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

---

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 definidos en la normativa internacional y nacional vigente:

- **Intervención humana directa:** Los árboles deben provenir de actividades de intervención humana, como la plantación, siembra o fomento de semilleros naturales.
- **Período mínimo de 30 años:** Para que un proyecto sea válido, debe garantizarse que los árboles permanezcan en el terreno durante un período mínimo de tiempo, generalmente 30 años, lo que excluye la absorción de carbono de cultivos estacionales, cuyo carbono es liberado nuevamente al ser cosechados.
- **Superficie mínima de 1 hectárea:** El proyecto debe abarcar al menos 1 hectárea de terreno para ser considerado.

- **Fracción mínima de cabida cubierta del 20 %:** Para que un área sea considerada como bosque, debe cubrir al menos el 20 % del área con especies arbóreas.
- **Altura mínima de los árboles maduros de 3 metros:** Los árboles deben alcanzar una altura mínima de 3 metros en su madurez, aunque no es necesario que alcancen esta altura al inicio de la plantación.

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?* Para responderla, se integran datos del **Inventario Forestal Nacional** (IFN3–IFN4, MITECO) [4], reanálisis **ERA5-Land** [5] e **índices espectrales Landsat** (Collection 2, L2) [6] en una base de datos relacional jerárquica descrita en un trabajo complementario [**greenwestdb**].

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) [7] 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) [8] 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) [9] 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) [10] o Dong et al. (2023) [11], 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 [12], 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 [13, 14, 15]. 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

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 [16].

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, volumen con y sin corteza, área basimétrica, 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 al clorofila.

5.1.1. Estrutura de la base de datos

Estos datos se organizan en las siguientes entidades troncales:

- **parcelas**: icontiene la información básica de localización y características edáficas de cada parcela.
- **parcela_inventario**: describe el estado de cada parcela en un inventario determinado (`parcela_id`, `inventario_id`), incluyendo atributos edáficos y de contexto (p. ej., `nivel1_id`, `textura_id`).
- **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* (`cd_id`): n.º de pies (`npies`), área basimétrica (`abas`), volúmenes (`vcc`, `vsc`, `vle`), incrementos (`iavc`) y carbono (`ca`, `cr`).
- **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 (`ht`), diámetros (`dn1` y `dn2`), ubicación respecto del centro de la parcela (`rumbo`, `distancia`), volumen (`vcc, vsc, vle`), incremento (`iavc`) y carbono (`ca, cr`).
- **parcela_inventario_estacion**: almacena agregados climático-biofísicos por estación (`estacion_id`) en la misma granularidad parcela–inventario, incluyendo variables como precipitación (`PR`) y temperatura (`T2M, SKT, STL*`), 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.

5.1.2. Diccionario resumido de variables

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

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			

Continúa en la siguiente página

Variable	Descripción	Unidad	Tipo de dato
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
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

Continúa en la siguiente página

Variable	Descripción	Unidad	Tipo de dato
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
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

Continúa en la siguiente página

Variable	Descripción	Unidad	Tipo de dato
`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 puede capturar en un horizonte temporal de 20–30 años, a partir de las condiciones observadas en inventarios previos. Para ello se definieron 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

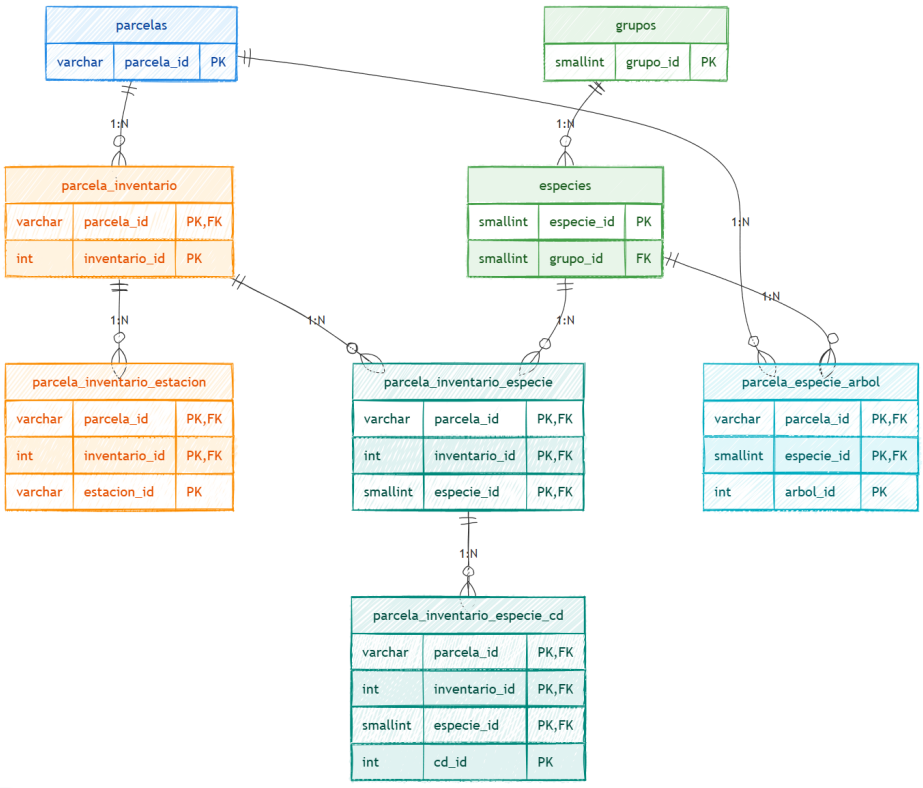

Figura 5.1. Esquema relacional de las tablas principales de la base de datos. Tabla extraida de [**greenwestdb**], donde se pueden consultar más detalles sobre las variables.

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 [**chave2014**] y las directrices del IPCC [**ipcc2006**], 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**).

5.3. Supuestos de elegibilidad y verificación externa

Para que un proyecto forestal sea elegible en programas de *créditos de carbono*, debe cumplir requisitos técnicos establecidos por marcos regulatorios internacionales [16, 15]. 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 de 30 años.** 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.
  No obstante, este planteamiento se enfrenta a la limitación impuesta por la **permanencia mínima de 30 años**, requisito fundamental en el contexto de los proyectos de compensación. El intervalo de tiempo entre los inventarios **IFN3** e **IFN4** es relativamente corto: no supera los 18 años.
  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.
  TODO: Actualizar figura y quitar título de la propia imagen
  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.
  TODO: Actualizar figura y quitar título de la propia imagen
  De esta forma, el modelo se entrena y valida sobre un conjunto de da-

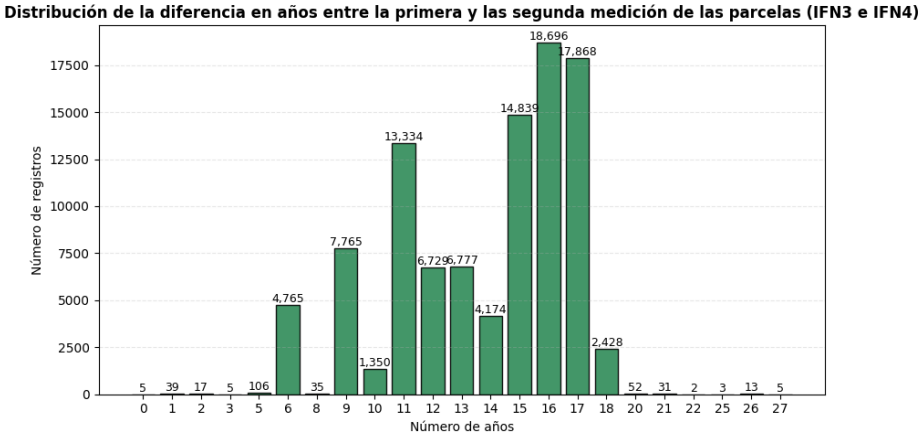

Figura 5.2. Distribución de la diferencia de años entre los inventarios IFN3 e IFN4.

