Universidad Pública de Navarra

ESCUELA TECNICA SUPERIOR DE INGENIEROS AGRONOMOS

Nafarroako Unibertsitate Publikoa

NEKAZARITZAKO INGENIARIEN GOI MAILAKO ESKOLA TEKNIKOA

Predicción de variables dasométricas del Inventario Forestal Nacional mediante datos LIDAR con técnicas de minera de datos.

presentado por

JORDI SEGÚ TELL

aurkeztua

MÁSTER EN SISTEMAS DE INFORMACIÓN GEOGRÁFICA Y TELEDETECCIÓN
MASTERRA INFORMAZIO SISTEMA GEOGRAFIKOETAN ETA TELEDETEKZIOAN

Septiembre de 2018

Abstract:
The management of forest resources is essential for the development of our society. This requires a forest management planning based on innovative studies, according to new technologies and seeking to lower costs. In this project, a methodology has been developed for the extraction of predictive regression models to determine the main dasometric variables of the beech forest layer with over 70% of the forest cover density in Navarre. For this purpose, data mining techniques and Python as programming language have been used. The inputs of the work are: data from the plots of the national forest inventory (dependent variables) and statistics derived from the LIDAR-PNOA flight for these same plots (independent variables) obtained with the LasTools software. The output is the model that best fits the input data, determined by the methodology used.

Keywords:
LIDAR, IFN, Python, Data Mining, Fagus sylvatica

Índice:

1. Introducción.

1.1. Antecedentes.

Son muchas las funciones y beneficios que aportan los bosques en la sociedad humana. Un estudio de situación y perspectivas de la conservación y desarrollo de los bosques realizado por la FAO (1980) concluye que los beneficios de los bosques se pueden dividir en funciones protectoras, funciones reguladoras y funciones productivas. Para mantener de una forma sostenible estas funciones, vitales para el desarrollo de nuestra sociedad, es necesaria una gestión activa del bosque basada en el estudio, planificación, gestión y explotación de los recursos del bosque. Esta gestión tiene que pasar por la obtención de datos fiables y relevantes que nos permitan conocer la cantidad de esos recursos naturales. Tradicionalmente el bosque se ha estudiado yendo a campo y realizando inventarios de forma manual. Estos métodos tienen un coste económico elevado, ya que se tiene que dedicar mucho tiempo y mano de obra para realizarlos. Actualmente, con la ayuda de nuevas tecnologías, se puede agilizar un poco más todo este proceso. Y precisamente este es uno de los objetivos del trabajo.

Según otro estudio realizado por la Departamento de Desarrollo Rural y Medio Ambiente de la Comunidad Foral de Navarra (2009) "En la Comunidad Foral de Navarra ha disminuido un 38% la superficie desarbolada mientras que la arbolada ha aumentado más del 40% desde el año 1971 al 2012" En la *figura 1,* extraída del mismo informe, se puede observar este hecho con la evolución de los datos del inventario forestal nacional (IFN).

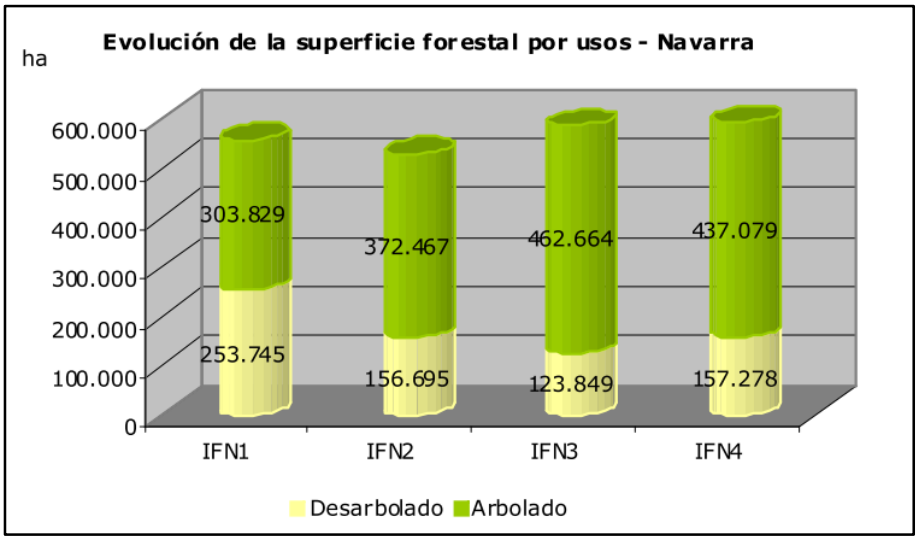

Figura 1: Evolución de la superficie forestal en Navarra (fuente: Gobierno de Navarra, 2009).

Cabe esperar que con el creciente aumento de la masa forestal las inversiones económicas destinadas a su gestión también aumenten. Pero esto no es así, según el Diario Público (2018), en referencia a los últimos presupuestos generales, el dinero que destinará el gobierno en temas de medioambiente es un 69% menor que antes de la crisis. Es por este motivo que las organizaciones y en general la comunidad científica, comprometida con el medio ambiente, tienen que buscar estrategias y metodologías de abaratamiento de costes para la realización de estudios que tengan como objetivo la gestión sostenible de los bosques. Este trabajo final de

máster se enmarca en este planteamiento. Los avances tecnológicos y científicos que van surgiendo en los últimos años ayudan a que esto pueda ser posible.

El LIDAR (Light Detection And Ranging) según el portal de ESRI (2016) "es una técnica de teledetección óptica que utiliza la luz de láser para obtener una muestra densa de la superficie de la tierra produciendo mediciones exactas de x, y y z". Si bien esta tecnología no la podríamos calificar como económica, ya que los sensores y los métodos principales de toma de datos (avionetas) tienen unos costes elevados. Luego una vez se han tomado los datos son muchas las aplicaciones y beneficios indirectos que puede tener. En España existe una cobertura completa de toda la península realizada por el PNOA (Plan Nacional de Ortofotografía Aérea) entre los años 2008 y 2015.

1.2. Objetivos.

Los distintos objetivos específicos del trabajo se han desglosado de la siguiente manera:

1) Objetivo general:

a. Crear modelos predictivos de regresión para determinar las principales variables dasométricas (volumen con corteza (VCC), área basimétrica (G) y número de pies (N)) de manera continua en el estrato forestal de hayedo en Navarra, a partir de métricas extraídas del LIDAR-PNOA (2011-2012) y de datos de campo del 4º inventario forestal nacional (IFN4).

2) Objetivos específicos:

a. Extraer métricas derivadas de los datos LIDAR con el software LAStools.

b. Modelizar y comparar los resultados obtenidos con distintas metodologías:

   i. Métodos de regresión con técnicas de minería de datos, escritas en Python y desarrolladas con la librería de scikit-learn.
   ii. Validación cruzada de los modelos, para determinar el modelo más robusto.

c. Automatizar los procesos para la extracción de datos:
   i. Datos IFN4: Creación de consultas y nuevas tablas con lenguaje SQL sobre la base de datos Access original del IFN4 para extraer los parámetros de las variables dependientes (VCC, G, N) de las parcelas del IFN4 en los modelos de regresión.
   ii. Métricas LIDAR: Creación de un archivo batch para encadenar las distintas herramientas de LAStools hasta llegar a la extracción de los parámetros de las variables independientes de cada parcela del IFN4.
   iii. Modelización en Python con la librería scikit-learn: Creación de un script con el intérprete Spyder y escrito en el lenguaje de programación de Python para obtener los modelos de regresión.

2. Material y métodos.

2.1. Área de estudio.

El área de estudio se ha definido a partir de las divisiones de estratos hechas por el cuarto Inventario Forestal Nacional (IFN4) de la Comunidad Foral de Navarra. En la *tabla 1* se muestran el listado de estratos con sus características.

Estrato	Definición				Cabida (ha)	Cantidad de parcelas
	Formación forestal dominante	Ocupación (%)	Estado de masa	Fracción de cabida cubierta (%)		
01	*Fagus sylvatica*	>=70	Fustal. Latizal	70 - 100	115.763,37	830
02	*Fagus sylvatica*	>=70	Fustal. Latizal	20 - 69	6.935,33	106
03	*Abies alba* y *Abies alba* con *Fagus sylvatica* y *Pinus sylvestris*	>=70; 30<=Esp.<70	Fustal. Latizal	20 - 100	10.128,40	85
04	*Fagus sylvatica* con *Quercus robur/ Q. petraea*	30<=Esp.<70	Fustal. Latizal	20 - 100	9.257,78	69
...						
21	*Castanea sativa* y *Castanea sativa* con *Quercus robur/Q. petraea* y *Fagus sylvatica*	>=70; 30<=Esp.<70	Fustal. Latizal Monte bravo. Repoblado	20 - 100 5 - 100	8.512,07	61
...						
24	*Fagus sylvatica, Corylus avellana, Crataegus monogyna, Acer campestre* y *Quercus pyrenaica*	>=70; 30<=Esp.<70	Monte bravo. Repoblado	5 - 100	2.731,22	35
...						
26	Matorral con arbolado ralo y disperso	>=70; 30<=Esp.<70	Fustal. Latizal	5 - 19	13.771,02	50
27	Árboles de ribera	>=70; 30<=Esp.<70	Todos	5 - 100	5.122,05	86
Todos					437.079,35	3.167

Tabla 1: Datos básicos de los estratos del IFN4 en Navarra. (Fuente: inventario forestal nacional)

Para el estudio se ha elegido el estrato 01. La formación forestal dominante es el *Fagus sylvatica* con una ocupación igual o mayor del 70% en masas de fustal/latizal con una fracción de cabida cubierta entre el 70% y el 100 %. La superficie total del estrato, y también de la zona de estudio, es de 115.763 ha y se albergan 830 parcelas de muestreo.

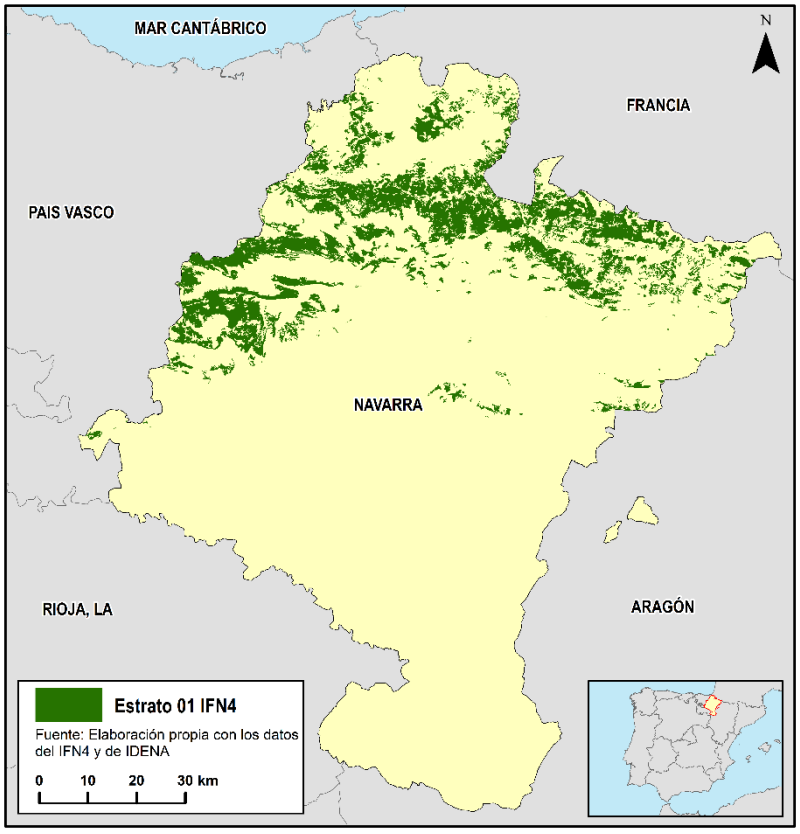

Figura 2: Ubicación del estrato 01 en la Comunidad Foral de Navarra. Fuente: Elaboración propia

En la *figura 2* se puede observar como el área de estudio se encuentra en el tercio norte de la Comunidad Foral de Navarra. Sus dimensiones equivalen al 11,3% de la superficie de la comunidad. Estos bosques se encuentran en las zonas más elevadas y forman una masa continua de este a oeste, solo interrumpida por los fondos de valle y zonas de pastizales donde hay otros tipos de masas. Dentro de esta extensa masa se pueden encontrar distintos tipos de hayedo en función de la fisonomía del bosque y su composición florística (*figura 3*), Departamento de desarrollo rural y medioambiente de la Comunidad Foral de Navarra (2009):

- **Hayedo Oriental:** Ubicados en la Sierra de Abodi, Valle del Roncal y Valle del Salazar. Se encuentran los hayedos de la serie montana pirenaica basófila y ombrófila del haya, caracterizados por crecer sobre sustratos ricos en bases bajo un ombroclima húmedo e hiperhúmedo. Aparecen en zonas altas, dejando paso al roble peloso y al pino silvestre cuando se desciende en cota.

