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

Correspondiente al grupo B1 (GreenWest) de las líneas de investigación de DemIA

Maider Araceli Urbón Jiménez[*1], Jaime Gabriel Vegas[1], Ana de Luis Reboredo[1], Belén Pérez Lancho[1] y Ana-Belén Gil-González[1]

[1]Grupo B1, Equipo de investigación BISITE, Universidad de Salamanca, Facultad de Ciencias, Salamanca, Castilla y León, España,

{murbon001, JaimeGabrielVegas, adeluis, lancho, abg}@usal.es

12 de noviembre de 2025

Resumen

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 descrita en un trabajo complementario [**?**], 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 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 [**? ?**].

Palabras clave: créditos de carbono, inteligencia artificial, forestación, reforestación, modelado predictivo, cambio climático.

[*]Autora de correspondencia: murbon001@usal.es

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 [**?**]. 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* [**? ?**] 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. Esto significa que los cultivos forestales naturales no son elegibles para la contabilización de carbono.
- **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. 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) [**?**], reanálisis **ERA5-Land** [**?**] e **índices espectrales Landsat** (Collection 2, L2) [**?**] en una base de datos relacional jerárquica descrita en un trabajo complementario [**?**].

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.

Contribución. GreenWest (i) modela la captura futura de carbono a resolución parcela–especie–clase diamétrica, (ii) demuestra *generalización temporal* mediante validación cruzada por grupos de periodo, y (iii) habilita su uso en marcos de optimización (selección de especie y emplazamiento) manteniendo compatibilidad con los requisitos de elegibilidad de créditos [**?**]. Estos resultados facilitan la planificación de proyectos de (re)forestación y la toma de decisiones para una gestión forestal alineada con una economía baja en carbono.

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.

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 forestación y reforestación.
- **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.

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 en el marco de la CMNUCC y el Protocolo de Kioto. Según estos acuerdos, cada país debe reportar sus emisiones y absorciones de gases de efecto invernadero, y puede utilizar actividades de forestación y reforestación 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 car-

bono 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 es la estimación del carbono almacenado en un momento dado a partir de datos de teledetección. Goetz et al. (2009) [**?**] revisan el uso de imágenes satelitales (MODIS, Landsat) en modelos empíricos de biomasa aérea, destacando su eficacia a escala regional en zonas boreales. Este tipo de estimaciones suele realizarse mediante regresiones lineales o algoritmos de mínimos cuadrados generalizados, con coeficientes de determinación (R^2) típicamente entre 0.6 y 0.8 según la resolución de entrada 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) [**?**] 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, Yang et al. (2023) [**?**] 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) [**?**] o Dong et al. (2023) [**?**], 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 [**?**], 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 [**? ? ?**]. Los detalles sobre la arquitectura del modelo, las variables utilizadas, los algoritmos implementados y las métricas de evaluación se desarrollan en la siguiente sección.

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 pue-

de 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 [**? ?**].

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 precipitaciones, 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.