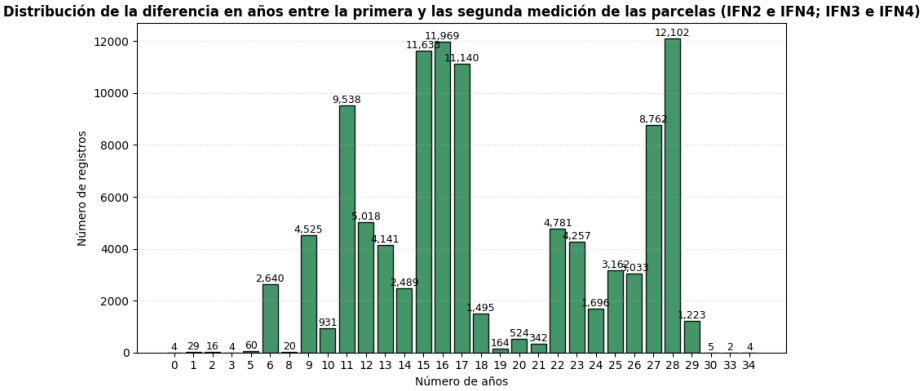

Figura 5.3. Distribución de la diferencia de años entre los inventarios IFN2–IFN3 e IFN3–IFN4.

tos más diverso y equilibrado, tanto en estructura como en amplitud temporal, manteniendo la coherencia metodológica y la trazabilidad de las estimaciones. Este enfoque no sólo mejora la robustez del aprendizaje, sino que también refuerza la capacidad del modelo para proyectar la captura de carbono en escenarios compatibles con los requisitos de permanencia de los proyectos de compensación.

- **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 prác-

tica, 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. En nuestro conjunto de datos, la altura no se registra explícitamente, por lo que 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*— más que del ajuste predictivo del modelo. 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: `c` de **IFN4** o `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 (variable objetivo por `c` o `carbono_bruto` y primera inventariación/ inventariación explicativa por **IFN3** o la unión de **IFN2** e **IFN3**).

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 **IFN3**. 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. Codificación y normalización

Las variables categóricas se codifican mediante *one-hot encoding*, generando variables binarias para cada clase. Las variables numéricas se escalan (normalización estándar o min-max, según el modelo) para asegurar que todas las magnitudes tengan el mismo orden de importancia durante el entrenamiento.

5.4.4. 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

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.	Anexos Apéndice A.18, 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.	Anexos 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.	Anexos varios
radio, orientacion, elevacion, pendiente	float	Topografía y geometría de parcela.	–
nivel1_id, nivel2_id, fccarb, fcctot	int/float	Niveles jerárquicos y cabida cubierta.	Anexos 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.

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.

5.5. Partición y validación

Para obtener una estimación imparcial del rendimiento y evitar *fugas de información* debidas a la correlación espacial dentro de cada parcela, la partición del conjunto de datos se realiza **por identificador de parcela** (`parcela_id`). Todas las observaciones asociadas a una misma parcela se asignan *íntegramente* a un único subconjunto, de modo que ninguna parcela aparece simultáneamente en entrenamiento y evaluación. **Validación interna y control de sesgo temporal.** Sobre el subconjunto de entrenamiento (80 %) se aplica *validación cruzada por grupos* utilizando como agrupador los *años transcurridos entre inventarios* (p. ej., 15, 16, 17, ...). Esta estrategia comprueba la *estabilidad* del modelo frente a cambios en el horizonte temporal y reduce el riesgo de sobreajuste específico de un periodo. La selección de hiperparámetros se realiza exclusivamente dentro de esta validación interna; el conjunto de evaluación (20 %) permanece *sellado* para la prueba final.

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.
- **Moda del Error**: valor más frecuente del error absoluto a nivel de parcela o individuo, útil para identificar el error típico en la predicción y detectar patrones dominantes en el comportamiento del modelo.

5.6. Selección de variables explicativas

La selección de predictores se abordó mediante cuatro estrategias complementarias: (1) selección automática mediante *Featurewiz*, (2) selección basada en el criterio de mínima redundancia y máxima relevancia (*mRMR*), (3) selección manual basada en criterios estadísticos y conceptuales, y (4) un procedimiento secuencial supervisado fundamentado en el rendimiento predictivo (SSSRP). El objetivo común de estas aproximaciones fue identificar un subconjunto parsimonioso de variables que maximizara la capacidad predictiva del modelo, redujera la colinealidad y mantuviera la coherencia ecológica de las relaciones.

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. Las configuraciones empleadas son:

```
['CatBoost', 'LightGBM', 'XGBoost', 'Random Forest', 'GBDT', 'BaggedDT'],
['CatBoost', 'LightGBM', 'Random Forest', 'GBDT'],
['LightGBM', 'XGBoost', 'GBDT'],
['CatBoost', 'Random Forest', 'GBDT'],
['LightGBM', 'Random Forest']
```

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 combinaciones 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 de los modelos

El desarrollo y evaluación de los modelos predictivos se realizó íntegramente en **Python**, utilizando librerías como `scikit-learn`, `cuML` y `PyTorch`, junto con implementaciones específicas de gradient boosting como `XGBoost`, `LightGBM` y `CatBoost`. El entrenamiento se llevó a cabo en dos fases. Para los modelos cuyos datos de entrenamiento provenían solo del IFN3 se uso un equipo local con procesador Intel Core i7 y 32 GB de RAM. Los modelos que usaban los datos del IFN2 e IFN3 se empleo el ordenador de HPC de la Universidad de Salamanca. Esto fue por la posibilidad de usar tarjetas gráficas Nvidia H100, que aceleraba mucho el entrenamiento para los modelos que tienen opción a ejecutarse en una GPU gracias a la posibilidad de paralelización. No obstante, el entrenamiento se podría hacer llevado a cabo en un equipo de escritorio equipado con una tarjeta gráfica comercial, ya que los requisitos no son tan altos.

6.1. Preparación del conjunto de datos

La variable objetivo a predecir es el **carbono bruto acumulado** por parcela (`carbono_bruto4`). Para garantizar la consistencia de la muestra, se aplicaron los siguientes filtros:

- Se eliminaron observaciones con `carbono_bruto4` nulo o igual a cero.
- Se retuvieron únicamente aquellas filas donde `carbono_bruto` es nulo o estrictamente menor que `carbono_bruto4`, evitando duplicidades o inconsistencias entre ambas medidas.
- Se restringió el análisis al inventario `ifn_id = 3`.

Las variables explicativas se seleccionaron a partir de una configuración manual (`features`) que combina:

- Índices de pies por hectárea a distintos diámetros (`npies_1`, `npies_2`, ..., `npies_70`).
- Variables climáticas y edáficas (por ejemplo, `martonneidx_id`, `tipsuelo2_id`).
- Variables topográficas (elevación, pendiente, orientación).
- Índices de vegetación y variables de temperatura superficial (`NDII`, `EVI`, `GNDVI`, `skt_mean`, `skt_std`) en distintas estaciones.
- Atributos forestales y estructurales de la masa (`tipo_especie`, `estado_id`, `orgmasa1_id`, `matorg_id`, etc.).

