Universidad de Valladolid

GRADO EN ESTADÍSTICA

TRABAJO FIN DE GRADO

Evaluación de modelos para la predicción de la plaga Roya

Autora
Dilyana Valerieva Ivanova

Tutores
María Teresa González Arteaga
Gustavo Río Briones

A mi madre por ser el mejor modelo a seguir

Agradecimientos

Agradecer a mi madre por ser mi apoyo principal, por transmitirme valores como la responsabilidad y la perseverancia, por educarme y hacer de mí la persona que soy. Por supuesto, agradecer a mi hermano y al resto de mi familia por ser un ejemplo de superación.

Agradecer a mi novio, por las escapadas de desconexión y las escapadas de estudio. Por estar en todo momento a mi lado y ofrecerme su apoyo.

Agradecer a mis compañeros y amigos de carrera por acompañarme en este viaje de crecimiento profesional y personal.

Agradecer al Instituto Tecnológico Agrario y en concreto a mi tutor Gustavo Río Briones por ofrecerme la oportunidad de trabajar con profesionales como ellos, por introducirme en un campo tan apasionante, por el conocimiento transmitido, por su tiempo y dedicación.

Agradecer a la Universidad de Valladolid, a los profesores del Grado en Estadística y en particular agradecer a mi tutora María Teresa González Arteaga por su paciencia, por la ayuda ofrecida, por su tiempo y amabilidad.

Resumen

El presente Trabajo Fin de Grado se realiza en colaboración con el Instituto Tecnológico Agrario de Castilla y León (ITACYL) en el Plan Director de lucha contra las plagas agrícolas. La plaga en concreto que se estudia es la roya amarilla, una enfermedad que afecta a tres de las producciones principales en Castilla y León: el trigo, la cebada y el tripticale.

El objetivo principal del proyecto es el desarrollo de un modelo *machine learning* exhaustivo que permita predecir de forma temprana la aparición de roya amarilla. Para alcanzar el mismo es necesario integrar información procedente de diversas fuentes que incluyen dos determinantes clave en el desarrollo de la plaga: las condiciones meteorológicas, recogidas por las estaciones meteorológicas de la Agencia Estatal de Meteorología de España y por las estaciones meteorológicas de la red Inforiego del Instituto Tecnológico Agrario y la evolución del cultivo en distintas parcelas de Castilla y León registrada por los técnicos del ITACYL.

El desarrollo del trabajo sigue la metodología de ciencias de datos propuesta por *International Business Machines Corporation* o IBM e incluye: la comprensión del problema a abordar, una descripción del enfoque analítico a seguir, la descripción, limpieza y tratamiento de los datos, el modelado y la evaluación de los modelos.

Como resultado se obtiene un modelo de regresión con árboles de decisión bastante acertado y razonable que captura la tendencia de desarrollo de la plaga en los últimos meses de la campaña del cultivo.

Abstract

This Final Degree Project is carried out in collaboration with the Agrarian Technological Institute of Castilla y León (ITACYL) in the Master Plan for the control of agricultural pests. The specific pest under study is yellow rust, a disease that affects three of the main crops in Castilla y León: wheat, barley and tripticale.

The main objective of the project is the development of a comprehensive machine learning model that allows early prediction of the appearance of yellow rust. To achieve this, it is necessary to integrate information from various sources, including two key determinants in the development of the pest: meteorological conditions, collected by the meteorological stations of the Spanish State Meteorological Agency and by the meteorological stations of the Inforiego network of the Agrarian Technological Institute, and the evolution of the crop in different plots in Castilla y León, recorded by ITACYL technicians.

The development of the work follows the data science methodology proposed by *International Business Machines Corporation* or IBM and includes: the understanding of the problem to be addressed, a description of the analytical approach to be followed, the description, cleaning and treatment of the data, modeling and evaluation of the models.

The result is a fairly accurate and reasonable. The decision tree regression model captures the trend of pest development in the last months of the crop season.

Índice general

Índice de figuras

Índice de cuadros

1. Introducción

El presente Trabajo Fin de Grado se realiza en medio de un crecimiento desmesurado de cantidad de datos cuyo registro, a nivel mundial, pasa de dos zetabytes en 2010 a 64 zetabytes en 2020. En la última el número de datos se ha multiplicado por más de treinta, aún así, estas cifras no suponen nada en comparación con el 40 % de aumento anual esperado hasta 2025, un crecimiento del volumen de datos cuya gestión sin lugar a duda supone un reto para la sociedad [1].

En este entorno de desarrollo surge el concepto *Big data* que hace referencia precisamente a datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a mayor velocidad. Los datos poseen un valor intrínseco, sin embargo, no tienen ninguna utilidad hasta que dicho valor se descubra. Para que estos datos puedan aportar nuevas perspectivas y abran paso a nuevas oportunidades y soluciones es necesario integrar, gestionar y analizar los mismos con tecnologías especificas [2].

Los principales impulsores de este gran incremento de datos son el auge del Internet de las Cosas, el desarrollo de la 5G y la industria 4.0 entre otras. Dentro de esta última se encuentra la agricultura 4.0, una agricultura basada en la integración y uso de nuevas técnicas y tecnologías, tanto para la recopilación de datos como realización de tareas que permitan una mayor eficiencia, rendimiento y automatización en el sector agrícola. Ente las tecnologías y técnicas más utilizadas en la agricultura 4.0 se encuentran los drones utilizados para sobrevolar los campos de cultivo, los sensores ambientales que obtienen información sobre la temperatura, suelo o humedad, las cámaras utilizada para la monitorización de los cultivos y la obtención de información y la robótica para la automatización de tareas como puede ser la poda de vides [3].

A continuación se describe el contexto agrario en el que se desarrolla este trabajo, sus objetivos y la estructura del mismo mediante el seguimiento de una metodología de ciencia de datos.

1.1. Contexto

El Instituto Tecnológico Agrario de Castilla y León (ITACYL) es un ente público de derecho privado adscrito a la Consejería de Agricultura y Ganadería de la Junta de Castilla y León. Por el Acuerdo 53/2009, de 14 de mayo de la Junta de Castilla y León, se aprueba el Plan Director de lucha contra las plagas agrícolas en Castilla y León (en adelante Plan Director).

Dentro del organigrama del ITACYL se integra el Observatorio de Vigilancia y Control de Plagas de Castilla y León (en adelante Observatorio), como encargado de la coordinación, supervisión y adopción de

medidas para la gestión de los riesgos derivados de la presencia de plagas y enfermedades en el territorio de la comunidad. El Observatorio organiza su trabajo entorno a estrategias de gestión integrada (GIP). La finalidad de estas estrategias de gestión preventiva, integrada y respetuosa con el medio ambiente, es establecer las actuaciones destinadas a monitorizar y contener el desarrollo de estas enfermedades, así como la provisión de información útil para los procesos de toma de decisiones de los profesionales agrícolas. En la actualidad, el Observatorio realiza tareas de seguimiento, control y divulgación de información de más de 30 plagas y enfermedades.

Una de las líneas de trabajo definidas dentro de la estrategia del Observatorio, es el desarrollo de modelos predictivos de plagas, los cuales tienen como objetivos:

- Anticipación a los momentos de riesgo.
- Elaboración de mapas espaciotemporales de graduación de riesgos.
- Concepción de medidas racionales dentro de la GIP.
- Apoyo en los procesos de toma de decisión.
- Mejora de las actividades de vigilancia mediante la racionalización de medios.
- Reducción y racionalización de las aplicaciones químicas (y/o aumento de su eficiencia).
- Aumento de la renta de los agricultores a través de la dupla "disminución de pérdidas" y "reducción de gastos".

En este Trabajo Fin de Grado se trabaja sobre la roya amarilla, una enfermedad causada por el hongo especie *Puccinia striiformis* que puede afectar a trigo, cebada y triticale. En la Figura 1.1 se pueden ver diferentes representaciones de la plaga.

Figura 1.1: Roya amarilla

Castilla y León es una región de producción principalmente cerealista. La cantidad de superficie declarada en las solicitudes unidas de la Política Agraria Comunitaria (PAC) el año 2022 de este tipo de cereales ha sido de cerca de 5 millones de hectáreas, lo que supone un 33 % de la superficie total cultivada en la Comunidad tal como se muestra en la tabla 1.1 obtenida de las solicitudes de ayudas PAC [4]. Estas cifras muestran la importancia de disponer de herramientas que sirvan para mitigar el coste debido a las pérdidas en la producción derivadas de la enfermedad, pero también para reducir el impacto medioambiental en la aplicación de productos fitosanitarios de manera preventiva solo en situaciones de riesgo real.

PRODUCTO	SUPERFICIE (ha)	% PAC
TRIGO DURO	1.650,22	0,03 %
TRITICALE	53.615,84	0,90 %
CEBADA	884.739,54	14,83 %
TRIGO BLANDO	1.014.343,78	17,00 %
SUP TOTAL PAC	5.966.304,03	32,76 %

Cuadro 1.1: Declaración PAC del año 2022

Este trabajo se desarrolla en un entorno de investigación en el que existen multitud de estudios que buscan relacionar las variables meteorológicas con la afectación de las plagas. Los dos estudios fundamentales de referencia son *Disease–Weather Relationships for Powdery Mildew and Yellow Rust on Winter Wheat* [5] y *A Threshold-Based Weather Model for Predicting Stripe Rust Infection in Winter Wheat* [6]. En ambos estudios se intenta determinar cómo está la plaga y se sabe que este estado depende fuertemente de dos cosas:

- **Condiciones meteorológicas.** Cada plaga tiene unas condiciones óptimas de desarrollo que se pueden medir en base a temperatura, humedad y precipitaciones principalmente. En el caso concreto de la Roya se ha comprobado que lluvias abundantes pueden limpiar la hoja y detener el proceso de desarrollo de la plaga.

- **Evolución del cultivo.** La plaga solo se puede desarrollar en determinadas fases del cultivo y cuanta más abundancia de cereal haya, es de esperar un mayor desarrollo de la plaga. La evolución del cultivo depende de las variables meteorológicas y la época del año, pero también de otros factores como la fecha de siembra, la variedad de trigo o las labores del agricultor entre otras.

En la actualidad, el ITACYL no dispone de un modelo de detección de plaga desarrollado e implementado utilizando técnicas de *machine learning*. Se parte de una aproximación inicial con histogramas de frecuencia y modelos poisson, binomial y binomial negativo que han sido elaborados como estudios previos, pero no han llegado a ser implantado.

1.2. Objetivos del trabajo

El objetivo principal de este trabajo es el desarrollo de un modelo de *machine learning* que permita predecir de forma temprana la aparición de roya amarilla. Los usuarios de este modelo serán los técnicos de campo encargados de la inspección de las parcelas de trigo de Castilla y León y los agricultores que trabajan estas mismas parcelas. Los agricultores, a través de la aplicación *Sativum* [7] podrán ser notificados del riesgo de aparición de plaga antes de que esta llegue a infectar los cultivos y los técnicos encargados de la inspección podrán priorizar el orden de evaluación de las parcelas. Esta evaluación prioriza las parcelas con una probabilidad de afectación de plaga alta permitiendo minimizar los costes de evaluación en un entorno de recursos limitado.

En el modelo objetivo se valora la exhaustividad y en menor medida la precisión, es decir, se desea detectar todas aquellas parcelas que están infectadas aunque esto suponga identificar parcelas no infectadas como tal. El motivo de ello es el uso que se pretende dar al modelo obtenido. No se desea predecir una observación en concreto en un momento concreto, si no seguir una evolución de la plaga durante un espacio temporal más amplio como podría ser una semana.

Para llevar acabo el trabajo y alcanzar el objetivo descrito se dispone de datos e información procedente de diferentes fuentes. Las condiciones meteorológicas son recogidas por 52 estaciones meteorológicas propias del ITACYL de la red de Inforiego y por aproximadamente 97 estaciones de la Agencia Estatal de Meteorología de España o AEMET.

Para la captura de los datos de la evolución del cultivo, el Observatorio de plagas organiza su trabajo en torno a zonas de vigilancia. Para cada una de estas zonas, se seleccionan una serie de parcelas que son monitorizadas periódicamente durante la campaña de la plaga que corresponda. En el caso de las plagas y enfermedades de cereal, como es el caso de roya amarilla, la campaña se desarrollo entre finales de enero y mediados de agosto. En este periodo, los técnicos de campo visitan periódicamente (entre 2-3 semanas) las parcelas para determinar el estado de cada plaga en la zona. Como resultado de este proceso, se obtiene la caracterización de la parcela mediante la variable estado fenológico y la observación de la plaga, la identificación positiva o negativa que el técnico hace sobre la plaga en la parcela en una fecha concreta.