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

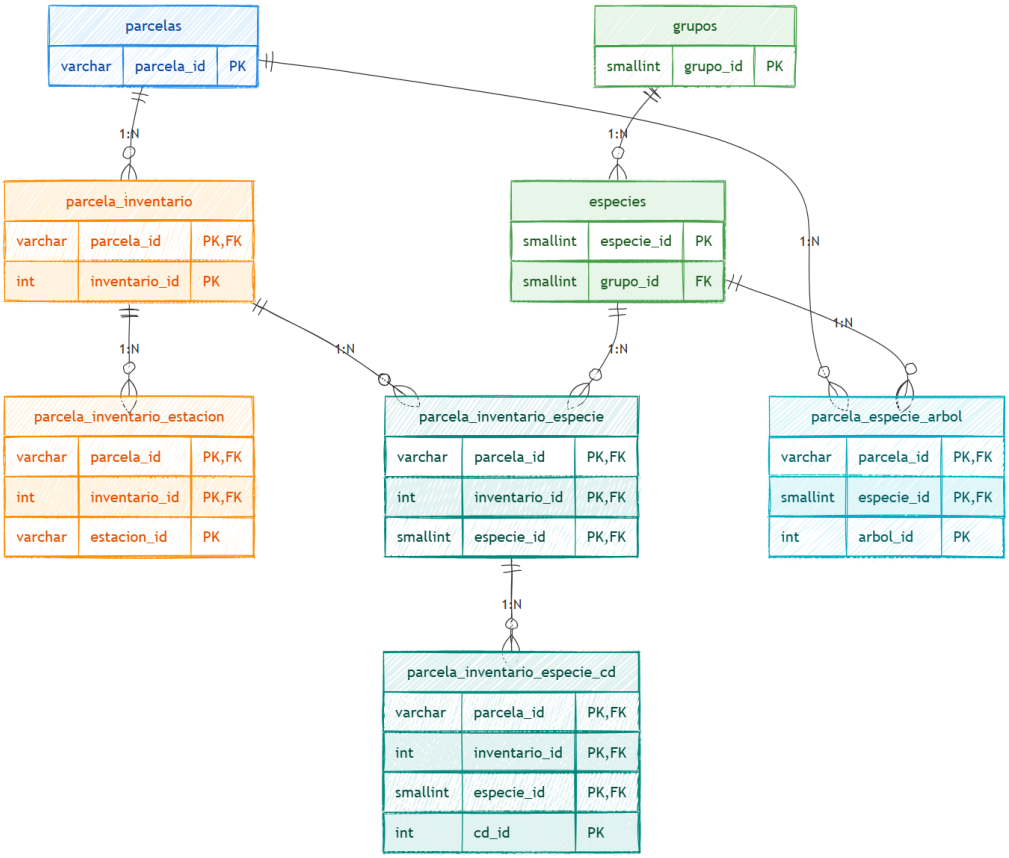

Figura 5.1: Esquema relacional de las tablas principales de la base de datos. Tabla extraida de [**?**], donde se pueden consultar más detalles sobre las variables.

Diccionario resumido de variables

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

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
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

Continúa en la siguiente página

Variable	Descripción	Unidad	Tipo de dato
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
grupo_id	Grupo funcional → grupos.	–	Identificador
grupos.nombregrupo	Nombre del grupo.	–	Texto

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 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 [**?**] y las directrices del IPCC [**?**], 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 [**? ?**]. 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 fo-

restales 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 **??** muestra la distribución de la diferencia de años entre las mediciones del IFN3 y el IFN4. Como puede observarse, la mayoría de las parcelas presentan intervalos comprendidos entre 6 y 17 años, un rango demasiado estrecho para evaluar la estabilidad del modelo en horizontes más amplios.

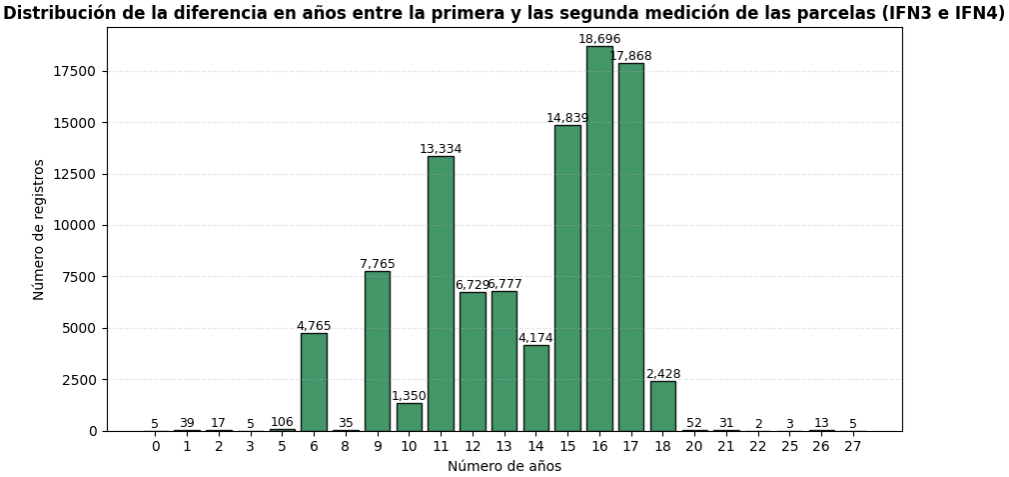

Figura 5.2: Distribución de la diferencia de años entre los inventarios IFN3 e IFN4.

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.

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

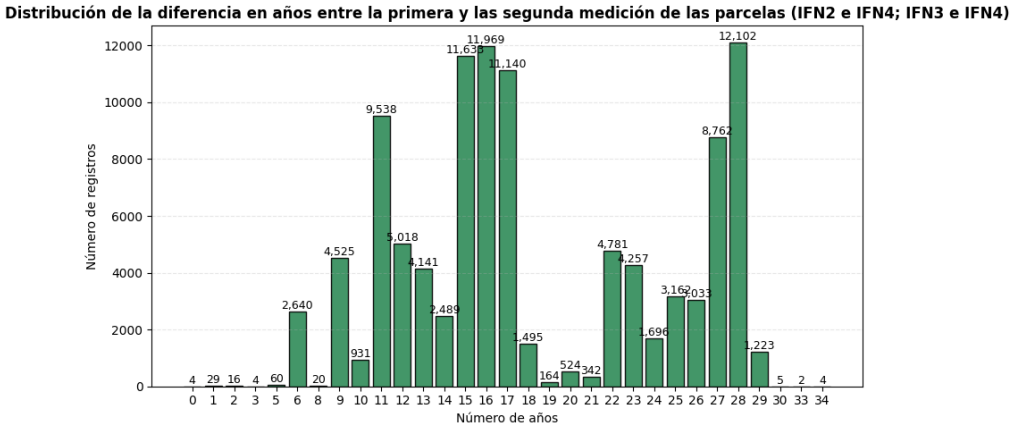

Figura 5.3: Distribución de la diferencia de años entre los inventarios IFN2–IFN3 e IFN3–IFN4.