- **Hayedo Septentrional:** Ubicados en la zona cantábrica, concretamente en las cuencas del Bidasoa y del Urumea, en el valle de Ultzama y en Quinto Real. Corresponden a la serie cántabro-euskalduna acidófila del haya, caracterizada porque las hayas aparecen a cotas incluso inferiores a los 400 m, formando unos bosques densos y umbríos con ejemplares de talla elevada y porte esbelto, en los que apenas crece sotobosque.

- **Hayedo Occidental:** Ubicados en las Sierras de Aralar y Urbasa. Pertenecen a la serie orocantábrica y cántabro-euskalduna basófila y ombrófila del haya, caracterizada por su adaptación a climas menos lluviosos. Aparecen a partir de los 600-700 m de altitud y son también bosques muy sombríos que si desaparecen dan paso a una vegetación espinosa cuya etapa madura está dominada por el espino *Crataegus monogyna*, especie muy típica en Urbasa.

- **Hayedo Meridional**: Estos son los hayedos que se encuentran más al sur, dados los requerimientos de humedad que tiene esta especie, se encaraman a las zonas más altas de las sierras, apareciendo a partir de los 900-1.000 m. Suelen crecer en replanos y laderas con cierta pedregosidad, formando bosques muy diferentes de los anteriores, mucho menos densos y con ejemplares de menor talla. El estrato arbustivo aquí es muy rico y está dominado por el boj.

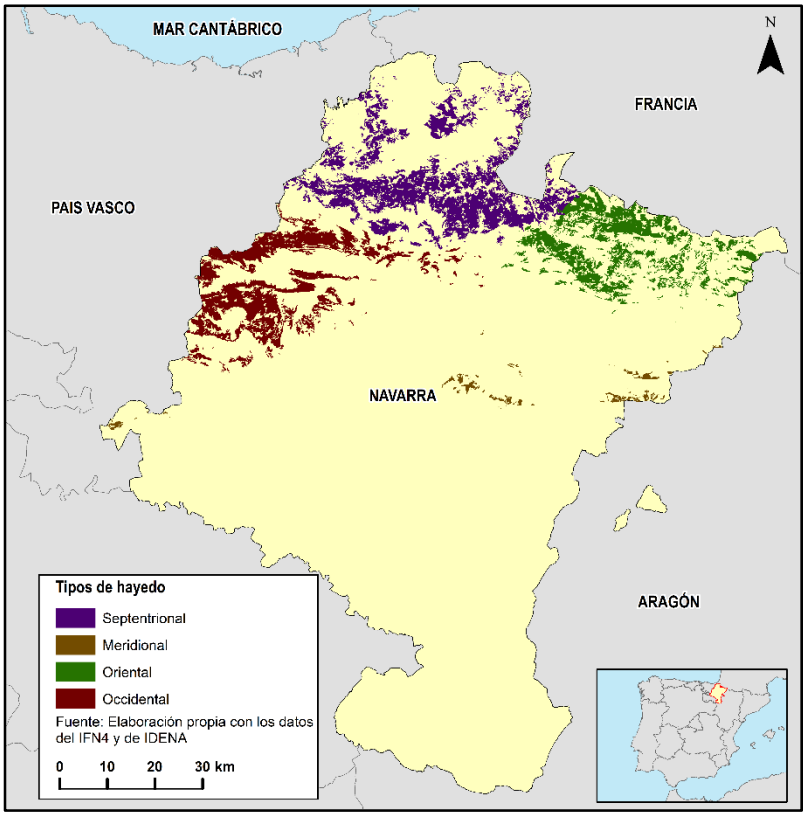

Figura 3: División del estrato L1 del IFN4 en Navarra por tipologías de hayedo que determina el Departamento de desarrollo rural y medioambiente de la Comunidad Foral de Navarra (2009)

2.2. Parcelas de campo.

En este apartado se describe el Inventario Forestal Nacional (IFN) y su metodología. Después se expone como se han extraído los datos de las parcelas de entrenamiento a partir de la base de datos original del IFN4. Y finalmente se detallan cuáles han sido los criterios de la selección de las parcelas válidas y su estratificación para modelizar los parámetros de interés.

2.2.1. Cuarto inventario forestal nacional (IFN4).

El inventario forestal nacional (IFN) es un proyecto que está dirigido por el Ministerio de Agricultura, Pesca y Alimentación del Gobierno de España. Sus impulsores definen sus objetivos generales como: "la suministración de información estadística homogénea y adecuada sobre el estado y evolución de los montes españoles, para servir como instrumento para la coordinación de las políticas forestales y de conservación de la naturaleza de las comunidades autónomas del Estado y de la Unión Europea. La unidad básica de trabajo es la provincia y, al ser un inventario continuo, se repiten las mismas mediciones cada 10 años, recorriéndose todo el territorio nacional en cada ciclo decenal." En el caso de Navarra el primer IFN fue publicado el año 1971, el IFN2 en el año 1986, IFN3 2000 y el IFN4 en el año 2012.

Para la realización de este estudio se han seleccionado los datos del IFN4 porque fue publicado en los años más próximos en que se hizo el vuelo LIDAR-PNOA de Navarra (2011-2012). La

información necesaria para poder crear los modelos de regresión son las métricas forestales de las parcelas de muestreo del inventario. La obtención de estos datos no ha sido una tarea fácil, ya que su disposición en la base de datos requiere de un cierto análisis e interpretación. Para ello se crearon una serie de consultas SQL que se pueden ver en el Anexo 1, en el próximo apartado se detalla de una manera más clara el proceso seguido.

2.2.2. Obtención de los datos de las parcelas.

La metodología usada en el IFN consiste en la delimitación de una malla de 1 x 1 km donde en cada vértice de la malla se sitúa una parcela de muestreo. La parcela es un círculo de 25 metros de radio. Y dentro de este se sitúan tres círculos concéntricos a 5m, 10m y 15m del centro de la parcela (*figura 4*). Donde en cada círculo se mide un tipo de árbol distinto en función de su diámetro:

- Círculo de 5m: arboles entre 75 y 125 mm Ø.
- Círculo de 10m: arboles entre 125 y 225 mm Ø.
- Círculo de 15m: arboles entre 225 y 425 mm Ø.
- Círculo de 25m: arboles > 425 mm Ø.

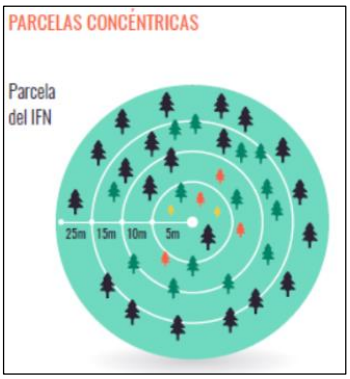

Figura 4: Organización de una parcela del IFN4. (Fuente: Inventario forestal nacional)

Para pasar el dato de árbol a valores por hectárea el IFN define unos factores de expansión. Los factores de expansión *(tabla 2)* son el resultado de dividir la superficie de una hectárea por la superficie de la subparcela. Para obtener los datos en las unidades por hectárea, se tiene que multiplicar el factor de expansión a cada valor de árbol individual de la base de datos del IFN y luego sumar todos los arboles que corresponden a esa parcela.

Rango diamétrico	Factor de expansión
7,5 a 12,5 cm	127,32
12,5 a 22,5 cm	31,83
22,5 a 42,5 cm	14,14
> 42,5 cm	5,0

Tabla 2: Factores de expansión definidos por el IFN. Fuente Inventario Forestal Nacional

Para la realización de este estudio se necesita extraer de la base de datos del IFN:

- Coordenadas del centro de la parcela X e Y (en el sistema de referencia ETRS89 UTM zona 30N.) Para poder localizar las parcelas de nuestra área de estudio.

Los siguientes datos que se tienen que extraer son las variables para predecir en el estudio:

- **Área basimétrica (G):** Es la suma de las áreas de las secciones a altura a la altura normal (1,3m) de todos los arboles de una hectárea. Su unidad de medida es el m^2/ha.

- **Volumen con corteza (VCC):** Es el volumen maderable de un tronco desde la base hasta su diámetro mínimo (7,5cm) incluyendo la corteza. Su unidad de medida es el m^3/ha.

- **Número de pies (N):** Es la cantidad de arboles que hay en una hectárea. Su unidad de medida es el numero de pies por hectárea.

Esta información se encuentra en distintas tablas de la base de datos (Access), que nos proporciona el IFN4 en su documentación.

A través de consultas SQL sobre la base de datos se han obtenido los datos por parcela. En el ANEXO 1 se presentan las distintas consultas hechas. Luego con el software de ArcMap se han proyectado dichas parcelas con la herramienta "*X Y table to point*" y se ha generado un "*buffer*" de 25 metros para la obtención de un archivo "*shapefile*" con el polígono circular de cada parcela.

El IFN4 para delimitar los centros de las parcelas de muestreo de la malla reticular de 1 x 1 km, comentada anteriormente, utiliza un GPS de la marca Garmin que tiene una precisión de 5-10 metros. Este dato tiene una gran importancia en nuestro estudio. Si el GPS para determinar el centro de la parcela tiene esta precisión, cabe esperar que la coincidencia entre la parcela de muestreo en el IFN y la nube de puntos LIDAR puede ser casi nula. Aun así, la tipología de un hayedo suele ser regular y también se puede esperar que los valores que da el IFN para los parámetros de G, N y VCC son muy similares en la zona colíndate, al centro real de la parcela, teniendo en cuenta el desplazamiento de la parcela. A lo largo del trabajo este dato tendrá una importancia especial y se irá mencionando. En el apartado de resultados se comprobará si realmente es determinante o no para nuestro estudio.