1.3. Estructura

Para la realización de este trabajo se sigue la metodología de ciencia de datos propuesta por *International Business Machines Corporation* o IBM . Esta metodología ofrece un marco sobre cómo proceder con métodos y procesos en la obtención de conocimiento a partir de distintos conjuntos de datos. La metodología es aplicable a cualquier problema en el que se disponga de datos y no depende de tecnologías ni herramientas específicas. En la figura 1.2 se muestran el proceso iterativo de la metodología y a continuación se describen brevemente las etapas que sigue esta metodología [8]:

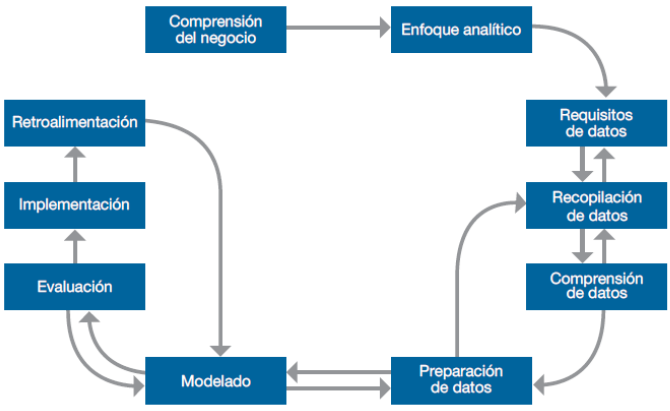


Figura 1.2: Metodología Fundamental para la Ciencia de Datos (extraída de [8])

- **Comprensión del negocio.** Todo proyecto comienza con la comprensión del negocio o problema a abordar. Consiste en estudiar el problema y definir sus objetivos.

- **Enfoque analítico.** Esta etapa implica expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático. Si objetivo es identificar la técnica más adecuadas para el resultado deseado.

- **Requisitos de datos.** Los métodos analíticos a utilizar requieren de determinados contenidos de datos, formatos y representaciones orientados por el conocimiento del contexto del problema.

- **Recopilación de datos.** Se trata de reunir los datos disponibles y relevantes para el problema. Cuantos más datos se dispongan del problema, los modelos predictivos mejor podrán representar los eventos.

- **Comprensión de los datos.** Una vez que se han recopilado los datos iniciales, es necesario comprender el contenido de los mismos, evaluar su calidad y descubrir conocimientos iniciales sobre ellos.

- **Preparación de los datos.** Consiste en construir el conjunto de datos que se utilizará para la siguiente etapa de modelado. Entre las actividades de la preparación de los datos se encuentra la combinación de los datos de múltiples fuentes, limpieza de los mismos y el proceso llamado ingeniería de características. Este último proceso implica la construcción de nuevas características a partir de las ya existentes.

- **Modelado.** En la etapa de modelado se utiliza el conjunto de datos preparado para el desarrollo de modelos predictivos o descriptivos según el enfoque analítico previamente definido. Para un enfoque determinado se pueden probar múltiples algoritmos con sus respectivos parámetros para encontrar el que mejores resultados ofrece para el problema.

- **Evaluación.** Consiste en obtener medidas que representen la calidad de los resultados y permitan evaluar al modelo y ajustarlo según las necesidades.

- **Implementación.** Una vez elaborado y evaluado el modelo predictivo es posible implementarlo en un entorno de producción. Esta implementación puede ser la generación de un informe o la inclusión del modelo en un flujo de trabajo que ofrezca algún servicio.

- **Retroalimentación.** La última etapa consiste en obtener información sobre el impacto que tiene el modelo en el entorno que se implementó. Su finalidad es actualizar el modelo y obtener mejores resultados en un futuro

El trabajo se estructura en capítulos que tratan de seguir el orden de la metodología descrita anteriormente. Seguidamente se describe el contenido de cada uno de estos capítulos:

- **Capítulo 1. Introducción.** En este primer capítulo se describe el contexto en el que se desarrolla el trabajo, los objetivos del mismo, la metodología utilizada para la ejecución y la presente descripción de la estructura. Se aborda la primera etapa de la metodología presentada en la figura 1.3

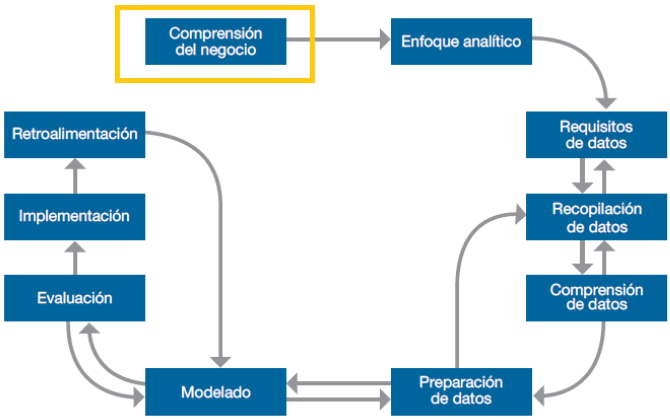


Figura 1.3: Metodología Fundamental para la Ciencia de Datos - Comprensión del negocio (extraída de [8])

- **Capítulo 2. Contexto tecnológico.** Se describe brevemente la tecnología que se utiliza para la planificación, coordinación y ejecución del trabajo. Se describen los programas utilizados, el lenguaje y sus librerías.
- **Capitulo 3. Árboles de regresión y clasificación.** Se especifica el enfoque analítico llevado a cabo para resolver el problema y las métricas que se utilizarán para su evaluación.
- **Capítulo 4. Descripción de los datos iniciales.** Se presentan los conjuntos de datos disponibles para el problema, los conjuntos de datos meteorológicos y de campo necesarios para alcanzar el objetivo junto con los ficheros de interés utilizados para la ampliación de la información disponible en los primeros.
- **Capítulo 5. Limpieza y tratamiento de los datos.** Incluye los procedimientos llevados a cabo para la imputación de datos faltantes junto con un estudio sobre la afectación de dicha imputación sobre la distribución de los mismos. Además se evalúan dos posibles soluciones al problema de asignación de variables meteorológicas a las parcelas.
- **Capítulo 6. Creación del fichero para el modelado.** Se lleva a cabo la ingeniería de características descrita anteriormente con el fin de crear características de interés para el problema. Finalmente se obtiene un fichero único que se utiliza para construir modelos en el capítulo siguiente.
- **Capítulo 7. Modelado y evaluación.** Se crean los conjuntos de datos necesarios para el entrenamiento de un modelo *machine leanring* (entrenamiento, validación y test), se realiza la selección de variables junto con el ajuste de algunos parámetros y se evalúa el modelo final elegido.
- **Capítulo 8. Conclusiones y lineas futuras.** Se añaden las conclusiones sobre el trabajo realizado y sus resultados y se describen posibles lineas futuras de trabajo.

2. Contexto tecnológico

En esta sección se describe la tecnología utilizada para la planificación, ejecución y control de versiones del trabajo. Respecto a la ejecución, se presenta el lenguaje de programación elegido, el entorno de desarrollo y los programas requeridos para el tratamiento de los datos. En todas las secciones se hace especial hincapié en las librerías y funcionalidades utilizadas, haciendo referencia a la paginas oficiales donde se encuentra disponible su documentación.

2.1. Python

Python es un lenguaje de programación de alto nivel de código abierto y propósito general utilizado principalmente para la automatización y elaboración de scripts, desarrollo de aplicaciones web con frameworks específicos como Django y para el análisis de datos. Esta última aplicación es posible gracias a librerías como [9]:

- **NumPy.** Biblioteca que da soporte para la creación de vectores y matrices grandes multidimensionales, junto con una colección de funciones matemáticas de alto nivel para operar con ellas de forma eficiente [10].

- **Pandas.** Librería escrita como extensión de NumPy para manipulación y análisis de datos. En particular ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales [11].

- **GeoPandas.** Extensión de la librería Pandas que permite trabajar con datos geoespaciales y realizar operaciones espaciales [12].

- **Matplotlib.** Extensión de NumPy diseñada para la generación de gráficos a partir de datos contenidos en listas o arrays [13].

- **Scipy.** Librería que contiene módulos para optimización, álgebra lineal, integración, interpolación y funciones espaciales entre otras [14].

- **Scikit-learn.** Biblioteca para aprendizaje automático que incluye varios algoritmos de clasificación y regresión como los árboles aleatorios utilizados en este trabajo, *ExtraTreeRegressor* y *ExtraTreeClassifier* [15].

2.2. PyCharm

PyCharm es una aplicación informática utilizado para la programación en lenguaje de programación Python que proporciona servicios integrales para facilitar al programador el desarrollo del código. Se utiliza la licencia académica *PyCharm Professional Edition* que cuenta con: asistencia y análisis de codificación con completado de código, navegación por el proyecto y código con vista de la estructura de archivos y saltos rápidos entre archivos, clases y métodos. Otra funcionalidad utilizada en la elaboración de este trabajo es la interfaz unificada con Git para la integración de control de versiones [16].

En la Figura 2.1 se muestra un captura de la interfaz del programa en un momento de depuración del código. Se puede ver a la izquierda la estructura de los scripts y ficheros, en el centro el código de uno de los scripts, a la derecha el contenido de una estructura de datos `DataFrame` de Python y por último, en la parte inferior una consola para realizar consultas.

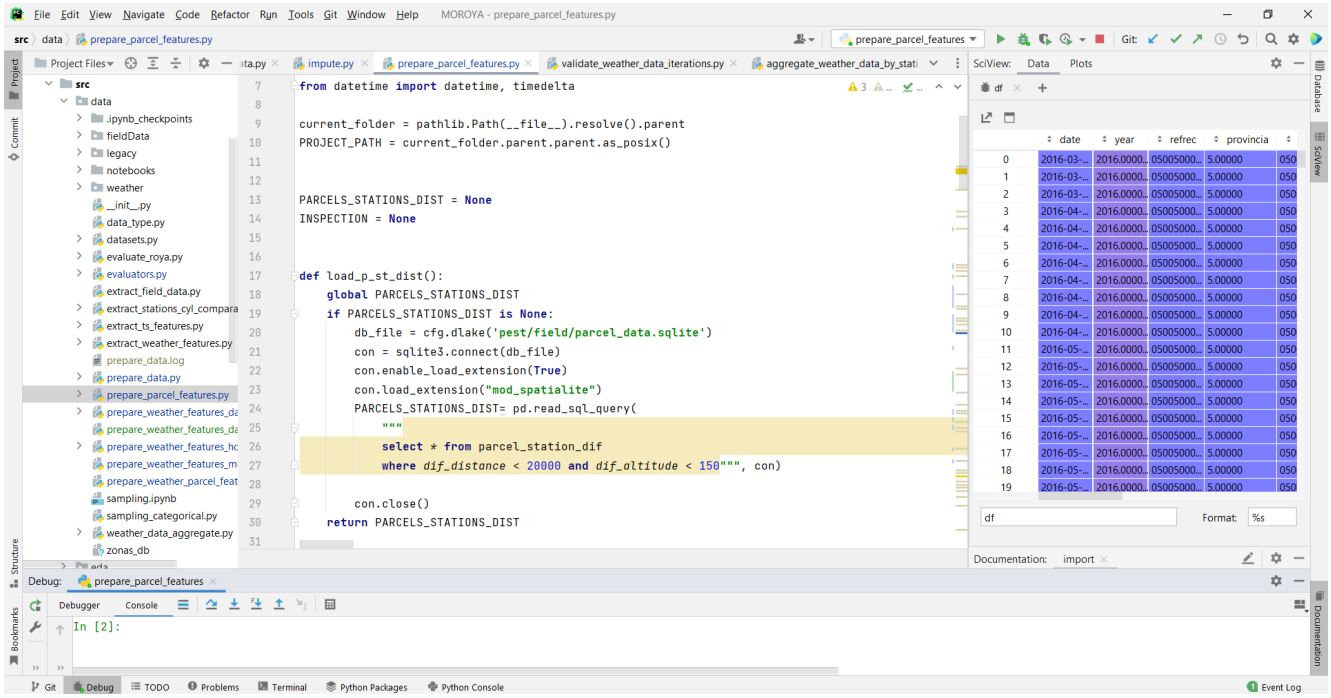


Figura 2.1: Interfaz Pycharm

2.3. Gitea

La gestión de los diversos cambios realizados sobre el código del proyecto se ha realizado con el sistema de control de versiones Gitea. Esta plataforma, de código abierto y desarrollada utilizando Git, ha permitido la elaboración del trabajo de forma colaborativa con Gustavo Río Briones, uno de los tutores del proyecto. En la Figura 2.2 se muestra la interfaz del programa.