- **Superficie mínima de 1 ha.** Este criterio se considera *externo* al alcance del modelo predictivo, ya que el aprendizaje se realiza a nivel de parcela e inventario y no sobre polígonos de superficie total. En la práctica, la verificación de la superficie se realiza *ex ante*, sobre la geometría declarada del proyecto forestal. En los terrenos forestales generados a partir de intervención humana directa —como plantaciones o repoblaciones—, la extensión suele presentar una estructura homogénea, con una especie dominante, edades coetáneas y densidades estandarizadas. Bajo estas condiciones, el carbono total es proporcional a la superficie: duplicar el área de una masa forestal homogénea implica aproximadamente duplicar su carbono almacenado. Por tanto, la variable de superficie no afecta al ajuste interno del modelo y su cumplimiento puede evaluarse fácilmente a nivel de proyecto, sin comprometer la validez de las predicciones.
- **Fracción mínima de cabida cubierta del 20 %.** La base de datos dispone de `fccarb` (arbórea) y `fcctot` (total). Este umbral se aplica como *filtro de elegibilidad* previo o posterior al modelado, sin modificar la arquitectura del modelo (`fccarb`> 20).
- **Altura mínima de 3 m en la madurez.** Este requisito se refiere a la altura que alcanzan los árboles en su fase de pleno desarrollo, y no a la altura inicial de los plantones. Por tanto, las mediciones realizadas durante las etapas tempranas de crecimiento no determinan la elegibilidad del proyecto, siempre que las especies seleccionadas sean capaces de superar los 3 metros en la madurez. 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**).

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`.

Los conteos de observaciones por inventario y condición se resumen a continuación:

- **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**

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

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 (código), tipo de especie y grupo taxonómico.	Anexos **??** y **??**
ocupa	int	Grado de ocupación/presencia de la especie en la parcela (0–10).	–
estado_id, fpmasa_id, tratmasa_id, orgmasa_1_id	int (CF)	Estado/fase de desarrollo, forma principal de masa, tratamiento de masa, organización de masa.	Anexos **??**, **??**, **??** y **??**
tipsuelo1_id, tipsuelo2_id, tipsuelo3_id	int (CF)	Tipos de suelo de la parcela (niveles jerárquicos).	Anexo **??**
rocosidad_id, textura_id, matorg_id, modcomb_id, disesp_id, comesp_id, merosiva_id	int (CF)	Rocosidad, textura, materia orgánica, modelo de combustibilidad, distribución/composición específica, manifestaciones erosivas.	Anexos **??**, **??**, **??**, **??**, **??**, **??** y **??**.
radio	float	Radio de la parcela (m).	–
orientacion, elevacion, pendiente	float	Orientación (grados), elevación (m s.n.m.), pendiente (%).	–
nivel1_id, nivel2_id	int (CF)	Niveles jerárquicos/estratos de inventario.	Anexos **??** y **??**.
fccarb, fcctot	float	Fracción de cabida cubierta arbórea y total.	–
npies_{1,2,5,10,15, 20,25,30,35,40,45, 50,55,60,65,70}	float	Número de pies por clase diamétrica CD (cm). Cada campo corresponde a la CD indicada.	–
periodo	int	Años transcurridos entre inventarios considerados en el modelo.	–
evi_{stat}_{est}	float	Índice EVI por estación; stat ∈ {max, mean, median, min, std}, est ∈ {invierno, otoño, primavera, verano}.	–
gndvi_{stat}_{est}	float	Índice GNDVI por estación (misma convención de stat y est).	–
ndii_{stat}_{est}	float	Índice NDII por estación (misma convención).	–
ndvi_{stat}_{est}	float	Índice NDVI por estación (misma convención).	–
pr_{stat}_{est}	float	Precipitación: stat ∈ {max, mean, min, std, sum} por estación est.	–
skt_{stat}_{est}	float	Temperatura de superficie (skin temperature): stat ∈ {max, mean, min, std}.	–
stl1-4_{stat}_{est}	float	Temperatura de suelo por capa (1–4): stat ∈ {max, mean, min, std}.	–
t2m_{stat}_{est}	float	Temperatura del aire a 2 m: stat ∈ {max, mean, min, std}.	–
c4	float	Carbono capturado en el cultivo en el IFN4 en t/ha.	–
carbono_bruto4	float	Carbono capturado en el cultivo en el IFN4 en t.	–

Tabla 5.2: Resumen de variables de entrada del modelo. Para variables estacionales se usa la notación variable_{stat}_{est}, con estadísticas stat y estaciones est en {invierno, otoño, primavera, verano}. Las variables npies_{CD} se repiten para cada clase diamétrica indicada. Las variables destacadas en verde se encuentran recogidas tanto para el IFN2 como para el IFN3, las no destacadas solo se recogen para el IFN3.

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.

6. SELECCIÓN DE VARIABLES EXPLICATIVAS

La selección de variables explicativas se abordó mediante tres estrategias complementarias: (1) selección automática basada en el algoritmo *Featurewiz*, (2) selección mediante el criterio de relevancia y mínima redundancia (*mRMR*), y (3) una selección manual fundamentada en criterios estadísticos, ecológicos y de interpretabilidad. El objetivo común fue identificar un subconjunto óptimo de predictores que maximice la capacidad explicativa del modelo sobre la variable dependiente $c4$ (carbono estimado), evitando colinealidad y preservando el sentido físico de las relaciones.

