coeficiente de determinación respecto a los mejores modelos individuales. No obstante, sí se aprecia una mejora sistemática en el error absoluto medio (MAE), con reducciones que oscilan aproximadamente entre 210 y 371 kg de carbono (o kg/ha), dependiendo del escenario considerado. Esta reducción, aunque moderada en términos relativos, resulta relevante desde un punto de vista práctico, ya que implica predicciones más precisas en el rango de error típico y justifica la consideración del *stacking* como una estrategia complementaria.

En cuanto a la estructura de los ensambles, los mejores resultados se obtienen cuando se combinan modelos base de alta calidad y naturaleza similar (principalmente variantes de *gradient boosting*) y se emplean metamodelos con complejidad moderada, como MLP o SVR lineal. Por el contrario, los *stacks* con pocos modelos base o aquellos que incorporan metamodelos excesivamente flexibles, como Random Forest en el segundo nivel, tienden a ofrecer un rendimiento inferior, probablemente debido a la baja dimensionalidad del espacio de meta-predictores o a un sobreajuste innecesario del ruido residual.

En síntesis, los resultados confirman que los modelos individuales basados en árboles constituyen una solución sólida y eficiente, mientras que el *stacking* aporta mejoras incrementales principalmente en términos de reducción del

error medio.

Tabla 8.3. Comparación sintética del rendimiento de los modelos según inventarios utilizados y variable objetivo.

IFN	Variable objetivo	Modelo	Modelos	R^2	RMSE	MAE
2 y 3	tC/ha	LightGBM	1	0.79	22.77	11.65
2 y 3	tC/ha	stack1 + MLP	6	0.79	22.39	11.32
2 y 3	tC	CatBoost	1	0.84	13.85	6.61
2 y 3	tC	stack1 + MLP	6	0.85	13.76	6.40
3	tC/ha	CatBoost	1	0.8598	17.7087	9.2504
3	tC/ha	stack1 + MLP	6	0.8656	17.3380	8.8789
3	tC	LightGBM	1	0.9091	10.6623	5.4774
3	tC	stack1 + MLP	6	0.9140	10.3723	5.2515

La Tabla 8.3 sintetiza el rendimiento de los mejores modelos identificados en cada una de las cuatro líneas de entrenamiento consideradas, permitiendo una comparación directa entre inventarios utilizados, variable objetivo y complejidad del modelo. En todos los escenarios se observa un patrón consistente: los modelos individuales basados en *gradient boosting* (LightGBM o CatBoost) ofrecen un rendimiento sólido, que se ve ligeramente mejorado mediante la incorporación de esquemas de *stacking*.

Cuando se emplean conjuntamente los inventarios IFN2 e IFN3 y se predice la variable normalizada en tC/ha, el rendimiento del modelo individual (LightGBM) y del *stack* es prácticamente equivalente en términos de R^2, si bien el *stacking* logra una reducción apreciable del MAE, pasando de 11.65 a 11.32 tC/ha. Un comportamiento análogo se observa al predecir carbono total (tC) con IFN2 e IFN3, donde CatBoost alcanza ya valores elevados de R^2 (0.84), y el *stack* introduce una mejora moderada pero consistente tanto en R^2 como en los errores (RMSE y MAE).

En los escenarios basados exclusivamente en IFN3, los niveles de rendimiento son, en general, superiores. Para la variable en tC/ha, CatBoost explica cerca del 86 % de la varianza observada, mientras que el *stack* incrementa ligeramente este valor y reduce el MAE en aproximadamente 0.37 tC/ha. De forma aún más clara, al predecir carbono total (tC), LightGBM alcanza un R^2 superior a 0.91, y el *stacking* vuelve a aportar una mejora incremental, reduciendo el error absoluto medio hasta valores en torno a 5.25 tC.

En conjunto, estos resultados confirman que la mayor ganancia del *stacking* no reside tanto en aumentos sustanciales del R^2, sino en una reducción sistemática del error medio, lo que se traduce en predicciones más precisas en términos absolutos. Al mismo tiempo, la tabla pone de manifiesto que los enfoques basados en árboles constituyen una base extremadamente robusta, sobre la que los ensambles apilados actúan como un refinamiento adicional más que como un cambio de paradigma.

8.4. Síntesis de resultados

A partir del análisis realizado, pueden resumirse las principales conclusiones en los siguientes puntos:

- El conjunto de datos depurado muestra una variables objetivos marcadas con gran variabilidad: `carbono_bruto4` presenta menor dispersión (SD $\approx$ 36 tC/ha) que `c4` (SD $\approx$ 47 tC/ha), lo que anticipa un problema predictivo más complejo para esta última.
- El análisis ANOVA confirma que el *periodo* tiene un efecto estadísticamente significativo sobre ambas variables de carbono, evidenciando la existencia de variaciones temporales sistemáticas relevantes para su modelización.
- Entre las estrategias de selección de variables evaluadas (manual, FeatureWiz y mRMR), la selección manual, basada en bloques temáticos con coherencia ecológica, ofrece el mejor equilibrio entre simplicidad y rendimiento, superando en precisión y error a las selecciones automáticas.
- Los bloques de variables más informativos son, en orden aproximado de importancia: estructura de la masa forestal, características de especie, condiciones edáficas y topográficas, índices de vegetación e información climática estacional. La mayor parte del poder predictivo se concentra en las características estructurales y de especie.
- Los modelos individuales muestran que los métodos basados en árboles y *gradient boosting* (CatBoost, LightGBM, XGBoost y GBDT) alcanzan el mejor rendimiento global, con valores de R^2 de hasta 0,85 y errores moderados (inferiores al 50 % de la desviación típica de la variable).
- CatBoost destaca como el mejor modelo individual, gracias a su capacidad para capturar relaciones no lineales y manejar adecuadamente la complejidad y heterogeneidad de los datos.

- Métodos como AdaBoost, KNN o BayesianNN muestran un rendimiento sustancialmente inferior, lo que los descarta como candidatos eficaces para este tipo de predicción.
- Las técnicas de *stacking* aportan mejoras sistemáticas en el error absoluto medio de los modelos, reduciendolos, en la mejor configuración, en torno a un 5 %.
- El rendimiento del *stacking* depende del meta-modelo: los modelos lineales (Regresión Lineal y Ridge) ofrecen combinaciones estables y robustas; los meta-modelos Random Forest tienden al sobreajuste; y los meta-modelos moderadamente no lineales (SVR y MLP) proporcionan las mayores mejoras.
- *TODO: En este item se habla de la capacidad de predecir del modelo para un horizonte temporal de "entre 5 y 30 años con un nivel elevado de precisión". En la discusión se menciona que esta precisión varía mucho dependiendo de la cantidad de carbono, así que igual habría que matizar. Podría ser buena idea que en el apartado de Resultados simplemente se expusieran los resultados obtenidos en el de Discusión se entrase ya en valoraciones.*

  El modelo desarrollado es capaz de predecir, a partir de las características estructurales, ecológicas y ambientales de un cultivo forestal, la cantidad de carbono almacenado en un horizonte temporal de entre 5 y 30 años con un nivel elevado de precisión. El mejor modelo obtenido se construye mediante un metamodelo `MLP` combinando los modelos `CatBoost,LightGBM,XGBoost, Random Forest, BaggedDT` y `GBDT` y alcanza un coeficiente de determinación de $R^2 = 0{,}85$, junto con un error típico de **RMSE = 13.76 tC** y un error medio absoluto de **MAE = 6.40 tC**.

9. Discusión

Por simplicidad se hará la discusión únicamente para los modelos entrenados con los datos del IFN2 y el IFN3. Esto es porque se puede comprobar que estos modelos se reducen al modelo con el IFN2 cuando los datos de entrada (se ha comprobado con el test) proceden del IFN3 (las métricas son ligeramente peores por la "contaminación" de los datos del IFN2, pero en esencia se comportan de la misma manera).

9.1. Variable c4 (en toneladas de carbono por hectárea)

9.1.1. Modelos base

La variable `periodo` (el número de años entre la medición y la predicción) tiene una gran importancia en el estudio de los modelos. Es por esto que en las Figuras 9.1 y 9.2 se incluyen métricas en función de esta variable. En la Figura 9.1 se muestra la evolución del RMSE en cuantiles en función de la variable `periodo`, junto con un histograma que nos permite visualizar la cantidad de datos en el conjunto de entrenamiento para cada valor de `periodo`. A su vez, la Figura 9.2 es muy similar pero mostrando el RMSE porcentual en cuantiles en función de la variable `periodo`.

9.2. Conjunto de datos de entrenamiento

TODO: Revisar teiendo en cuenta los resultados completos Uno de los resultados más consistentes del presente trabajo es que los modelos entrenados utilizando exclusivamente el IFN3 como conjunto de variables explicativas alcanzan un rendimiento sistemáticamente superior al de aquellos entrenados de forma conjunta con IFN2 e IFN3, tanto para la predicción del carbono total (tC) como del carbono normalizado por superficie (tC/ha). Esta mejora se manifiesta de forma clara en valores más elevados de R^2 y en reducciones apreciables de las métricas de error (RMSE y MAE), y se observa de manera estable en todos los algoritmos evaluados.

Una explicación plausible de este comportamiento está relacionada con la calidad y homogeneidad de los datos. El IFN3 fue realizado aproximadamente una década después del IFN2, incorporando avances metodológicos y tecnológicos relevantes en la recogida de información de campo, así como protocolos más refinados para la medición de variables estructurales y de

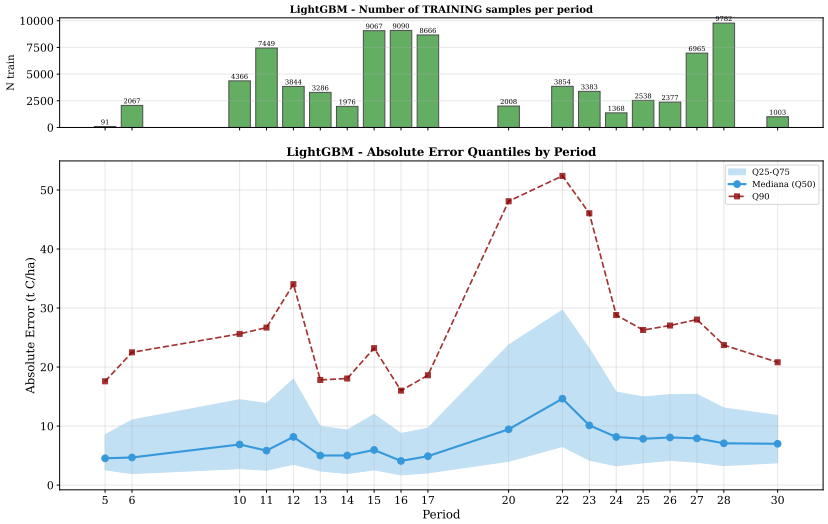

Figura 9.1. Evolución del RMSE en cuantiles en función de la variable `periodo` para el modelo LightGBM con IFN2 e IFN3 como explicativos.