2.2.3. Método de selección de parcelas válidas.

Una vez proyectadas las parcelas se ha realizado una validación de éstas, para detectar posibles errores que puedan afectar en la calidad de los modelos. Se ha comprobado si la posición de las parcelas coincidía con un hayedo de fracción de cabida cubierta mayor del 70%, como indica el IFN, en la descripción de este estrato. Para preseleccionar las parcelas y agilizar el proceso de comprobación visual, se han utilizado los propios datos LIDAR de la extracción de métricas forestales (*metodología explicada en el apartado 2.3.3*), en función de:

- **Parámetro de densidad (dns):** Indica qué parcelas tienen poca densidad. Es de esperar que, si la definición del estrato es de una densidad mayor del 70%, este valor tiene que ser alto.
    - Criterio utilizado: Parcelas válidas = dns < 40 (Antes se comprobó con una dns < 70, pero discriminaba muy poco y se cambió a un valor de 40 para agilizar el proceso de búsqueda de parcelas de una manera más automática.)

- **Parámetro de curtosis (kur):** Nos dice si la distribución de las alturas de los puntos LIDAR se acerca a una distribución normal. Se espera que en un hayedo con FCC mayor de 70 este parámetro sea alto debido a la homogeneidad de este tipo de masas.
    - Criterio utilizado: Parcelas válidas = kur > 2.5

De las 830 parcelas que tiene el estrato 01 del IFN4 salieron en la preselección 86 parcelas no válidas. Que se comprobaron de manera visual, para saber si realmente sus características coincidían con las del estrato 01. Para ello se utilizó la ortofoto del mismo año del inventario

(''Navarra2012_Ortofoto25cm_ETRS89.ecw''), superponiendo el buffer de parcelas encima. El criterio visual utilizado fue que las parcelas tenían que estar encima de una masa forestal densa que no coincidieran con zonas abiertas, caminos o carreteras. Finalmente, de las 84 parcelas preseleccionadas se descartaron totalmente 19 parcelas por no cumplir dichos criterios (*figura 5*). Con un resultado final de **811 parcelas válidas**.

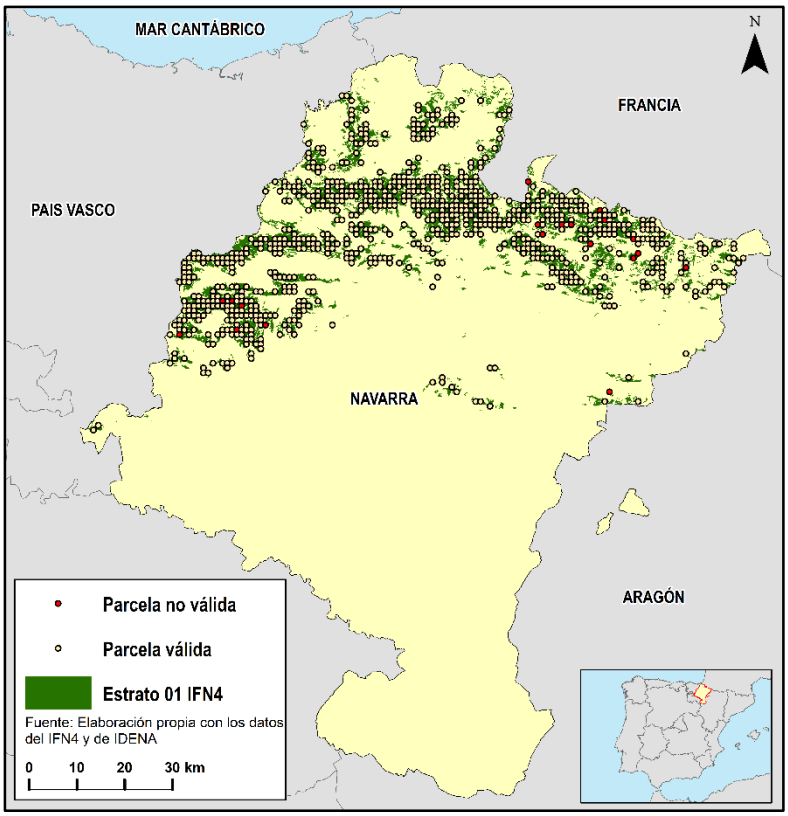

Figura 5: Discriminación de parcelas válidas y no válidas. Fuente: Elaboración propia.

2.2.4. Estratificación de los datos.

Inicialmente, se empezó a realizar la modelización con todas las parcelas válidas. Pero rápidamente se observó que los modelos que salían no eran buenos. Entonces se decidió realizar una estratificación de datos, creando regiones con características parecidas entre sí para ajustar mejor los modelos.

La primera estratificación se basa en uno de los propios parámetros que nos proporciona la herramienta "lascanopy" del software de LasTools, cobertura de la masa (cov) (*metodología explicada en el apartado 2.3.3*).

- Densidad alta: cov > 80
- Densidad media: cov > 50 y cov < 80
- Densidad baja: cov < 50

La segunda estratificación se basa en parámetros biológicos y paisajísticos. Básicamente se divide el conjunto de datos en función de las tipologías de hayedo que existen en navarra. Definidos en el apartado 2.1.

En la primera selección de parcelas, directamente de la base de datos del IFN se encuentran 830 parcelas. Luego con la discriminación de parcelas válidas y no validas, explicada en el apartado 2.2.3., se comprobaron individualmente 86 parcelas. De las cuales se descartaron 19, quedando finalmente 811 Parcelas válidas.

Las dos estratificaciones que se hicieron, para adaptar mejor los modelos, han dado los siguientes resultados:

- Estratificación por el parámetro (cov):
    - Densidad alta: 321 parcelas
    - Densidad media: 211 parcelas
    - Densidad baja: 278 parcelas

- Estratificación por tipología de hayedo:
    - Hayedo occidental: 261 parcelas
    - Hayedo oriental: 184 parcelas
    - Hayedo meridional: 16 parcelas
    - Hayedo septentrional: 349 parcelas

En la modelización de G, N y VCC se han utilizado estas dos estratificaciones y también el conjunto de datos entero, para cada una de las tres variables.

2.3. Datos LIDAR.

2.3.1. Introducción al LIDAR-PNOA.

El Plan Nacional de Ortofotografía Aérea (PNOA) tiene como objetivo obtener ortofotografías aéreas y modelos digitales de elevación de todo el territorio español con una periodicidad de 2 o 3 años. Los vuelos con cámaras digitales fotogramétricas se iniciaron en el año 2004 y luego ya en el año 2008 se incorporaron vuelos con sensores LIDAR. Los datos de los vuelos LIDAR-PNOA son de baja densidad (0,5 puntos/m^2). El PNOA introdujo por primera vez la cartografía derivada de la tecnología LIDAR en su plan de 'Sistema Nacional de Cartografía de Zonas Inundables'. Las principales necesidades que querían resolver introduciendo la tecnología LIDAR eran la actualización de modelos digitales del terreno (MDT), con datos de una precisión mayor, la generación de modelos 3D de vegetación y edificaciones. Y también satisfacer las necesidades de usuarios y organismos que demandaban información altimétrica de altas precisiones (IGN, 2016).

2.3.2. Vuelo LIDAR Navarra 2011-2012.

Para este estudio se han utilizado los datos del vuelo que se hizo en Navarra durante los años 2011 y 2012. Las características de este vuelo se resumen en la siguiente tabla elaborada a partir del pliego de vuelo:

Sistema de referencia	ETRS89 con proyección UTM 30N. Para cambiar de cotas elipsoidales a ortométricas se utilizará el modelo de geoide EGM2008-REDNAP (Adaptación del geoide mundial EGM08 a España)
Tamaño de la hoja LIDAR	2 x 2 km
Sensor	ALS60 de Leica
FOV	Máximo de 50º efectivos
Resolución espacial. Densidad promedio	0,5 puntos del primer retorno por metro cuadrado. Sin considerar los puntos de solape entre pasadas.
Calibración del sensor	Antigüedad ≤ 12 meses.
Resolución radiométrica	8 bits
Retornos	Capacidad de captar hasta 4 retornos
Fechas de vuelo, horario y condiciones meteorológicas	Las condiciones meteorológicas no pueden afectar la operatividad del sistema, sobre todo si se realiza el vuelo LIDAR juntamente con el vuelo fotogramétrico.
Recubrimiento transversal	Margen mínimo de 15% en el extremo superior e inferior.
Precisión altimétrica RMSE	≤ 0,20 m
Clasificación de los puntos	Realizada con el software de TerraScan

Tabla 3: Características técnicas del vuelo LIDAR-PNOA de Navarra.

La descarga de los datos de este vuelo se puede hacer desde dos sitios distintos la *ftp* del gobierno de Navarra y el portal del CNIG del instituto geográfico nacional. En este caso se han descargado del portal del CNIG. Cada hoja LIDAR tiene un tamaño medio de unos 35 MB. La obtención del número de serie de cada hoja se ha hecho con una selección por localización con el software de ArcMap, entre la capa de puntos de los centros de las parcelas del estrato uno del IFN y la malla de datos LIDAR 2x2 km. Finalmente se han tratado 914 hojas con un peso total de 30,1 GB.

2.3.3. Extracción de las métricas forestales.

Para la extracción de las métricas forestales se ha utilizado el software LAStools, creado por Martin Isenburg y escrito en el lenguaje C++. Esta librería se puede ejecutar mediante línea de comandos y también se puede integrar como un paquete de herramientas en ArcGis y Qgis. Para utilizarla no hace falta comprar la licencia, pero se tiene que tener en cuenta que a partir de un tratamiento de dos millones de puntos los resultados que obtenemos tendrán una marca de agua, por ejemplo, si exportamos un modelo digital del terreno. En el trabajo se utiliza de forma gratuita asumiendo que en los resultados tendremos la marca de agua.

El proceso de datos se ha hecho mediante un archivo batch que va encadenando distintas herramientas de LAStools con sus respectivos resultados. En la *figura 6* se representa este proceso de los datos LIDAR en forma de diagrama de flujo. La dinámica general del archivo batch es que para cada herramienta que se utiliza se va creando una carpeta nueva donde se almacena el resultado y la siguiente herramienta utiliza la carpeta del resultado de la anterior como entrada de datos y así sucesivamente hasta exportar como CSV el resultado final. El tiempo de procesado ha sido de **18 horas y 44 minutos** con un PC de 2 CORES de 3.00GHz y 8 GM de RAM, generando un volumen total de datos de **255 Gb**.

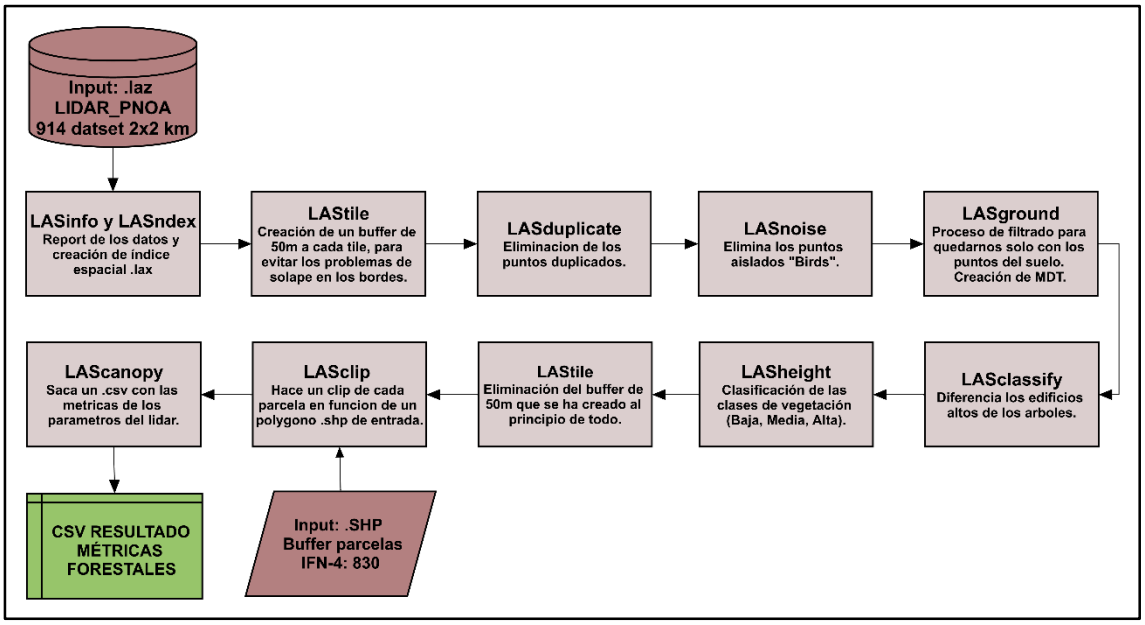

Figura 6: Diagrama de flujo del tratamiento de los datos LIDAR para sacar las métricas forestales con el software de LAStools (elaboración propia).