6.1. Selección automática mediante Featurewiz

El método *Featurewiz* se basa en un enfoque de selección de características guiado por importancia predictiva. El procedimiento combina dos etapas principales: (i) un filtrado inicial por correlación, en el que se eliminan variables altamente colineales (en este caso, con un umbral de $|r| > 0{,}70$); y (ii) un refinamiento mediante modelos de *Gradient Boosting* que estiman la importancia relativa de cada variable en la predicción del objetivo. De esta manera, *Featurewiz* conserva únicamente aquellas variables con una contribución significativa a la mejora del rendimiento predictivo, proporcionando un conjunto compacto y eficiente de predictores.

6.2. Selección mediante mRMR

El enfoque *mRMR* (minimum Redundancy - maximum Relevance) selecciona las variables que maximizan su relevancia estadística respecto a la variable objetivo, minimizando al mismo tiempo la redundancia entre ellas. Este método utiliza medidas de información mutua para cuantificar la dependencia no lineal entre las variables. En la práctica, el algoritmo mRMR prioriza aquellas variables que aportan información nueva y no redundante sobre el fenómeno modelado (en este caso, la acumulación de carbono), favoreciendo la diversidad informativa frente a la mera fuerza de correlación. Este enfoque permite obtener un conjunto equilibrado de predictores que explican diferentes dimensiones del sistema ecológico.

6.3. Selección manual basada en criterios estadísticos y conceptuales

La selección manual de variables se realizó de forma guiada por criterios tanto estadísticos como conceptuales. En primer lugar, se evaluó la significancia de cada variable mediante pruebas univariantes (ANOVA y correlaciones), eliminando aquellas sin influencia estadísticamente significativa sobre la variable objetivo. Posteriormente, se

analizaron las correlaciones entre predictores para reducir la colinealidad, manteniendo únicamente una variable representativa de cada grupo altamente correlacionado. Además, se consideraron criterios ecológicos y de interpretación física, asegurando que las variables retenidas representasen aspectos estructurales, edáficos, topográficos, climáticos y espectrales relevantes para el proceso de acumulación de carbono. El objetivo fue equilibrar la robustez estadística con la coherencia ecológica, obteniendo un conjunto final de predictores que mantuviera un compromiso entre precisión, interpretabilidad y sentido biogeográfico.

6.4. 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 dos 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.
- 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.

Protocolo de reporte. Para cada modelo se reportan: (i) el rendimiento medio y la dispersión en la validación cruzada por grupos (entrenamiento), y (ii) el desempeño final en el conjunto de evaluación independiente (20 %). Este protocolo garantiza comparabilidad entre modelos, control del sesgo espacial por parcela y verificación explícita de la robustez temporal.

6.5. Modelos evaluados

A continuación, se detalla el diseño general y las estrategias empleadas para la selección y optimización de modelos.

Entrenamiento y optimización

Se aplicó *RandomizedSearchCV*, una técnica de búsqueda aleatoria de hiperparámetros que evalúa distintas combinaciones utilizando validación cruzada. Este procedimiento permite optimizar el rendimiento de cada modelo sin incurrir en un coste computacional tan elevado como el de una búsqueda exhaustiva.

VALIDACIÓN CRUZADA POR GRUPOS. En algunas configuraciones probadas, se emplea la validación cruzada por grupos (*Group k-Fold Cross Validation*). Este método divide el área de estudio en k bloques, basados en una característica común de los datos, asegurando que los datos dentro de cada bloque estén relacionados. El modelo se entrena con $k-1$ bloques y se valida con el bloque restante, repitiendo este proceso k veces.

Este enfoque es útil cuando los datos tienen agrupaciones naturales, como por ejemplo, diferentes parcelas o periodos de tiempo. Al mantener los datos relacionados en un mismo bloque, se evita la filtración de información entre los conjuntos de entrenamiento y validación, lo que permite una mejor evaluación de la capacidad de generalización del modelo. Así, se asegura que el modelo no se sobreajuste y sea robusto al ser evaluado en contextos no vistos previamente.

Modelos ensemble

Para mejorar la precisión y robustez, se emplearon diversos métodos de *ensemble learning*, que combinan múltiples modelos base (*base learners*) para generar una predicción agregada. Esta estrategia se inspira en la teoría de la sabiduría colectiva: la combinación de estimaciones independientes tiende a superar a cualquier estimador individual.

TÉCNICAS UTILIZADAS:

- **Voting y Averaging:** combinan modelos ya entrenados mediante votación mayoritaria o promedio.
- **Bagging y Boosting:** construyen modelos desde cero y los combinan. Bagging reduce la varianza al entrenar modelos en subconjuntos aleatorios; Boosting mejora el sesgo al entrenar secuencialmente, corrigiendo errores anteriores.
- **Stacking:** combina modelos optimizados usando un metamodelo que aprende a integrar sus predicciones.