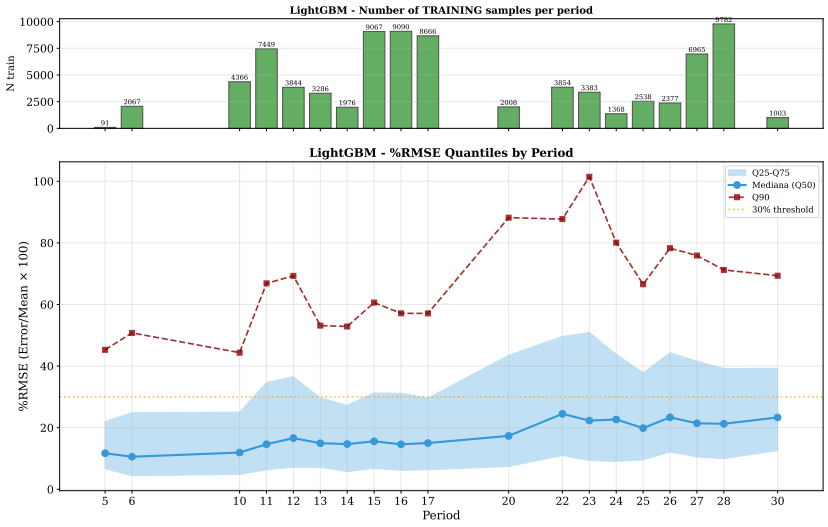

Figura 9.2. Evolución del RMSE porcentual en cuantiles en función de la variable `periodo` para el modelo LightGBM con IFN2 e IFN3 como explicativos.

estado de las masas forestales. A esto se une el aumento de variables que se miden en cada apeo de las parcelas. Esta mayor precisión y cantidad en las variables explicativas reduce el ruido inherente al proceso de modelización y facilita el aprendizaje de relaciones más consistentes entre predictores y variable objetivo. En este contexto, la inclusión de datos procedentes del IFN2 podría introducir heterogeneidad adicional asociada a diferencias metodológicas entre inventarios, lo que penaliza el rendimiento predictivo global.

Por otro lado, el número de observaciones disponibles para el entrenamiento difiere entre configuraciones. El entrenamiento exclusivo con datos del IFN3 se realiza sobre un conjunto de menor tamaño y que, además, incorpora filtros de calidad más estrictos, (filtro por `fccarb`> 20), que no está disponible en el IFN2. A pesar de estas diferencias en tamaño muestral y criterios de selección, los resultados no muestran indicios de sobreajuste en ninguno de los casos.

En conjunto, estos resultados ponen de manifiesto que, en este caso de estudio, la calidad y coherencia temporal del inventario parecen tener un impacto más relevante en el rendimiento predictivo que la simple agregación de información procedente de inventarios previos. Este hallazgo es especialmente relevante de cara a futuras aplicaciones operativas, ya que sugiere que modelos entrenados sobre inventarios recientes y metodológicamente homogéneos pueden ofrecer estimaciones más precisas y fiables del carbono forestal, incluso cuando se dispone de menos fuentes de información histórica.

A modo de conclusión, los resultados obtenidos indican que la elección del conjunto de entrenamiento debe adaptarse al horizonte temporal de predicción considerado. En particular, el modelo entrenado exclusivamente con datos del IFN3 resulta más adecuado para predicciones a medio plazo, en un horizonte temporal aproximado de entre 9 y 17 años, donde la mayor calidad y coherencia metodológica de este inventario se traduce en estimaciones más precisas y estables del carbono forestal. Por el contrario, cuando el objetivo es realizar predicciones a más largo plazo, con horizontes temporales superiores y que pueden extenderse hasta los 30 años, el uso combinado de datos procedentes del IFN2 y del IFN3 se hace obligatorio, ya que ofrece una base temporal más amplia que permite capturar mejor la evolución de las masas forestales en periodos prolongados. En este contexto, aunque el rendimiento predictivo sea ligeramente inferior, la integración de ambos inventarios aporta robustez frente a escenarios de extrapolación temporal, lo que hace recomendable su empleo para proyecciones de largo plazo.

9.3. Variable objetivo

En la interpretación de los resultados es fundamental contextualizar la diferencia entre las dos variables objetivo empleadas en el estudio: el carbono expresado en toneladas absolutas (`tC`) y el carbono normalizado por superficie (`tC/ha`). Tal y como se observa en la Tabla 6.1, la variable `carbono_bruto4` (tC) muestra una media inferior, pero una elevada dispersión relativa, con un rango amplio y valores extremos asociados a parcelas con estructuras muy heterogéneas. Por su parte, la variable `c4` (tC/ha) presenta una media más alta y una variabilidad absoluta mayor.

Estas diferencias estructurales tienen implicaciones directas sobre la capacidad predictiva de los modelos. En términos generales, la variable expresada en tC resulta más sencilla de modelizar, ya que integra implícitamente la superficie y reduce parte de la variabilidad introducida por la normalización por hectárea. Como consecuencia, los modelos entrenados para predecir carbono total alcanzan sistemáticamente valores más altos de R^2 y errores más bajos (RMSE y MAE) que aquellos orientados a la predicción de tC/ha. Esto indica que una mayor fracción de la varianza es explicada por las variables explicativas disponibles cuando la respuesta se expresa en términos absolutos.

En cambio, la predicción en tC/ha constituye un problema más exigente desde el punto de vista estadístico, al amplificar la heterogeneidad intraparcela y la influencia de factores locales no completamente capturados por los predictores. No obstante, esta variable resulta especialmente relevante para aplicaciones de comparación espacial, evaluación de productividad y análisis de eficiencia en el secuestro de carbono, lo que justifica su inclusión a pesar de presentar métricas de ajuste ligeramente inferiores. En conjunto, los resultados ponen de manifiesto que la elección de la variable objetivo debe alinearse con el objetivo final del análisis, asumiendo el compromiso existente entre interpretabilidad ecológica y rendimiento predictivo.

9.4. Distribución del error

El análisis conjunto de las métricas globales de los modelos predictivos muestra de forma consistente que el valor del RMSE es aproximadamente el doble del MAE. Dado que el RMSE penaliza de manera más severa los errores de gran magnitud, esta diferencia indica que, aunque el error medio absoluto se mantiene en niveles moderados, existen observaciones concretas en las que los modelos cometen desviaciones significativamente mayores.

Este patrón sugiere que la dificultad predictiva no es homogénea a lo largo de todo el rango de la variable objetivo, sino que se concentra en de-

terminados valores o contextos específicos. En este sentido, resulta necesario complementar la evaluación con métricas relativas como el SMAPE, que permiten analizar el error en proporción a la magnitud de la variable, así como con representaciones gráficas —como los diagramas de dispersión entre valores observados y predichos— que facilitan la identificación visual de sesgos, heterocedasticidad o rangos problemáticos del modelo.

En la Figura 9.3 se representa la dispersión entre los valores observados y las predicciones obtenidas por el modelo LightGBM (el de mayor R^2) para la variable objetivo `c4` (tC/ha)empleando como explicativos los inventarios IFN2 y IFN3. También se incluye un histograma de la densidad de puntos.

TODO: hace falta otra igual para carbono_bruto4. No se que modelo es el de la foto. Actualizar la imagen. Poner el caption de la foto en CASTELLANO.

Aunque `c4` presenta un rango muy amplio, desde valores próximos a cero hasta aproximadamente 880 tC/ha, la figura pone de manifiesto que la mayor concentración de observaciones se sitúa en el intervalo comprendido entre 0 y 200 tC/ha. En este rango, que además concentra la mayor parte de la masa de datos, la nube de puntos se alinea de forma clara en torno a la diagonal identidad.

Si bien existen casos puntuales en los que las predicciones se desvían notablemente de los valores reales, especialmente en los extremos superiores del rango, la estructura general del gráfico indica que el modelo reproduce de manera consistente la relación media entre valores observados y predichos. Este comportamiento es coherente con la diferencia observada entre RMSE y MAE, y refuerza la idea de que los errores más elevados se concentran en un subconjunto reducido de observaciones, mientras que el ajuste es sólido en las regiones donde se acumula la mayor densidad de datos.

Este comportamiento se analiza de forma explícita en la Figura 9.4, donde se representan, por intervalos de la variable objetivo `c4`, los valores de SMAPE y RMSE obtenidos por los distintos modelos base, junto con un histograma que muestra el número de observaciones disponibles en cada rango. Esta figura permite evaluar simultáneamente la magnitud del error y su dependencia de la densidad de datos a lo largo del dominio de la variable.

Los resultados confirman que el error relativo no es homogéneo. En los rangos intermedios de carbono, donde existe una combinación favorable de volumen de datos, comportamiento más estable del sistema y una cantidad de carbono suficientemente grande como para permitir un margen de error razonable (no es lo mismo un error de una toneladas en una parcela de 2

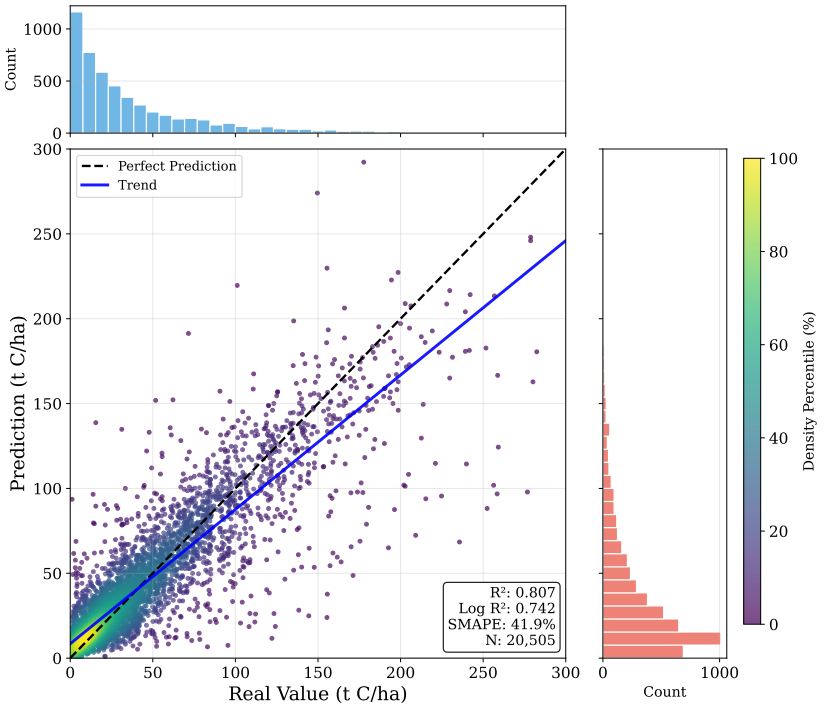

Figura 9.3. Dispersión de las predicciones frente a los valores reales para la variable objetivo `c4`. Se representa únicamente el rango $[0, 300]$ tC.