6.2. División entrenamiento–prueba y validación cruzada

Cuando la tabla de datos incluye el identificador de parcela (`parcela_id`), la partición entrenamiento–prueba se realiza mediante `GroupShuffleSplit`, reservando un 20 % de las parcelas para prueba y empleando el 80 % restante para entrenamiento. De este modo, todas las observaciones de una misma parcela quedan necesariamente en el mismo conjunto, evitando fuga de información.

Sobre el conjunto de entrenamiento se aplica validación cruzada por grupos mediante `GroupKFold` con 5 particiones, utilizando también `parcela_id` como variable de agrupación. En ausencia de esta columna, se recurre a un `KFold` clásico estratificado en 5 particiones con barajado.

Este esquema permite evaluar la capacidad de generalización del modelo en parcelas no vistas, reduciendo el riesgo de sobreajuste estructural.

6.3. Preprocesamiento de variables

El preprocesamiento se implementó con un `ColumnTransformer` que distingue tres bloques de variables:

- **Variables numéricas generales**: imputación mediante la mediana y escalado con `StandardScaler`.
- **Variables de densidad de pies** (`npies_1`, ..., `npies_70`): imputación mediante un valor constante (0) y posterior escalado. Se tratan como un bloque numérico específico por su naturaleza y rango de valores.
- **Variables categóricas**: imputación por la moda, conversión explícita a cadena de texto y codificación *one-hot* con `OneHotEncoder`, ignorando categorías desconocidas en el conjunto de prueba.

Todo el preprocesado se integra en un `Pipeline` junto con el modelo de regresión, de manera que la imputación, escalado y codificación se ajustan únicamente sobre el conjunto de entrenamiento y se aplican de forma coherente a validación y prueba.

6.4. Modelos entrenados y ajuste de hiperparámetros

Se entrenó un conjunto de modelos base que cubren distintas familias de algoritmos de regresión supervisada:

- **Modelos basados en árboles y ensambles**:
  - `RandomForestRegressor`

    - `GradientBoostingRegressor` (GBDT)
    - `AdaBoostRegressor`
    - `BaggingRegressor` (Bagged Decision Trees)
    - `XGBRegressor` (`XGBoost`)
    - `LGBMRegressor` (`LightGBM`)
    - `CatBoostRegressor` (`CatBoost`)
- **Modelos basados en instancias**:
    - `KNeighborsRegressor` (KNN)
- **Redes neuronales**:
    - `MLPRegressor` (perceptrón multicapa)
- **Máquinas de soporte vectorial**:
    - `LinearSVR` (SVR con kernel lineal)
- **Modelos probabilísticos / bayesianos**:
    - `BayesianRidge` (Bayesian Neural Network en sentido amplio)

Para cada modelo se definió un espacio de hiperparámetros específico (número de árboles, profundidad máxima, tasa de aprendizaje, tamaño de vecindad, arquitectura de la red, etc.), y se realizó una búsqueda en rejilla mediante `GridSearchCV`, usando como métrica de optimización el coeficiente de determinación R^2. La validación interna se llevó a cabo con el mismo esquema de validación cruzada descrito anteriormente (`GroupKFold` o `KFold`, según disponibilidad de grupos), y se empleó `n_jobs=-1` para explotar el paralelismo multinúcleo.

De cada ajuste se almacenaron:

- El mejor estimador (pipeline completo) según el R^2 medio en validación cruzada.
- Las métricas en el conjunto de prueba: R^2, RMSE, MAE, así como la mediana del error absoluto y la moda del error absoluto redondeado.

6.5. Ensamblado tipo stacking

Con el fin de explotar la posible complementariedad entre modelos, se implementó un esquema de *stacking* manual basado en predicciones *out-of-fold* (OOF). A partir de los mejores modelos ajustados en la fase anterior (`slow_best_models`), se definieron varias configuraciones de modelos base:

- `['CatBoost', 'LightGBM', 'XGBoost', 'Random Forest', 'GBDT', 'BaggedDT']`

- ['CatBoost', 'LightGBM', 'Random Forest', 'GBDT']
- ['LightGBM', 'XGBoost', 'GBDT']
- ['CatBoost', 'Random Forest', 'GBDT']
- ['LightGBM', 'Random Forest']

Para cada configuración:

1. Se generaron predicciones OOF para cada modelo base utilizando el mismo esquema de validación cruzada que en el entrenamiento individual. En cada pliegue, se ajusta un clon del modelo sobre las particiones de entrenamiento y se predice sobre la partición de validación, construyendo así una matriz de meta-características de tamaño $n_{\text{train}} \times M$, donde M es el número de modelos base.
2. Se entrenó de nuevo cada modelo base sobre todo el conjunto de entrenamiento y se obtuvieron sus predicciones sobre el conjunto de prueba, generando una matriz de meta-características de tamaño $n_{\text{test}} \times M$.

Sobre estas meta-características se ajustaron distintos meta-modelos (`meta_models`):

- **Modelos lineales**: `LinearRegression`, `Ridge`.
- **Modelos basados en árboles**: `GradientBoostingRegressor`, `RandomForestRegressor`.
- **Máquina de vectores soporte**: `SVR` con kernel lineal.
- **Red neuronal**: `MLPRegressor` con una capa oculta.

Antes del meta-modelo se incluyó un `StandardScaler` aplicado sobre las meta-características, de forma que se estabiliza el entrenamiento cuando se combinan modelos con escalas de salida diferentes.

Cada combinación *(configuración de bases, meta-modelo)* define un *stack* distinto. Para cada uno de ellos se evaluó el rendimiento en el conjunto de prueba en términos de R^2, RMSE y MAE, seleccionando finalmente el **stack** con mejor R^2 en prueba (y, en caso de empate, menor RMSE) como modelo final de referencia.

6.6. Datos finales de entrenamiento

Tras aplicar los criterios de elegibilidad y filtrado descritos en la Sección 5.4.1, 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

A partir de los estadísticos descriptivos de la Tabla 6.1 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).

Esta diferencia implica que `c4` es una variable 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`.

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.

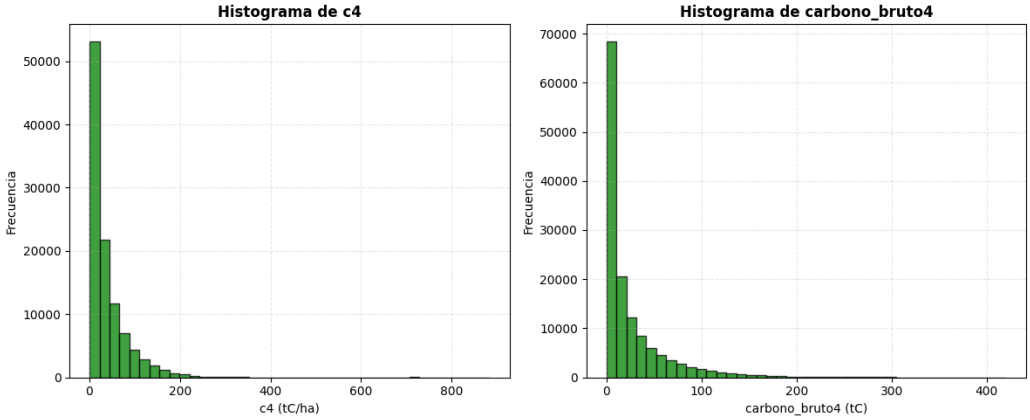

Figura 6.1. Distribución de las variables `c4` y `carbono_bruto4` en el conjunto depurado.

6.6.1. 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.