A continuación, se explica detalladamente cada herramienta utilizada del LAStools:

- *LASinfo:* saca archivo de texto con un resumen de las propiedades técnicas de cada hoja original del LIDAR-PNOA. En total son 914 hojas que se han procesado mediante un script de Python para sacar un resumen general de los valores medios más relevantes (*tabla 4*).

Sumatorio de todos los puntos	11.532.415.744
Media de los puntos de cada hoja	1.2617.522,7
Densidad media del conjunto de puntos	1,42 puntos/m^2
Densidad media de los últimos rebotes (suelo)	1,15 puntos/m2
Porcentaje de puntos de primero retorno	35,60%
Porcentaje de puntos de retornos intermedio	0,84%
Porcentaje de puntos de último retorno	35,59%
Porcentaje de puntos con un solo retorno	27,98%

Tabla 4: Resumen de los 914 archivos LIDAR de entrada.

- *LASindex*: Crea un archivo .lax de indexación espacial. Esto sirve para agilizar los procesos internos del LAStools. Con este archivo el programa indexa en memoria cada parcela LIDAR y hace más agiles los procesos posteriores, sobre todo cuando se trabaja con un volumen de datos tan grande.

- *LAStile*: Redimensiona los archivos de entrada con un buffer de un tamaño especificado. Para realizarlo coge los puntos de las hojas colindantes a la hoja que se esta procesando. La finalidad de esta herramienta es evitar la imprecisión de las zonas de frontera entre hojas. El buffer creado ha sido de 50 metros.

- *LASduplicate*: Elimina los puntos duplicados que coinciden en XYZ debido a la superposición de las pasadas. Los puntos eliminados se pueden almacenar en un nuevo archivo.

- *LASnoise*: Elimina o clasifica los puntos considerados como ruido (los llamados "*birds"*). Por defecto los puntos de ruido quedan clasificados en el nivel 7. En este caso el algoritmo de la herramienta comprueba los puntos que sobresalen de la media, mediante un análisis por vecinos más cercanos. Si se encuentra puntos que se desvían mucho de la altura media y no hay otros puntos cercanos a los que le pasa lo mismo, se clasifica el punto como ruido.

- *LASground*: Clasifica los puntos de suelo como clase dos y los puntos restantes como clase uno. El parámetro más importante que configurar en esta herramienta es el paso de malla. En la documentación de LAstools se aconseja que el paso de malla sea de 5 metros en bosques, 10 metros en terrenos sin vegetación y de 25 metros en ciudades. Otro parámetro para tener en cuenta son el '*fine*' y el '*extra_fine*' que incrementan la búsqueda inicial de puntos de suelo en zonas de montaña muy inclinadas. Para el trabajo se ha configurado un paso de malla de 5 metros y con el parámetro 'extra_fine'.

- *LASclassify:* Discrimina entre edificios altos y árboles. La herramienta necesita que previamente se hayan clasificado los puntos del suelo, en función de la diferencia de altitudes entre los puntos de suelo y los restantes (sin incluir los '*birds*') discrimina los edificios.

- *LASheight:* Clasifica cada punto LIDAR en función de la altura respecto a su punto de suelo más cercano, definido por un paso de malla. Como *LASclassify* para poderla usar primero se tiene que haber calculado *LASground*. En el trabajo se ha usado para clasificar los distintos tipos de vegetación: Vegetación baja (0,1 – 0,5 m), vegetación media (0,6 – 2 m) y vegetación alta (> 2 m).

- *LAStile:* En este paso quitamos el buffer de 50 metros que se había creado en el tercer paso del batch.

- *LASclip:* Recorta la superficie correspondiente a cada parcela del IFN que entra en el estudio. Para ello se le introduce como entrada el archivo *shapefile* con las parcelas del área de estudio. El resultado son 830 archivos *laz* correspondientes a cada parcela.

- *LAScanopy:* Cálculo de las métricas de la nube de puntos que hemos normalizado anteriormente en cada parcela. Las métricas que extraemos son las siguientes: altura mínima (min), altura máxima (max), promedio de las alturas (avg), promedio de altura cuadrática (qav), desviación estándar (std), curtosis (kur), los percentiles: (p01), (p05), (p10), (p25), (p50), (p75), (p90), (p95) y (p99). La fracción de cabida cubierta (cov) y la densidad de la masa (dns). El resultado de este proceso es un archivo CSV con el número identificador de la parcela (extraído del shapefile) y los valores de las métricas forestales.

2.4. Regresión estadística.

La regresión es una técnica estadística para estimar las relaciones entre una variable dependiente (Y) y una o más variables independientes (X). Su objetivo principal es determinar una función matemática que prediga los valores de Y (incógnita) mediante los valores de X (predictoras). En este trabajo la Y son las variables extraídas del inventario forestal nacional: Volumen con corteza (VCC), número de pies por hectárea (N) y área basimétrica (G). Las variables de X son las métricas extraídas del LIDAR-PNOA mediante el tratamiento de los datos con LAStools.

2.4.1. Modelos de regresión con técnicas de minería de datos.

La minería de datos se define como la exploración y análisis de grandes cantidades de datos para descubrir patrones significativos utilizando medios automáticos o semiautomáticos (Berry & Linoff, 1997).

Se creyó oportuno modelizar con estas técnicas debido a la naturaleza del estudio. Como se ha ido comentando a lo largo de trabajo, se parte de un error muy grande en los datos de entrada por culpa principalmente de la precisión del GPS en la determinación del centro de las parcelas de campo del IFN. Por el contrario, disponemos de gran cantidad de datos para entrenar los algoritmos de regresión, en total 811 parcelas de muestro donde de cada una hay 18 variables independientes. Esto nos lleva a formular la hipótesis de que es posible sacar modelos óptimos con estas técnicas.

La implementación de este método se ha realizado mediante el lenguaje de programación Python *3.6* con la distribución de *Anaconda* y el intérprete *Spyder*. Para ello ha sido necesario utilizar la librería *Scikit-Learn* que incorpora los algoritmos de *Machine Learning* utilizados en el estudio: KNearest Neighbords (KNN), Support Vector Regressor (SVR), Random forest y Adaboost.

Los algoritmos utilizados parten de tres naturalezas distintas. KNN se basa en los vecinos cercanos, SVR en las maquinas de soporte vectorial y RandomForest y Adaboost son *ensembles*.

- **KNN, k-Nearest Neighbords:** Se basa en el aprendizaje por analogía, encuadrado en las metodologías del *lazy learning*. El algoritmo se aprende todos los datos de entrenamiento y realiza la predicción de los nuevos datos buscando los valores de los datos más similares. Una de las características más importantes es que este algoritmo no crea ninguna función o formula en su entrenamiento, solo almacena los datos y cuando se realiza una predicción calcula las distancias de los datos almacenados al nuevo dato. En su configuración se tiene que tener en cuenta:
    - El tipo de datos de entrenamiento que se le introducen, el tipo de medida de distancia (euclídea, manhatan y minkowski)
    - Valor de K que es el número de vecinos que se utiliza en la predicción

En este estudio los parámetros utilizados son los siguientes: distancia de minkowski y cinco vecinos cercanos y con el método automático.

- **SVR Support Vector Regressor:** se basan en los modelos de aprendizaje supervisado. Este método se encuentra dentro de los métodos más robustos y precisos de todos los conocidos en minería de datos. Su funcionamiento se basa en aproximar su predicción utilizando el mejor margen posible. El algoritmo construye un hiperplano en un espacio de N dimensiones para poder predecir los nuevos datos de entrada. En los parámetros de configuración que se tienen que tener en cuenta son los siguientes:

    - Epsilion: especifica el *epsilion-tube* dentro del cual no se asocia ninguna penalización si la función de perdida de entrenamiento con los puntos predichos queda dentro de una distancia determinada del valor real.
    - Kernel: Puede ser lineal, polinómico, sigmoide o rbf.

  En este estudio se ha usado un Epsilion de 0,1 con un kernel rbf.

- **Random Forest:** Este método pertenece a los ensembles. Un ensemble es una combinación de distintas modificaciones del mismo algoritmo de regresión, en este caso un árbol de decisión. Es uno de los modelos más efectivos para las clasificaciones, pero en este caso se utilizará su versión para regresión. Este es un modelo aditivo que realiza las predicciones combinando decisiones de una secuencia de modelos base. En su configuración se tienen que tener en cuenta:

    - El número de árboles máximo que tendrá el modelo final, cuantos más arboles mayor será la precisión de los datos de entrenamiento, pero es muy fácil que caiga en sobre-entrenamiento, y que luego al hacer una predicción sobre otros datos su precisión baje muchísimo.
    - La profundidad de nodos máxima de cada árbol individual. Como en el parámetro anterior se regula para no caer en sobre-entrenamiento.
    - Peso mínimo del ‘’hijo’’: son las observaciones mínimas que tiene que tener el resultado de un nodo. Como más grande sea este valor más simple será cada árbol.
    - Función para medir la calidad de la división de los nodos del árbol de decisión.

  En este estudio se han utilizado los siguientes parámetros: Máximo de 10 árboles de decisión, sin restricción de profundidad, con un peso mínimo de hijo de una sola unidad y con el error cuadrático medio (mse) para determinar la calidad de cada división.

- **Adaboost:** Al igual que el Random Forest también es un ensemble y como tal es una combinación de distintos métodos de aprendizaje. En este caso el algoritmo comienza ajustando una regresión a un conjunto de datos, inicialmente cada ejemplo tiene el mismo peso y conforme va aprendiendo modelos de regresión los pesos de los ejemplos se modifican en función de su error de predicción. Cuando más error tenga cada ejemplo mayor será el peso que tendrá dentro del algortimo y cuanto menos error tenga más pequeño será el peso. Entonces los regresores posteriores se centran en predecir los casos más difíciles. Es fundamental que el algoritmo que se implementa en el Adaboost pueda manejar pesos. Para configurarlo hay que tener en cuenta:

- Estimador base a partir del que se implementa el conjunto.
- Máximo de repeticiones hasta detener el entrenamiento.
- Función de perdida que actualizar los pesos después de cada iteración (lineal, cuadrática y exponencial)

Para el estudio se ha configurado el Adaboost con un estimador base de árbol de decisión, con un máximo de 50 repeticiones y una función de perdida lineal.