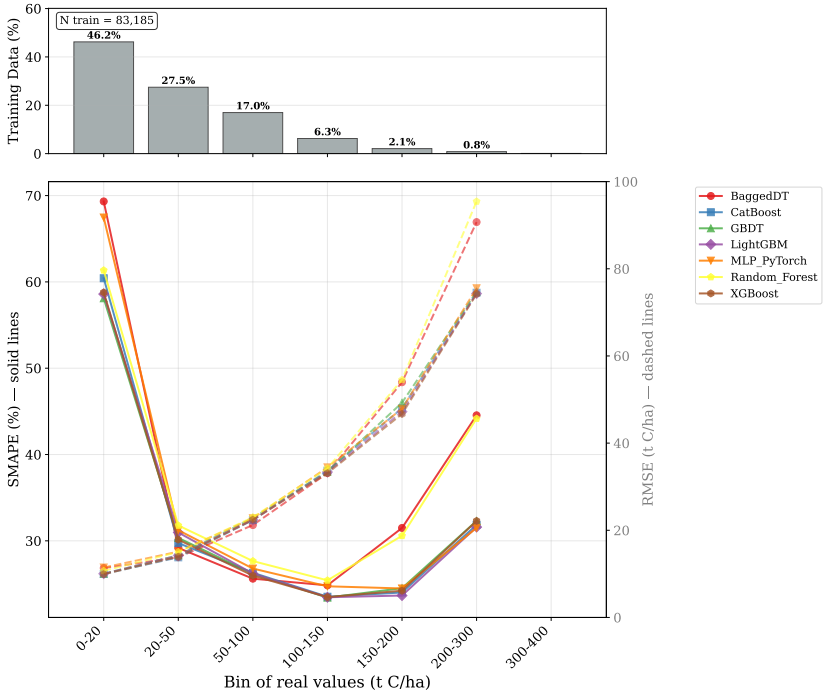

Figura 9.4. SMAPE y RMSE por rangos de la variable objetivo `c4` para los modelos base, junto con la distribución del número de observaciones en cada intervalo.

toneladas totales que en otra de 100 toneladas), el SMAPE alcanza valores mínimos, bajando del 20 % de error. En cambio, para valores bajos de `c4`, el SMAPE es elevado pese a la abundancia de observaciones, lo que refleja una alta variabilidad relativa en este régimen y una mala precisión de los modelos. Finalmente, en los valores más altos de carbono, el incremento del SMAPE coincide con una fuerte reducción del número de datos, evidenciando que la escasez de observaciones en los extremos de la distribución limita la capacidad de generalización del modelo.

TODO: Tal vez conviene hablar aqui sobre como de bien predice según PERIODO. Un grafico y tal.

10. Conclusiones

El objetivo principal de este trabajo es desarrollar un modelo de inteligencia artificial capaz de predecir de forma precisa la capacidad de captura de dióxido de carbono en cultivos forestales españoles, a partir de información estructural, edáfica, climática y espectral disponible en los Inventarios Forestales Nacionales y en fuentes de datos auxiliares. Los resultados obtenidos permiten afirmar que dicho objetivo se ha cumplido de manera satisfactoria.

En primer lugar, se ha demostrado que es posible construir modelos predictivos robustos y generalizables para estimar el carbono forestal a medio y largo plazo. Entre las distintas configuraciones evaluadas, el mejor rendimiento global se alcanzó mediante un esquema de stacking entrenado con datos del IFN2 e IFN3 como variables explicativas y del IFN4 como variable objetivo, utilizando como modelos base CatBoost, LightGBM, XGBoost, Random Forest, GBDT y BaggedDT, y una red neuronal multicapa (MLP) como metamodelo. Esta configuración permite predecir el carbono total en toneladas para horizontes temporales comprendidos entre 5 y 30 años, alcanzando un coeficiente de determinación en test de aproximadamente $R^2 = 0{,}85$ y un error absoluto medio del orden de 6.4 toneladas de carbono. Estos valores indican una elevada capacidad explicativa y una precisión compatible con aplicaciones prácticas en planificación forestal y estimación de créditos de carbono.

Así mismo, el análisis comparativo entre modelos individuales y esquemas de stacking ha puesto de manifiesto que los algoritmos basados en árboles de decisión son los que presentan un mejor comportamiento de forma consistente, destacando especialmente CatBoost y LightGBM. Aunque el stacking no produce incrementos sustanciales en el coeficiente de determinación, sí aporta mejoras sistemáticas en el MAE, lo que resulta especialmente relevante desde un punto de vista operativo, al reducir el error medio en las estimaciones finales.

Un segundo resultado relevante del estudio es la identificación de los factores que condicionan en mayor medida la capacidad de captura de carbono. El proceso de selección manual de variables permitió reducir un conjunto inicial de 445 predictores a un subconjunto compacto de 44 variables, manteniendo una representación equilibrada de todos los ámbitos ecológicos implicados. Los resultados muestran que la mayor parte de la capacidad predictiva del modelo se explica por variables estructurales y de composición de la masa forestal, en particular el número de pies por clase diamétrica y las varia-

bles asociadas a la especie y su estado. Las variables edáficas, topográficas y de manejo aportan información adicional relevante, mientras que las variables climáticas y los índices de vegetación actúan como moduladores del crecimiento y la acumulación de carbono, refinando las predicciones especialmente a escala estacional.

En conjunto, este trabajo contribuye con una metodología reproducible y basada en datos reales para la predicción de la captura de carbono forestal a futuro. La integración de información multifuente, el uso de técnicas avanzadas de aprendizaje automático y la validación rigurosa del rendimiento permiten disponer de una herramienta con potencial aplicación en la planificación de proyectos de forestación, la optimización del secuestro de carbono y la evaluación técnica de iniciativas vinculadas al mercado de créditos de carbono. Estos resultados refuerzan el valor de la inteligencia artificial como apoyo a la toma de decisiones ambientales y abren la puerta a futuras extensiones del modelo, como su adaptación a otros contextos geográficos o su integración en sistemas operativos de gestión forestal.

El objetivo de este trabajo es la obtención de un modelo de Inteligencia Artificial capaz de predecir el carbono que una cierta parcela de terreno forestada o reforestada capturará en un cierto periodo de tiempo. Para ello se han recogido datos de tierra (Inventario Forestal Nacional [7]), datos meteorológicos [8] e imágenes satelitales [17] con los que se han entrenado varios modelos para intentar predecir el carbono capturado por las parcelas presenten en las iteraciones 2 y 3 del Inventario Forestal Nacional, comparando el resultado con la última de las iteraciones, la 4. Las predicciones se hicieron para dos configuraciones distintas: usando como datos explicativos únicamente los del inventario 2 y usando como datos explicativos los de los inventarios 3 y 4. A su vez, para cada caso se realizó la predicción de dos variables objetivo: la predicción del carbono en toneladas por hectárea (tC/ha) y la predicción del carbono en toneladas (tC). Los resultados para los mejores modelo en cada caso están recogidos en la Tabla **??**.

Con estos resultados podemos afirmar que disponemos de datos suficientes y de suficiente calidad para entrenar modelos capaces de predecir el carbono capturado con un error aceptable.

11. Recomendaciones para Futuras Investigaciones

A partir de los resultados obtenidos y de las limitaciones identificadas durante el desarrollo de este trabajo, se proponen a continuación varias líneas de investigación que podrían contribuir a mejorar y ampliar el alcance del modelo desarrollado.

En primer lugar, sería recomendable ampliar y diversificar la base de datos empleada. La incorporación de futuras ediciones del Inventario Forestal Nacional permitiría reforzar la dimensión temporal del conjunto de entrenamiento y evaluar con mayor detalle la estabilidad del modelo ante horizontes temporales más largos. Así mismo, la extensión del estudio a otras regiones bioclimáticas, tanto dentro como fuera del ámbito nacional, permitiría analizar la capacidad de generalización del modelo y su adaptabilidad a contextos ecológicos distintos.

En relación con las variables explicativas, futuras investigaciones podrían explorar la inclusión de nuevas fuentes de información, como datos de teledetección de mayor resolución espacial o temporal (por ejemplo, LIDAR aéreo o satelital). Del mismo modo, la incorporación explícita de variables relacionadas con perturbaciones (incendios, plagas, siembras, talas...) podría mejorar la capacidad del modelo para capturar dinámicas no lineales en la acumulación de carbono.

Desde el punto de vista metodológico, sería de interés evaluar arquitecturas de aprendizaje más avanzadas, como modelos de deep learning especializados en series temporales o enfoques híbridos que combinen modelos mecanicistas de crecimiento forestal con técnicas de aprendizaje automático. Así mismo, el análisis sistemático de la incertidumbre asociada a las predicciones, por ejemplo, mediante enfoques bayesianos o técnicas de quantile regression, permitiría proporcionar intervalos de confianza, un aspecto especialmente relevante para aplicaciones vinculadas a la certificación de créditos de carbono.

Otra línea de trabajo prometedora consiste en profundizar en la interpretabilidad de los modelos. El uso de técnicas explicativas avanzadas podría facilitar una comprensión más detallada del papel de cada variable en la predicción final, reforzando la confianza de técnicos y gestores en el uso del modelo y favoreciendo su adopción en contextos operativos.

Por último, desde una perspectiva aplicada, sería recomendable desarrollar herramientas que faciliten la transferencia del modelo a usuarios finales. Esto podría materializarse en una interfaz gráfica o plataforma web que per-

mita introducir escenarios de plantación y obtener estimaciones de captura de carbono de forma directa. En este contexto, también podría explorarse la integración del modelo con sistemas de registro y trazabilidad, como tecnologías de blockchain, para apoyar la gestión y certificación de créditos de carbono de manera transparente y verificable.