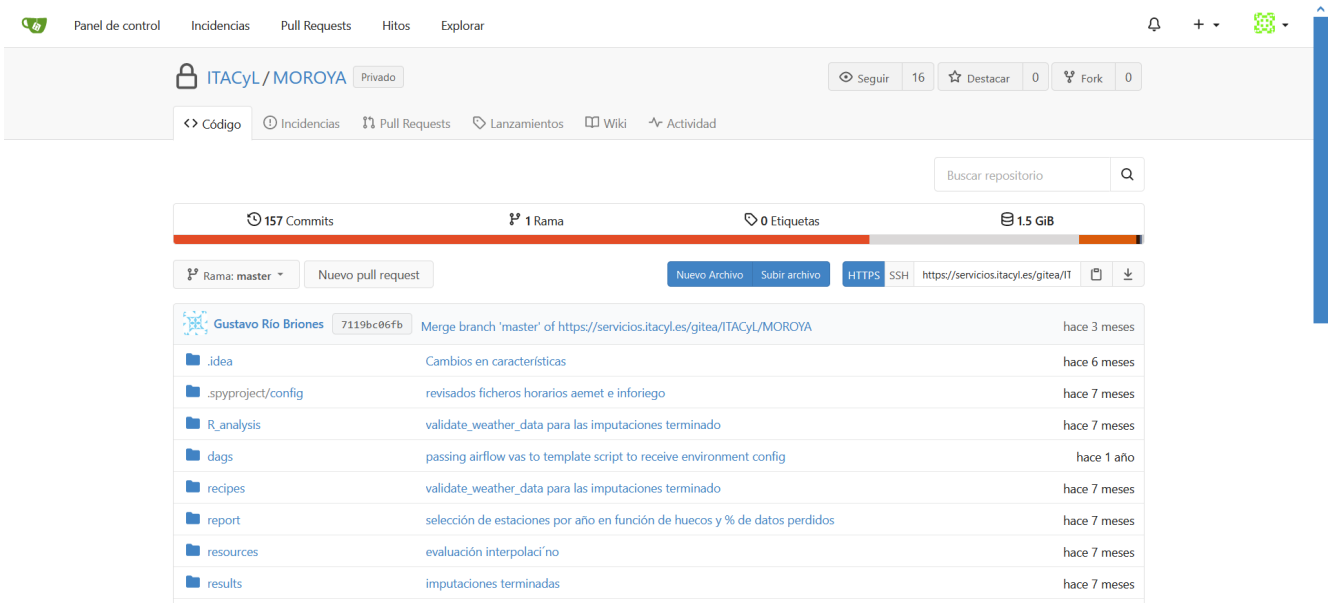


Figura 2.2: Interfaz Gitea

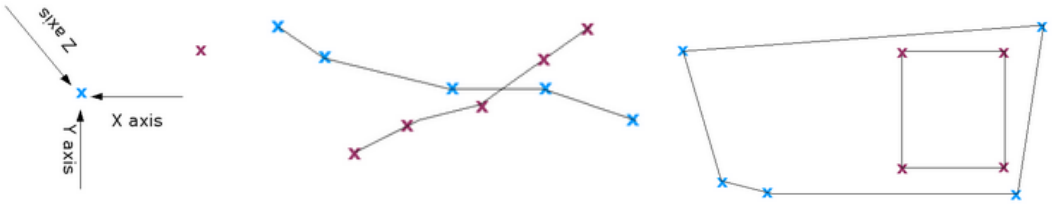
2.4. QGIS

QGIS es un sistema de información geográfica, SIG (o GIS de su nombre en inglés *Geographical Information System*). Es un programa de software libre que integra herramientas que permiten organizar, almacenar, manipular, analizar y modelizar grandes cantidades de datos procedentes del mundo que real que están vinculados a una referencia espacial. Facilitan la incorporación de aspectos ambientales que conducen a la toma de decisiones de una manera más eficaz [17].

QGIS facilita la interconexión con muchas bases de datos geoespaciales, GeoPackage, SpatiaLite, PostgreSQL/PostGIS y Oracle Database entre otras. Además ofrece a los usuarios la posibilidad de automatizar tareas en lenguaje C++ o Python.

QGIS puede manejar datos en formato raster y vectorial gracias a la librería GDAL. A continuación se describen estos datos.

- **Datos vectoriales.** Los datos vectoriales permiten representar objetos espaciales como casas, carreteras, arboles, ríos y parcelas del mundo real dentro de un ambiente SIG. Los objetos espaciales vectoriales tienen atributos, que consiste en texto o información numérica que los describe, su forma se representa mediante geometrías. La geometría se compone de uno o más vértices con posiciones x, y, z interconectados. Cuando la geometría de un objeto espacial consiste en un solo vértice, como puede ser un árbol o una farola, se conoce como un elemento punto, (ver Figura 2.3a). Cuando la geometría consiste en dos o más vértices y el primer y último vértice no son iguales, se forma una polilínea como la polilínea de la Figura 2.3b, que podría representar un río. Cuando tres o más vértices están presentes, y el último vértice es igual a la primero, se forma un polígono como el de la Figura 2.3c, que podría representar una parcela de campo [18].



(a) Punto (b) Polilínea (c) Polígono

Figura 2.3: Geometrías que representan objetos espaciales

- **Datos ráster.** Un conjunto de datos ráster esta compuesto por una matriz de píxeles llamadas celdas. Cada píxel contiene un valor que representa las condiciones de la zona cubierta por dicha celda. Se utilizan estos datos en una aplicación SIG cuando se desea mostrar información que es continua a través de un área y no puede ser dividida fácilmente en entidades vectoriales. En la Figura 2.4 se representa un ejemplo de estos datos, datos del relieve de España y una capa correspondiente a la Comunidad Autónoma de Extremadura superpuesta [19].

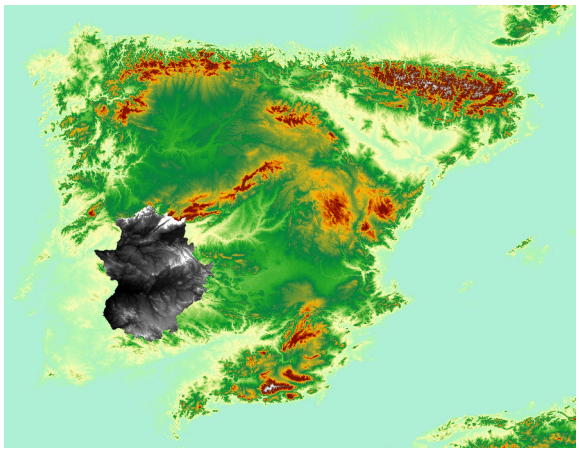


Figura 2.4: Matriz datos ráster

2.5. SQLite

Para realizar este Trabajo Fin de grado se accede a diversas bases de datos mediante el sistema de gestión de bases de datos SQLite, proyecto de dominio público. SQLite ofrece facilidad para la conexión desde el programa QGIS y para el manejo de datos geoespaciales mediante la extensión Spatialite [20].

2.6. TPOT

Tree-Based Pipeline Optimization Tool o TPOT es uno de los primeros métodos de aprendizaje automático automatizado o AutoML de código abierto programado en Python y desarrollado para la comunidad de ciencia de datos. El objetivo de un AutoML es facilitar la selección de un algoritmo *machine learning* mediante la selección de parámetros y/o características por medio de métodos que permiten detectar patrones complejos en el campo de *big data* [21].

El objetivo de TPOT es automatizar la construcción de modelos de *machine learning* mediante la combinaciones de árboles de expresión que pueden ser recorridos por algoritmos de programación genética. TPOT hace uso de las librerías de `scikit-learn` para probar múltiples algoritmos de *machine learning*.

2.7. Enterprise Architect

Sparx Systems Enterprise Architect es una herramienta de pago utilizada para la modelización y el diseño de productos *software*. En este trabajo ha sido utilizada para la creación del diagrama incluido en la figura 2.5. Este diagrama contiene la estructura de los ficheros y scripts generados para la elaboración de esta trabajo [22].

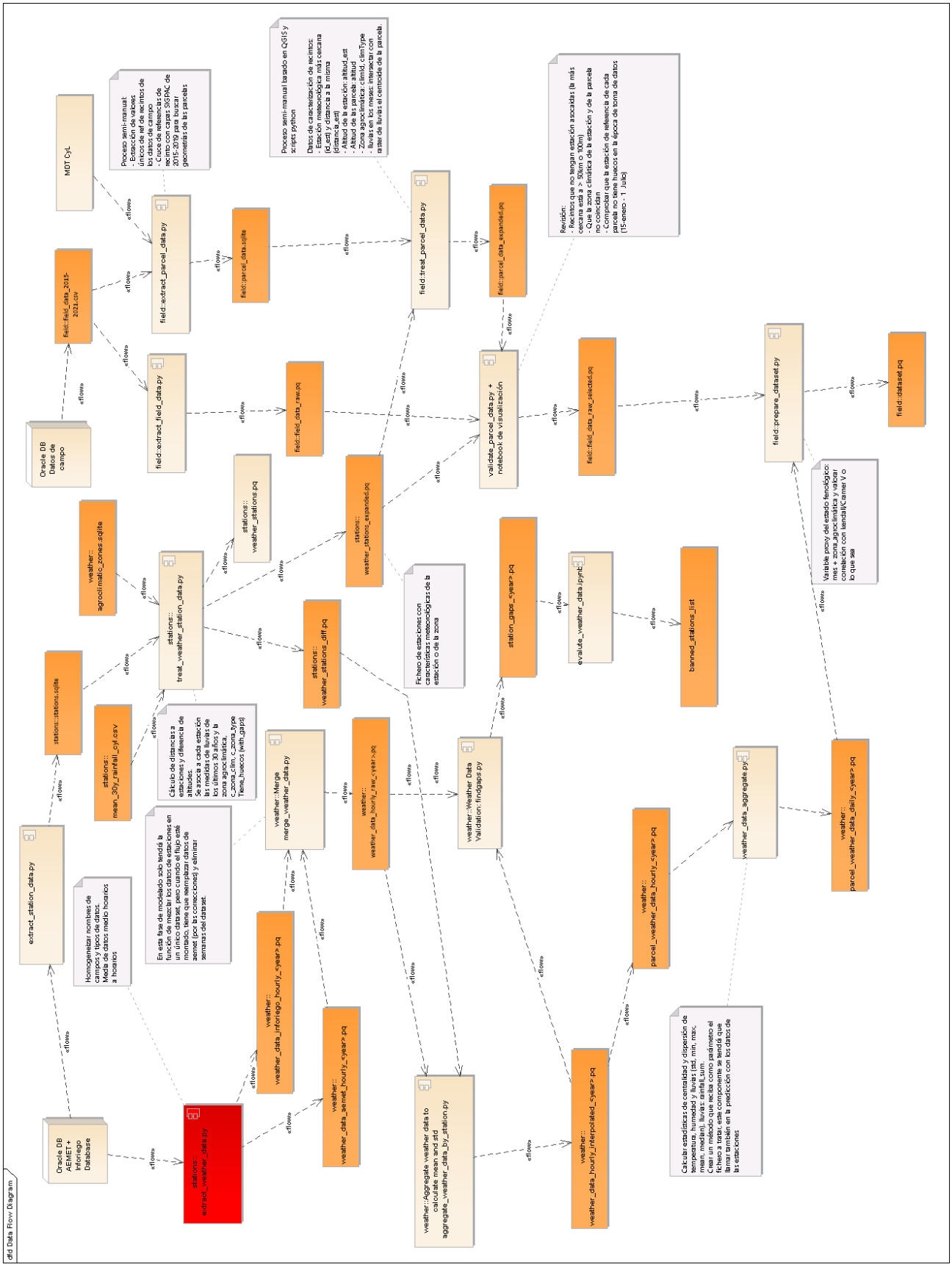


Figura 2.5: Diagrama realizado en Entreprise Architect

3. Árboles de regresión y clasificación

En esta sección se aborda la segunda etapa de la metodología representada en la figura 3.1. Consiste en expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático. Para ello se describe inicialmente el entorno de trabajo y se detalla la técnica a utilizar para la resolución del problema.

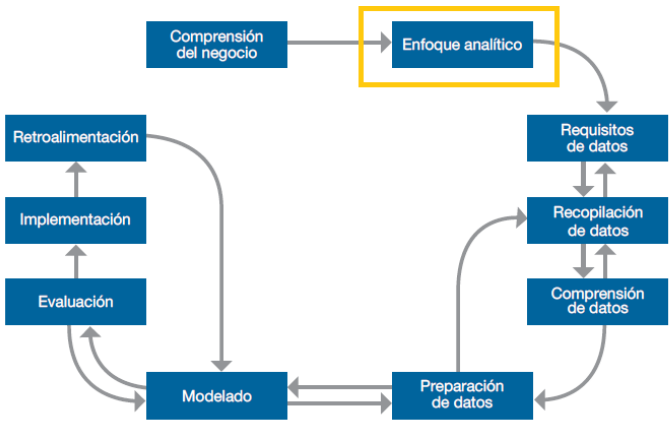


Figura 3.1: Metodología Fundamental para la Ciencia de Datos - Enfoque analítico (extraída de [8])

La Inteligencias Artificial o IA se define como la inteligencia llevada a cabo por *agentes inteligentes*, programas con la capacidad de aprender y razonar como un humano [23]. Los principales problemas a los que se enfrentan inicialmente los sistemas inteligentes incluyen tareas como la planificación, el procesamiento del lenguaje natural, la habilidad de movimiento y la manipulación de objetos siguiendo una serie de reglas o algoritmos. La llegada de internet produce un aumento masivo en la cantidad de datos, y el aumento de la potencia de cálculo de los ordenadores permite extraer información de estos datos y por lo tanto conocimiento aprovechable por las máquinas. En estas condiciones surge el Machine Learning, una rama de la inteligencia artificial que acorde a Arthur Samuel, pionero en la inteligencia artificial, en 1959 ,'capacita a las computadoras a aprender sin ser explícitamente programadas' [24], es decir, son las máquinas las que extraen patrones de la información presente algoritmos de aprendizaje automático.