2.4.2. Selección de variables con el método *stepwise*.

La selección de las variables independientes que entrenaran a cada se realiza con el método de "stepwise". Este método es un procedimiento automático que selecciona las mejores variables independientes (LIDAR - PNOA) en función de la variable dependiente (IFN). Su procedimiento consiste en ir comprobando paso a paso como influye una variable en la predicción de una regresión. En este caso se comprueba mediante la colinealidad de cada variable independiente.

Existen dos métodos de *stepwise*:

- *Forward*: Se trata de un método aditivo de selección paso a paso de variables. Se empieza con un modelo de una sola variable, la que presenta mayor colinealidad con la variable dependiente. A cada paso, se identifica la variable que presenta la mayor colinealidad con los errores del modelo del paso anterior y se la añade al modelo. Se comprueba que la variable añadida es significativa y se procede al siguiente paso. Si la variable añadida no es significativa, se para el proceso y se conserva el modelo del paso anterior.

  Puede ocurrir que, al añadir una nueva variable al modelo, otras variables presentes dejen de ser significativas. En aquel caso, estas variables deben ser eliminadas del modelo antes de seguir añadiendo nuevas variables.

- *Backward:* Es un método discriminante de selección paso a paso de variables. El punto de inicio es el cálculo de un modelo incluyendo la totalidad de las variables independientes. A cada paso, se elimina la variable independiente menos significativa (de mayor p-value) y se recalcula el modelo. Este proceso sigue hasta que todas las variables del modelo sean significativas (p-value inferior a 5%).

En este caso se ha seleccionado una función extraída del portal (*https://datascience.stackexchange.com, 2018*) que realiza los dos tipos de selección quedándose con la mejor combinación de variables, comparando los dos métodos de selección.

2.4.3. Coeficiente de determinación

La elección de la técnica de minera de datos que se ajusta más a los datos de entrada se realizará mediante el coeficiente de determinación R^2_{aj}. A continuación, se muestra su fórmula.

$$R^2 = \frac{scE}{scG} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Este indicador sirve para medir la bondad de ajuste del modelo. Es la proporción de varianza de la variable dependiente que es predecible a partir de la variable o variables independientes que entran en el modelo. Indica la proporción de variabilidad de la variable dependiente respecto a la media que explica el modelo de regresión. Es un valor que oscila entre 0 y 1, como más se aproxime a 1 mejor explica el modelo la variable dependiente.

2.4.4. Validación cruzada de K-particiones con repetición

Para comprobar si los modelos de regresión genrados se ajustan a los datos de entrada, se utliza la validación cruzada de K-particiones con repetición. En la *figura 7* se resume mediante un diagrama de flujo el funcionamiento. A continuación se explica de una manera más detallda.

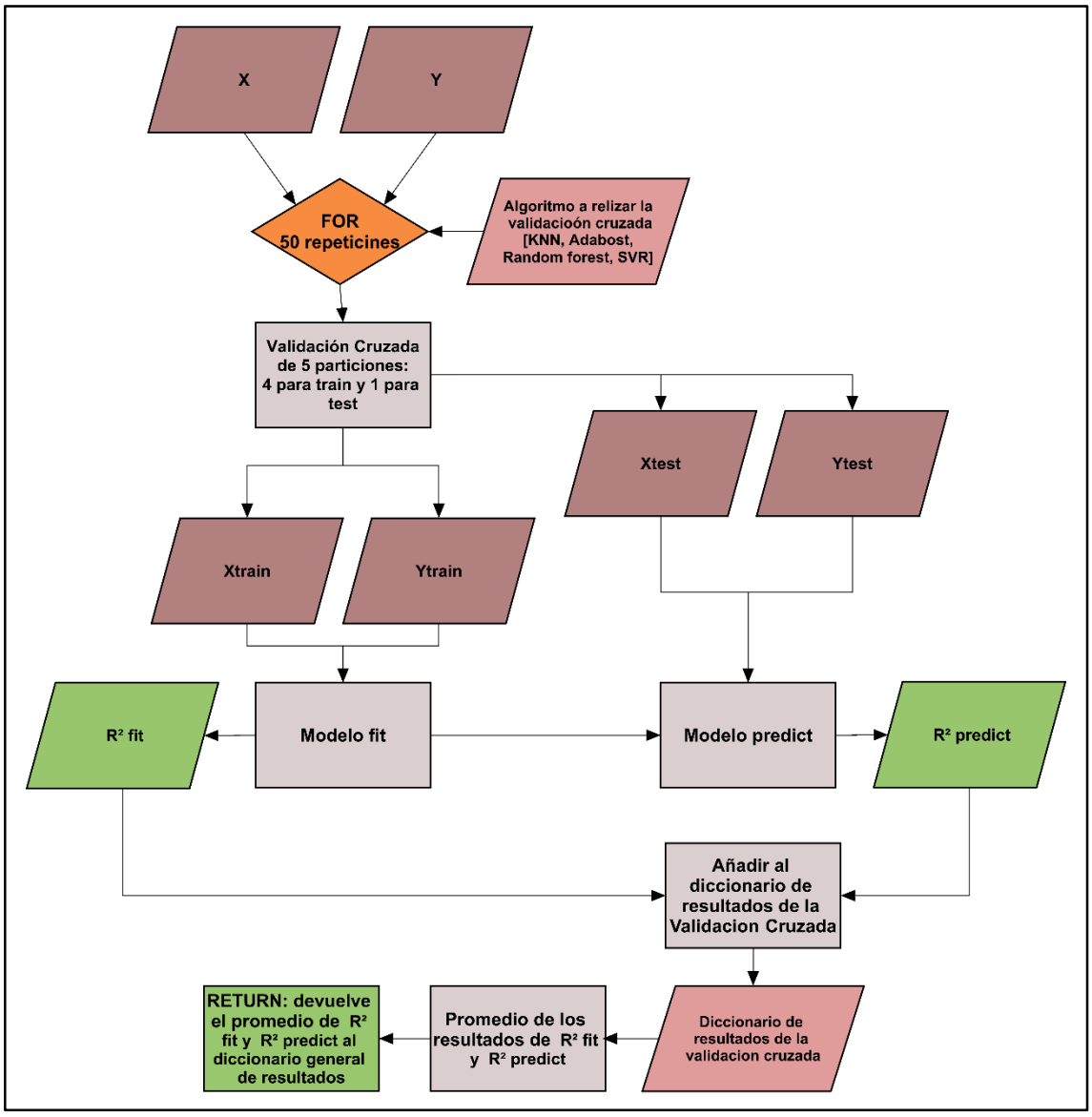

Figura 7: Diagrama de flujo de la validación cruzada de K-particiones repeticiones.

Para cada modelo generado se parte de manera aleatoria los datos de X e Y en cinco partes. Cuatro de ellas sirven para entrenar el algortimo y una para testear dicho algoritimo. En cada iteración se obtine el R^2_{aj} tanto de entrenamiento como de test. Este último dato es el que servirá

para determinar cuál es la calidad del modelo. Para eliminar la variabilidad que pueda ocasionar este cálculo en función de la aleatoriedad que pueda tener cada iteración, se calcula el promedio de R^2_{aj} de las 50 iteraciones para el resultado de TEST de la validación cruzada.

2.5. Explicación genérica del script.

El cuerpo principal de este trabajo se basa en la generación de un script con el lenguaje de programación Python 3.6 para la automatización del proceso de regresión. Nos encontramos con un error muy grande en la precisión del GPS del IFN para la geolocalización de las parcelas forestales, esto determina las características de este estudio. Antes de crear este script se probaron otras metodologías, como la regresión lineal mediante la función '*lm*' del software R, que no dieron buenos resultados. También la modelización mediante *machine learning* con Python utilizando cada uno de los algoritmos explicados en el apartado anterior de manera individual. La conclusión que se obtuvo de todas estas pruebas es que es prácticamente imposible modelar los datos de Y con semejante error de partida.

Es por este motivo que se tuvo que adaptar una metodología a medida para poder modelar estos datos con un ajuste adecuado. En la *figura 8* se muestra el diagrama de flujo que resume el funcionamiento del script. Su finalidad es probar varios tipos de técnicas de minería de datos, para cada variable a predecir (VCC, N y G), donde el script se queda con el mejor modelo en cada caso:

- Partimos de 3 variables a predecir:
    - VCC, N y G

- Se aplican 8 estratificaciones del área de estudio:
    - Área total, hayedo con densidad alta, hayedo con densidad media, hayedo con densidad baja, hayedo occidental, hayedo oriental, hayedo meridional y hayedo septentrional.

- Se realizan 2 tipos de procesado de las variables:
    - Transformación logarítmica y sin transformación.

- Se seleccionan las variables independientes de 2 maneras:
    - Selección stepwise y sin selección.

- Modelización con cuatro técnicas de minería de datos:
    - Random Forest, SVR, KNN y Adaboost

- Y finalmente se validan con una validación cruzada de 5 particiones que se repite 50 veces.

Para cada combinación de variable dependiente (VCC, G y N) y su respectiva estratificación de datos. El script selecciona el algoritmo de minería de datos que su modelo se ajusta más a los datos de entrada, según el mejor valor de R^2_{aj} en los datos de Test de la validación cruzada. Como hay cuatro tipos de combinaciones de datos de entrada:

- Variable dependiente **sin** transformación y **sin** selección de variables independientes.
- Variable dependiente **sin** transformación y **con** selección de variables independientes
- Variable dependiente **con** transformación logarítmica y **sin** selección de variables independientes.
- Variable dependiente **con** transformación logarítmica y **con** selección de variables independientes.

El resultado de este paso serán los cuatro algoritmos de minería de datos que se ajustan más a los datos de entrada para cada combinación de variable dependiente y estratificación de datos. A este resultado a lo largo del trabajo le llamaremos "resultado intermedio". Si multiplicamos los cuatro resultados para cada combinación de variable dependiente y estratificación da un resultado intermedio de **90 modelos**[1] (3 variables independientes * 8 estratificaciones * 4 combinaciones de datos de entrada). El script imprime este resultado en un *report* en formato CSV para poder ser interpretado fácilmente con una tabla de Excel.

Para obtener el resultado intermedio el script ha tenido que generar **19.200 modelos,** lo que serían **800 modelos** para cada combinación de variable dependiente y estratificación de datos. Ya que para cada combinación de variable dependiente y estratificación de datos el script realiza todos los pasos comentados anteriormente (2 tipos de procesado de variable dependiente * 2 tipos de selección de variables * 4 técnicas de minería de datos * 50 repeticiones de la validación cruzada = 800 modelos).

El *report* del resultado intermedio almacenan para cada algoritmo seleccionado tres R^2_{aj}: el primero deriva del resultado del modelo con todas las parcelas de entrenamiento. El segundo es el promedio de los R^2_{aj} de cada iteración de la validación cruzada para los datos de *Train*. Y el tercero es el promedio R^2_{aj} de cada iteración de la validación cruzada para los datos de *Test*. El script selecciona la mejor técnica de minera de datos en función de este último R^2_{aj}, ya que así se evita seleccionar los modelos que sobreentrenan demasiado con los datos de entrenamiento y al aplicar ese modelo generado con los datos de entrenamiento a los datos de test tiende ajustarse muy poco.

El **tiempo total de ejecución** del script con un PC de 2 CORES de 3.00GHz y 8 GM de RAM es de **1h 44 minutos**.

Es interesante guardar la preselección del resultado de los 90 modelos para observar cual es el comportamiento de los algoritmos seleccionados en la modelización y en la validación cruzada. Pero el objetivo de este trabajo es decidir el mejor modelo para cada variable dependiente (VCC, G, N) y estratificación de datos (sin la combinación de datos de entrada). Por lo que el resultado final será el modelo que tiene el mejor R^2_{aj} de validación cruzada en los datos de Test de las cuatro posibilidades en la combinación de datos de entrada. Un resultado final de **24 modelos**, uno para cada variable de dependiente (VCC, G, N) y estratificación de datos de entrada.