En conjunto, estas líneas de investigación futura permitirían consolidar y ampliar el impacto del modelo propuesto, reforzando su utilidad científica, técnica y aplicada en el ámbito de la gestión forestal y la mitigación del cambio climático.

Agradecimientos

Investigación financiada por la subvención **TSI-100933-2023-1** de la **Convocatoria de Cátedras Universidad-Empresa (Cátedras ENIA 2022), Ministerio de Transformación Digital y Función Pública de España**, y el **Plan de Recuperación y Resiliencia de la UE** (*NextGenerationEU/PRTR*).

Referencias

[1] Intergovernmental Panel on Climate Change. *Climate Change 2007: Mitigation of Climate Change*. Cambridge, UK: Cambridge University Press, 2007.

[2] United Nations Framework Convention on Climate Change. *The Kyoto Protocol*. 1997. URL: https://unfccc.int/resource/docs/convkp/kpeng.pdf.

[3] United Nations Framework Convention on Climate Change. *Paris Agreement*. 2015. URL: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

[4] United Nations Framework Convention on Climate Change. *Decision 16/CMP.1: Land use, land-use change and forestry*. Conference of the Parties serving as the meeting of the Parties to the Kyoto Protocol. Acuerdos de Marrakech. 2005. URL: https://unfccc.int/resource/docs/2005/cmp1/eng/08a03.pdf.

[5] United Nations Framework Convention on Climate Change. *Report of the individual review of the initial report of Spain under the Kyoto Protocol*. UNFCCC Secretariat. Incluye la definición nacional de bosque de España: 1 ha, 20 % de cabida cubierta, 3 m de altura. 2010. URL: https://unfccc.int/files/kyoto_protocol/compliance/plenary/application/pdf/cc-ert-irr-2007-14__report_of_the_review_of_ir_of_spain.pdf.

[6] European Union. *Regulation (EU) 2018/841 on the inclusion of greenhouse gas emissions and removals from land use, land use change and forestry*. Official Journal of the European Union. Marco LULUCF y principios de permanencia y contabilidad de sumideros. 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R0841.

[7] MITECO. *Inventario Forestal Nacional (IFN2, IFN3, IFN4): metodología y bases de datos*. Ministerio para la Transición Ecológica y el Reto Demográfico (España). 2023. URL: https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/ifn.aspx.

[8] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Giorgio Arduini, Gianpaolo Balsamo et al. "ERA5-Land: A state-of-the-art global reanalysis at land surfaces". En: *EGU General Assembly / ECMWF (Copernicus Climate Data Store)* (2021). DOI: 10.24381/cds.e2161bac. URL: https://doi.org/10.24381/cds.e2161bac.

[9] USGS. *Landsat Collection 2 Level-2 Science Products: Surface Reflectance*. U.S. Geological Survey. 2021. URL: https://www.usgs.gov/landsat-missions/landsat-collection-2-level-2-science-products.

[10] Maider Araceli Urbón Jiménez, Jaime Gabriel Vegas, Ana de Luis Reboredo y Belén Pérez Lancho. "GreenWest-DB: Base de datos integrada de atributos forestales, climáticos y espectrales para España". Manuscrito en preparación. Universidad de Salamanca, Grupo BISITE. 2025.

[11] Scott J. Goetz, Alessandro Baccini, Nadine T. Laporte, Tanya Johns, Walter Walker, Josef Kellndorfer, Richard A. Houghton y M. Sun. "Mapping and monitoring carbon stocks with satellite observations: a comparison of methods". En: *Carbon Balance and Management* 4.2 (2009), pág. 2. DOI: 10.1186/1750-0680-4-2. URL: https://doi.org/10.1186/1750-0680-4-2.

[12] Jintong Ren, Lizhi Liu, You Wu, Lijian Ouyang y Zhenyu Yu. "Estimating Forest Carbon Stock Using Enhanced ResNet and Sentinel-2 Imagery". En: *Forests* 16.7 (2025). Submission received: 13 June 2025 / Revised: 15 July 2025 / Accepted: 18 July 2025 / Published: 20 July 2025, pág. 1198. DOI: 10.3390/f16071198. URL: https://doi.org/10.3390/f16071198.

[13] Fugen Jiang, Muli Deng, Jie Tang, Liyong Fu y Hua Sun. "Integrating spaceborne LiDAR and Sentinel-2 images to estimate forest aboveground biomass in Northern China". En: *Carbon Balance and Management* 17 (2022), pág. 12. DOI: 10.1186/s13021-022-00212-y.

[14] Gyri Reiersen, David Dao, Björn Lütjens, Konstantin Klemmer, Kenza Amara, Attila Steinegger, Ce Zhang y Xiaoxiang Zhu. "ReforesTree: A dataset for estimating tropical forest carbon stock with deep learning and aerial imagery". En: *arXiv preprint arXiv:2201.11192* (2022). URL: https://arxiv.org/abs/2201.11192.

[15] Wenquan Dong, Edward T.A. Mitchard, Hao Yu, Steven Hancock y Casey M. Ryan. "Forest aboveground biomass estimation using GEDI and earth observation data through attention-based deep learning". En: *arXiv preprint arXiv:2311.03067* (2023). URL: https://arxiv.org/abs/2311.03067.

[16] Ministerio para la Transición Ecológica y el Reto Demográfico. *INSTRUCCIONES DE USO DE LA CALCULADORA DE ABSORCIONES DE CO2 EX ANTE DE LAS ESPECIES FORESTALES ARBÓREAS ESPAÑOLAS DEL MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO*. Accedido: 2025-07-16. 2023. URL: https://www.miteco.gob.es/content/dam/miteco/es/cambio-climatico/temas/mitigacion-politicas-y-medidas/instruccionescalculadoraabexante_tcm30-485629.pdf.

[17] USGS. *USGS Landsat 5 Level 2, Collection 2, Tier 1*. Accedido: 2025-07-08. 2025. URL: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LT05_C02_T1_L2.

[18] J. Muñoz Sabater. *ERA5-Land hourly data from 1950 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Accedido: 07-07-2025. 2019. DOI: 10.24381/cds.e2161bac. URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview.

[19] Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO). *Guía para la estimación de absorciones de dióxido de carbono*. 2021. URL: https://www.miteco.gob.es/content/dam/miteco/es/cambio-climatico/temas/mitigacion-politicas-y-medidas/guiapa_tcm30-479094.pdf.

[20] Intergovernmental Panel on Climate Change. *2006 IPCC Guidelines for National Greenhouse Gas Inventories*. Geneva, Switzerland: IPCC, 2006.

[21] Jérôme Chave, Maxime Réjou-Méchain, Alberto Búrquez, Emmanuel Chidumayo, Matthew S. Colgan, Welington B. C. Delitti, Alvaro Duque, Tron Eid, Philip M. Fearnside, Rosa C. Goodman, Mark Henry, Angelina Martínez-Yrízar, Wilson A. Mugasha, Helene C. Muller-Landau, Maurizio Mencuccini, Brian W. Nelson, Alfred Ngomanda, Eurípedes M. Nogueira, Edgar Ortiz-Malavassi, Raphaël Pélissier, Pierre Ploton,

Casey M. Ryan, Juan G. Saldarriaga y Ghislain Vieilledent. "Improved allometric models to estimate the aboveground biomass of tropical trees". En: *Global Change Biology* 20.10 (2014), págs. 3177-3190. DOI: 10.1111/gcb.12629.

[22] Gregorio Montero Ricardo Ruiz-Peinado y M. Muñoz. *Producción de biomasa y fijación de CO_2 por los bosques españoles*. Serie Forestal, 23. Madrid, España: Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), 2009.

[23] Miguel del Río y Gregorio Montero Ricardo Ruiz-Peinado. "New models for estimating the carbon sink capacity of Spanish softwood species". En: *Forest Systems* 20.1 (2011), págs. 176-188. DOI: 10.5424/fs/2011201-11643. URL: https://doi.org/10.5424/fs/2011201-11643.

[24] Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO). *Manual de campo y base de datos del Cuarto Inventario Forestal Nacional (IFN4)*. Subdirección General de Política Forestal y Lucha contra la Desertificación. Madrid, España, 2017. URL: https://www.miteco.gob.es/es/biodiversidad/temas/inventarios-nacionales/inventario-forestal-nacional/cuarto_inventario.html.

Apéndice A. Apéndices

Apéndice A.1. Origen y cálculo de las variables `ca` y `cr`

Las variables `ca` (carbono arbóreo) y `cr` (carbono radical) incluidas en la base de datos del *Inventario Forestal Nacional* (IFN4) derivan de las ecuaciones alométricas de biomasa desarrolladas por el *Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria* (INIA), en particular por *Gregorio Montero* y *Ricardo Ruiz-Peinado* [22, 23]. Estas ecuaciones fueron elaboradas a partir de datos de campo obtenidos mediante talas y pesadas directas de árboles de distintas especies representativas de la flora forestal española.

Cada ecuación estima la biomasa seca (en kilogramos) de los diferentes componentes del árbol en función del diámetro normal (D, en cm, medido a 1,3 m del suelo) y la altura total (H, en m). Para cada especie o grupo de especies similares se dispone de ecuaciones específicas de la forma:

$$W_i = a_i \cdot D^{b_i} \cdot H^{c_i}$$

donde W_i representa la biomasa del componente i (fuste, corteza, ramas, hojas, raíces, etc.), y a_i, b_i y c_i son coeficientes empíricos obtenidos mediante regresión no lineal. En los casos en que una especie no dispone de ecuación propia, se utiliza la de otra especie considerada análoga por similitud morfológica o ecológica.

Los componentes de biomasa definidos en el IFN4 incluyen [24]:

- W_s: biomasa del fuste (kg),
- W_c: biomasa de la corteza del fuste (kg),
- W_{b7}: biomasa de ramas mayores de 7 cm de diámetro (kg),
- W_{b2-7}: biomasa de ramas entre 2 y 7 cm de diámetro (kg),
- $W_{b0,5-2}$: biomasa de ramas entre 0,5 y 2 cm de diámetro (kg),
- W_t: biomasa de ramas menores de 0,5 cm de diámetro (kg),
- W_h: biomasa de hojas (kg),
- W_{db}: biomasa de ramas muertas (kg),
- $W_T = W_s + W_c + W_{b7} + W_{b2-7} + W_{b0,5-2} + W_t + W_h$: biomasa aérea total (kg),
- W_r: biomasa radical (raíces, kg).