Boosting y aprendizaje gradual

El *boosting* se basa en el aprendizaje secuencial, donde cada nuevo modelo intenta corregir los errores residuales del anterior. Esta técnica permite construir modelos fuertes a partir de modelos débiles, alcanzando gran precisión. Sin embargo, requiere una cuidadosa configuración de hiperparámetros para evitar el sobreajuste.

Entre las implementaciones destacadas se incluyen:

- **XGBoost:** modelo GBM que optimiza rendimiento con gradientes de primer y segundo orden, regularización L1/L2, manejo automático de valores faltantes, y técnicas de generalización como *shrinkage* y *column subsampling*.

- **LightGBM:** algoritmo eficiente para grandes volúmenes de datos, con crecimiento *leaf-wise* y soporte nativo para variables categóricas.
- **AdaBoost:** ajusta modelos simples secuencialmente, enfocando el aprendizaje en observaciones mal clasificadas.
- **CatBoost:** especializado en variables categóricas y robusto frente a datos ruidosos, usando codificación por orden aleatorio.
- **Gradient Boosting Decision Trees (GBDT):** construye árboles secuenciales ajustados a residuos, optimizando mediante descenso por gradiente.

Bagging

El *bagging* (Bootstrap Aggregating) entrena múltiples modelos independientes sobre subconjuntos de datos generados por muestreo con reemplazo. Las predicciones se combinan por promedio o votación. Esta técnica reduce la varianza y mejora la estabilidad de modelos inestables.

- **Random Forest:** combina árboles de decisión (CART) con selección aleatoria de características en cada división. Es escalable, robusto a datos faltantes, y menos propenso al sobreajuste.
- **Bagged Decision Trees (BaggedDT):** genera árboles sin poda entrenados en muestras bootstrap. Promedia sus predicciones para reducir la varianza.

Otros modelos utilizados

Además de los métodos ensemble, se evaluaron modelos representativos de distintos paradigmas de aprendizaje supervisado:

- **K-Nearest Neighbors (KNN):** modelo basado en instancia que predice a partir de los vecinos más cercanos. Sensible a la escala y a *outliers*.
- **Multi-Layer Perceptron (MLP):** red neuronal con una o más capas ocultas, capaz de modelar relaciones no lineales complejas.
- **Support Vector Regression (SVR):** modelo de márgenes para regresión, con soporte para kernels no lineales.
- **SVM con kernel:** modelo poderoso para clasificación y regresión no lineal, aunque costoso y sensible a hiperparámetros.
- **Bayesian Neural Network:** enfoque probabilístico que estima incertidumbre en las predicciones. Incluye variantes como la *Bayesian Ridge Regression*.
- **Naive Bayes:** clasificador probabilístico rápido y simple, útil en texto y alta dimensionalidad. Se evaluaron variantes:
    - *Gaussian Naive Bayes:* para datos continuos.
    - *Multinomial Naive Bayes:* para conteos y texto.
    - *Bernoulli Naive Bayes:* para variables binarias.

Comparación y justificación de modelos

La evaluación de múltiples modelos responde a la necesidad de identificar no solo el de mejor rendimiento, sino también el más adecuado según la naturaleza del problema y los datos disponibles. Se compararon algoritmos lineales, no lineales, basados en vecinos, redes neuronales, modelos probabilísticos y diferentes técnicas de *ensemble*. Vemos un resumen de los modelos aplicados en la tabla **??**.

Modelo	Tipo / Técnica	Características destacadas	Observaciones
Random Forest	Bagging (Árboles)	Uso de bootstrap, selección aleatoria de atributos, reducción de varianza	Robusto y escalable; menor interpretabilidad
Bagged Decision Trees (BaggedDT)	Bagging	Árboles sin poda, entrenados en paralelo sobre muestras con reemplazo	Preciso pero costoso computacionalmente
XGBoost	Boosting (GBM)	Regularización L1/L2, manejo de valores faltantes, poda anticipada	Alto rendimiento, sensible a hiperparámetros
LightGBM	Boosting (Leaf-wise)	Crecimiento hoja a hoja, eficiente en grandes volúmenes	Rápido y preciso; riesgo de sobreajuste
AdaBoost	Boosting (Stumps)	Aumenta peso de errores, pondera modelos por precisión	Sencillo y efectivo con datos limpios
CatBoost	Boosting especializado	Codificación avanzada de variables categóricas, robustez a ruido	Ideal para datos heterogéneos
Gradient Boosting Decision Trees (GBDT)	Boosting	Árboles secuenciales ajustados a residuos	Buen rendimiento; mayor coste de entrenamiento
K-Nearest Neighbors (KNN)	Basado en instancia	No requiere entrenamiento, predice por proximidad	Sensible a escala y outliers
Multi-Layer Perceptron (MLP)	Red neuronal	Modela relaciones no lineales complejas	Requiere normalización y regularización
Support Vector Regression (SVR)	Kernel y márgenes	Predicción dentro de tolerancia ε, uso de kernels no lineales	Robusto; elevado coste computacional
SVM con kernel	SVM no lineal	Maximiza margen, admite distintos kernels (RBF, polinomial, etc.)	Alta precisión; sensible a hiperparámetros
Bayesian Neural Network / Ridge Regression	Probabilístico / Bayesiano	Predicción con incertidumbre, estimación automática de hiperparámetros	Útil para inferencia y regularización
Naive Bayes (Gaussian, Multinomial, Bernoulli)	Probabilístico	Asume independencia condicional, rápido y simple	Eficaz en texto y alta dimensionalidad