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. Entre los estimadores evaluados se incluyeron `LightGBM`, `Random Forest`, `XGBoost`, `CatBoost`, `Gradient Boosting`, `Bagging Regressor`, `AdaBoost`, `KNN`, `MLP`, `SVR` y `Bayesian Ridge`. 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.

Finalmente, el modelo seleccionado se **reentrenó con validación cruzada estratificada por grupos**, definidos según el periodo temporal de la

observación. Este esquema de validación cruzada por grupos permitió evaluar la estabilidad del modelo frente a periodos no observados durante el entrenamiento, garantizando así su capacidad de generalización temporal y la fiabilidad de las predicciones en escenarios futuros.

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 también 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.

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.

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.

TODO: pensar en quitar esta tabla

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

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

8.1. Resultados

En esta sección se presentan los resultados de los modelos descritos en la Sección **??**.

8.1.1. Toneladas de carbono por hectárea

La Tabla 8.1 resume las métricas principales para el conjunto de datos que usa el IFN2 y el IFN3 como explicativo y trata de predecir la variable `c4`, esto es, aquella en tC/ha. En la Tabla 8.2 se presentan las métricas para los modelos de *stacking*.

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	0.7872	22.7674	11.6504
XGBoost	0.7837	22.9524	11.5904
CatBoost	0.7830	22.9899	11.6069
GBDT	0.7825	23.0136	11.6580
MLP_Torch	0.7712	23.6065	12.2874
BaggedDT	0.7404	25.1423	13.0214
RandomForest	0.7320	25.5473	12.9079
BayesianNN	0.6776	28.0211	14.6893
SVR	0.5511	33.0652	13.7077

8.1.2. Toneladas de carbono

La Tabla 8.3 resume las métricas principales para el conjunto de datos que usa el IFN2 y el IFN3 como explicativo y trata de predecir la variable `carbono_bruto4`, esto es, el carbono en toneladas absolutas (no normalizado por hectárea). En la Tabla 8.4 se encentra lo propio para los modelos stacking.

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

Tabla 8.3. 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	0.8448	13.8457	6.6148
LightGBM	0.8412	14.0055	6.6539
XGBoost	0.8400	14.0544	6.6545
GBDT	0.8376	14.1591	6.7220
MLP_Torch	0.8318	14.4103	6.9314
BaggedDT	0.8212	14.8581	7.2820
RandomForest	0.8190	14.9498	7.1353
BayesianNN	0.7749	16.6740	8.9064
SVR	0.6794	19.8968	8.1374

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

9. Discusión

Dividiremos las discusión es dos partes muy parecidos, para cada una de las variables objetivo. Luego compararemos brevemente los resultados en conjunto.

9.1. Discusión sobre la variable objetivo c4

Comenzando por los modelos individuales, observamos que el modelo LightGBM es el que presenta un mayor valor de R^2 (0,79) y un menor RMSE (22,77 tC/ha), no así un menor MAE (11,65 tC/ha), caso en el que los modelos XGBoost (11,59 tC/ha) y CatBoost (11,61 tC/ha) mejoran ligeramente. Los resultados de estos modelos son muy similares, logrando captar cerca de un 80 % de la varianilidad de los datos.

Mirar a las estadísticas globales de los resultados nos da una idea de la calidad de los modelos, pero en este caso, donde los datos tienen tanta variabilidad, podemos comprobar que esto no resulta del todo útil. En la Figura 9.1 podemos observar la dispersión de las predicciones respecto a los valores reales, junto con una visualización de la densidad de puntos. Observando la figura nos damos cuenta de varias cosas:

- Los datos abarcan un gran rango de valores, de 0 a 900 tC/ha.
- La gran mayoría de los datos se encuentran en el rango de 0 a 200 tC/ha.
- Podemos ver por los valores de densidad que los valores más abundantes son aquellos cercanos a cero
- Mientras que hay casos en los que los valores predichos difieren notablemente de los reales, la línea de mayor densdad se encuentra siguiendo la línea diagonal, lo que indica que el modelo efectivamente captura la tendencia general de los datos.

Debido a la elevada variabilidad de los datos y a la irregularidad de la distribución respecto a los años (ver Figura 5.3), el modelo no tiene la misma precisión para todos los rangos de valores. En al Figura 9.2 podemos observar el SMAPE para cada uno de los modelos base, junto con un histograma que muestra la distribución del número de valores del conjunto de entrenamiento para cada rango de valores elegidos. Los rangos en los que no hay datos es porque no se disponen de suficientes para hacer el cálculo.

TODO: Meter versión nueva de esta imagen (hay línea que se tapan entre sí)

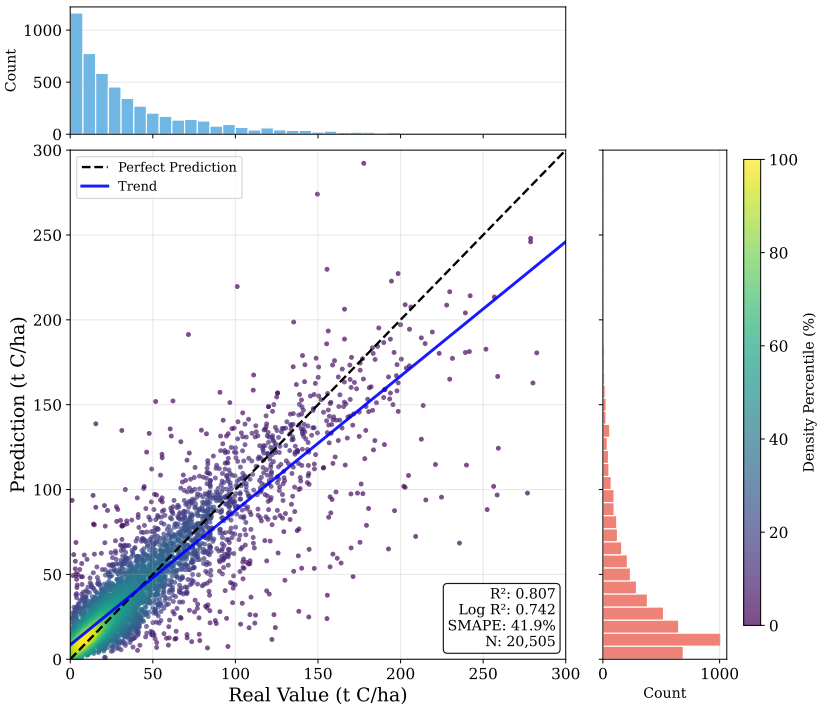

Figura 9.1. Dispersión del modelo LightGBM de las predicciones respecto a los valores reales para la variable objetivo `c4`. Solo se incluye el rango de valores [0−300] tC/ha.