1 – La multiplicación da un valor de 96, pero solo ha dado lugar a 90 modelos porque el hayedo meridional solo cuenta con 16 parcelas de muestreo y la función de selección de las variables no se ha podido implementar con tan pocas parcelas.

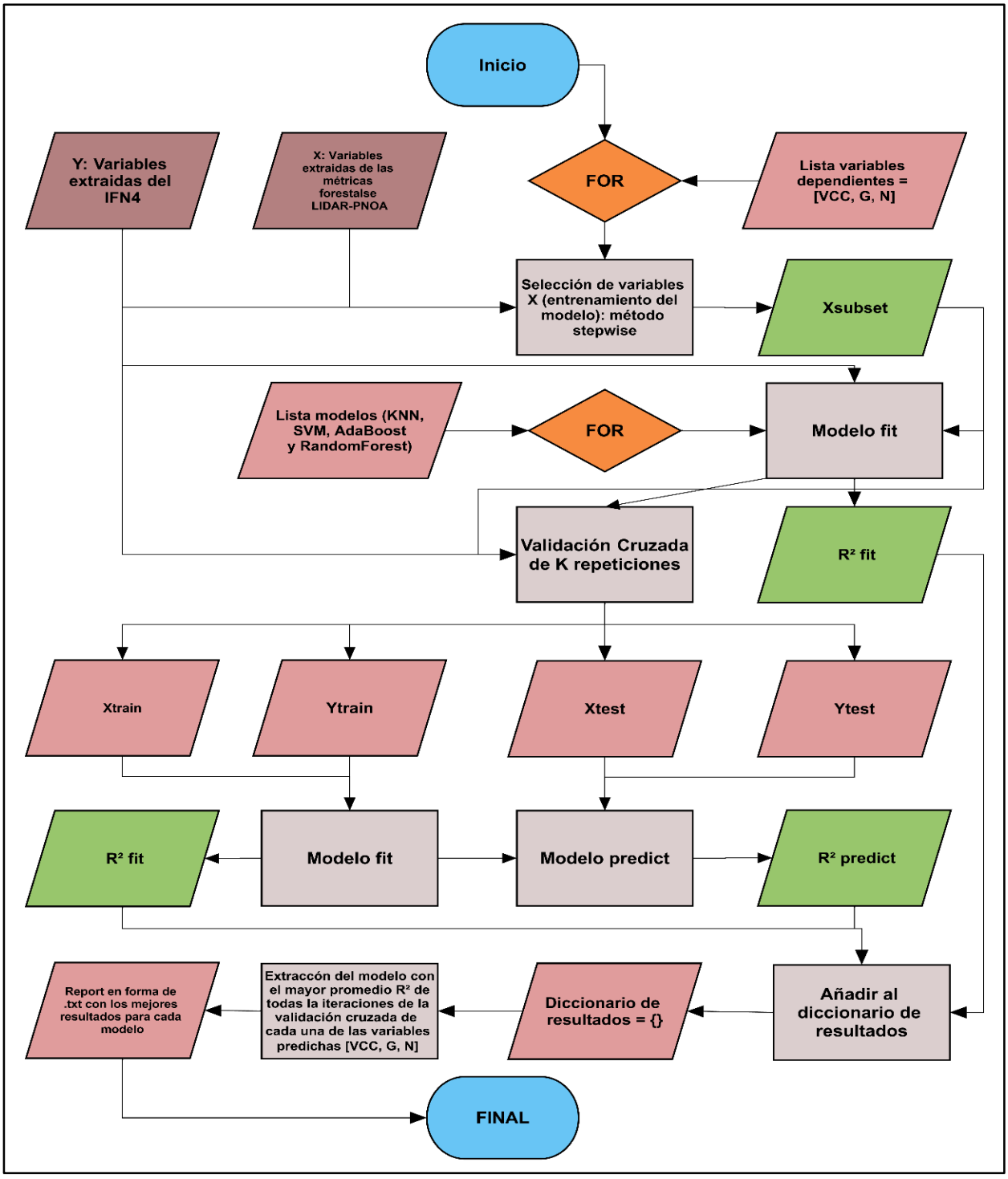

Figura 8: Diagrama de flujo que resume el funcionamiento del script de Python para la automatización del proceso de regresión.

3. Resultados y discusión.

3.1. Resultados de las métricas LIDAR.

El resultado de las métricas LIDAR es un CSV de 830 filas, donde cada fila corresponde a los resultados de cada parcela del IFN que entra en el estudio. Luego se quitan las parcelas que se han descartado buscando el número de identificación. Las columnas del CSV son el resultado de cada métrica (min, max, avg, qav, ske, kur, p01, p05, p10, p25, p50, p75, p90, p95, p99, cov y dens).

A continuación, en la *tabla 5* se resumen los resultados de las parcelas. Para poderlo hacer se creo un script de Python para que calculara los valores de media y desviación estándar de cada parámetro.

Variable	Media	Deviación estándar
min	2,36 m	1,61
max	23,72 m	5,57
avg	14,06 m	5,31
qav	250,22 m	163,56
std	4,64 m	1,59
ske	-0,44	0,91
kur	3,75	2,88
p01	3,68 m	3,07
p05	6,07 m	4,68
p10	7,73 m	5,25
p25	10,91 m	5,89
p50	14,4 m	5,94
p75	17,5 m	5,77
p90	19,65 m	5,54
p95	20,71 m	5,58
p99	22,32 m	5,52
cov	66,93 %	23,67
dns	56,14 %	24,41

Tabla 5: resumen de los promedios y desviaciones estándares de los datos LIDAR para el conjunto de las parcelas de muestro

Observando los resultados se puede ver que en general son bastante lógicos. Existe una correlación alta entre los valores de los percentiles, desde el p01 al p99 la métrica de alturas va en aumento. Además, la desviación estándar mantiene unos valores bastante homogéneos en todos los casos. Llama la atención los parámetros de cov y dns (fracción de cabida cubierta y densidad), el IFN define el estrato de estudio como un hayedo de más de 70% de fracción de cabida cubierta. Según estos resultados esto no sería del todo cierto, porque tanto el parámetro de cov como el de dns, tienen unos valores más bajos de los esperados.

3.2. Resultado de los modelos.

Para analizar los resultados de los modelos de regresión se hará en dos partes. En primer lugar, se analizarán los resultados del resultado intermedio donde entran todos los modelos que superaron a los otros algoritmos en la validación cruzada, 90 modelos. Después se analizará los 24 modelos finales, donde solo quedan los que explican más cada variable dependiente por estratificación de área de estudio.

3.2.1. Resultado intermedio.

A continuación, en la *tabla 6*, se muestra un ejemplo de la estructura del resultado intermedio para una de las variables dependientes y uno de los estratos. Con la finalidad de poner el lector en contexto y que le sea más fácil la interpretación posterior. En este caso se muestran los resultados de la variable dependiente de VCC para el hayedo oriental.

Algoritmo	Y	Subset	Transformación	X	R2aj Modelo	R2aj test CV	R2aj train CV	(R2ajTrain - R2ajTest)
AdaBoostRegressor	VCC	hayedo_oriental	Sin transformación	variablesSinSeleccion	0.26	0.73	0.69	0.47
AdaBoostRegressor	VCC	hayedo_oriental	Sin transformación	['std','kur','p50','cov']	0.25	0.67	0.65	0.42
KNeighborsRegressor	VCC	hayedo_oriental	Log	variablesSinSeleccion	0.22	0.51	0.50	0.28
KNeighborsRegressor	VCC	hayedo_oriental	Log	['qav','kur','p50','p90','cov']	0.21	0.50	0.50	0.29

Tabla 6: Ejemplo de la estructura de los resultados para la variable dependiente de VCC en el hayedo oriental.

Para cada una de las tres variables dependientes estudiadas (VCC, G y N) y para cada una de las ocho estratificaciones se analizan cuatro modelos como el de la *tabla 6*. En la primera columna se muestra el algoritmo que ha predicho los resultados para la variable "Y" de la segunda columna. Luego se indica cual ha sido la estratificación de los resultados y su transformación. La columna de "X" indica las métricas del LIDAR que entran en la modelización cuando pone "variablesSinSeleccion" son los modelos donde no se ha hecho selección stepwise, y por lo tanto entran en el modelo todas las variables del LIDAR. Las cuatro últimas columnas pertenecen a los datos numéricos del resultado. Se muestra en el R^2_{aj} de la validación cruzada (CV) en test y train, el R^2_{aj} del modelo sin CV y finalmente la diferencia entre el R^2_{aj} de train y el R^2_{aj} de test de la CV. Este ultimo dato se ha calculado para saber los algoritmos que sobreentrenan más los modelos con los datos de entrenamiento y luego en la posterior validación con los datos de test dan unos peores resultados, como más alto es este valor más sobreentrenado esta el algoritmo.

En la *tabla 7* se resume cuales han sido los mejores algoritmos, elegidos por el script, para cada tipo de variable dependiente estudiada:

Algoritmo	VCC	VCC %	G	G %	N	N%	Total	Total %
AdaBoostRegressor	22	73.3	19	63.3	16	53.3	57	63.3
SVR	3	10	10	33.3	5	16.7	18	20
KNeighborsRegressor	4	13.3	1	3.3	7	23.3	12	13.3
RandomForestRegressor	1	3.3	0	0	2	6.7	3	3.3

Tabla 7: Resumen de los algoritmos que ha elegido el script para cada variable dependiente.

Adaboost ha sido el algoritmo que más veces ha modelado mejor los resultados. Destacando sobre todo en la variable de VCC donde ha sido el mejor predictor en el 73,3% de los casos. En la variable G también destaca el Adaboost con un 63,3% de los casos, seguido del SVR que tiene ha sido elegido una de cada tres veces. La variable de N es la que las diferencias son más

pequeñas, adaboost ha sido el mejor en un poco mas del 50% de las veces, mientras que en segundo lugar ha quedado el KNN seguido de cerca por el SVR. El algoritmo Random Forest solo ha sido elegido tres veces como el mejor.

A continuación, en la *tabla 8* se intenta explicar porque Random Forest ha sido el peor algoritmo en este estudio.

Algoritmo	Promedio de la diferencia entre (R^2_{aj} train - R^2_{aj} test)	Desviación estándar de la diferencia (R^2_{aj} train - R^2_{aj} test)
AdaBoostRegressor	0.48	0.12
SVR	0.35	0.37
KNeighborsRegressor	0.39	0.13
RandomForestRegressor	0.84	0.02

Tabla 8: Resumen de las diferencias de resultados entre los datos de Train y Test en la CV.