A partir de estas ecuaciones, el cálculo de biomasa y carbono en el IFN4 se realiza de la siguiente forma:

1. **Biomasa por árbol (kg):** en la tabla `Mayores_exs` se incluyen las medidas de diámetro y altura de cada pie. Aplicando las ecuaciones alométricas correspondientes se obtiene la biomasa aérea (W_T) y radical (W_r) para cada árbol.
2. **Conversión a carbono (kg):** se aplica un factor de conversión estándar de 0.5, según las directrices del IPCC [20], de forma que:

$$\text{CA} = 0{,}5 \times W_T, \quad \text{CR} = 0{,}5 \times W_r$$

3. **Expansión a valores por hectárea (t/ha):** los valores por árbol se convierten a toneladas por hectárea mediante un *factor de expansión* (*Fac*), que refleja la densidad de árboles por unidad de superficie dentro de cada clase diamétrica y especie. Este factor se calcula en función del número de pies inventariados y la superficie de muestreo, permitiendo expresar los resultados en términos comparables de biomasa o carbono por hectárea.
4. **Agregación por clases diamétricas y especie:** finalmente, en la tabla `Parcelas_exs` se agrupan los valores por parcela, especie y clase diamétrica (CD), sumando las contribuciones individuales ya expandidas. El resultado son los valores medios de biomasa y carbono por hectárea (t/ha) para cada combinación de parcela y especie.

El mismo procedimiento se aplica tanto a la biomasa aérea (para obtener `ca`) como a la biomasa radical (para `cr`). De esta forma, **`ca` Y `cr` representan el**

carbono almacenado en la biomasa viva, aérea y subterránea respectivamente, expresado en toneladas de carbono por hectárea (t/ha).

Este enfoque metodológico se ajusta a las recomendaciones del *IPCC Guidelines for National Greenhouse Gas Inventories* [20], garantizando la coherencia con los métodos de reporte de carbono a nivel internacional y facilitando la comparación de los resultados con otros estudios y marcos regulatorios.

Apéndice A.2. Estado de las Poblaciones (`estado_id`)

Se determinará las fases de desarrollo de las *poblaciones* codificándose de la siguiente forma:

1. **Repoblado**. Conjunto de pies que desde el estrato herbáceo llega hasta el subarbustivo y los pies inician la tangencia de copas.
2. **Monte bravo**. Comprende desde el estrato y clase de edad anterior hasta el momento en que por efecto del crecimiento, los pies empiezan a perder las ramas inferiores; es decir que en esta clase de edad, las ramas se encuentran a lo largo de todo el fuste.
3. **Latizal**. Comprende desde la clase anterior hasta que los pies tienen 20 cm de diámetro normal; es decir, el diámetro de su fuste, medido a la altura de 1,30 m del suelo.
4. **Fustal**. Se caracteriza esta clase de edad, porque sus pies tienen diámetros normales superiores a 20 cm.

Apéndice A.3. Forma Principal de Masa (IFN3 e IFN4: `fpmasa_id`)

1. **Coetánea**. Cuando al menos el 90 % de sus pies tienen la misma edad individual. Ejemplo típico: las repoblaciones.
2. **Regular**. Cuando al menos el 90 % de sus pies pertenecen a la misma clase artificial de edad o misma clase diamétrica en su defecto.
3. **Semirregular**. Cuando al menos el 90 % de sus pies pertenecen a dos clases artificiales de edad cíclicamente contiguas o dos clases diamétricas contiguas en su defecto.
4. **Irregular**. Cuando no se cumplen las condiciones anteriores, es decir, cuando en cualquier parte de la masa existen pies más o menos mezclados, de todas las clases de edad que tiene la masa o de varias clases diamétricas en su defecto.

Apéndice A.4. Tratamiento de la Masa (IFN3 e IFN4: `tratmasa_id`)

1. **Monte alto**. Cuando todos los pies proceden de semilla.
2. **Monte medio**. Cuando coexisten pies de la misma especie, unos procedentes de semilla (brinzales) y otros de brote (chirpiales).
3. **Monte bajo**. Cuando todos los pies proceden de brote de cepa o de raíz.

Apéndice A.5. Origen de la Masa (IFN3 e IFN4: `orgmasa_id`)

1. **Natural**. Bosque desarrollado espontáneamente, sin intervención humana directa.
2. **Artificial**. Plantado intencionadamente por el ser humano.
3. **Naturalizado**. Bosque originalmente plantado pero que ha evolucionado hacia una estructura más similar a un bosque natural.

Apéndice A.6. Tipo de Suelo (`tipsuelo1_id`, `tipsuelo2_id`, `tipsuelo3_id`)

Se utilizará la siguiente codificación para el tipo de suelo, diferenciando tres variables:

Tipo de suelo (I): Presencia de sales, yesos o hidromorfía

1. **No se observan sales, yesos ni procesos de fidromorfía.**
2. **Suelo salino.** Si presenta al menos dos de las siguientes características:
   - Presencia de eflorescencias en la superficie o a distintas profundidades.
   - Existencia de plantas halófitas.
   - Zonas llanas o endorreicas con climas secos que provocan gran evaporación.
3. **Suelo yesífero.** Si presenta alguna de las siguientes características:
   - Presencia de materia yesífera en superficie o a distintas profundidades.
   - Existencia de plantas gipsófilas.
4. **Suelo hidromorfo.** Si el suelo presenta síntomas de hidromorfía acusada, cumpliendo al menos dos de las siguientes:
   - Zona encharcada permanente o casi permanentemente de forma natural.
   - Zona llana o endorreica con climas húmedos.
   - Grietas en verano si no hay encharcamiento.
   - Presencia de vegetación indicadora de hidromorfismo.

   Identificandose las siguientes:
   - Formaciones vegetales indicadoras de hidromorfía:
     - Ribereñas: *saucedas, mimbreras, alisedas.*
     - Brezales con *Erica ciliaris, Erica tetralix.*
     - Turberas arboladas (excepto Cornisa Cantábrica y Pirineos).
     - Turberas de montaña con *Sphagnum, Erica tetralix.*
     - Cervunales con *Nardus stricta.*
     - Carrizales y espadañares (*Phragmites, Tipha, Cladium*).
     - Juncales (*Scirpus, Juncus*).

    - Pastizales con cárices (*Carex spp.*).
    - Marismas.
- Formaciones vegetales gipsófilas:
    - Aznallar: matorral de *Ononis tridentata*.
    - Tomillares gipsófilos con:
        - *Lepidium subulatum*
        - *Gypsophila spp.*
        - *Matthiola fruticulosa*
- Formaciones vegetales indicadoras de suelos salinos:
    - Salicorniales: matas leñosas crasas (Salicornia, Arthrocnemum, Halozylon).
    - Bosques halófitos del género *Tamarix*.
    - Saladar o sosar: predominio de *Suaeda vera*.
    - Saladar blanco: predominio de *Atriplex halimus*.

Tipo de suelo (II y III): Composición del suelo (calizo o silíceo)

1. **Suelo calizo.** Más del 50 % de la vertical del perfil da efervescencia con ácido clorhídrico.
    - **Moderadamente básico:** pH en superficie $\leq$ 8.5.
    - **Fuertemente básico:** pH en superficie >8.5.
2. **Suelo silíceo.** Menos del 50 % de la vertical del perfil da efervescencia.
    - **Moderadamente ácido:** pH $\geq$ 5.5.
    - **Fuertemente ácido:** pH <5.5.

Apéndice A.7. Rocosidad (`rocosidad_id`)

Se considerará el conjunto de la parcela clasificando la rocosidad según la siguiente codificación:

1. **Sin pedregosidad**: la superficie de la parcela está completamente cubierta de vegetación.
2. **Poco pedregoso**: cuando la superficie de la parcela cubierta por rocas coherentes es menor del 25 %.
3. **Pedregoso**: cuando la superficie rocosa está comprendida entre el 25 % y el 50 %.
4. **Muy pedregoso**: cuando la superficie rocosa se sitúa entre el 50 % y el 75 %.
5. **Roquedo**: cuando la superficie de rocas es mayor del 75 %. En este caso, no se tomará ningún dato adicional correspondiente a suelos.

Apéndice A.8. Textura del Suelo (`textura_id`)

Se clasificará en función de la siguiente codificación:

1. **Suelo arenoso.** Si los cilindros se deshacen sin apenas formarse.
2. **Suelo franco.** Es posible hacer cilindros gruesos pero no delgados.
3. **Suelo arcilloso.** Se consiguen cilindros de unos 5 mm de diámetro.

Apéndice A.9. Contenido en Materia Orgánica (IFN3 e IFN4: `matorg_id`)

Según la siguiente clasificación:

1. **Suelo muy humífero.** Cuando a 15 cm la pureza es menor de 4, o cuando la capa de broza sea de espesor mayor de 5 cm y a 15 cm de profundidad la pureza sea menor de 6.
2. **Suelo moderadamente humífero.** Cuando a 15 cm la pureza sea menor de 6 con capa de broza nula o de escaso espesor, o cuando dicha capa tenga espesor mayor de 5 cm y a 15 cm de profundidad la pureza sea igual o mayor de 6.
3. **Suelo poco humífero.** En los restantes casos.

Apéndice A.10. Modelo de Combustible (IFN3 e IFN4: `modcomb_id`)

Se determinará la clase de combustible que es más probable que propague el fuego si hubiese un incendio en la zona, hasta un máximo de 60m: pasto, matorral, hojarasca de bosque o deshechos o restos de corta. Se determinará el modelo de combustible a partir de la siguiente clave:

Tabla A.1. Descripción de los modelos de combustible del Inventario Forestal Nacional, clasificados por grupo funcional.

GRUPO	MOD.	DESCRIPCIÓN DEL MODELO
Pastos	1	Pasto fino, seco y bajo, que recubre completamente el suelo. Puede aparecer algunas plantas leñosas dispersas ocupando menos de 1/3 de la superficie.
	2	Pasto fino, seco y bajo, que recubre completamente el suelo. Las plantas leñosas dispersas cubren de 1/3 a 2/3 de la superficie; pero la propagación del fuego se realiza por el pasto.
	3	Pasto grueso, denso, seco y alto (>1 m). Puede haber algunas plantas leñosas dispersas. Los campos de cereales son representativos de este modelo.