Podemos observar que el SMAPE tiene un valle entre los datos $[20-50]$ y $[200-300]$ tC/ha, donde llega a ser inferior al 30 % para los mejores modelos. Atendiendo a la distribución de los datos de entrenamiento en la Figura 9.2 nos damos cuenta de que para valores pequeños ($[0-20]$ tC/ha) el SMAPE es alto pese a que en ese rango se encuentra prácticamente la mitad de los datos de entrenamiento, lo que indica que el error medio del modelo no es lo suficientemente pequeño como para hacer predicciones confiables para cultivos pequeños. Tras el punto inicial el SMAPE disminuye rápidamente mientras el RMSE aumenta ligeramente. Este comportamiento se mantiene durante la zona valle, lo que indica que, si bien el error absoluto aumenta, no lo hace tanto el error relativo. Esto es un indicativo de que es esta zona el modelo comprende mejor el comportamiento de los datos. El SMAPE comienza a aumentar rápidamente a partir del rango $[150-200]$ tC/ha, valor que coincide con una gran escasez de datos comparado con el resto de

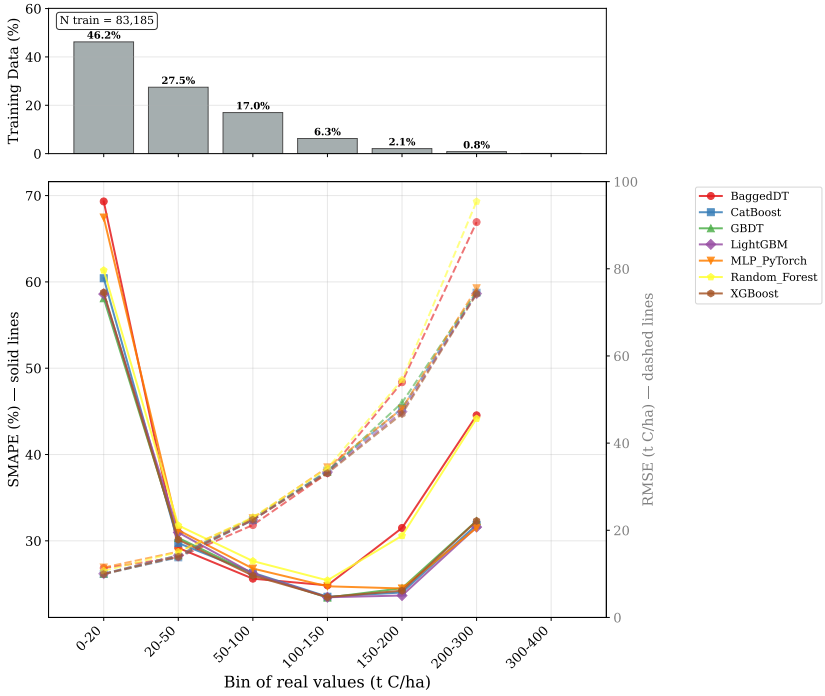

Figura 9.2. SMAPE y RMSE de los modelos base para la variable objetivo `c4`.

rangos.

Para comprobar que los modelos están aprendiendo la lógica del crecimiento de los árboles (cuanto más tiempo, más carbono absobido) podemos hacer la simulación de predecir el carbono para un mismo cultivo a los largo de todos los años disponibles. Esta prueba se ha realizado en la Figura 9.3 para el caso del modelo LightGBM. Se encogieron las cinco especies más comunes del dataset junto con sus valores más comunes de los datos de entrada (la media para los parámetros numéricos, la moda para los categóricos) y se realizó la predicción cambiando únicamente el año de predicción. Observamos que el carbono aumenta a medida que lo hace el tiempo, lo que indica que la lógica del modelo es la esperada.

Los resultados recogidos en la Tabla 8.2 muestran que el *stacking* ofrece un rendimiento comparable al de los mejores modelos individuales basados en árboles y *gradient boosting*. En concreto, mientras que CatBoost obtiene un R^2 de 0.78 y un RMSE de 22.99 tC/ha, las configuraciones de *stacking* alcanzan valores similares, con R^2 en torno a 0.79 y RMSE cercano a 22.39 tC/ha en el mejor caso (`stack5_MLP`).

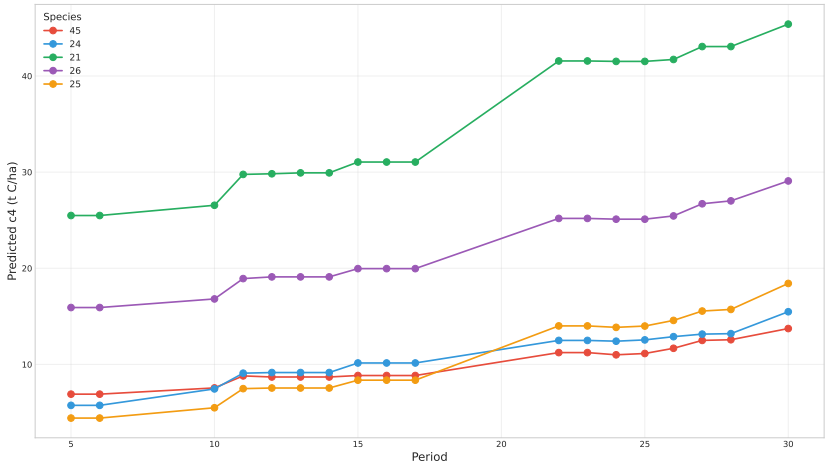

Figura 9.3. Valor de carbono en función del tiempo para los cinco cultivos más comunes del dataset para el modelo LightGBM.

En general, la agregación de modelos mediante *GradientBoosting* como meta-modelo no produce mejora en los resultados, lo que sugiere que la mezcla de arquitecturas distintas es más beneficiosa.

El rendimiento del *stacking* depende de manera importante del meta-modelo empleado. En primer lugar, los modelos lineales (Regresión Lineal y Ridge) ofrecen, de forma sistemática, un rendimiento sólido y muy estable en todas las configuraciones, situándose casi siempre entre las mejores alternativas dentro de cada grupo de `stack_configs`. Esto sugiere que, dado el reducido número de meta-predictores (salidas de los modelos base), una combinación esencialmente lineal es suficiente para explotar gran parte de la información disponible sin incurrir en sobreajuste.

En contraste, los meta-modelos basados en *Random Forest* muestran de manera consistente los peores resultados dentro de cada configuración. Este comportamiento indica que, sobre un espacio de baja dimensión, una capacidad excesiva de modelado no aporta beneficios y tiende más bien a ajustar ruido en las predicciones de los modelos base.

Los meta-modelos no lineales más sencillos, como MLP y SVR, proporcionan mejoras puntuales sobre los lineales. Destaca especialmente la configuración `stack5_MLP`, que alcanza el valor más alto de R^2 (0.79) junto con el menor RMSE (22.39 tC/ha) de todas las combinaciones evaluadas. En conjunto, estos resultados indican que existe una ligera ganancia al introducir cierta no linealidad en la combinación de las predicciones base, pero que di-

cha ganancia se produce sólo cuando el modelo de segundo nivel mantiene una complejidad moderada y bien regularizada.

9.2. Discusión sobre la variable objetivo `carbono_bruto_4`

Para el caso de los modelos individuales, para la variable `carbono_bruto_4` el modelo CatBoost con un $R^2 = 0{,}8448$ supera ligeramente a LightGBM, que era el que mejor R^2 obtenía para la variable `c4`. Así mismo, el modelo CatBoost también es el que presenta un menor RMSE (13,85 tC) y un menor MAE (6,6148 tC). Si hacemos la comparación de las métricas entre los modelos que predices `c4` y `carbono_bruto_4` observamos que los R^2 mejoran de forma sistemática para la variable `carbono_bruto_4`, lo que sugiere que es más sencillo predecir la variable objetivo en toneladas de carbono.

Podemos observar los valores predichos contra los reales en la figura 9.4. Observamos un caso muy similar a aquel observado para la variable `c4`, con un rango de valores posibles muy grande (de 0 a 300 tC) situandose la gran mayoría en el rango entre 0 y 100 tC. También ocurre que la mayor densidad de puntos se encuentra en valores pequeños, además de que la línea de mayor densidad se encuentra cercana a la línea de predicción correcta, lo que indica que el modelo capta de forma correcta la tendencia de los datos.