Tabla 6.1: Resumen de modelos de aprendizaje supervisado aplicados

7. IMPLEMENTACIÓN DE LOS MODELOS

El desarrollo y evaluación de los modelos predictivos se realizó íntegramente en **Python** y el entorno de ejecución fue local, en un equipo con procesador Intel Core i7 y 32 GB de RAM, lo que permitió realizar experimentos de forma eficiente con un conjunto de datos de tamaño considerable (~80.000 muestras).

El preprocesamiento de datos se llevó a cabo mediante la librería `scikit-learn`, utilizando `Pipeline` y `ColumnTransformer` para combinar transformaciones numéricas y categóricas. En particular, las variables numéricas se imputaron con la mediana y se escalaron con `StandardScaler`, mientras que las variables categóricas se trataron mediante imputación por moda y codificación *one-hot*. La función objetivo a predecir fue el carbono total (`CZ`) acumulado en cada parcela.

Se implementaron y optimizaron diversos modelos de regresión supervisada, incluyendo:

- **Modelos basados en árboles:** `RandomForestRegressor`, `XGBoost`, `LightGBM`, `CatBoost`, `GradientBoosting`, `AdaBoost` y `Bagging`.
- **Modelos basados en instancias:** `KNeighborsRegressor`.
- **Modelos de redes neuronales:** `MLPRegressor`.
- **Modelos de soporte vectorial:** `SVR`.
- **Modelos probabilísticos:** `BayesianRidge`.

La optimización de hiperparámetros se realizó mediante `RandomizedSearchCV`, con validación cruzada de 5 particiones y búsqueda en espacios definidos manualmente para cada modelo. Los modelos fueron evaluados en términos de R^2 y **RMSE**, tanto en el conjunto de entrenamiento como en el de prueba.

Con el objetivo de mejorar el rendimiento predictivo, se evaluaron además varias configuraciones de `StackingRegressor`, combinando distintos subconjuntos de modelos base (previamente entrenados) con diversos meta-modelos (`LinearRegression`, `Ridge`, `GradientBoosting`, `SVR`, `MLP`, entre otros). Estas combinaciones permitieron comparar sinergias entre modelos complementarios.

El tiempo de entrenamiento varió según el modelo y la configuración de hiperparámetros. Gracias al uso de `n_jobs=-1` se aprovechó el paralelismo multinúcleo para acelerar la optimización.

8. RESULTADOS

Presentar los resultados obtenidos al aplicar el modelo a los datos de entrada. Incluir gráficos y tablas que ayuden a ilustrar el rendimiento del modelo.

PROCESO DE ENTRENAMIENTO Y VALIDACIÓN DEL MODELO

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.

8.1. Elección de variables

Para la selección de variables se adoptó un enfoque sistemático basado en la evaluación del desempeño predictivo de distintos subconjuntos de características, definidos por bloques temáticos con significado ecológico y funcional. Cada combinación de variables se entrenó mediante el algoritmo `CatBoost`, manteniendo constantes la variable objetivo (carbono capturado) y los parámetros de modelado, de forma 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 incluyeron progresivamente grupos de variables relacionadas con las características de la especie, las propiedades edáficas, el terreno, las condiciones climáticas (temperatura y precipitación) y los índices de vegetación. A partir de la comparación de los resultados, se identificaron los bloques con mayor aporte marginal al rendimiento del modelo, priorizando aquellos cuya incorporación mejoró consistentemente el R^2 sin incrementar de forma significativa la complejidad o redundancia del conjunto de predictores.

El análisis comparativo de configuraciones de variables mediante el modelo `CatBoost` permitió estimar la contribución marginal de cada bloque al poder predictivo del modelo (medido a través de R^2 y RMSE). A partir de los resultados, se establecieron las siguientes conclusiones:

1. **Bloque especies.** Incluye variables sobre estado, forma, tratamiento, origen, distribución, composición y fracción de cabida cubierta. Es el bloque con mayor impacto, con incrementos de hasta +0,0054 en R^2. Su efecto es consistente en todas las combinaciones, por lo que se considera esencial.
2. **Bloque temperaturas y precipitaciones.** Aporta una mejora sistemática y estable, especialmente en presencia de variables edáficas o de terreno. Los incrementos típicos oscilan entre +0,002 y +0,005 en R^2. Se recomienda su inclusión por su bajo coste computacional y su valor informativo adicional.
3. **Bloque terreno.** Comprende tipo de suelo, rocosidad, orientación, elevación y pendiente. Su efecto es moderado (variaciones de ±0,001 en R^2) y, aunque aporta cierta estabilidad al modelo, su relevancia es secundaria. Puede incorporarse cuando la disponibilidad de datos lo permita.
4. **Bloque índices de vegetación.** (NDII, GNDVI) Produce una ganancia leve y no siempre significativa ($\Delta R^2 \approx 0{,}002$–$0{,}004$), pero tiende a mejorar ligeramente el ajuste medio. Su inclusión es recomendable si los datos están disponibles sin aumentar el coste del pipeline.
5. **Bloque soil.** (erosividad, textura, materia orgánica, combustibilidad) Presenta la menor rentabilidad predictiva. Aunque puede mejorar en combinación con variables climáticas, su efecto es inferior al del bloque especies. Se recomienda mantenerlo solo por motivos interpretativos o contextuales.