En la segunda columna de la *tabla 8* se muestra el promedio de la diferencia entre R^2_{aj} de train y de test. Este calculo muestra como cambia el R^2_{aj} en los datos de testeo respecto a su resultado en los datos de entrenamiento. Un algoritmo que tiene un R^2_{aj} muy alto con los datos de train luego tiene un valor muy bajo en los datos de test. Esto es lo que le pasa al algoritmo de Random Forest, el modelo generado tiene un ajuste muy bueno para los datos de train pero luego al hacer la validación de dicho modelo con los datos de test se ajusta muy poco, debido al sobreentrenamiento. Además, la desviación estándar de la diferencia entre R^2_{aj} train y R^2_{aj} test es la más baja de todos los algoritmos, esto indica que la variación del sobreentrenamiento es muy baja y que por lo tanto es constante y repetitiva. Fijando la observación en los otros algoritmos, el SVR es el que menor diferencia de R^2_{aj} tiene, aunque es el que mayor en desviación estándar. Esto es porque tiene modelos donde la diferencia entre R^2_{aj} en train y test es muy poca, pero luego tiene modelos donde esta diferencia es muy grande. En resumen, se podría afirmar que es el algoritmo más irregular. El KNN también tiene una diferencia baja de R^2_{aj}, normal por la propia tipología del algoritmo, que básicamente almacena los resultados de entrenamiento y no crea ninguna formula matemática para predecir. La desviación de la diferencia de R^2_{aj} es baja por lo que se puede afirmar que es bastante regular. Finalmente, el Adaboost tiene un promedio de diferencia de R^2_{aj} bastante medio-alto y se confirma también es regular por la desviación por el valor que tiene la desviación estándar.

3.2.2. Resultado final.

De los cuatro valores que se obtienen para cada combinación de variable independiente y estratificación del aparatado anterior. Se selecciona el mejor modelo en función del que tenga un mayor ajuste de R^2_{aj} (cercano a uno) en la comprobación de la CV para los datos de test. En la *tabla 9,* se muestran los 24 resultados finales y sus características.

Para analizar los resultados primero nos fijaremos en la columna del R^2_{aj} test, la columna que ha marcado el criterio de selección de los mejores modelos. A modo de visión general los modelos resultantes no son significativamente buenos, no hay ninguno que supere el valor de 0,5 en el resultado del R^2_{aj} para los datos de test. Aún así se pueden derivar un seguido de observaciones interesantes.

La variable dependiente que mejor se explica con los resultados obtenidos por los modelos generados por el script es el VCC. Los 5 mejores modelos pertenecen a esta variable en 5 estratificaciones de datos distintas. La segunda variable mejor explicada es la N, pero ya tiene unos resultados bastante peores que los de VCC. Y finalmente, muy cerca de los resultados de N, pero un tanto peor encontramos la variable G.

La estratificación de datos de entrada que ha dado mejores resultados es el hayedo occidental (*tabla 10*). En las tres variables dependientes es la que tiene el valor más alto de R^2_{aj}. La segunda mejor estratificación ha sido la del hayedo de densidad media. Fijándonos en la propia tipología de esta estratificación podemos decir que va a concorde con el hayedo occidental. Las dos estratificaciones provienen de dos tipologías distintas, la primera se hizo con los criterios biológicos de las tipologías del hayedo en Navarra y la segunda con los criterios de los propios datos del LIDAR. El hayedo occidental, por sus características es un hayedo de densidad media ya que su distribución se encuentra, en gran parte, en zonas menos húmedas que el resto de los hayedos de Navarra. Por lo que hace que las características de las dos estratificaciones sean parecidas. Seguidamente vine la estratificación de datos donde entran todas las parcelas (*todo*), esta estratificación ya tiene unos valores de R^2_{aj} bastante peores, pero cabe destacar que en el estrato están las 811 parcelas y que el tercer mejor resultado por estratificaciones sea este nos da un dato un tanto revelador. Donde solo los hayedos de una densidad media se explican mejor con esta metodología, los otros hayedos de densidades más altas y más bajas no superan los resultados de todas las tipologías de hayedo juntas. De las otras estratificaciones lo más destacable es que el hayedo meridional es la estratificación que perores resultados ha dado, y con diferencia, quedando en última posición en todos los casos. Este hecho puede ser causado sobre todo porque solo tiene 16 parcelas de muestro. Y nos indica que para obtener buenos resultados en estos estudios se necesita un numero elevado de parcelas de muestreo.

Cambiando la mirada a la columna de R^2_{aj} del modelo (*tabla 9)*, donde entran todas las parcelas de muestreo sin la partición de datos que hace la CV. El mejor modelo es uno que ha quedado sobreentrenado, SVR para el parámetro G del hayedo denso, porque tiene un muy buen ajuste con los datos de Train, pero luego baja mucho su ajuste en la validación con los datos de Test. Siguiendo esta línea de análisis se puede observar como en todos los casos el ajuste de los datos de Train es mayor que el ajuste en los datos de Test.

Algoritmo	Y	Subset	Transformación	X	R2aj Modelo	R2aj test CV	R2aj train CV	(R2ajTrain - R2ajTest)
AdaBoostRegressor	G	hayedo_occidental	Sin transformacion	variablesSinSeleccion	0.539	0.174	0.585	0.411
AdaBoostRegressor	N	hayedo_occidental	Log	['max','std','ske','kur','p95','cov']	0.473	0.143	0.503	0.361
AdaBoostRegressor	VCC	hayedo_occidental	Log	variablesSinSeleccion	0.679	0.426	0.741	0.316
AdaBoostRegressor	G	hayedo_oriental	Sin transformacion	variablesSinSeleccion	0.583	0.106	0.650	0.543
KNeighborsRegressor	N	hayedo_oriental	Log	['max','p99','cov','dns']	0.362	0.031	0.374	0.343
AdaBoostRegressor	VCC	hayedo_oriental	Sin transformacion	variablesSinSeleccion	0.694	0.257	0.730	0.473
SVR	G	hayedo_septentrional	Sin transformacion	variablesSinSeleccion	0.141	-0.028	0.140	0.168
AdaBoostRegressor	N	hayedo_septentrional	Log	['qav','std','p01','p10','p95','cov','dns']	0.405	0.089	0.455	0.365
AdaBoostRegressor	VCC	hayedo_septentrional	Sin transformacion	['avg','qav','std','p01','p05','p25','p90','p95','cov']	0.429	0.124	0.451	0.327
SVR	G	high_density	Log	variablesSinSeleccion	0.914	-0.021	0.914	0.935
AdaBoostRegressor	N	high_density	Log	variablesSinSeleccion	0.521	0.135	0.559	0.424
AdaBoostRegressor	VCC	high_density	Sin transformacion	variablesSinSeleccion	0.524	0.129	0.548	0.419
SVR	G	low_density	Sin transformacion	['std','p10','p99']	0.206	0.057	0.197	0.140
AdaBoostRegressor	N	low_density	Log	['avg','std','p50']	0.392	0.046	0.438	0.392
AdaBoostRegressor	VCC	low_density	Sin transformacion	variablesSinSeleccion	0.628	0.241	0.666	0.424
AdaBoostRegressor	G	medium_density	Sin transformacion	['avg','std','p10','p75','p95']	0.419	0.131	0.427	0.296
AdaBoostRegressor	N	medium_density	Log	variablesSinSeleccion	0.450	0.056	0.520	0.464
AdaBoostRegressor	VCC	medium_density	Sin transformacion	['qav','std','kur','p75','p90','p95']	0.546	0.308	0.570	0.261
AdaBoostRegressor	G	todo	Sin transformacion	variablesSinSeleccion	0.259	0.068	0.283	0.215
AdaBoostRegressor	N	todo	Log	variablesSinSeleccion	0.290	0.093	0.311	0.217
AdaBoostRegressor	VCC	todo	Sin transformacion	['min','avg','qav','std','ske','kur','p25','p90','p95','cov','dns']	0.402	0.232	0.420	0.188
SVR	G	hayedo_meridional	Sin transformacion	variablesSinSeleccion	0.136	-2.581	0.127	2.708
KNeighborsRegressor	N	hayedo_meridional	Sin transformacion	variablesSinSeleccion	0.097	-4.067	0.115	4.182
SVR	VCC	hayedo_meridional	Sin transformacion	variablesSinSeleccion	-0.120	-2.215	-0.108	2.107

Tabla 9: Resultado final de los 24 mejores modelos. Cada modelo corresponde al que tiene un valor de R^2_{aj} en los datos de Test de la validación cruzada para cada variable dependiente y estratificación de datos.

Posición	Variable G	Variable N	Variable VCC
1	hayedo occidental	hayedo occidental	hayedo occidental
2	medium density	high density	medium density
3	hayedo oriental	todo	hayedo oriental
4	todo	hayedo septentrional	low density
5	low density	medium density	todo
6	high density	low density	high density
7	hayedo septentrional	hayedo oriental	hayedo septentrional
8	hayedo meridional	hayedo meridional	hayedo meridional

Tabla 10: Resumen de las posiciones que toma cada estratificación de datos para cada variable dependiente estilo ranking de mejores resultados.

En la *tabla 11* se muestra el resumen de la distribución de los mejores algoritmos para cada variable dependiente por número de estratificaciones. El algoritmo que más veces es elegido como el mejor es el Adaboost Regresor un 70,8% de las veces. Destacando por encima de los otros en la variable dependiente de VCC con un 87,5% de las veces. El algoritmo que menos veces ha sido elegido como el mejor ha sido el KNN, teniendo en cuenta que el Random Forest no ha entrado en ninguna ocasión como el mejor y sí que estaba en los resultados intermedios. La variable dependiente más disputada es la G donde se reparten las elecciones en un 50% de los casos entre el Adaboost Regresor y el SVR.

Modelo	VCC	VCC %	G	G %	N	N%	Total	Total %
AdaBoostRegressor	7	87.5	4	50	6	75	17	70.8
SVR	1	12.5	4	50	0	0	5	20.8
KNeighborsRegressor	0	0	0	0	2	25	2	8.3

Tabla 11: Resumen de la distribución de los mejores algoritmos por cada variable dependiente.

Por último, se analizarán las características de los datos de entrada que han determinado la configuración de los mejores modelos de predicción para cada combinación de variable y estrato. Las características analizadas son la selección de variables (selección stepwise o sin seleccion) y la transformación de ellas (transformación logarítmica o sin transformación). Curiosamente y sin una aparente relación han dado lugar a unos resultados muy similares. El 62,5% de los mejores modelos han sido sin la selección de variables y estas sin ser transformadas. Este dato sobresale para el algoritmo SVR donde le pasa en un 80% de las veces. Mientras que el KNN con solo dos modelos finales reparte a partes iguales sus resultados, al entrar solo dos veces como el mejor, no se puede considerar como un dato representativo. El Adaboost Regresor, el que ha aportado los mejores modelos al estudio, tiene un porcentaje de variables sin selección y sin transformación de 58,8%.

	Sin selección		Con selección			Sin transformación		Transformación log	
	Absoluto	%	Absoluto	%		Absoluto	%	Absoluto	%
AdaBoostRegressor	10	58.8	7	41.2		10	58.8	7	41.2
SVR	4	80.0	1	20.0		4	80.0	1	20.0
KNeighborsRegressor	1	50.0	1	50.0		1	50.0	1	50.0
Todos	15	62.5	9	37.5		15	62.5	9	37.5

Tabla 12: Resumen de las características de los datos de entrada para la configuración de cada algoritmo y del conjunto de ellos.

4. Conclusiones.

- En Navarra el bosque de hayedo denso conforma el 11,3% de la superficie de la comunidad foral. Es de vital importancia estudiar las dinámicas forestales que sufren este tipo de bosques con mecanismos innovadores y automatizados. Para poder hacer planes de gestión adaptados a las características tipológicas de cada bosque.