De igual forma, en la Figura 9.5 se muestran los valores de SMAPE y RMSE de los modelos individuales para distintos rangos de valores de la variable objetico `carbono_bruto_4`. De nuevo, el comportamiento es similar al observado para la variable `c4`, pudiendo observar un valle en el SMAPE entre los rangos $[20 - 50]$ tC y $[200 - 300]$ tC, mientras que en el RMSE aumenta a medida que el valor de la variable objetivo aumenta. Esto sugiere que las mejores predicciones son aquellas dentro del rango de valores del valle, donde el error porcentual es menor (llegando a cerca del 25 % de error para el rango $[100 - 150]$ tC). Un mayor valor del error en valores extremos de la variable objetico sugiere que los modelos no pueden alcanzar la suficiente precisión como para ser útil en valores de carbono pequeños (donde un error "pequeño" penaliza en mayor medida que para valores de carbono altos), bien porque los modelos no son suficeintemente bueno o bien por la elevada variabilidad de los datos. Para valores de carbono altos el error elevado se debe muy posiblemente a la poca cantidad de datos disponibles.

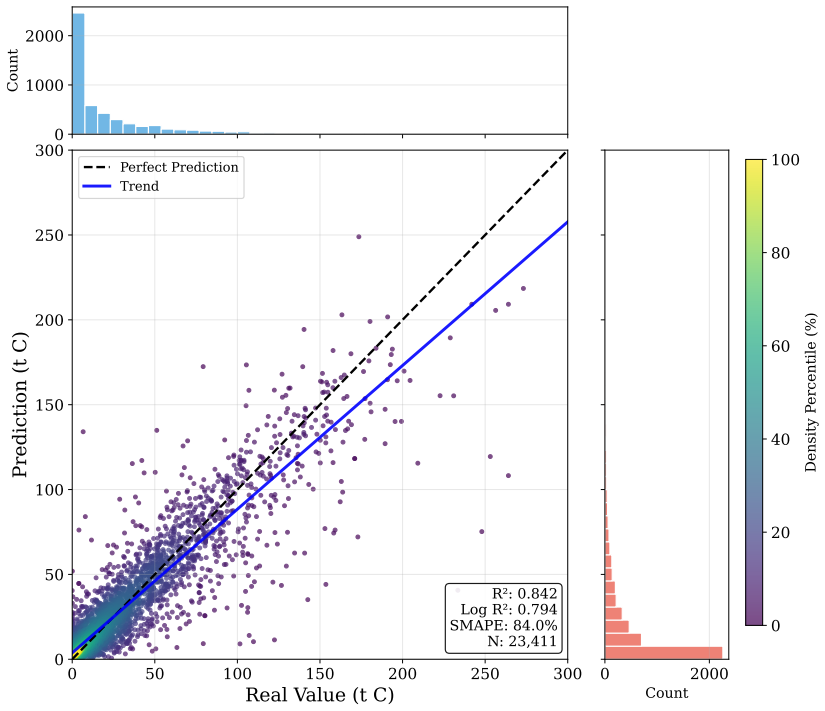

Figura 9.4. Densidad de la variable `carbono_bruto_4` para el modelo CatBoost. Solo se muestran los valores entre 0 y 300 tC.

9.3. 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 automá-

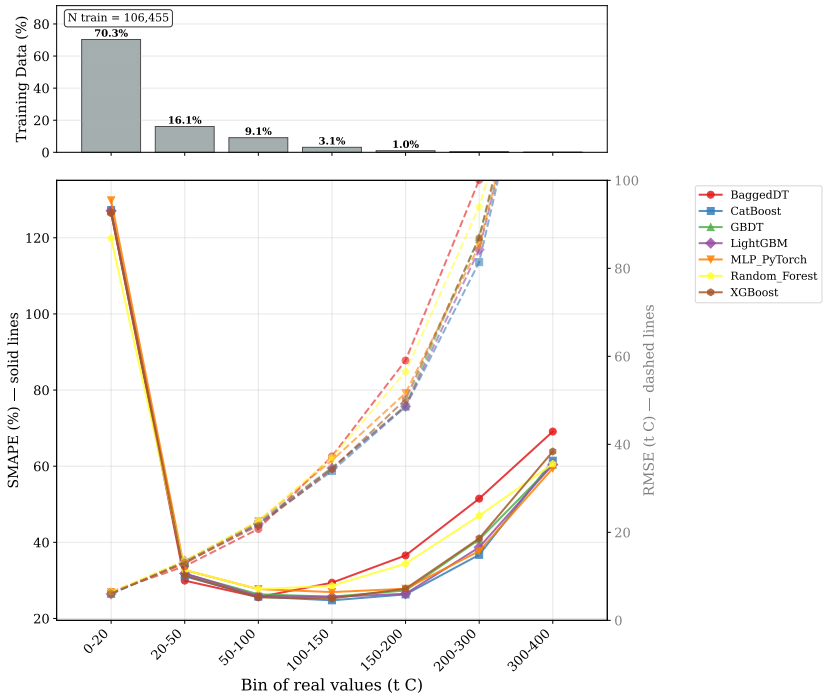

Figura 9.5. SMAPE y RMSE de los modelos base para la variable `carbono_bruto_4`.

ticas.

- 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 superiores a 0.79 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* mejoran de forma consistente el rendimiento de los modelos individuales, alcanzando la mejor configuración (el metamodelo SVR con los modelos `CatBoost, Random Forest, GBDT`) un R^2 de 0.86 y reduciendo el RMSE y el MAE frente a CatBoost en un 20,26 % y en un 23,16 % respectivamente.
- El rendimiento del ensamble 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, destacando SVR en las configuraciones `stack3` y `stack4`.
- En conjunto, los resultados muestran que la combinación de modelos mediante *stacking*, aplicada con meta-modelos bien regularizados, permite aprovechar la complementariedad entre los distintos algoritmos y alcanzar una capacidad predictiva superior a la de cualquier modelo individual.
- 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 por hectárea en un horizonte temporal de entre 4 y 35 años con un nivel elevado de precisión. El mejor modelo obtenido se construye mediante un metamodelo `SVR` combinando los modelos `CatBoost, Random Forest` y `GBDT` y alcanza un coeficiente de determinación de $R^2 = 0{,}8604$, junto con un error típico de **RMSE = 17.22 tC/ha** y un error medio absoluto de **MAE = 8.78 tC/ha**.

10. Conclusiones

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 [4]), datos meteorológicos [5] e imágenes satelitales [13] 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 casos: en unidades de toneladas de carbono por hectárea (tC/ha) y en unidades de toneladas de carbono (tC).

11. Recomendaciones para Futuras Investigaciones

Sugerir áreas que podrían beneficiarse de estudios adicionales o mejoras en la metodología.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

Apéndice A. Anexos

Apéndice A.1. Anexo: 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* [**montero2009**, 17]. 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 [**miteco_ifn4_manual**]:

- 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 [**ipcc2006**], 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* [**ipcc2006**], 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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.
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.

Continúa en la siguiente página

GRUPO	MOD.	DESCRIPCIÓN DEL MODELO
	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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: 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. Anexo: Resultados

Apéndice A.17.1. Ifn2 e Ifn3 como explicativo para `carbono_bruto4` (tC)

Tabla A.3. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en tC con el conjunto de datos que emplea IFN2 e IFN3 como explicativos.