Un modelo de aprendizaje automático cuenta con variables y algoritmos que permiten extraer información de estos datos. Atendiendo a los datos, en la actualidad se emplean principalmente tres corrientes para entrenar a los algoritmos:

- **Aprendizaje supervisado.** En este paradigma se utilizan las etiquetas de los datos. En problemas

de clasificación esta etiqueta es una variable categórica que representa una clase de pertenencia, un valor discreto. Un ejemplo concreto es la segmentación, problemas en los que la clasificación se realiza a nivel de píxel. Por otro lado, en los problemas de regresión la etiqueta es una variable numérica que toma un valor continuo, como puede ser la predicción de número de fallos. Algunas de las técnicas utilizadas para la solución de este tipo de problemas incluyen los clasificadores Naive Bayes, *K-NN* o k-vecinos más próximos, los árboles de decisión, la regresión y las máquinas de soporte de vectores, entre otras.

- **Aprendizaje no supervisado.** En este tipo de aprendizaje se dispone únicamente de los datos, sin las etiquetas. La finalidad es la agrupación y la obtención de información atendiendo a las similitudes, diferencias y patrones. En problemas en los que se dispone del número de grupos o clusters la técnica más habitual es *K-means clustering*, si no se dispone del número exacto, la técnica más utiliza es el agrupamiento jerárquico,

- **Aprendizaje por refuerzo.** En este paradigma existe cierta supervisión pero esta no viene en forma de etiquetas conocidas de antemano. El aprendizaje por refuerzo se aplica en problemas en los que el agente que aprende interacciona con el entorno que observa, con las entradas, tomando unas acciones o decisiones que son las salidas. La supervisión se recibe a medida que el algoritmo toma estas decisiones en forma de recompensa o *feedback*. El *feedback* indica el grado con el cual la salida del modelo permite alcanzar el objetivo.

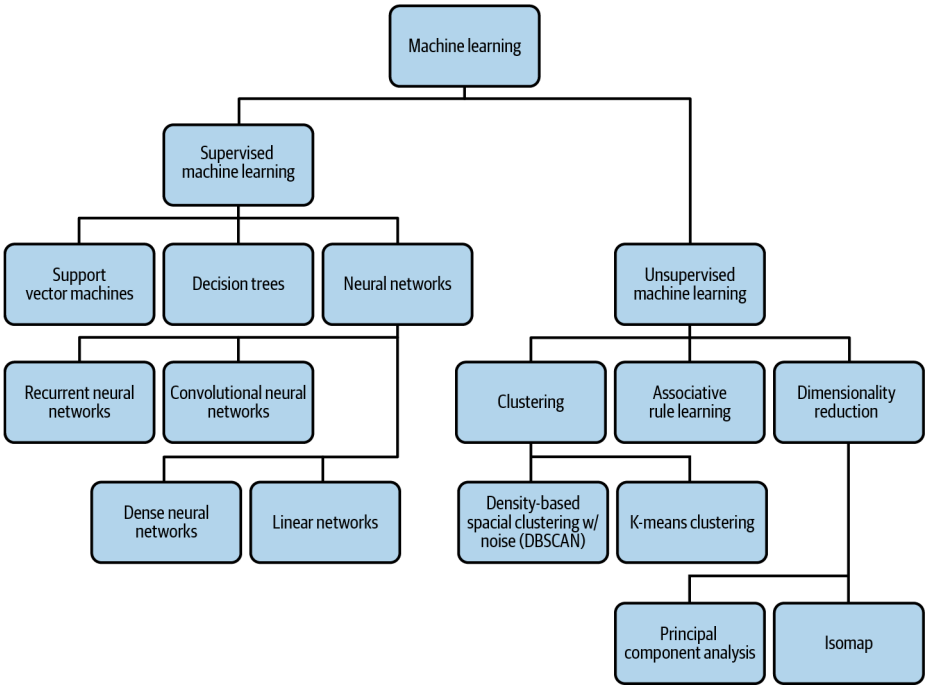


Figura 3.2: Diagrama con tipos de *machine learning* (extraída de [25])

Artículos de investigación recientes destacan la posibilidad de utilizar técnicas de *Decicion trees* incluidas en el diagrama de la figura 3.2, para la predicción de la plaga Roya. Un ejemplo de estas técnicas son los *Random Forest* o bosques aleatorios [26].

3.1. Árboles de decisión

Los métodos basados en árboles para regresión y clasificación o CART del inglés (*Classification and Regression Trees*) involucran la estratificación o segmentación del espacio predictivo en un número finito de regiones, tal como aparece representado en las imágenes (a) y (b) de la Figura 3.3. En la misma figura se puede ver en la imagen (c) el conjunto de reglas para la segmentación del espacio resumido en forma de árbol.

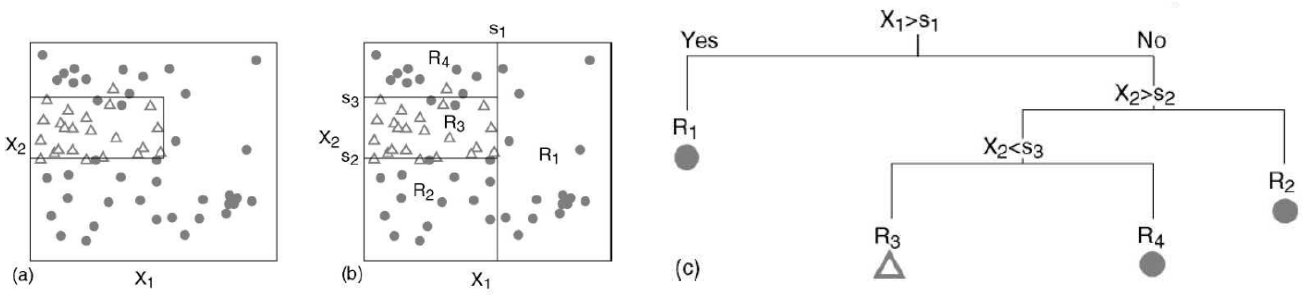


Figura 3.3: Árboles de clasificación

En el procedimiento a seguir se parte de un conjunto de n observaciones $\{(x_i, g_i)\}_{i=1}^{n}$ con $g_i \in \{1, ..., q\}$ y se dan los siguientes pasos [27]:

1. El algoritmo comienza eligiendo una variable X_i y obtiene un punto de corte s_i tal que las observaciones con $X_i \leq s_i$ van a un nodo y $X_i > s_i$ van a otro.

2. Cada uno de los nodos se puede volver a dividir eligiendo una variable y un nuevo punto de corte. Se pueden usar variables anteriormente utilizadas.

3. Un nodo se considera terminal si ya no es dividido más.

En cada iteración crece la profundidad del árbol, definida como la trayectoria más larga entre la raíz y uno de los nodos terminales. El criterio para la segmentación del espacio en regiones $R_1, ..., R_j$ es ligeramente diferente según sea un problema de de regresión o de clasificación.

3.1.1. Árboles de regresión

En los árboles de regresión el objetivo es encontrar las regiones $R_1, ..., R_j$ que minimizan la suma residual de cuadrados o RSS de la fórmula 3.1 donde $\hat{y}_{R_j}$ es la media de la respuesta de las observaciones de la región j-ésima [28].

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{3.1}$$

Debido a que resulta computacionalmente inviable la consideración de todas las posible particiones del espacio de características en j regiones, se considera la aproximación *top down greedy* conocida como *recursive binary splitting* o división binaria recursiva. Esta aproximación se denomina *top-down* porque

comienza en la parte superior del árbol donde todas las observaciones pertenecen a una única región y en cada iteración se originan dos ramas que la dividen. El procedimiento es *greedy* porque en cada paso la división se realiza tomando la mejor división en ese mismo paso y sin tener en cuenta futuros pasos.

Para llevar a acabo la división binaria recursiva, se selecciona el predictor X_j y el punto de corte s que divide el espacio de predicción en dos regiones $R_1(j,s) = \{X|X_j < s\}$ y $R_2(j,s) = \{X|X_j \geq s\}$ que reducen el RSS. La ecuación a minimizar es la ecuación 3.2 y el proceso se repite hasta alcanzar el criterio de parada que podría ser alcanzar una región que no tiene más de un número dado de observaciones o alcanzar un número máximo de nodos terminales.

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{3.2}$$

Para obtener la predicción de una observación se utiliza la media o la moda de las observaciones de entrenamiento en la región en la que la observación a predecir pertenece.

3.1.2. Árboles de clasificación

Los árboles de clasificación son muy similares a los árboles de regresión, la principal diferencia reside en la respuesta cualitativa que predicen. El criterio utilizado para realizar la división se basa en la pureza u obtención de los nodos más homogéneos posible. Una medida de la impureza razonable es el error de mal clasificación o tasa de mal clasificados de la fórmula 3.3. Esta tasa mide la fracción de observaciones del conjunto de entrenamiento contenidos en la región que no pertenecen a la clase más común. $\hat{p}_{mk}$ representa la proporción de observaciones de la región m-ésima que pertenecen a la clase k-ésima.

$$E = 1 - max_k(\hat{p}_{mk}) \tag{3.3}$$

En la práctica es preferible el uso de medidas sensibles al crecimiento de los árboles como es el índice de Gini definido en la fórmula 3.4. Este índice se define como una medida de la varianza total en el conjunto de las K clases o pureza de los nodos. Las medidas de impureza toman valores grandes cuando los valores p_{mk} están equilibrados. Interesan nodos puros con una medida pequeña. Un valor bajo indica que el nodo contiene observaciones de una sola clase pues todos los valores $\hat{p}_{mk}$ están próximos a cero o uno.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.4}$$

Una alternativa del índice de Gini utilizada como medida de la pureza de los nodos es la entropía de la fórmula 3.5. Tanto el índice de Gini como la entropía obtienen resultados similares.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{3.5}$$

En los árboles de clasificación la observación x que se quiere predecir se asigna a la clase más común o moda de las observaciones de entrenamiento que se encuentran en la región a la que pertenece la observación.

3.2. Poda de los árboles

El proceso de construcción de los árboles puede llevar a árboles complejos que sobreajustan los datos, es decir, se comportan muy bien para los datos de entrenamiento, con un error aparente bajo, calculado sobre el conjunto de entrenamiento, pero con un error de generalización alto, error calculado sobre los datos de test. En esta sección se describen alguna de las técnicas utilizadas para disminuir el posible sobreajuste.

3.2.1. Poda de árboles de regresión

Una posible solución al problema de sobreajuste en los árboles de regresión es la construcción de los mismos hasta que el RSS exceda un umbral. Sin embargo esto no siempre produce buenos resultados, pues puede darse el caso de tener tener una división aparentemente mala en un principio que en sucesivas divisiones obtenga una reducción considerable del RSS.

Una mejor estrategia para solventar el problema de sobreajuste consiste en obtener un árbol muy complejo, con muchas divisiones T_0 y a continuación podarlo con el fin de obtener subárboles que consigan el menor error para un conjunto de observaciones test. Como resulta inviable computacionalmente considerar tódos los posibles subárboles, se considera una secuencia de subárboles registrados con un parámetro α. Para cada valor de α existe un subárbol $T \in T_0$ tal que la expresión 3.6 es lo más pequeña posible. En esta expresión $|T|$ indica el número de nodos terminales, R_m es la región correspondiente al nodo terminal m y $\hat{y}_{R_m}$ es la respuesta predicha para este nodo. Esta técnica se conoce como *Cost complexity pruning* traducida en español como poda de la complejidad de costos.

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T| \tag{3.6}$$

El parámetro α controla la complejidad del árbol y su ajuste a los datos de entrenamiento. A medida que α aumenta, existe un precio a pagar por el aumento del número de nodos y por lo tanto la función 3.6 se minimizará con árboles pequeños. El parámetro α se asemeja al parámetro λ de la regularización *lasso* y *ridge* que controlan la complejidad de los modelos lineales [29]. El parámetro α se puede obtener utilizando validación cruzada en el algoritmo descrito seguidamente:

1. Se obtiene el árbol de regresión utilizando el algoritmo de división binaria recursiva explicado en la sección 3.1.1.
2. Aplicar la técnica de poda que considera la minimización de la fórmula 3.6 para obtener una secuencia de subárboles en función del parámetro α.
3. Utilizar validación cruzada con K-fold para obtener el parámetro α. Se dividen las observaciones de entrenamiento en K folds o particiones y para cada partición $k = 1, ..., K$:
    - Repetir los pasos uno y dos con los datos del K fold

- Evaluar el error de predicción con los fols restantes.