Continúa en la siguiente página

GRUPO	MOD.	DESCRIPCIÓN DEL MODELO
Matorral	4	Matorral o plantación joven muy densa; de más de 2 m de altura; con ramas muertas en su interior. Propagación del fuego por las copas de las plantas.
	5	Matorral disperso, denso y verde, de menos de 1 m de altura. Propagación del fuego por la hojarasca, el pasto, las ramillas y el matorral.
	6	Parecido al modelo 5, pero con especies más inflamables, de mayor talla, pudiéndose encontrar ramas gruesas en el suelo. Propagación del fuego con vientos moderados a fuertes.
	7	Matorral de especies muy inflamables; de 0.5 a 2 m de altura, situado como sotobosque en masas de coníferas.
Hojarasca bajo arbolado	8	Bosque denso, sin matorral. Propagación del fuego por la hojarasca muy compacta, formada por acículas cortas (5 cm o menos) o por hojas planas no muy grandes.
	9	Parecido al modelo 8, pero con hojarasca menos compacta, formada por acículas largas y rígidas (P. pinaster) o follaje de frondosas de hoja grande, caducas (castaño o robles).
	10	Bosque con gran cantidad de leña y árboles caídos, como consecuencia de vendavales, plagas intensas, etc.
Restos de corta y operaciones selvícolas	11	Bosque claro y fuertemente aclarado. Restos de poda o aclareo ligeros (diámetro <7.5 cm).
	12	Predominio de los restos sobre el arbolado. La hojarasca y el matorral presente ayudarán a la propagación del fuego.
	13	Grandes acumulaciones de restos gruesos y pesados, cubriendo todo el suelo.

Apéndice A.11. Distribución Espacial (`disesp_id`)

La disposición de la vegetación en el espacio se clasificará según la siguiente codificación:

1. **Uniforme.** Cuando el estrato arbóreo presenta continuidad en el espacio.
2. **Diseminada en bosquetes aislados.** Cuando la masa arbórea se encuentra dividida en porciones que tienen una superficie inferior a 0,5 ha.
3. **Diseminada en individuos aislados.** Cuando se trata de dehesas.
9. **Otras o no se sabe.** En caso diferente a los anteriores o si se desconoce el dato exacto.

Apéndice A.12. Composición Específica (`comesp_id`)

En función de las especies presentes:

1. **Masas homogéneas o puras**. Masas monoespecíficas con una única especie arbórea. La normativa española precisa que una masa es monoespecífica o pura cuando al menos el 90 % de los pies pertenecen a la misma especie.
2. **Masas heterogéneas o mezcladas pie a pie**. Masas de diferentes especies que se juntan o bien se entremezclan por golpes o grupos, siempre que tengan una altura similar.
3. **Masas heterogéneas o mezcladas con subpiso**. Las dos o más especies mezcladas, cuando alcancen el estado adulto y la estabilidad, presentarán alturas diferentes.
9. **Otras o no se sabe**. En caso diferente a los anteriores o desconocer el dato exacto.

Apéndice A.13. Manifestaciones Erosivas (`merosiva_id`)

Se observará la parcela y sus alrededores hasta una distancia de 60 metros desde el centro, y se codificará la existencia de manifestaciones erosivas según la siguiente clave:

1. **No hay ninguna manifestación.**
2. **Cuellos de raíces al descubierto:** los cuellos de las raíces están visibles, con acumulación de residuos aguas arriba de los tallos y obstáculos, así como abundancia superficial de piedras.
3. **Presencia de regueros:** canales paralelos de erosión con una profundidad máxima de un palmo (aproximadamente 20 cm).
4. **Cárcavas y barrancos en V:** erosión lineal más profunda que los regueros, con forma de "V".
5. **Cárcavas y barrancos en U:** erosión avanzada con formas suavizadas y amplias en "U".
6. **Deslizamientos del terreno:** desplazamientos de masas de tierra, ladera o materiales del suelo.

Apéndice A.14. Nivel de usos del suelo (IFN3 e IFN4: `nivel1_id`)

1. **Monte.** Toda superficie en la que vegetan especies arbóreas, arbustivas, de matorral o herbáceas, ya sea espontáneamente o procedan de siembra o plantación, siempre que no sean características de cultivo agrícola o fueran objeto del mismo.

2. **Agrícola.** Territorio o ecosistema poblado con siembras o plantaciones de herbáceas y/o leñosas, anuales o plurianuales que se laborea con una fuerte intervención humana, puede estar poblado por especies forestales de fruto (flor, hojas o en el futuro biomasa) siempre que la intervención humana sea importante. Incluye las dehesas, montes huecos o montes adehesados de base cultivo, siempre que la fracción de cabida cubierta de los árboles sea inferior al 5 %.
3. **Artificial.** Territorio o ecosistemas dominado por edificios, parques urbanos (aunque estén poblados de árboles), viveros fuera de los montes (aunque sean de especies forestales), carreteras (salvo las vías de servicio de los montes) u otras construcciones humanas que tengan superficies continuas.
4. **Humedal.** Lo constituyen las lagunas, charcas, zonas húmedas, marismas y corrientes discontinuas de agua en las que, al menos durante 6 meses del año, esté presente dicho líquido.
5. **Agua.** Es la parte de la tierra constituida por ríos, lagos, embalses, canales o estanques con superficies continuas de más de 0.26 ha y con agua prácticamente todo el año.

Apéndice A.15. Nivel morfoestructural (IFN3 e IFN4: `nivel2_id`*)*

Para el nivel de usos del suelo Monte se definirán los siguientes niveles morfoestructurales.

1. **Monte arbolado.** Territorio o ecosistema con especies forestales arbóreas como manifestación vegetal de estructura vertical dominante y con una fracción de cabida cubierta igual o superior al 20 %; incluye dehesas con base cultivo o pastizal con labores siempre que la fracción arbolada supere el 20 %, y excluye terrenos con fuerte intervención humana para obtener frutos, hojas, flores o varas.
2. **Monte arbolado ralo.** Terreno de uso forestal con especies arbóreas forestales dominantes y fracción de cabida cubierta entre el 10 % y 20 % (incluido el 10 %, excluido el 20 %); también aplica a terrenos con matorral o pastizal natural como dominantes, pero con presencia importante de árboles forestales, incluyendo dehesas de base de cultivo.
3. **Monte temporalmente desarbolado.** Terreno que fue monte arbolado recientemente y que casi con seguridad volverá a estar cubierto de árboles en un futuro próximo.
4. **Monte desarbolado.** Terreno con matorral y/o pastizal natural o

débil intervención humana como cobertura dominante, con fracción de cabida cubierta por árboles forestales inferior al 5 %.

5. **Monte sin vegetación superior.** Terreno de uso forestal que no está poblado por vegetales superiores debido a condiciones actuales de suelo, clima o topografía, aunque podría estarlo en otras circunstancias.
6. **Árboles fuera del monte.** Incluye riberas arboladas no estructuradas con los montes, bosquetes de menos de 2.500 m^2, alineaciones de especies arbóreas o arbustivas de menos de 25 m de anchura, y árboles sueltos en terreno forestal.
7. **Monte arbolado disperso.** Terreno forestal con especies arbóreas dominantes y fracción de cabida cubierta entre el 5 % y el 10 % (incluido el 5 %, excluido el 10 %); también terrenos con matorral o pastizal como cobertura dominante pero con presencia significativa de árboles forestales, incluyendo dehesas de base cultivo.

Apéndice A.16. Código de los grupos taxonómicos de las especies (`grupo_id`)

Tabla A.2. Relación de códigos de grupo taxonómico utilizados en la variable `grupo_id`.

Código	Grupo taxonómico	Código	Grupo taxonómico
7	Acacia	69	Phoenix
15	Crataegus	73	Betula
19	Coníferas	77	Tilia
20	Pinos	78	Sorbus
31	Abies	79	Platanus
35	Larix	80	Laurisilva
40	Quercus	91	Buxus
53	Tamarix	93	Pistacia
57	Salix	94	Laurus
58	Populus	95	Prunus
60	Eucalyptus	99	Frondosas
65	Ilex	399	Morus
68	Arbutus	455	Fraxinus
917	Cedrus	936	Cupressus
937	Juniperus	956	Ulmus
975	Juglans	976	Acer
997	Sambucus		

Apéndice A.17. Código de las especies (`especie_id`)

Tabla A.3. Relación de especies empleadas en el estudio y metadatos asociados.

Cód.	Nombre	Sinonimia	Tipo	Grupo
307	Acacia dealbata	Acacia dealbata	1	7
207	Acacia melanoxylon	Acacia melanoxylon	1	7
7	Acacia spp.	-	1	7
392	Gleditsia triacanthos	Acacia gleditsia	1	7
92	Robinia pseudoacacia	Acacia robinia	1	7
292	Sophora japonica	Acacia sofora	1	7

Continúa en la siguiente página

Tabla A.3. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
515	Crataegus azarolus	Espino	1	15
415	Crataegus laciniata	Majoleto	1	15
315	Crataegus laevigata	Espino majuelo	1	15
215	Crataegus monogyna	Majuelo	1	15
15	Crataegus spp.	-	1	15
30	Mezcla de coníferas	Coníferas | excepto pinos	0	19
19	Otras coníferas	-	0	19
29	Otros pinos	-	0	20
20	Pinos	-	0	20
27	Pinus canariensis	-	0	20
24	Pinus halepensis	-	0	20
25	Pinus nigra	Pinus laricio | Pinus clusiana	0	20
26	Pinus pinaster	Pinus maritima	0	20
23	Pinus pinea	-	0	20
28	Pinus radiata	Pinus insignis	0	20
21	Pinus sylvestris	-	0	20
22	Pinus uncinata	Pinus montana | Pinus mugo	0	20
31	Abies alba	Abies pectinata	0	31
32	Abies pinsapo	-	0	31
235	Larix decidua	Alerce común	0	35
335	Larix leptolepis	Larix kaempferi | Alerce leptolepis	0	35
35	Larix spp.	-	0	35
435	Larix x eurolepis	Alerce híbrido	0	35
49	Otros quercus	-	1	40
344	Quercus alpestris	-	1	40
47	Quercus canariensis	Quercus lusitanica var. baetica	1	40
44	Quercus faginea	Quercus lusitanica var. faginea	1	40
45	Quercus ilex ssp. ballota	Quercus rotundifolia	1	40
245	Quercus ilex ssp. ilex	-	1	40
244	Quercus lusitanica	Quercus fruticosa | Quejigueta	1	40
42	Quercus petraea	Quercus sessiliflora	1	40
243	Quercus pubescens	Quercus pubescens | Quercus humilis	1	40
43	Quercus pyrenaica	Quercus toza	1	40
41	Quercus robur	Quercus pedunculata	1	40