Aunque el bloque *terreno* no mostró inicialmente una contribución destacada al rendimiento global del modelo, se realizó un análisis adicional para evaluar el efecto individual de cada una de sus variables. Los resultados indicaron que la variable *elevación* presenta una influencia positiva sobre la capacidad predictiva del modelo, mejorando ligeramente las métricas de ajuste. En concreto, la inclusión de *elevación* elevó el coeficiente de determinación a $R^2 = 0{,}8772$ y redujo el error cuadrático medio a RMSE = 14,36, frente a los valores obtenidos sin dicha variable ($R^2 = 0{,}8766$, RMSE = 14,40).

Este incremento, aunque moderado, resulta relevante al tratarse de una única variable adicional (41 predictores frente a 40), lo que sugiere que la altitud puede captar gradientes ambientales asociados a la variabilidad en la captura de carbono que no están plenamente representados por los demás bloques.

Recomendación práctica: la configuración óptima corresponde a fijas + especies + temperaturas/precipitaciones + índices, con $R^2 = 0{,}8766$ y 40 variables. Como alternativa ligera, fijas + especies + terreno, $R^2 = 0{,}8745$ y fijas + especies + índices, $R^2 = 0{,}8744$ ofrecen un equilibrio adecuado entre rendimiento y complejidad. En términos de prioridad, los bloques deben incluirse en el siguiente orden: **especies** $\gg$ **temperaturas y precipitaciones** > **terreno** $\gtrsim$ **índices** > **soil**.

En conjunto, las variables fijas explican la mayor parte de la varianza, mientras que los bloques adicionales permiten afinar la estimación del carbono capturado. Dado el alto rendimiento base del modelo `CatBoost`, las mejoras marginales son pequeñas pero consistentes, destacando el valor de las variables de especie y climáticas como componentes clave en la predicción.

Finalmente, se evaluó la sustitución del bloque de variables climáticas (temperaturas y precipitaciones) por el Índice de Martonne, una métrica sintética que integra de forma conjunta la información térmica e hídrica en un solo parámetro. Esta modificación permitió reducir significativamente la dimensionalidad del bloque climático (de ocho variables a una), manteniendo un desempeño prácticamente equivalente. El modelo resultante (`CatBoost`, 33 variables) alcanzó un $R^2 = 0{,}8721$ y un RMSE de 14,66, valores muy similares a los obtenidos con las variables climáticas explícitas ($R^2 = 0,8766$, RMSE = 14,40). Esta equivalencia, junto con la reducción en complejidad y carga computacional, respalda el uso del Índice de Martonne como sustituto eficiente del conjunto de variables de temperatura y precipitación en la modelización de la captura de carbono.

9. DISCUSIÓN

Interpretar los resultados obtenidos, comparándolos con investigaciones previas. Discutir las limitaciones del modelo y las posibles áreas de mejora.

10. CONCLUSIONES

Recapitular las conclusiones más importantes del estudio. Resaltar la relevancia del modelo desarrollado y su aplicación en proyectos de forestación y reforestación.

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*).

12. ANEXOS

12.1. 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* [**? ?**]. 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 [**?**]:

- 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 [**?**], de forma que:

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

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* [**?**], 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.

12.2. 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.

12.3. 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.

12.4. 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.

12.5. 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.

12.6. 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 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 5.5.
   - **Fuertemente ácido:** pH <5.5.

12.7. 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.

12.8. 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.

12.9. 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.

12.10. 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:

GRUPO	MOD. COMBUSTIBLE	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.
	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.

Tabla 12.1: Descripción de los modelos de combustible del Inventario Forestal Nacional, clasificados por grupo funcional.

12.11. 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.

12.12. 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.

12.13. 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.

12.14. 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.

12.15. 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.

12.16. Anexo: Código de los grupos taxonómicos de las especies (`grupo_id`)

Tabla 12.2: 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		

12.17. Anexo: Código de las especies (`especie_id`)

Tabla 12.3: Relación de especies empleadas en el estudio y metadatos asociados.