Promediar los resultados para cada valor de α y escoger el valor α que minimice la media del error.

4. Devolver el subárbol del paso dos que corresponde al valor α escogido

3.2.2. Poda de árboles de clasificación

La poda para árboles de clasificación resulta muy similar, la única diferencia está en la función de error utilizada. La impureza de un nodo m del árbol T se expresa como $Q_m(T)$ y se puede medir con las medidas descritas en la sección 3.1.2. La impureza total del árbol T se expresa en la ecuación 3.7 donde $n(m)$ es el número de observaciones en el nodo m [27].

$$Q(T) = \sum_{m \in T} \frac{n(m)}{n} Q_m(T) \tag{3.7}$$

Se busca que la impureza total del árbol $Q(T)$ sea pequeña. Si se divide el nodo m en dos nodos nuevos m_L y m_R para tener un nuevo árbol T' se trata de obtener el nodo m, la variable X_j y el corte s_j que obtengan un mayor valor para la fórmula 3.8.

$$Q(T) - Q(T') = Q_m(T) - \left(\frac{n(m_L)}{n(m)} Q_{m_L}(T') + \frac{n(m_R)}{n(m)} Q_{m_R}(T') \right) \tag{3.8}$$

El criterio penalizado en los árboles de clasificación se puede expresar con la fórmula 3.9 donde $|T|$ hace referencia al número total de nodos terminales y T al árbol.

$$CP_\alpha(T) = \sum_{m=1}^{|T|} \frac{n_m}{n} Q_m(T) + \alpha |T| \tag{3.9}$$

Al mover los valores de α en un intervalo de cero a infinito, solo hay un número reducido de árboles que minimiza 3.9 en algún α. Si $\alpha = 0$ se tienen nodos terminales totalmente "puros" y un erro aparente cero por otro lado, si α tiende a infinito, se tiene un árbol con todas las observaciones en un único nodo terminal [30].

Para elegir el árbol de clasificación, se utiliza validación cruzada y se monitorizan los errores estimados de generalización. Se utilizan representaciones como las del ejemplo mostrado en la Figura 3.4. La regla 1-SE indica elegir el modelo más simple, con menos nodos terminales, cuya precisión sea comparable con la del mejor modelo. En el mismo ejemplo de la figura 3.4 se puede ver que el error de generalización es similar con 18 nodos y con tres nodos.

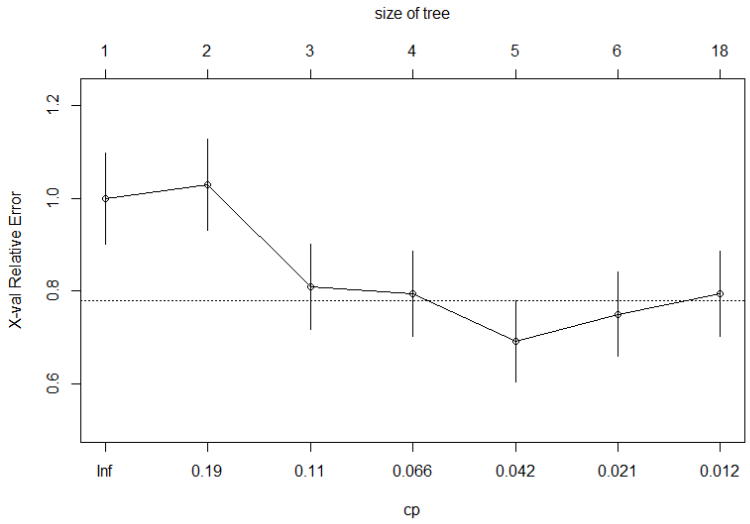


Figura 3.4: Error sobre el conjunto de test según el criterio de penalización

3.3. Métodos de ensemble

Los árboles que no se encuentran muy desarrollados tienen una representación gráfica muy intuitiva y son fáciles de interpretar, además, permiten trabajar con variables cualitativas sin la necesidad de utilizar variables dummy. A pesar de estas ventajas, el uso de árboles aleatorios tienen dos inconvenientes principales. La primer desventaja es la inestabilidad de los mismos o baja robustez debida a la alta varianza. Esta inestabilidad hace que pequeños cambios en el conjunto de entrenamiento cambien drásticamente los resultados obtenidos. La segunda desventaja es su bajo poder predictivo, generalmente inferior a otros modelos de regresión y clasificación. Con el fin de mejorar el poder predictivo de los árboles y crear modelo más robustos surgen las técnicas *bagging*, *random forests* y *boosting* y los métodos de *ensemble*.

Un *ensemble* es un método que combina varios modelos simples con el fin de obtener un modelo que mejore el poder de generalización y la robustez de los modelos individuales. Se pueden distinguir dos familias de *ensembles* [31]:

- **Métodos de promedio.** Del inglés. *averaging methods*, son métodos que utilizan una serie de estimadores independientes y obtienen como predicción la media de estos. Este técnica permite reducir la varianza. Como ejemplos se tienen lo métodos *bagging* y los *random forest* o bosques aleatorios.
- **Métodos boosting.** Los estimadores base se disponen en secuencia y tratan de reducir el sesgo de los estimadores previos. Ejemplos de estos métodos son *AdaBoost* y *Gradient Tree Boosting*.

3.3.1. Bagging

El método *bagging* o *bootstrap aggregation* es un procedimiento frecuentemente utilizado en el contexto de árboles de decisión que permite reducir la variabilidad de los modelos. Para ello, se obtiene B subconjuntos de datos de entrenamiento y con cada uno se ajusta un árbol predictivo o modelo. A la hora de predecir una instancia nueva, se obtiene la media de las predicciones individuales en los problemas de

regresión y la moda para los problemas de clasificación. Para la obtención de los subconjuntos de datos se aplica el método de remuestreo *Bootstrap*.

3.3.2. Random Forests

La técnica *Random Forests* ofrece una mejora al método *Bagging*. Este último a pesar de proporcionar una predicción promedio de varios árboles aleatorios no reduce sustancialmente la varianza, pues los árboles generados se encuentran altamente correlacionados. *Random Forests* le añade una aleatoriedad a la técnica de predicción *Bagging* en la elección de m predictores de los p predictores totales con un valor m elegido normalmente como raíz cuadrada de p.

3.4. Métricas utilizadas para la evaluación de los modelo

Las métricas permiten evaluar la capacidad predictiva del modelo cuantificando el grado con el que las predicciones realizadas se acercan a los verdaderos valores. Las métricas a utilizar para evaluar un modelo dependen del problema abordado, es decir, dependen de si se trata de un problema de clasificación o un problema de regresión. A continuación se describe cada una de las métricas empleadas en este trabajo.

3.4.1. Métricas en regresión

En los problemas de regresión se tiene como variable respuesta una variables continua. En este trabajo esta variable corresponde a la proporción de plaga detectada. Las métricas utilizadas son:

- **Error cuadrático medio**. MSE o *Mean Squared Error* es la media de los errores al cuadrado, es decir, la media de la diferencia de los valores reales y los valores predichos al cuadrado.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.10}$$

- **Error absoluto medio**. MAE o *Mean absolute error*, es la media de los errores absolutos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.11}$$

Ambas métricas miden la distancia entre las predicciones realizadas y los verdadero valores, pero el MSE es más susceptible a valores extremos o outliers al elevar al cuadrado sus errores. Por ello el MAE resulta ser una medida más robusta, más estable ante pequeños cambios de la entrada.

3.4.2. Métricas en clasificación

Los problemas de clasificación se caracterizan por tener una variable respuesta categórica o cualitativa. En el problema de clasificación imbalanceado planteado en este trabajo se tiene como variable respuesta una variable binaria que indica la presencia o ausencia de plaga con valores cero (o negativo) y uno (o positivo) respectivamente, donde la clase mayoritaria es cero. En la tabla 3.1 se representan términos necesarios para el cálculo de las siguientes métricas:

		Predicted	
Actual		**Negative**	**Positive**
	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Cuadro 3.1: Matriz de confusión para una clasificación binaria

- **Precisión.** Proporción de instancias positivas predichas correctamente, es decir, número total de instancias positivos frente al total de instancias predichas como positivas. Intuitivamente es la capacidad de no clasificar una parcela sin plaga como parcela con plaga.

$$Precision = \frac{T_P}{T_P + F_P} \tag{3.12}$$

- **Recall o sensibilidad.** Proporción de instancias clasificadas como positivas del número total de instancias que son verdaderamente positivas. Intuitivamente es la capacidad de detectar aquellas parcelas que están infectadas.

$$Recall = \frac{T_P}{T_P + F_N} \tag{3.13}$$

- **F-score.** Establece un compromiso entre la precisión y la sensibilidad calculando la media armónica de ambas métricas. El factor β es un valor positivo que permite dar mayor importancia a una de las métricas, la sensibilidad es considerado β veces más importante que la precisión. Cuando β es uno, se da la misma importancia a ambas métricas.

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta \cdot precision) + recall} \tag{3.14}$$

En el problema de clasificación de este trabajo se tiene especial interés en evitar los falsos negativos o el error de tipo II, es decir, no detectar la plaga cuando esta está presente y por lo tanto se pondera más la métrica *recall* o sensibilidad.

4. Descripción de los datos iniciales

En este trabajo se utilizan principalmente tres conjuntos de datos. Los dos primeros conjuntos contienen datos de las condiciones meteorológicas registradas por las estaciones meteorológicas propias del ITACYL y las estaciones de medición de la AEMET presentadas en la sección 1. El tercer conjunto de datos recoge la información de las inspecciones realizadas por los técnicos del Observatorio de Plagas de Castilla y León para la detección temprana de la roya amarilla en parcelas cultivadas de trigo. A continuación se describe cada uno de ellos en detalle, así como ficheros de interés que se han utilizado para la ampliación de estos tres conjuntos iniciales. Este capítulo representa las etapas tres y cuatro de la Metodología Fundamental para la Ciencia de Datos presentada por IBM y representada en la figura 4.1.

Figura 4.1: Metodología Fundamental para la Ciencia de Datos - Requisitos y recopilación de datos (extraída de [8])

4.1. Datos meteorológicos

El desarrollo de la roya amarilla en el trigo esta principalmente influenciado por las circunstancias meteorológicas presentes durante el crecimiento de la planta. Una humedad alta con lluvia o rocío y un rango de temperatura entre 10 y 15 grados centígrados resulta ser la condición idónea para su proliferación [5]. Este conocimiento y la disponibilidad de las variables meteorológicas recogidas por las estaciones de medida del ITACYL y de la AEMET motivan la selección de las características meteorológicas como variables explicativas de la aparición de la plaga. En esta sección se presentan dichas variables e información sobre las estaciones que han anotado las mismas.

4.1.1. Variables meteorológicas

Las variables meteorológicas se encuentran recogidas en dos conjuntos diferentes por el hecho de ser registradas por estaciones pertenecientes a dos agencias diferentes. En el capítulo 5, limpieza y tratamiento de los datos, se describen las diferencias de estos dos conjuntos así como las técnicas utilizadas para su unión y para la imputación de valores ausentes debidos a la desconexión de las estaciones meteorológicas. A pesar de las diferencias, las variables que recoge cada estación son las mismas:

- **Date.** Fecha y hora de la toma de las medidas.
- **Idest.** Código identificador de la estación meteorológica que registra las variables meteorológicas.
- **Temperatura.** Medida en grados centígrados.
- **Precipitaciones.** Medidas en litros por metro cuadrado.
- **Humedad.** Medida en porcentaje de vapor de agua.
- **Velocidad del viento.** Medida en metros por segundo.
- **Dirección del viento.** Dirección del viento medida en metros por segundo (m/s).
- **Radiación.** Medida en watios por metro cuadrado (W/m^2).

4.1.2. Información de estaciones meteorológicas

Se dispone de una base de datos relacional a la que se accede mediante el sistema de gestión de bases de datos SQLite. Esta base contiene información relevante sobre las distintas estaciones meteorológicas que permitirá principalmente localizar las estaciones en el mapa, calcular distancias entre las mismas y calcular la distancia entre las estaciones y las parcelas inspeccionadas. Las características disponibles son:

- **Identificador.** Identificador de la estación meteorológica
- **Provincia.** Código de la provincia en la que se encuentra la estación.
- **Municipio.** Código del municipio en el que se encuentra la estación.
- **Comunidad autónoma.** Código de la comunidad autónoma en la que se encuentra la estación.
- **Nombre.** Nombre de la estación meteorológica.
- **X.** Coordenada proyectada X de la estación.
- **Y.** Coordenada proyectada Y de la estación.
- **Altitud.** Altitud en metros.
- **Baja.** Fecha a partir de la cual la estación no toma medidas, si es que existe tal fecha.
- **Indicador estación.** Valor binario que indica si la estación es propia del ITACYL o es propia de la AEMET.