Continúa en la siguiente página

Tabla A.3. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
48	Quercus rubra	Quercus borealis	1	40
46	Quercus suber	-	1	40
253	Tamarix canariensis	Tarajal	1	53
53	Tamarix spp.	-	1	53
257	Salix alba	Sauce blanco	1	57
357	Salix atrocinerea	Bardaguera	1	57
858	Salix canariensis	Sauce canario	1	57
557	Salix cantabrica	Sauce cantábrico	1	57
657	Salix caprea	Sauce cabruno	1	57
757	Salix elaeagnos	Sarga	1	57
857	Salix fragilis	Mimbre	1	57
957	Salix purpurea	Mimbrera	1	57
57	Salix spp.	-	1	57
51	Populus alba	-	1	58
58	Populus nigra	-	1	58
52	Populus tremula	-	1	58
258	Populus x canadensis	Populus x euroamericana	1	58
62	Eucalyptus camaldulensis	Eucalyptus rostrata	1	60
61	Eucalyptus globulus	-	1	60
364	Eucalyptus gompho-cephalus	Eucalipto gonfo	1	60
64	Eucalyptus nitens	-	1	60
464	Eucalyptus robusta	-	1	60
264	Eucalyptus viminalis	Eucalipto viminalis	1	60
63	Otros eucaliptos	-	1	60
65	Ilex aquifolium	-	1	65
82	Ilex canariensis	-	1	65
282	Ilex platyphylla	Naranjero	1	65
268	Arbutus canariensis	Madroño canario	1	68
68	Arbutus unedo	-	1	68
469	Phoenix canariensis	Palmera	1	69
69	Phoenix spp.	-	1	69
273	Betula alba	Betula verrucosa | Abedul pubescens	1	73
373	Betula pendula	Betula hispanica | Abedul péndula	1	73

Continúa en la siguiente página

Tabla A.3. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
73	Betula spp.	-	1	73
277	Tilia cordata	Tilo cordata	1	77
377	Tilia platyphyllos	Tilo común	1	77
77	Tilia spp.	-	1	77
278	Sorbus aria	Mostajo	1	78
378	Sorbus aucuparia	Serbal de cazadores	1	78
778	Sorbus chamaemespilus	Serbal chame	1	78
478	Sorbus domestica	Serbal común	1	78
678	Sorbus latifolia	Serbal de hoja ancha	1	78
78	Sorbus spp.	-	1	78
578	Sorbus torminalis	Serbal torminal	1	78
79	Platanus hispanica	Platanus hybrida	1	79
279	Platanus orientalis	Plátano oriental	1	79
80	Laurisilva	-	1	80
89	Otras laurisilvas	-	1	80
291	Buxus balearica	Boj de Baleares	1	91
91	Buxus sempervirens	-	1	91
293	Pistacia atlantica	Cornicabra canaria	1	93
93	Pistacia terebinthus	Cornicabra	1	93
294	Laurus azorica	Laurel canario	1	94
94	Laurus nobilis	Laurel	1	94
395	Prunus avium	Cerezo silvestre	1	95
495	Prunus lusitanica	Loro | hija	1	95
595	Prunus padus	Prunus	1	95
295	Prunus spinosa	Espino negro	1	95
95	Prunus spp.	Prunus	1	95
70	Mezcla de frondosas de gran porte	Frondosas de gran porte (H.t. >10 m)	1	99
90	Mezcla de pequeñas frondosas	Frondosas de pequeño porte (H.t. $\leq$ 10 m)	1	99
99	Otras frondosas	Otras frondosas	1	99
499	Morus alba	Morera	1	399
599	Morus nigra	Morera	1	399
399	Morus spp.	Morera	1	399
55	Fraxinus angustifolia	-	1	455
255	Fraxinus excelsior	Fresno excelsior	1	455

Continúa en la siguiente página

Tabla A.3. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
355	Fraxinus ornus	Fresno orno	1	455
955	Fraxinus spp.	Fresnos	1	455
17	Cedrus atlantica	-	0	917
217	Cedrus deodara	Cedrus deodara	0	917
317	Cedrus libani	Cedrus libani	0	917
917	Cedrus spp.	Cedrus spp.	0	917
337	Juniperus cedrus	Enebro canario	0	917
236	Cupressus arizonica	Ciprés arizónica	0	936
336	Cupressus lusitanica	Ciprés lambertiana	0	936
436	Cupressus macrocarpa	Ciprés americano	0	936
36	Cupressus sempervirens	-	0	936
936	Cupressus spp.	Cipres	0	936
37	Juniperus communis	-	0	937
237	Juniperus oxycedrus	Enebro oxicedro	0	937
39	Juniperus phoenicea	-	0	937
239	Juniperus sabina	Sabina rastrera	0	937
937	Juniperus spp.	Enebros y sabinas	0	937
38	Juniperus thurifera	-	0	937
238	Juniperus turbinata	Sabina canaria	0	937
256	Ulmus glabra	Ulmus montana	1	956
56	Ulmus minor	Ulmus campestris	1	956
356	Ulmus pumila	Olmo pumilo	1	956
956	Ulmus spp.	Olmo	1	956
275	Juglans nigra	Nogal	1	975
75	Juglans regia	-	1	975
975	Juglans spp.	-	1	975
76	Acer campestre	-	1	976
276	Acer monspessulanum	Arce de Montpelier	1	976
376	Acer negundo	Negundo fraxinifolia | Arce negundo	1	976
476	Acer opalus	Arce ópalus	1	976
676	Acer platanoides	Arce platanoide	1	976
576	Acer pseudoplatanus	Arce seudoplátano	1	976
976	Acer spp.	Arces	1	976
97	Sambucus nigra	Saúco negro	1	997

Continúa en la siguiente página

Tabla A.3. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
297	Sambucus racemosa	Saúco racemosa	1	997
997	Sambucus spp.	-	1	997
11	Ailanthus altissima	Ailanthus glandulosa	1	-
54	Alnus glutinosa	-	1	-
2	Amelanchier ovalis	Guillomo	1	-
88	Apollonias barbujana	Apollonias canariensis	1	-
98	Carpinus betulus	Carpe	1	-
72	Castanea sativa	Castanea vesca	1	-
13	Celtis australis	-	1	-
67	Ceratonia siliqua	-	1	-
18	Chamaecyparis lawsoniana	-	0	-
369	Chamaerops humilis	Palmito	1	-
9	Cornus sanguinea	-	1	-
74	Corylus avellana	-	1	-
569	Dracaena draco	Drago	1	-
83	Erica arborea	-	1	-
283	Erica scoparia	Tejo | brezo arbóreo escopario	1	-
5	Euonymus europaeus	-	1	-
71	Fagus sylvatica	-	1	-
299	Ficus carica	Higuera	1	-
3	Frangula alnus	Rhamnus frangula	1	-
1	Heberdenia bahamensis	Heberdenia excelsa	1	-
12	Malus sylvestris	-	1	-
60	Mezcla de eucaliptos	Eucaliptos	1	-
50	Mezcla de árboles de ribera	Árboles ripícolas	1	-
81	Myrica faya	-	1	-
281	Myrica rivasmartinezii	-	1	-
6	Myrtus communis	-	1	-
87	Ocotea phoetens	-	1	-
66	Olea europaea	Olea oleaster	1	-
59	Otros árboles ripícolas	-	1	-
84	Persea indica	-	1	-
8	Phillyrea latifolia	-	1	-
86	Picconia excelsa	Notelaea excelsa	1	-

Continúa en la siguiente página

Tabla A.3. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
33	Picea abies	Picea excelsa	0	-
289	Pleiomeris canariensis	Delfino	1	-
34	Pseudotsuga menziesii	Pseudotsuga douglasii	0	-
16	Pyrus spp.	-	1	-
40	Quercus	-	1	-
4	Rhamnus alaternus	Aladierno	1	-
389	Rhamnus glandulosa	Sanguino	1	-
96	Rhus coriaria	Zumaque	1	-
457	Salix babylonica	Sauce llorón	1	-
85	Sideroxylon marmulano	-	1	-
10	Sin asignar	Sin asignar	1	-
14	Taxus baccata	-	0	-
219	Tetraclinis articulata	Tetraclinis articulata	0	-
319	Thuja spp.	Thuja	0	-
489	Visnea mocanera	Mocan	1	-

Apéndice A.18. Resultados

Apéndice A.18.1. IFN2 e IFN3 como explicativos para `carbono_bruto4` *(tC)*

Tabla A.4. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en toneladas (carbono_bruto4) con el conjunto de datos que emplea IFN2 e IFN3 como explicativos.

Modelo	R^2_{test}	$\text{RMSE}_{\text{test}}$	MAE_{test}
CatBoost	**0.845**	**13.846**	**6.615**
LightGBM	0.841	14.006	6.654
XGBoost	0.840	14.054	6.655
GBDT	0.838	14.159	6.722
MLP	0.832	14.410	6.931
BaggedDT	0.821	14.858	7.282
Random Forest	0.819	14.950	7.135
BayesianNN	0.775	16.674	8.906
SVR	0.679	19.897	8.137

TODO: corregir las combinaciones para el stack. No coinciden con las descritas arriba ni con las entrenadas para IFN3 como explicativo. Debería ser stack1-6 modelos, satck2-4 modelos, stack3-3 modelos, stack4-3 modelos, stack5-2 modelos TODO: comprobar que sean los valores finales

Stack	**Metamodelo**	**Bases**	**Test R^2**	**RMSE**	**MAE**
stack1	GradientBoosting	2	0.84	14.04	6.53
stack1	LinearRegression	2	0.84	13.97	6.62
stack1	Ridge	2	0.84	13.97	6.62
stack1	RandomForest	2	0.81	15.13	7.27
stack1	SVR	2	0.84	14.15	6.53
stack1	**MLP**	**2**	**0.84**	**13.97**	**6.48**
stack2	GradientBoosting	3	0.84	13.98	6.52
stack2	LinearRegression	3	0.84	13.91	6.59
stack2	Ridge	3	0.84	13.91	6.59
stack2	RandomForest	3	0.83	14.65	7.02
stack2	SVR	3	0.84	14.06	6.51
stack2	**MLP**	**3**	**0.84**	**13.91**	**6.51**
stack3	GradientBoosting	4	0.84	13.88	6.45
stack3	LinearRegression	4	0.85	13.81	6.56
stack3	Ridge	4	0.85	13.81	6.56
stack3	RandomForest	4	0.83	14.54	6.91
stack3	SVR	4	0.84	13.97	6.47
stack3	**MLP**	**4**	**0.85**	**13.78**	**6.41**
stack4	GradientBoosting	5	0.85	13.82	6.42
stack4	LinearRegression	5	0.85	13.78	6.53
stack4	Ridge	5	0.85	13.78	6.53
stack4	RandomForest	5	0.84	14.24	6.75
stack4	SVR	5	0.84	13.94	6.45
stack4	**MLP**	**5**	**0.85**	**13.77**	**6.43**
stack5	GradientBoosting	6	0.85	13.81	6.42
stack5	LinearRegression	6	0.85	13.78	6.54
stack5	Ridge	6	0.85	13.78	6.54
stack5	RandomForest	6	0.84	14.21	6.71
stack5	SVR	6	0.84	13.94	6.45
stack5	**MLP**	**6**	**0.85**	**13.76**	**6.40**

Tabla A.5. Resultados de las diferentes configuraciones de stacking utilizando IFN2 e IFN3 como explicativos de la variable en toneladas de carbono.