- El Inventario Forestal Nacional realiza medidas dasométricas periódicas en el total de la cobertura boscosa de España. La utilización de estos datos para la automatización de los cálculos dasométricos con técnicas de regresion mediante machine learning utilizando las variables independientes que podemos extraer de estadísticas LIDAR, puede ser un buen punto de partida para el abaratamiento de los costes de los estudios. Ya que los datos del IFN son públicos y de fácil acceso. Eso sí, se tiene que tener en cuenta que las precisiones de los centros de las parcelas son bajas, de cinco a diez metros de error. Con lo que comporta este hecho, sería interesante en futuros estudios realizar una validación previa más exhaustiva para poder implementar en el algoritmo algún tipo de corrección de este grave error.

- En España existe una cobertura LIDAR de baja densidad que puede ser utilizada para el desarrollo de estudios de gestión forestal con finalidades de abaratar y automatizar cálculos dasométricos. El tratamiento de estos datos con el software de LasTools permite gran automatización de los procesos mediante archivos batch, con unos resultados óptimos. Teniendo en cuenta que si sobrepasamos los dos millones de puntos el programa nos genera una marca de agua.

- Las herramientas de machine learning implementadas con el lenguaje de programación de Python 3 son de gran utilidad cuando se tienen que tratar tal volumen de datos con problemáticas tan claras, ya comentadas, en los datos de partida. Su configuración permite poder generar gran cantidad de modelos para ir validándolos de manera automática hasta obtener el modelo óptimo para nuestro objeto de estudio.

- Los distintos mecanismos utilizados en la búsqueda del mejor modelo de regresión: Estratificación de los datos de entrada, selección de variables, transformación de ellas, modelización con cuatro algoritmos y validación cruzada de ellos. Según los resultados obtenidos, son todos necesarios. Porque no existe un patrón claro de estipulación de los mejores parámetros. Aún así se pueden extraer algunas observaciones:

    - Para obtener unos resultados óptimos con esta metodología se tiene que tener un volumen alto de parcelas de muestreo.

    - El algoritmo Adaboost regresor es mejor predictor en un 70% de los casos.

    - La estratificación del hayedo occidental es la que da mejores resultados, probablemente porque es un hayedo menos denso y las pulsaciones del laser del LIDAR penetran con más facilidad.

    - Los métodos de selección y transformación de variables pueden ayudar a mejorar los modelos, aunque en la mayoría de los casos no son determinantes.

5. Bibliografía.

Ahmed, R., Siqueira, P., & Hensley, S. (2013). A study of forest biomass estimates from lidar in the northern temperate forests of New England. *Remote Sensing of Environment*, *130*, 121–135. https://doi.org/10.1016/j.rse.2012.11.015

Alexander, C., Korstjens, A. H., & Hill, R. A. (2017). Structural attributes of individual trees for identifying homogeneous patches in a tropical rainforest. *International Journal of Applied Earth Observation and Geoinformation*, *55*, 68–72. https://doi.org/10.1016/j.jag.2016.11.004

Barreiro-Fernández, L., Buján, S., Miranda, D., Diéguez-Aranda, U., & González-Ferreiro, E. (2016). Accuracy assessment of LiDAR-derived digital elevation models in a rural landscape with complex terrain. *Journal of Applied Remote Sensing*, *10*, 016014. https://doi.org/10.1117/1.JRS.10.016014

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. *Neuronal Information Processing - Letters and Reviews*, *11*(10), 203–224. https://doi.org/10.4258/hir.2010.16.4.224

Bullinaria, J. A. (2005). IAI: Machine Learning What is Machine Learning, 1–20.

Crespo-Peremarch, P., Ruiz, L. A., & Balaguer-Beser, A. (2016). A comparative study of regression methods to predict forest structure and canopy fuel variables from LiDAR full-waveform data. *Revista de Teledetección*, (45), 27. https://doi.org/10.4995/raet.2016.4066

Dale, D. (2017). Selección de variables stepwise. Retrieved from https://datascience.stackexchange.com/questions/24405/how-to-do-stepwise-regression-using-sklearn

Departamento de Desarrollo Rural y Medio Ambiente, C. foral D. N. (2009). ESTADO DEL MEDIO AMBIENTE EN NAVARRA 2009. *Gobierno de Navarra. Departamento de Desarrollo Rural y Medio Ambiente*. Navarra.

ESRI. (2016). What is Lidar Data. *ArcGIS Online*. Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/manage-data/las-dataset/what-is-lidar-data-.htm

Figueiredo, E. O., d'Oliveira, M. V. N., Braz, E. M., de Almeida Papa, D., & Fearnside, P. M. (2016). LIDAR-based estimation of bole biomass for precision management of an Amazonian forest: Comparisons of ground-based and remotely sensed estimates. *Remote Sensing of Environment*, *187*, 281–293. https://doi.org/10.1016/j.rse.2016.10.026

Garc, J., Mateos-garc, D., Riquelme-santos, J. C., & Miranda, D. (2011). A Comparative Study between Two Regression Methods on LiDAR Data: A Case Study, 311–318.

García-Gutiérrez, J., Martínez-Álvarez, F., Troncoso, A., & Riquelme, J. C. (2015). A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. *Neurocomputing*, *167*, 24–31. https://doi.org/10.1016/j.neucom.2014.09.091

García-Gutiérrez, J., Martínez-Álvarez, F., Troncoso, A., & Riquelme, J. C. (2015). A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. *Neurocomputing*, *167*, 24–31. https://doi.org/10.1016/j.neucom.2014.09.091

Gleason, C. J., & Im, J. (2012). Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment*, *125*, 80–91. https://doi.org/10.1016/j.rse.2012.07.006

Ii, N., Focus, F., & Clim, C. (2003). Salud en los bosques Salud en los bosques, 1–11.

Instituto Geográfico Nacional. (2014). Presentación PNOA-LiDAR. *Ministerio de Fomento de España*. Retrieved from http://www.ign.es/ign/layout/datosGeodesicos.do%5Cnhttp://www.ign.es/ign/layoutIn/actividadesGeodesiaStmagd.do

Instituto Geográfico Nacional. (2011). Especificaciones Técnicas para la realización del vuelo lidar que permita la obtención de información altimétrica en el ámbito de la Comunidad Foral de Navarra, 1–9.

Instituto Geográfico Nacional. (2012). Especificaciones Técnicas para la realización del vuelo lidar que permita la obtención de información altimétrica, 1–9.

Isenburg, M. (2018). LAStools. Retrieved from https://rapidlasso.com/lastools/

Jakubowski, M. K., Guo, Q., & Kelly, M. (2013). Tradeoffs between lidar pulse density and forest measurement accuracy. *Remote Sensing of Environment*, *130*, 245–253. https://doi.org/10.1016/j.rse.2012.11.024

Lo, C.-S., & Lin, C. (2013). Growth-Competition-Based Stem Diameter and Volume Modeling for Tree-Level Forest Inventory Using Airborne LiDAR Data. *Ieee Transactions on Geoscience and Remote Sensing*, *51*(4), 2216–2226. https://doi.org/10.1109/tgrs.2012.2211023

McRoberts, R. E., Gobakken, T., & Næsset, E. (2012). Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. *Remote Sensing of Environment*, *125*, 157–166. https://doi.org/10.1016/j.rse.2012.07.002

Montealegre, A. L., Lamelas, M. T., & De La Riva, J. (2015). Interpolation routines assessment in ALS-derived Digital Elevation Models for forestry applications. *Remote Sensing*, *7*(7), 8631–8654. https://doi.org/10.3390/rs70708631

Navarra, C. F. de. (2012). Documentación del mapa de usos del suelo en Navarra. El Hayedo en Navarra. Retrieved from http://www.cfnavarra.es/agricultura/informacion_agraria/MapaCultivos/usosfrondosashaya.html

Palminteri, S., Powell, G. V. N., Asner, G. P., & Peres, C. A. (2012). LiDAR measurements of canopy structure predict spatial distribution of a tropical mature forest primate. *Remote Sensing of Environment*, *127*, 98–105. https://doi.org/10.1016/j.rse.2012.08.014

Pearse, G. D., Dash, J. P., Persson, H. J., & Watt, M. S. (2018). Comparison of high-density LiDAR and satellite photogrammetry for forest inventory. *ISPRS Journal of Photogrammetry and Remote Sensing*, *142*(March), 257–267. https://doi.org/10.1016/j.isprsjprs.2018.06.006

Público, D. (2018). El gasto previsto en 2018 en medio ambiente es un 69% menor que antes de la crisis. Retrieved from https://www.publico.es/sociedad/presupuestos-gasto-previsto-2018-medio-ambiente-69-menor-crisis.html

Resop, J. P., Kozarek, J. L., & Hession, W. C. (2012). In Pr es s Pr, *78*(4), 1–9. https://doi.org/10.14358/PERS.81.12.21

ScikitLearn. (2018). Libreria de Python para míneria de bases de datos. Retrieved from http://scikit-learn.org

Silva, C. A., Klauberg, C., Hentz, A. M. K., Corte, A. P. D., Ribeiro, U., & Liesenberg, V. (2018). Comparing the performance of ground filtering algorithms for terrain modeling in a forest environment using airborne LiDAR data. *Floresta e Ambiente*, *25*(2). https://doi.org/10.1590/2179-8087.015016

Simpson, J. E., Smith, T. E. L., & Wooster, M. J. (2017). Assessment of errors caused by forest vegetation structure in airborne LiDAR-derived DTMs. *Remote Sensing*, *9*(11). https://doi.org/10.3390/rs9111101

UNEP-UNESCO-FAO. (1980). SITUACIÓN Y PERSPECTIVAS DE LA CONSERVACIÓN Y DESARROLLO DE LOS BOSQUES. Roma. Retrieved from http://www.fao.org/3/w9950s04.htm

Valbuena Rabadán, M. Á. (2013). Determinación de variables forestales de masa y de árboles individuales mediante delineación de copas a partir de datos LIDAR Aerotransportado. Aplicación a las masas de Pinus sylvestris L. en Álava, 342.

Vincent, G., Sabatier, D., Blanc, L., Chave, J., Weissenbacher, E., Pélissier, R., … Couteron, P. (2012). Accuracy of small footprint airborne LiDAR in its predictions of tropical moist forest stand structure. *Remote Sensing of Environment*, *125*, 23–33. https://doi.org/10.1016/j.rse.2012.06.019

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining* (1st ed.). Chapman & Hall/CRC.

Zhao, F., Strahler, A. H., Schaaf, C. L., Yao, T., Yang, X., Wang, Z., … Newnham, G. J. (2012). Measuring gap fraction, element clumping index and LAI in Sierra Forest stands using a full-waveform ground-based lidar. *Remote Sensing of Environment*, *125*, 73–79. https://doi.org/10.1016/j.rse.2012.07.007

6. Anexos

Anexo 1: Consultas SQL para la extracción de los datos de la base de datos del IFN4.

Consulta SQL para obtener los datos por árbol:

```
SELECT DISTINCT m.Estadillo, m.Cla, m.Subclase, m.Estrato, m.Especie, de.NOMBREIFN4, de.NOMBREIFN, p.CX, p.CY, m.VCC, m.G, m.nArbol, (m.Dn1+m.Dn2)/2 AS Diametro_normal
FROM Mayores_exs AS m, DescEspecie AS de, Parcela AS p
WHERE m.Estrato = '01' AND m.Especie = de.Especie AND p.Estadillo = m.Estadillo;
```

Consulta SQL para obtener los datos por parcela:

```
SELECT Estadillo, SUM(d.VCC_cor) AS VCC, SUM(d.G_cor) AS G, SUM(d.Parametro_correcion) AS nArbol, Cla, Subclase, Estrato, CX, CY
FROM Mayores_estrato01_corregidos AS d
GROUP BY Estadillo, Cla, Subclase, Estrato, CX, CY;
```