- **Frecuencia.** Frecuencia con que se toman las medidas. Las estaciones que anotan las medidas cada hora tiene una frecuencia horaria que se indica con una H, el resto tienen un valor NULL.
- **Geometría.** Geometría de tipo puntual que indica la localización de la estación en coordenadas proyectadas (sistema de referencia EPSG:25830).

4.2. Datos de campo

En esta sección se presenta el conjuntos de datos obtenido por el Observatorio de plagas del ITACYL descrito en la introducción del Trabajo Fin de Grado, sección 1, así como la información disponible sobre las distintas parcelas de cultivo de Castilla y León. Esta información permitirá localizar las inspecciones en el mapa, relacionarlas con estaciones meteorológicas cercanas, estudiar posibles influencias en las predicciones de variables y examinar correlaciones con la variable respuesta.

4.2.1. Inspecciones

Para la creación de los modelos predictivos se dispone de los resultados de las inspecciones realizadas entre 2015 y 2021 en distintas parcelas de trigo de Castilla y León en riesgo de ser infectadas por la plaga roya por técnics del ITACYL. El conjunto de datos contiene 5912 instancias que identifican y describen a cada una de las parcelas mediante:

- **Date.** Fecha en la que se realizó la inspección.
- **Year.** Año en en el que se realizó la inspección.
- **Referencia.** Número de referencia de la parcela.
- **Provincia.** Código de la provincia en la que se encuentra la parcela.
- **Municipio.** Código del municipio en el que se encuentra la parcela.
- **Estado fenológico.** Variable que indica la etapa del periodo vegetativo en el que se encuentra la plantación en el momento de ser inspeccionada: semilla seca, comienza la imbibición de la semilla, imbibición completa de la semilla, radícula emergida de la semilla ...
- **Número de unidades muéstrales infectadas.** En cada inspección se evalúan cinco muestras de cultivo de diferentes zonas de la parcela. Esta variable registra las unidades muéstrales en las que se ha detectado el desarrollo de roya y toma un valor entero comprendido entre 0 y 5.

4.2.2. Información de parcelas

Al igual que la información de las estaciones meteorológicas, este conjunto de datos se presenta en una base de datos relacional accesible mediante SQLite. Contiene todas las parcelas registradas en Castilla y León con la siguiente información:

- **Referencia.** Código de referencia de la parcela.
- **Provincia**.Código de la provincia en la que se encuentra la parcela.

- **Municipio.** Código del municipio en el que se encuentra la parcela.
- **Agregado.** Código del agregado.
- **Zona.** Código identificador de la zona.
- **Poligono.** Código identificador del polígono.
- **Parcela.** Código de la parcela.
- **Recinto.** Código del recinto.
- **SIGPAC.** Código que le otorga la SIGPAC a la parcela.
- **Superficie.** Superficie medida en metros cuadrados.
- **Perímetro.** Perímetro medido en metros.
- **Geometry.** Identificador de referencia espacial con coordenadas geográficas en en sistema de referencia 25830.

Las variables de mayor interés son el código de referencia de la parcela, para poder relacionar esta información con las parcelas inspeccionadas y la variable *Geometry* para poder localizarlas en el mapa, calcular las distancias a las que se encuentran de las estaciones meteorológicas y asignar su zona agroclimática.

4.2.3. Tabla con estados fenológicos codificados

La variable estado fenológico que anota el inspector en el conjunto de datos presentado en la sección 4.2.1 se encuentra descrita con una cadena de caracteres poco práctica para el estudio. Una codificación numérica ofrece un mejor manejo de la variable y la tabla de datos en formato `CSV` incluida en el anexo A.1 se utiliza precisamente para realizar el mapeo entre las siguientes dos variables:

- **Estado fenológico.** Cadena de caracteres que representa un estado fenológico.
- **Número.** Valor entre cero y nueve.

4.2.4. Información sobre el territorio de Castilla y León

Se dispone de información relevante sobre las condiciones del territorio de Castilla y León que permiten ampliar las características de las parcelas. Esta información se encuentra accesible en formato ráster tratable por QGIS, en el *Atlas Agroclimático de Castilla y León*, un compendio de mapas que trata de describir el clima, la agricultura y la ganadería de Castilla y León [32].

Los rásteres son fotografías aéreas digitales, imágenes de satélite, imágenes digitales o incluso mapas escaneados. Como se ha indicado en la sección 2.4, un conjunto de datos ráster esta compuesto de filas y columnas de píxeles, conocidas como celdas. Cada píxel representa una región geográfica, y el valor en ese píxel representa alguna característica de dicha región [33]. Las variables a las que se accede se encuentran disponibles en el *Atlas Agroclimático de Castilla y León* y son:

- **Altitud.** La altitud de cada punto de Castilla y León permite obtener la altitud de las parcelas, calcular la diferencia de altitud de estas parcelas con las estaciones y evaluar el ajuste de los modelos. El ráster que se utiliza tiene el aspecto de la Figura 4.2.

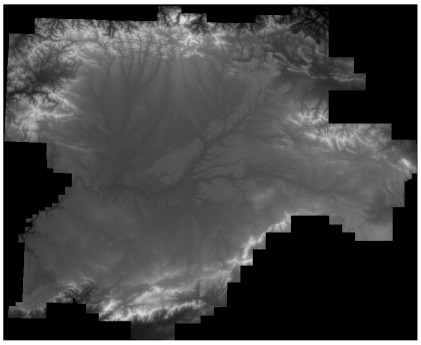


Figura 4.2: Raster de las altitudes de Castilla y León

- **Zonas agroclimáticas.** Las zonas agroclimáticas se caracterizan por tener características interrelacionadas entre el clima y los sistemas de cultivo. El entendimiento de estas interrelaciones ayuda a tomar mejores decisiones en el manejo agronómico de los cultivos [34]. El mapa es fácilmente representable con QGIS y la variable *Geometry* permite realizar una intersección con las parcelas y obtener la zona agroclimática correspondiente a cada parcela.

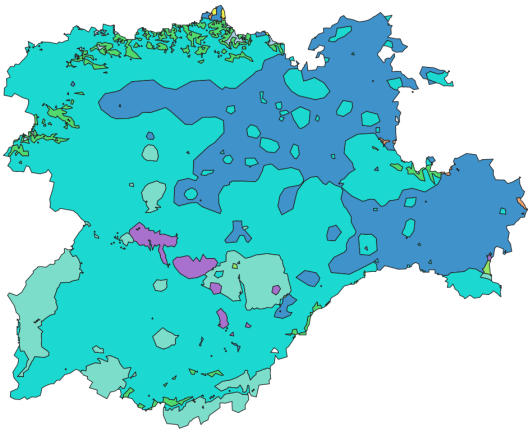


Figura 4.3: Zonas agroclimáticas de Castilla y León

- **Precipitaciones medias en los últimos 30 años.** Se dispone de las precipitaciones medias por mes registradas por las estaciones de AEMET en el treintenio 1981-2010. El formato de los valores es un ráster cuyo aspecto es similar al de la figura 4.2.

5. Limpieza y tratamiento de los datos

La metodología fundamental para la ciencia de datos diseñada por *International Business Machines Corporation* o IBM hace referencias a las etapas de comprensión y preparación de los datos representadas en la figura 5.1. Entre las actividades de esta etapa se encuentra la limpieza de los datos, la combinación de datos de múltiples fuentes y la transformación de variables más útiles [8].

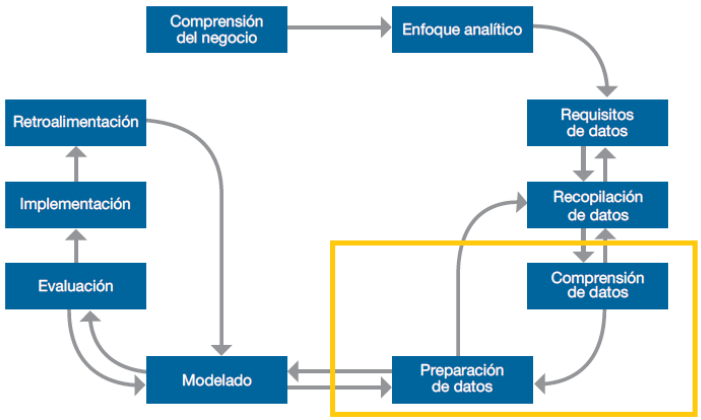


Figura 5.1: Metodología Fundamental para la Ciencia de Datos - Comprensión y preparación de los datos (extraída de [8])

En este capítulo se incluyen los procedimientos realizados para la unión de los ficheros de datos meteorológicos procedentes de diferentes fuentes y los procedimientos de limpieza de estos mismos datos y del conjunto de datos de campo. Se describe la forma de imputar valores faltantes y la forma de ampliar las variables de ambos conjuntos con variables existentes en los ficheros descritos en el capítulo 4. Por último se realiza una evaluación de las imputaciones realizadas, un análisis de la calidad de los datos y una evaluación de la distancia entre las parcelas y las estaciones.

5.1. Tratamiento variables meteorológicas

En los datos meteorológicos se tratan dos aspectos. El primero de ellos está relacionado con la frecuencia de las mediciones. Los datos meteorológicos provenientes de las estaciones del ITACYL se encuentran registrados con frecuencia medio-horaria, es decir, para cada hora se tienen dos mediciones, sin embargo, los variables registradas por las estaciones meteorológicas de la AEMET cuentan con una medición por hora. El segundo aspecto hace referencia a los datos faltantes debidos a fallos de las estaciones meteo-

rológicas o desconexiones que generan periodos de tiempo sin mediciones para las variables meteorológicas.

A continuación se detallan tanto las técnicas utilizadas para la minimización de los huecos por datos faltantes como el tratamiento previo para el ajuste de las frecuencias de los datos en cuatro pasos.

5.1.1. Primer paso: homogeneización de la escala temporal e imputación temporal

Se parte de las bases de datos proporcionadas por el ITACYL con la frecuencia semi-horaria y la base de datos de la AEMET con frecuencia horaria. Se accede a las bases de datos y se generan dos ficheros para cada conjunto de estaciones y para cada uno de los años estudiados 2015-2021.

- **Ficheros downloaded.** Se muestran en la Figura 5.2 y contiene los datos en bruto, sin ningún tratamiento, para poder estudiar los huecos o espacios temporales sin registro de la variables meteorológicas.

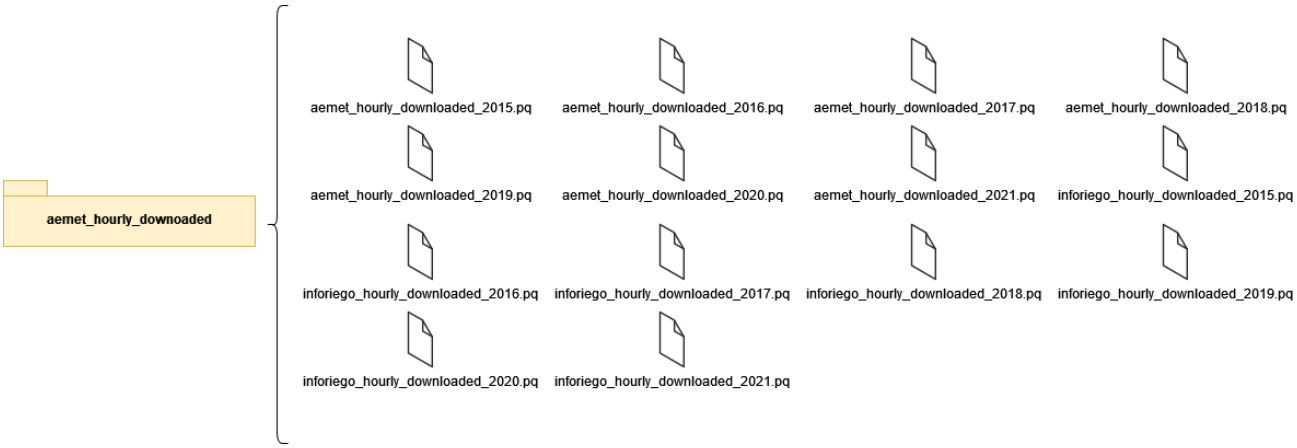


Figura 5.2: Ficheros de datos *downloaded*

- **Ficheros resampled.** Se muestran en la Figura 5.3. Para cada conjunto de estaciones, ITACYL y AEMET y para cada año se obtiene un fichero *resampled* tras aplicar una transformación de las series temporales a escala horaria, en el caso de los datos provenientes de las estaciones del ITACYL y un rellenado de los huecos de un máximo de cuatro horas para ambos conjuntos de estaciones mediante interpolación temporal.