Modelo	CV R^2	Test R^2	Test RMSE (tC)	Test MAE (tC)
CatBoost	0.8405	0.8454	13.7701	6.7100
LightGBM	0.8399	0.8418	13.9258	6.6542
XGBoost	0.8394	0.8403	13.9937	6.6695
GBDT	0.8372	0.8372	14.1270	6.6973
MLP	0.8294	0.8343	14.2546	7.0064
BaggedDT	0.8184	0.8218	14.7801	7.2181
Random Forest	0.8125	0.8163	15.0091	7.1317
SVR	0.7987	0.7988	15.7059	6.6313
BayesianNN	0.7703	0.7701	16.7897	8.8954
KNN	0.7380	0.7415	17.8022	8.1109
AdaBoost	0.4814	0.4802	25.2456	20.8181

Se mantienen las conclusiones extraidas en el análisis de resultados realizado para el modelo entrenado con los mismos datos pero variable objetivo `c4` (tC/ha), esto es:

- Los modelos presentan una buena capacidad de generalización.
- Los modelos basados en árboles de decisión y en *gradient boosting* son los que ofrecen, en general, el mejor equilibrio entre capacidad predictiva y estabilidad.
- Algoritmos como AdaBoost o KNN muestran un rendimiento claramente inferior

En particular, **CatBoost** destaca como el modelo con mejor rendimiento global, alcanzando un $R^2 = 0{,}8454$ y un RMSE de 13,77 tC/ha. Estos valores implican que el modelo es capaz de explicar una proporción sustancial de la variabilidad del carbono en las parcelas, reduciendo el error típico de predicción a menos de la mitad de la variabilidad natural de la variable (SD $\approx$ 36 tC). Esto indica que, dentro de la complejidad inherente al problema, Cat-

Boost logra capturar de manera más eficaz las relaciones no lineales presentes en los datos.

Apéndice A.17.2. Resultados del stacking frente a los modelos individuales

Los resultados recogidos en la Tabla A.4 muestran que el *stacking* no alcanza resultados tan satisfactorios como en 8.2. Esta técnica permite mejorar ligeramente el rendimiento respecto a los mejores modelos individuales basados en árboles y *gradient boosting*. En concreto, mientras que CatBoost obtiene un R^2 de 0.8454 y un RMSE de 13.77 tC, *stacking* alcanza R^2 en torno a 0.8484 y reduce el RMSE hasta 13.6336 tC.

Se observa aun patrón claro: en todas las configuraciones, los mejores resultados se obtienen cuando el meta-modelo es una red neuronal MLP (`stack3__MLP`, `stack4__MLP`, `stack5__MLP`), seguido muy de cerca por los meta-modelos basados en *gradient boosting*.

Tomamos como mejor combinación el modelo `stack5__MLP`, esto es, la combinación `MLP` de los modelos LightGB y Random Forest. Aunque la mejora en términos de R2 de este modelo respecto de CatBoost es ligera (Δ0,0003), la variación el MAE alcanza Δ0,0667 unidades, lo cual se traduce en una diferencia de 67kg de error medio. Ofrece un buen equilibrio entre una mejora mediocre y un aumento ligero de la complejidad del problema.

Apéndice A.17.3. Ifn2 como explicativo para `c4` (tC/ha)

Se mantienen las conclusiones extraidas en el análisis de resultados realizado para el modelo entrenado con la misma variable objetivo pero los datos del IFN2 e IFN3 como explicativos (A.3):

- Los modelos presentan una buena capacidad de generalización.
- Los modelos basados en árboles de decisión y en *gradient boosting* son los que ofrecen, en general, el mejor equilibrio entre capacidad predictiva y estabilidad.
- Algoritmos como AdaBoost o KNN muestran un rendimiento claramente inferior

En particular, **CatBoost** destaca como el modelo con mejor rendimiento global, alcanzando un $R^2 = 0{,}8614$ y un RMSE de 17,1604 tC/ha. Estos valores implican que el modelo es capaz de explicar una proporción sustancial de la variabilidad del carbono en las parcelas, reduciendo el error típico de predicción a menos de la mitad de la variabilidad natural de la variable (SD $\approx$ 47 tC/ha).

Stack	**Bases**	**Test R^2**	**RMSE**	**MAE**
stack1__GradientBoosting	6	0.8423	13.9071	6.5327
stack1__LinearRegression	6	0.8424	13.9005	6.6250
stack1__Ridge	6	0.8424	13.9005	6.6250
stack1__RandomForest	6	0.8113	15.2118	7.2916
stack1__SVR	6	0.8386	14.0695	6.5204
stack1__MLP	6	0.8432	13.8638	6.4806
stack2__GradientBoosting	4	0.8435	13.8538	6.5280
stack2__LinearRegression	4	0.8439	13.8326	6.5844
stack2__Ridge	4	0.8439	13.8326	6.5844
stack2__RandomForest	4	0.8267	14.5758	7.0146
stack2__SVR	4	0.8405	13.9833	6.4897
stack2__MLP	4	0.8442	13.8206	6.5289
stack3__GradientBoosting	3	0.8479	13.6581	6.4281
stack3__LinearRegression	3	0.8463	13.7272	6.5942
stack3__Ridge	3	0.8463	13.7272	6.5942
stack3__RandomForest	3	0.8301	14.4327	6.9187
stack3__SVR	3	0.8417	13.9308	6.4787
stack3__MLP	3	0.8481	13.6451	6.3674
stack4__GradientBoosting	3	0.8481	13.6470	6.4214
stack4__LinearRegression	3	0.8471	13.6915	6.5631
stack4__Ridge	3	0.8471	13.6915	6.5630
stack4__RandomForest	3	0.8345	14.2450	6.8175
stack4__SVR	3	0.8428	13.8817	6.4555
stack4__MLP	**3**	**0.8484**	**13.6336**	**6.4327**
stack5__GradientBoosting	2	0.8479	13.6540	6.4258
stack5__LinearRegression	2	0.8471	13.6899	6.5610
stack5__Ridge	2	0.8471	13.6899	6.5610
stack5__RandomForest	2	0.8382	14.0828	6.7136
stack5__SVR	2	0.8428	13.8832	6.4518
stack5__MLP	**2**	**0.8482**	**13.6420**	**6.5120**

Tabla A.4. Resultados de las diferentes configuraciones de stacking utilizando IFN2 e IFN3 como explicativos de la variable en tC.

Tabla A.5. 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 como explicativo.

Modelo	CV R^2	Test R^2	Test RMSE (tC/ha)	Test MAE (tC/ha)
Random Forest	0.8081	0.8268	19.1845	10.1126
XGBoost	0.8400	0.8581	17.3623	9.0307
CatBoost	0.8410	0.8614	17.1604	9.0277
LightGBM	0.8406	0.8563	17.4713	8.9638
GBDT	0.8368	0.8598	17.2612	9.2273
BaggedDT	0.8105	0.8293	19.0472	10.0722
AdaBoost	0.4580	0.4974	32.6789	25.2755
KNN	0.7433	0.7567	22.7382	11.9747
MLP	0.8158	0.8266	19.1954	10.7453
SVR	0.7318	0.7374	23.6220	10.9064
BayesianNN	0.7672	0.7766	21.7868	11.9026

En este caso (Tabla [**tab:stack_ifn2carb**]) la técnica de *stacking* no ofrece mejoras que compensen el incremento en la complejidad del modelo. Destaca el uso de MLP como metamodelo.