Apéndice A.18.2. IFN2 e IFN3 como explicativos para c4 (tC/ha)

Tabla A.6. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en tC/ha con el conjunto de datos que emplea IFN2 e IFN3 como explicativos.

Modelo	R^2_{test}	$\text{RMSE}_{\text{test}}$	MAE_{test}
LightGBM	**0.787**	**22.767**	**11.650**
XGBoost	0.784	22.952	11.590
CatBoost	0.783	22.990	11.607
GBDT	0.783	23.014	11.658
MLP	0.771	23.607	12.287
BaggedDT	0.740	25.142	13.021
Random Forest	0.732	25.547	12.908
BayesianNN	0.678	28.021	14.689
SVR	0.551	33.065	13.708

TODO: corregir las combinaciones para el stack. No coinciden con las descritas arriba ni con las entrenadas para IFN3 como explicativo. Debería ser stack1-6 modelos, satck2-4 modelos, stack3-3 modelos, stack4-3 modelos, stack5-2 modelos

TODO: comprobar que sean los valores finales

Stack	Metamodelo	Bases	Test R^2	RMSE	MAE
stack1	GradientBoosting	2	0.78	23.33	11.63
stack1	LinearRegression	2	0.79	22.77	11.61
stack1	Ridge	2	0.79	22.77	11.61
stack1	RandomForest	2	0.75	24.91	12.88
stack1	SVR	2	0.78	23.11	**11.37**
stack1	**MLP**	**2**	**0.79**	**22.63**	**11.58**
stack2	GradientBoosting	3	0.78	23.18	11.53
stack2	LinearRegression	3	0.79	22.56	11.43
stack2	Ridge	3	0.79	22.57	11.43
stack2	RandomForest	3	0.75	24.45	12.35
stack2	SVR	3	0.79	22.85	**11.21**
stack2	**MLP**	**3**	**0.79**	**22.40**	**11.40**
stack3	GradientBoosting	4	0.78	23.21	11.45
stack3	LinearRegression	4	0.79	22.73	11.46
stack3	Ridge	4	0.79	22.74	11.45
stack3	RandomForest	4	0.75	24.78	12.41
stack3	SVR	4	0.78	23.00	**11.22**
stack3	**MLP**	**4**	**0.79**	**22.78**	**11.34**
stack4	GradientBoosting	5	0.77	23.53	11.48
stack4	LinearRegression	5	0.79	22.60	11.42
stack4	Ridge	5	0.79	22.60	11.41
stack4	RandomForest	5	0.75	24.73	12.22
stack4	SVR	5	0.79	22.87	**11.18**
stack4	**MLP**	**5**	**0.79**	**22.53**	**11.31**
stack5	GradientBoosting	6	0.78	23.28	11.42
stack5	LinearRegression	6	0.79	22.57	11.39
stack5	Ridge	6	0.79	22.57	11.38
stack5	RandomForest	6	0.76	24.15	12.03
stack5	SVR	6	0.79	22.86	**11.16**
stack5	**MLP**	**6**	**0.79**	**22.39**	**11.32**

Tabla A.7. Resultados de las diferentes configuraciones de stacking utilizando IFN2 e IFN3 como explicativos de la variable en tC/ha.

Apéndice A.18.3. IFN3 como explicativo para `carbono_bruto4` *(tC)*

Tabla A.8. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en tC con el conjunto de datos que emplea IFN3 como explicativo.

Modelo	R^2_{test}	$\text{RMSE}_{\text{test}}$	MAE_{test}
LightGBM	**0.909**	**10.662**	**5.477**
XGBoost	0.907	10.772	5.580
CatBoost	0.907	10.807	5.570
GBDT	0.904	10.942	5.732
MLP	0.896	11.392	6.382
BaggedDT	0.882	12.154	6.420
Random Forest	0.872	12.643	6.533
BayesianNN	0.842	14.079	7.890
SVR	0.825	14.797	7.124
KNN	0.788	16.270	8.166
AdaBoost	0.575	23.056	19.535

Stack	Metamodelo	Bases	Test R^2	RMSE	MAE
stack1	GradientBoosting	6	0.9121	10.4852	5.2841
stack1	LinearRegression	6	0.9122	10.4816	5.3798
stack1	Ridge	6	0.9122	10.4815	5.3798
stack1	RandomForest	6	0.9057	10.8580	5.5726
stack1	SVR	6	0.9098	10.6226	5.3112
stack1	**MLP**	**6**	**0.9140**	**10.3723**	**5.2515**
stack2	GradientBoosting	4	0.9124	10.4693	5.2930
stack2	LinearRegression	4	0.9120	10.4914	5.3853
stack2	Ridge	4	0.9120	10.4914	5.3853
stack2	RandomForest	4	0.9050	10.9015	5.6023
stack2	SVR	4	0.9096	10.6341	5.3147
stack2	**MLP**	**4**	**0.9136**	**10.3941**	**5.2625**
stack3	GradientBoosting	3	0.9105	10.5796	5.3948
stack3	LinearRegression	3	0.9112	10.5411	5.4317
stack3	Ridge	3	0.9112	10.5411	5.4317
stack3	RandomForest	3	0.8999	11.1916	5.8307
stack3	SVR	3	0.9089	10.6775	5.3694
stack3	**MLP**	**3**	**0.9122**	**10.4789**	**5.3739**
stack4	GradientBoosting	3	0.9084	10.7041	5.4036
stack4	LinearRegression	3	0.9088	10.6822	5.5379
stack4	Ridge	3	0.9088	10.6822	5.5379
stack4	RandomForest	3	0.8983	11.2777	5.8314
stack4	SVR	3	0.9060	10.8425	5.4644
stack4	**MLP**	**3**	**0.9103**	**10.5951**	**5.3681**
stack5	GradientBoosting	2	0.9098	10.6245	5.4007
stack5	LinearRegression	2	0.9092	10.6546	5.4719
stack5	Ridge	2	0.9092	10.6546	5.4718
stack5	RandomForest	2	0.8920	11.6247	6.1032
stack5	SVR	2	0.9069	10.7932	5.4151
stack5	**MLP**	**2**	**0.9101**	**10.6019**	**5.3545**

Tabla A.9. Resultados de las diferentes configuraciones de stacking con el conjunto que emplea IFN3 como explicativo de la variable en tC.

Apéndice A.18.4. IFN3 como explicativo para c4 (tC/ha)

Tabla A.10. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en tC/ha con el conjunto de datos que emplea IFN3 como explicativo.

Modelo	R^2_{test}	$\text{RMSE}_{\text{test}}$	MAE_{test}
CatBoost	**0.860**	**17.709**	**9.250**
XGBoost	0.858	17.828	9.207
LightGBM	0.858	17.841	9.159
GBDT	0.853	18.141	9.463
MLP	0.837	19.086	10.917
BaggedDT	0.826	19.726	10.402
Random Forest	0.826	19.730	10.489
BayesianNN	0.775	22.454	12.260
KNN	0.769	22.755	12.252
SVR	0.734	24.403	11.219
AdaBoost	0.473	34.336	26.295

Stack	Metamodelo	Bases	Test R^2	RMSE	MAE
stack1	GradientBoosting	6	0.8682	17.1742	8.9546
stack1	LinearRegression	6	0.8639	17.4508	8.9676
stack1	Ridge	6	0.8639	17.4510	8.9677
stack1	RandomForest	6	0.8592	17.7459	9.3749
stack1	SVR	6	0.8612	17.6207	8.8012
stack1	**MLP**	**6**	**0.8656**	**17.3380**	**8.8789**
stack2	GradientBoosting	4	0.8599	17.7069	8.9856
stack2	LinearRegression	4	0.8635	17.4725	8.9908
stack2	Ridge	4	0.8635	17.4727	8.9908
stack2	RandomForest	4	0.8523	18.1766	9.5079
stack2	SVR	4	0.8613	17.6142	8.8209
stack2	**MLP**	**4**	**0.8645**	**17.4083**	**8.8888**
stack3	GradientBoosting	3	0.8590	17.7638	9.0799
stack3	LinearRegression	3	0.8615	17.6005	9.0450
stack3	Ridge	3	0.8615	17.6005	9.0450
stack3	RandomForest	3	0.8449	18.6308	9.7939
stack3	SVR	3	0.8592	17.7490	8.8749
stack3	**MLP**	**3**	**0.8619**	**17.5765**	**8.9734**
stack4	GradientBoosting	3	0.8520	18.1962	9.2188
stack4	LinearRegression	3	0.8604	17.6722	9.1764
stack4	Ridge	3	0.8604	17.6723	9.1765
stack4	RandomForest	3	0.8396	18.9435	9.8789
stack4	SVR	3	0.8574	17.8597	9.0159
stack4	**MLP**	**3**	**0.8620**	**17.5712**	**9.0898**
stack5	GradientBoosting	2	0.8446	18.6435	9.1995
stack5	LinearRegression	2	0.8578	17.8360	9.1252
stack5	Ridge	2	0.8578	17.8361	9.1252
stack5	RandomForest	2	0.8332	19.3163	10.2404
stack5	SVR	2	0.8552	17.9992	9.0003
stack5	**MLP**	**2**	**0.8579**	**17.8333**	**9.1459**

Tabla A.11. Resultados de las diferentes configuraciones de stacking con el conjunto que emplea IFN3 como explicativo de la variable en tC/ha.