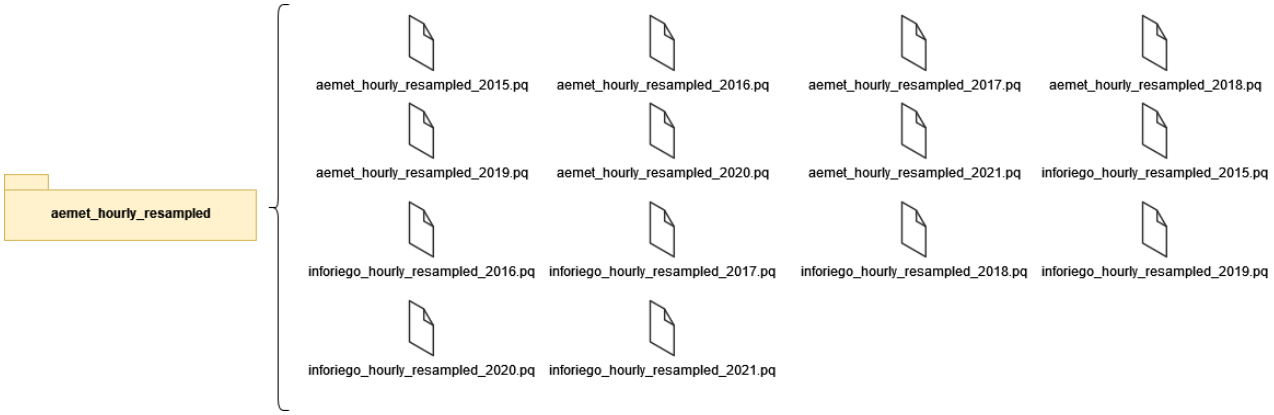


Figura 5.3: Ficheros de datos *resampled*

La imputación de valores en huecos de menos de cuatro horas hace referencia al tiempo porque toma como referencia valores de las variables de las propias estaciones en distintos puntos del tiempo. En concreto se realiza una interpolación lineal, un caso particular de la interpolación general de Newton que utiliza un polinomio de interpolación de grado uno. Se ajusta a los valores en los puntos x_1 o valor que toma la variable antes del hueco y el valor x_2 correspondiente al valor siguiente al hueco. La interpolación se calcula de la siguiente manera [35]:

$$f(x|x_1; x_2) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) \tag{5.1}$$

5.1.2. Segundo paso: agrupación de los datos

En los ficheros *resampled* se dispone de los datos con las series temporales de todas las estaciones en la misma escala horaria y los huecos de menos de cuatro horas rellenados mediante interpolación lineal. El siguiente paso es unir los ficheros de los datos meteorológicos registrados por las estaciones del ITACYL con los datos meteorológicos registrados por las estaciones de la AEMET en un único fichero por año llamado *raw* (ver esquema en la Figura 5.4).

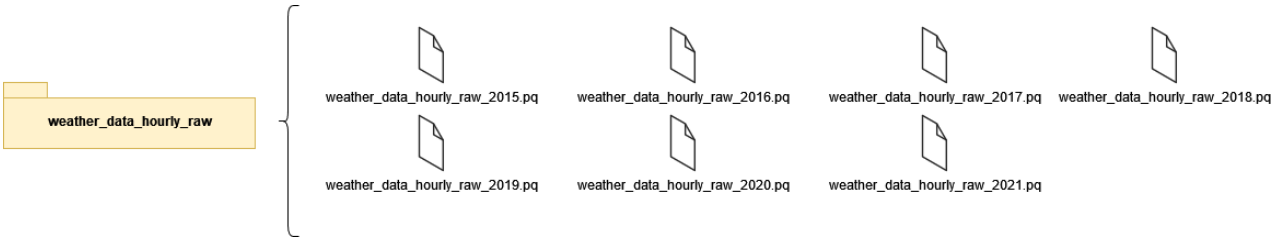


Figura 5.4: Ficheros de datos *raw*

5.1.3. Tercer paso: interpolación geográfica

La interpolación geográfica en este trabajo se utiliza para rellenar los valores ausentes de las variables meteorológicas en un periodo mayor de cuatro horas y menores de siete días. Esta imputación adquiere el adjetivo geográfica por el hecho de recurrir a valores tomados por estaciones meteorológicas cercanas y la técnica utilizada es *Inverse distance weighting* (IDW) o inverso de la distancia ponderado. IDW toma los datos de estaciones cercanas y considera la distancia a la que se encuentran estas, de manera que, estaciones más alejadas contribuyen menos que las estaciones cercanas.

El inverso de la distancia ponderado es un método determinista para interpolación multivariante. Es un método determinista porque ante unas entradas conocidas siempre proporciona las mismas salidas y es un método para interpolación multivariante porque la interpolación la realiza sobre funciones de más de una variable. La estimación de un valor z en una localización x es una media ponderada de las observaciones cercanas [36]:

$$\hat{z}(x) = \frac{\sum_i^n w_i z_i}{\sum_i^n w_i} \tag{5.2}$$

donde

$$w_i = |x - x_i|^{-\beta} \qquad \beta \geq 0 \tag{5.3}$$

$|.|$ corresponden a la distancia euclídea y β determina el grado con el que los puntos cercanos son preferidos sobre los puntos alejados, de manera que $\beta = 1$ correspondería a una relación inversa y $\beta = 2$ a una relación inversa al cuadrado.

Las imputaciones geográficas se realizan sobre los datos originales, es decir, sin las imputaciones lineales. Se consideran las coordenadas x, y, z o altitud de las estaciones, se toman como estaciones cercanas aquellas que se encuentran a una distancia menor de 50 kilómetros y una altitud menor de 150 metros y solo se realizan si hay un mínimo de tres estaciones cercanas.

5.1.4. Cuarto paso: unificación ficheros

Finalmente se agrupan las imputaciones temporales con las geográficas y se obtiene un único fichero por año con las variables tomadas por todas las estaciones, estaciones del ITACYL y estaciones de la AEMET. Los ficheros resultantes se muestran en la Figura 5.5

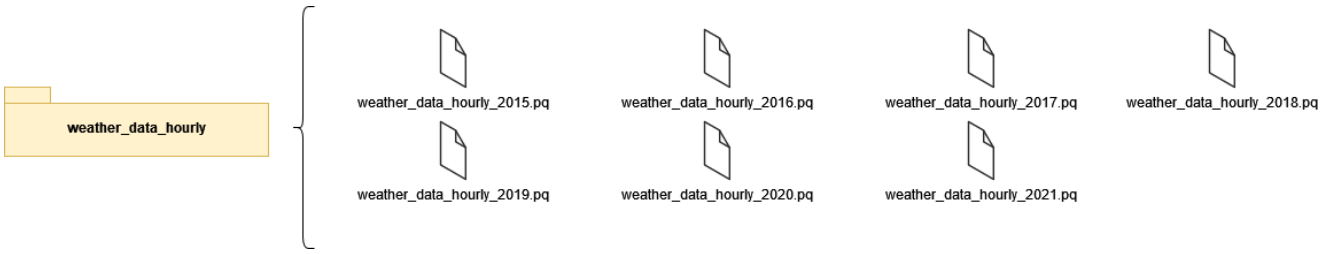


Figura 5.5: Ficheros de datos *hourly*

5.2. Tratamiento de los datos de campo

El tratamiento de los datos de campo consiste en la ampliación de los datos de las parcelas con el fin de ampliar los datos de las inspecciones realizadas y la creación de un nuevo conjunto de datos que permita guardar una relación entre parcelas y estaciones.

5.2.1. Ampliación de los datos de parcelas

Se amplían los datos presentados en la sección 4.2.2 con las variables zona agroclimática y altitud. Para asignar los valores de estas variables a las parcelas se realiza una intersección de las mismas con los ráster presentados en la sección 4.2.4. Se obtiene el valor de la altitud mediante la intersección representada en la figura 5.7 con la interfaz de `QGIS` y seguidamente se añade el valor de la zona agroclimática mediante la siguiente consulta `SQL` que hace uso de la librería `Spatialite`:

```
CREATE TABLE parcels_altcl AS
SELECT ogc_fid, refrec, c_provincia, c_municipio, c_agregado, c_poligono,
c_parcela, c_recinto, anio, c_uso_sigpac, c_coef_reg, m_pendiente, l_sup,
l_perimetro, altitud, z.climId,
z.climType
FROM parcels_alt  AS p, agroclimatic_zones AS z
WHERE st_intersects(st_centroid(p.geometry),z.geometry)=1
```

En el anterior código, la función `st_centroid` permiten sacar el centro de la parcela y la función `st_intersects` realizar la intersección de este punto que representa el polígono geométrico de la parcela con el mapa de las zonas agroclimáticas. Esta misma intersección se podría realizar con la interfaz que ofrece `QGIS` de la forma que aparece representada en la figura 5.6, siendo un procedimiento algo más costoso.

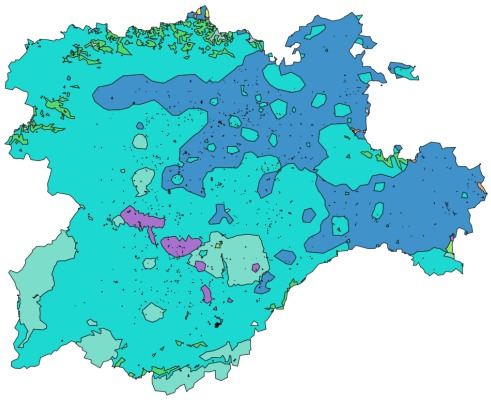


Figura 5.6: Intersección parcelas con zonas agroclimáticas obtenida con QGIS

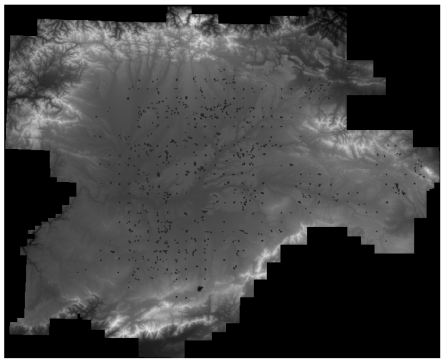


Figura 5.7: Intersección parcelas con altitudes obtenida con QGIS

Tanto en la figura 5.6 como en la figura 5.7 las regiones que aparecen como pequeños puntos son las correspondientes a todas las parcelas de CYL.

5.2.2. Localización de las parcelas respecto a las estaciones

Se crea una base de datos que relaciona las parcelas de Castilla y León con las estaciones meteorológicas mediante una consulta `SQL`. Esta base de datos permite un acceso inmediato a las distancias y a las diferencias en altitud entre las parcelas y las estaciones meteorológicas, variables necesarias para la futura caracterización de las inspecciones con los datos meteorológicos tomados por las estaciones cercanas. Se guardan las siguientes variables:

- **Referencia.** Número de referencia de la parcela.
- **Identificador.** Identificador de la estación meteorológica.
- **Diferencia de distancia.** Diferencia de la distancia de la estación al centro de la parcela en metros.
- **Diferencia de altitud.** Diferencia en valor absoluto de la altitud de la estación meteorológica a la altitud de la parcela en metros.
- **Altitud.** La altitud de la parcela.

5.2.3. Tratamiento de la variable zonas agroclimáticas

Los datos relativos a las zonas agroclimáticas presentados en la sección 4.1.2 contienen los nombres de las zonas agroclimáticas con espacios y tildes. Para un mejor uso de los mismos se crean las abreviaturas que se describen a continuación de la forma "abreviatura - nombre correspondiente" mediante una actualización a la base de datos que contiene dicha variable. Además se crea un identificador numérico para cada una de las abreviaturas.

- CONHF - dfc Cont. verano húmedo y frío
- CONHS - dfb: Cont, verano húmedo y suave
- ATLAN- cfb: Atlántico
- CONSS - dsb: Cont. verano seco y suave
- CONSF - dsc: Cont. verano seco y frio
- OCEN- csb: Oceanico verano seco
- ESTEF - BSk: Estepario Frío
- MEDIT - csa: Mediterraneo
- SUBTH - cfa: Subtropical húmedo

La consulta `SQL` que permite renombrar las variables y crear un identificador único es la siguiente:

```
alter table Clasificación_Köppen_26082013 rename to zonas

update zonas set clasificac='MEDIT' where clasificac= 'Csa:Mediterráneo'
update zonas set clasificac='MEDIT' where clasificac= 'Csa:Mediterránaeo'
update zonas set clasificac='OCEAN' where clasificac= 'Csb:Oceánico verano seco'
update zonas set clasificac='ATLAN' where clasificac='Cfb:Atlántico'
update zonas set clasificac='SUBTH' where clasificac='Cfa:Subtropical húmedo'
update zonas set clasificac='ESTEF' where clasificac= 'BSk:Estepario Frío'
update zonas set clasificac='CONSS' where clasificac= 'Dsb:Cont. verano seco y suave'
update zonas set clasificac='CONSF' where clasificac= 'Dsc:Cont. verano seco y frío'
update zonas set clasificac='CONHF' where clasificac= 'Dfc:Cont. verano húmedo y frío'
update zonas set clasificac='CONHS' where clasificac= 'Dfb:Cont. verano húmedo y suave'

create table agroclimatic_zones(
   climId integer primary key,
   climType TEXT (5),
   geometry geometry
);

insert into zonas_climaticas
select *
from zonas
```

5.3. Evaluación de la interpolación temporal

La interpolación temporal realizada en la sección 5.1.1 rellena los valores que faltan en un espacio de cuatro horas en las variables meteorológicas. Para observar el efecto que tienen se utilizan los datos en bruto, sin imputaciones temporales, *downloaded* y los datos con imputaciones temporales *resampled*. Se agrupan los valores con una frecuencia diaria y se calcula el valor absoluto de las diferencias para cada una de las variables.