Apéndice A.17.4. Ifn2 como explicativo para `carbono_bruto4` *(tC)*

Una vez más, se mantienen las conclusiones extraidas en el análisis de resultados realizado para el modelo entrenado con la misma variable objetivo pero los datos del IFN2 e IFN3 como explicativos (A.7):

- Los modelos presentan una buena capacidad de generalización.
- Los modelos basados en árboles de decisión y en *gradient boosting* son los que ofrecen, en general, el mejor equilibrio entre capacidad predictiva y estabilidad.
- Algoritmos como AdaBoost o KNN muestran un rendimiento claramente inferior

En particular, **CatBoost** destaca como el modelo con mejor rendimiento global, alcanzando un $R^2 = 0{,}8976$ y un RMSE de 11,1825 tC. Estos valores implican que el modelo es capaz de explicar una proporción sustancial de la variabilidad del carbono en las parcelas, reduciendo el error típico de predic-

Stack	**Bases**	**Test R^2**	**RMSE**	**MAE**
stack1__GradientBoosting	6	0.8649	16.9454	8.7640
stack1__LinearRegression	6	0.8650	16.9342	8.7602
stack1__Ridge	6	0.8650	16.9342	8.7602
stack1__RandomForest	6	0.8633	17.0456	9.1193
stack1__SVR	6	0.8608	17.1971	8.5929
stack1__MLP	6	**0.8673**	**16.7925**	**8.6644**
stack2__GradientBoosting	4	0.8641	16.9945	8.7752
stack2__LinearRegression	4	0.8653	16.9197	8.7680
stack2__Ridge	4	0.8653	16.9196	8.7680
stack2__RandomForest	4	0.8573	17.4123	9.1919
stack2__SVR	4	0.8614	17.1598	8.6112
stack2__MLP	4	**0.8675**	**16.7764**	**8.7338**
stack3__GradientBoosting	3	0.8594	17.2836	8.8767
stack3__LinearRegression	3	0.8625	17.0935	8.8214
stack3__Ridge	3	0.8625	17.0935	8.8214
stack3__RandomForest	3	0.8527	17.6940	9.5033
stack3__SVR	3	0.8591	17.2998	8.6736
stack3__MLP	3	**0.8639**	17.0060	8.8602
stack4__GradientBoosting	3	0.8646	16.9618	8.9358
stack4__LinearRegression	3	0.8645	16.9659	8.9309
stack4__Ridge	3	0.8645	16.9660	8.9309
stack4__RandomForest	3	0.8495	17.8810	9.6129
stack4__SVR	3	0.8604	17.2203	8.7753
stack4__MLP	3	**0.8668**	16.8260	8.8613
stack5__GradientBoosting	2	0.8534	17.6522	8.9497
stack5__LinearRegression	2	0.8569	17.4362	8.9133
stack5__Ridge	2	0.8569	17.4362	8.9133
stack5__RandomForest	2	0.8233	19.3788	10.1032
stack5__SVR	2	0.8545	17.5846	8.7686
stack5__MLP	2	**0.8561**	17.4875	8.8770

Tabla A.6. Resultados de las diferentes configuraciones de stacking utilizando IFN2 como conjunto explicativo de la variable en tC/ha.

Tabla A.7. Resumen del rendimiento de los modelos para la predicción de la variable de carbono en tC con el conjunto de datos que emplea IFN2 como explicativo.

Modelo	CV R^2	Test R^2	Test RMSE (tC)	MAE (tC)
Random Forest	0.8587	0.8607	13.0457	6.4514
XGBoost	0.8966	0.8957	11.2890	5.5822
CatBoost	**0.8974**	**0.8976**	**11.1825**	**5.5840**
LightGBM	0.8974	0.9007	11.0128	5.4472
GBDT	0.8940	0.8912	11.5309	5.7745
BaggedDT	0.8700	0.8721	12.5015	6.3180
AdaBoost	0.5702	0.5765	22.7439	19.2913
KNN	0.7803	0.7892	16.0461	7.9430
MLP	0.8860	0.8890	11.6427	6.3052
SVR	0.8155	0.8159	14.9943	7.1427
BayesianNN	0.8348	0.8344	14.2244	7.8642

ción a menos de la mitad de la variabilidad natural de la variable (SD $\approx$ 36 tC).

En este caso (Tabla [**tab:stack_ifn2_tc**]) la técnica de *stacking* no ofrece mejoras que compensen el incremento en la complejidad del modelo, aunque es cierto que se rompe la barrera del $R^2 > 0{,}9$. Destaca el uso de MLP como metamodelo.

De entre los modelos entrenados para predecir `carbono_bruto4` (tC) con IFN3 como explicativo destaca aquel entrenado con MLP como metamodelo para combinar `CatBoost, LightGBM, Random Forest` y `GBDT`, con un $R^2 =$ 0,9039, un $RMSE = 10{,}8387$ y un $MAE = 5{,}2856$.

Apéndice A.18. Anexo: Código de las especies (`especie_id`)

Tabla A.9. 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

Continúa en la siguiente página

Tabla A.9. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
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
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

Continúa en la siguiente página

Tabla A.9. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
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
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 gomphocephalus	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

Continúa en la siguiente página

Tabla A.9. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
69	Phoenix spp.	-	1	69
273	Betula alba	Betula verrucosa | Abedul pubescens	1	73
373	Betula pendula	Betula hispanica | Abedul péndula	1	73
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

Continúa en la siguiente página

Tabla A.9. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
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
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

Continúa en la siguiente página

Tabla A.9. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
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
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	-

Continúa en la siguiente página

Tabla A.9. Relación de especies (continuación).

Cód.	Nombre	Sinonimia	Tipo	Grupo
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	-
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	-

Stack	**Bases**	**Test R^2**	**RMSE**	**MAE**
stack1__GradientBoosting	6	0.9013	10.9803	5.2917
stack1__LinearRegression	6	0.9028	10.8960	5.3982
stack1__Ridge	6	0.9028	10.8962	5.3982
stack1__RandomForest	6	0.8946	11.3469	5.5493
stack1__SVR	6	0.9017	10.9600	5.2741
stack1__MLP	6	**0.9043**	**10.8148**	**5.2523**
stack2__GradientBoosting	4	0.9016	10.9648	5.2924
stack2__LinearRegression	4	0.9027	10.9000	5.4073
stack2__Ridge	4	0.9027	10.9002	5.4074
stack2__RandomForest	4	0.8938	11.3894	5.6335
stack2__SVR	4	0.9016	10.9633	5.2807
stack2__MLP	4	**0.9039**	**10.8337**	5.2856
stack3__GradientBoosting	3	0.8989	11.1106	5.4026
stack3__LinearRegression	3	0.9006	11.0195	5.4354
stack3__Ridge	3	0.9006	11.0196	5.4354
stack3__RandomForest	3	0.8868	11.7604	5.8714
stack3__SVR	3	0.8996	11.0728	5.3318
stack3__MLP	3	**0.9002**	**11.0388**	5.4138
stack4__GradientBoosting	3	0.8965	11.2463	5.4273
stack4__LinearRegression	3	0.8979	11.1665	5.5801
stack4__Ridge	3	0.8979	11.1665	5.5801
stack4__RandomForest	3	0.8861	11.7940	5.8454
stack4__SVR	3	0.8960	11.2707	5.4679
stack4__MLP	3	**0.8993**	**11.0889**	5.4012
stack5__GradientBoosting	2	0.8995	11.0811	5.3741
stack5__LinearRegression	2	0.9007	11.0163	5.4479
stack5__Ridge	2	0.9006	11.0164	5.4479
stack5__RandomForest	2	0.8813	12.0431	6.0737
stack5__SVR	2	0.8988	11.1210	5.3609
stack5__MLP	2	**0.9011**	**10.9941**	**5.3603**

Tabla A.8. Resultados de las diferentes configuraciones de stacking utilizando IFN2 como conjunto explicativo de la variable en tC.