Código	Nombre	Sinonimia	Tipo	Grupo
307	Acacia dealbata	Acacia dealbata	1	7
207	Acacia melanoxylon	Acacia melanoxylon	1	7
7	Acacia spp.	-	1	7
392	Gleditsia triacanthos	Acacia gleditsia	1	7
92	Robinia pseudoacacia	Acacia robinia	1	7
292	Sophora japonica	Acacia sofora	1	7
515	Crataegus azarolus	Espino	1	15
415	Crataegus laciniata	Majoleto	1	15
315	Crataegus laevigata	Espino majuelo	1	15
215	Crataegus monogyna	Majuelo	1	15
15	Crataegus spp.	-	1	15
30	Mezcla de coníferas	Coníferas | excepto pinos	0	19
19	Otras coníferas	-	0	19
29	Otros pinos	-	0	20
20	Pinos	-	0	20
27	Pinus canariensis	-	0	20
24	Pinus halepensis	-	0	20
25	Pinus nigra	Pinus laricio | Pinus clusiana	0	20
26	Pinus pinaster	Pinus maritima	0	20
23	Pinus pinea	-	0	20
28	Pinus radiata	Pinus insignis	0	20
21	Pinus sylvestris	-	0	20
22	Pinus uncinata	Pinus montana | Pinus mugo	0	20
31	Abies alba	Abies pectinata	0	31
32	Abies pinsapo	-	0	31
235	Larix decidua	Alerce común	0	35
335	Larix leptolepis	Larix kaempferi | Alerce leptolepis	0	35
35	Larix spp.	-	0	35
435	Larix x eurolepis	Alerce híbrido	0	35
49	Otros quercus	-	1	40
344	Quercus alpestris	-	1	40
47	Quercus canariensis	Quercus lusitanica var. baetica	1	40
44	Quercus faginea	Quercus lusitanica var. faginea	1	40
45	Quercus ilex ssp. ballota	Quercus rotundifolia	1	40
245	Quercus ilex ssp. ilex	-	1	40
244	Quercus lusitanica	Quercus fruticosa | Quejigueta	1	40
42	Quercus petraea	Quercus sessiliflora	1	40

Continúa en la siguiente página

Tabla 12.3: Relación de especies empleadas en el estudio y metadatos asociados.

Código	Nombre	Sinonimia	Tipo	Grupo
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
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

Continúa en la siguiente página

Tabla 12.3: Relación de especies empleadas en el estudio y metadatos asociados.

Código	Nombre	Sinonimia	Tipo	Grupo
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. 10 m)	1	99
99	Otras frondosas	Otras frondosas	1	99
499	Morus alba	Morera	1	399
599	Morus nigra	Morera	1	399
399	Morus spp.	Morera	1	399
55	Fraxinus angustifolia	-	1	455
255	Fraxinus excelsior	Fresno excelsior	1	455
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

Continúa en la siguiente página

Tabla 12.3: Relación de especies empleadas en el estudio y metadatos asociados.

Código	Nombre	Sinonimia	Tipo	Grupo
336	Cupressus lusitanica	Ciprés lambertiana	0	936
436	Cupressus macrocarpa	Ciprés americano	0	936
36	Cupressus sempervirens	-	0	936
936	Cupressus spp.	Cipres	0	936
37	Juniperus communis	-	0	937
237	Juniperus oxycedrus	Enebro oxicedro	0	937
39	Juniperus phoenicea	-	0	937
239	Juniperus sabina	Sabina rastrera	0	937
937	Juniperus spp.	Enebros y sabinas	0	937
38	Juniperus thurifera	-	0	937
238	Juniperus turbinata	Sabina canaria	0	937
256	Ulmus glabra	Ulmus montana	1	956
56	Ulmus minor	Ulmus campestris	1	956
356	Ulmus pumila	Olmo pumilo	1	956
956	Ulmus spp.	Olmo	1	956
275	Juglans nigra	Nogal	1	975
75	Juglans regia	-	1	975
975	Juglans spp.	-	1	975
76	Acer campestre	-	1	976
276	Acer monspessulanum	Arce de Montpelier	1	976
376	Acer negundo	Negundo fraxinifolia | Arce negundo	1	976
476	Acer opalus	Arce ópalus	1	976
676	Acer platanoides	Arce platanoide	1	976
576	Acer pseudoplatanus	Arce seudoplátano	1	976
976	Acer spp.	Arces	1	976
97	Sambucus nigra	Saúco negro	1	997
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	-

Continúa en la siguiente página

Tabla 12.3: Relación de especies empleadas en el estudio y metadatos asociados.

Código	Nombre	Sinonimia	Tipo	Grupo
74	Corylus avellana	-	1	-
569	Dracaena draco	Drago	1	-
83	Erica arborea	-	1	-
283	Erica scoparia	Tejo | brezo arbóreo escopario	1	-
5	Euonymus europaeus	-	1	-
71	Fagus sylvatica	-	1	-
299	Ficus carica	Higuera	1	-
3	Frangula alnus	Rhamnus frangula	1	-
1	Heberdenia bahamensis	Heberdenia excelsa	1	-
12	Malus sylvestris	-	1	-
60	Mezcla de eucaliptos	Eucaliptos	1	-
50	Mezcla de árboles de ribera	Árboles ripícolas	1	-
81	Myrica faya	-	1	-
281	Myrica rivasmartinezii	-	1	-
6	Myrtus communis	-	1	-
87	Ocotea phoetens	-	1	-
66	Olea europaea	Olea oleaster	1	-
59	Otros árboles ripícolas	-	1	-
84	Persea indica	-	1	-
8	Phillyrea latifolia	-	1	-
86	Picconia excelsa	Notelaea excelsa	1	-
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	-