En la figura 5.8 se puede ver que la diferencia de temperatura apenas alcanza el grado centígrado. En cuanto a la humedad, la diferencia de porcentaje de vapor de agua tiene una mayor dispersión entre los valores cero y uno . La diferencia de precipitación diaria es mínima. En consecuencia podemos concluir que las imputaciones temporales no afectan a la distribución de las variables, la diferencia de los valores antes de la imputación y después de la imputación es mínima.

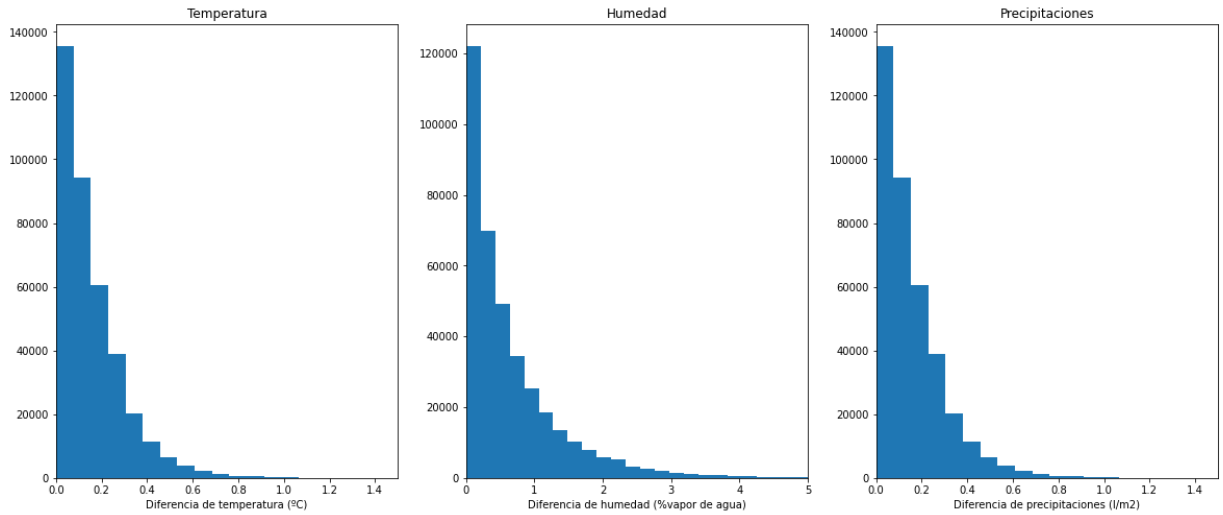


Figura 5.8: Distribución de las diferencias diarias de los datos con interpolación temporal y en bruto

5.4. Evaluación de la interpolación geográfica

La interpolación geográfica realizada en la sección 5.1.3 en las variables meteorológicas imputa en los huecos con un espacio temporal de entre cuatro horas y siete días la media de los valores que han registrado las estaciones cercanas. En esta sección se evalúa la diferencia semanal de los valores de las variables meteorológicas entre los ficheros en bruto, sin imputaciones *downloaded* y los ficheros con las imputaciones geográficas.

En la representación de las diferencias de las Figuras 5.9, 5.10 y 5.11 cada punto corresponde a una estación. En el eje x se representa el valor absoluto de la diferencia del valor de la variable meteorológica en los datos en bruto y en los datos con imputación temporal en un día de la semana. En el eje y se representa la completitud de los datos que ha habido esa semana para esa variable y esa estación, es decir, el porcentaje de horas en las que se ha registrado un valor para cada variable. Se tiene en cuenta

el porcentaje de datos completos para estudiar si este porcentaje influye en las diferencias de los valores de las variables. Existe la posibilidad de obtener mayores diferencias en aquellas semanas en las que ha habido pocos datos y la mayoría han sido imputaciones o medias calculadas.

En la figura 5.9 se representa la diferencia de la variable temperatura. Se puede ver que gran proporción de estas diferencias se encuentran entre cero y dos grados centígrados. En la figura 5.10 se observa la humedad medida en porcentaje de vapor de agua con una diferencia en valor absoluto entorno al cero y diez por ciento.

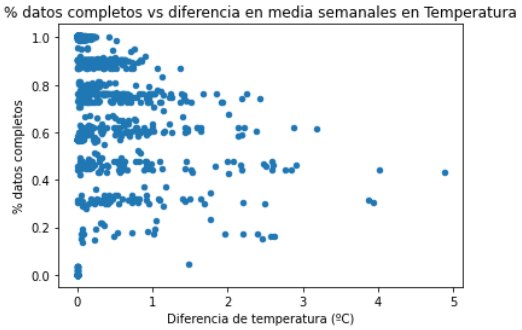


Figura 5.9: Diferencia de la temperatura tras la imputación geográfica

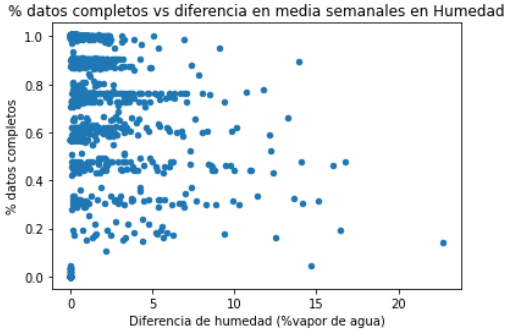


Figura 5.10: Diferencia de la humedad tras la imputación geográfica

Por último, en la tercera variable estudiada, las diferencias de precipitaciones, representadas en la figura 5.11 se encuentran entorno a 0 y 0.25 litros por metro cuadrado. Para las tres variables meteorológicas estudiadas las diferencias no son desmesuradas y son parecidas en los distintos niveles de completitud de datos, es decir, las imputaciones geográficas no producen resultados disparatados en los espacios temporales con más huecos.

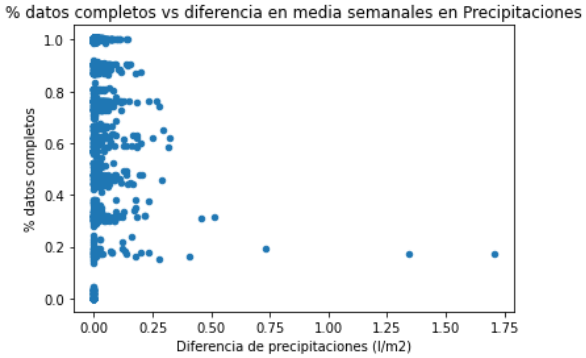


Figura 5.11: Diferencia de las precipitaciones tras la imputación geográfica

5.5. Evaluación de la calidad de los datos meteorológicos

Una vez realizadas y analizadas las imputaciones temporales y geográficas de los datos meteorológicos interesa conocer la distribución de los huecos o espacios temporales sin mediciones para las variables

restantes. Se obtiene para cada estación y día tres variables, una por cada variable meteorológica, que indican el número de huecos u horas sin medición que se tiene para esa variable en ese día. Si en un día dado la estación sí ha registrado las medidas, las variables hueco tendrán el valor cero y si por lo contrario, no ha habido un registro durante x horas, el valor de la variable hueco será x. El número total de estaciones meteorológicas que se tienen son 167.

En la figura 5.12 se representa para las variables temperatura, humedad y precipitación la distribución de las horas en las que las estaciones no han registrado algún valor. Se observa que tanto para la variable temperatura como para la variable humedad la mayor fecuencia se encuentra entre cero y cien horas. Sin embargo, para la variable precipitación existen más huecos y lo más frecuente es que estos huecos sean entre cien y doscientas horas. Tras la obtención de la distribución de los huecos de la figura 5.12 se decide excluir aquellas estaciones que tienen algún hueco de más de un mes o 720 horas. Con este filtro las estaciones restantes son 161.

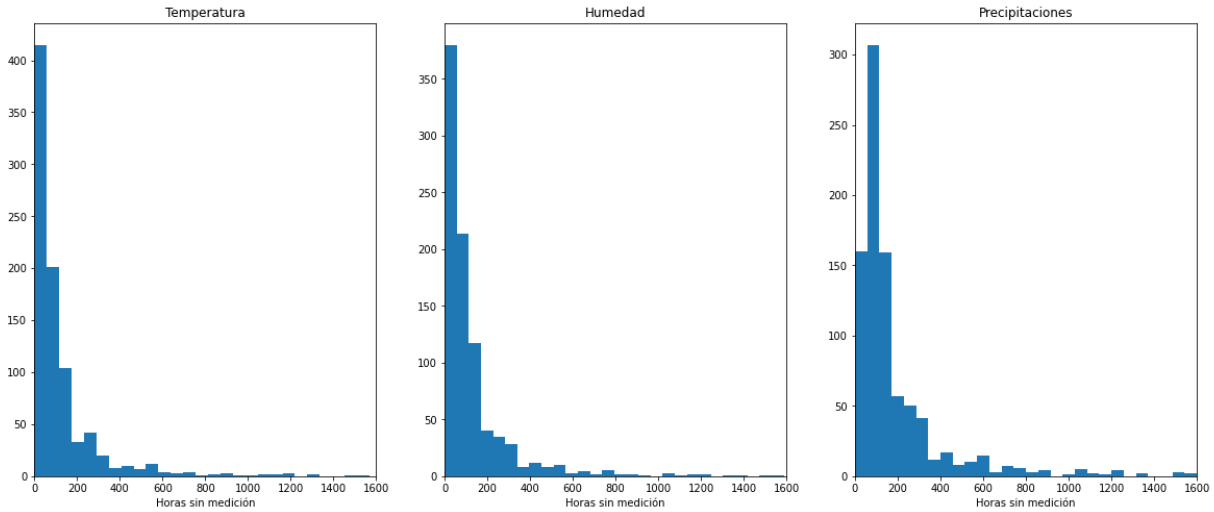


Figura 5.12: Distribución de las horas en los que no hay registro de cada una de las variables meteorológicas

5.6. Evaluación de las distancias entre parcelas y estaciones

En una primera aproximación se asigna a cada parcela la estación más cercana con la finalidad de caracterizar a estas parcelas con las variables meteorológicas registradas por las estaciones de medición correspondientes. Una vez realizada dicha asignación, se evalúa la distancia y la diferencia en altitud entre cada parcela y la estación asignada.

En la figura 5.13 se representa la diferencia de altitud en metros y en la figura 5.14 la distancia, también en metros, entre las distintas parcelas y la estación más cercana que tiene asignada cada una. Se calcula que un 12,18 % de las parcelas tienen asignada una estación que se encuentra a una diferencia de altitud de más de 150 metros y un 3.23 % de las parcelas tienen asignada una estación a más de 20 kilómetros de distancia.

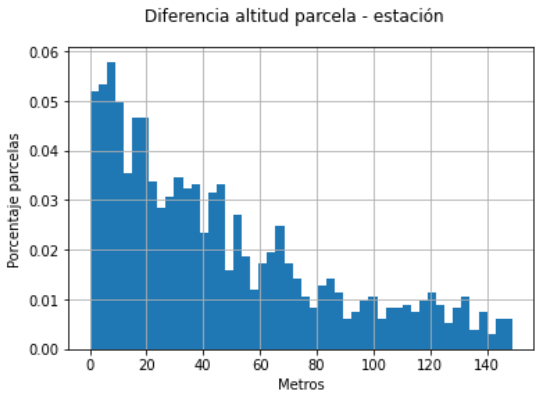


Figura 5.13: Diferencia de altitud parcela - estación

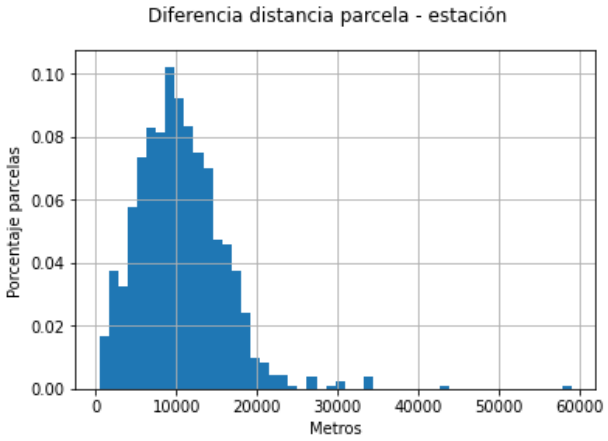


Figura 5.14: Distancia parcela - estación

Todas las estaciones que se encuentran a más de veinte kilómetros y a una diferencia de altitud mayor de ciento veinte metros, sobrepasan los límites marcados en la figura 5.15 con líneas discontinuas, se consideran problemáticas por el hecho de no caracterizar las condiciones meteorológicas de las parcelas con precisión. Se plantean dos soluciones para la asignación de estaciones.

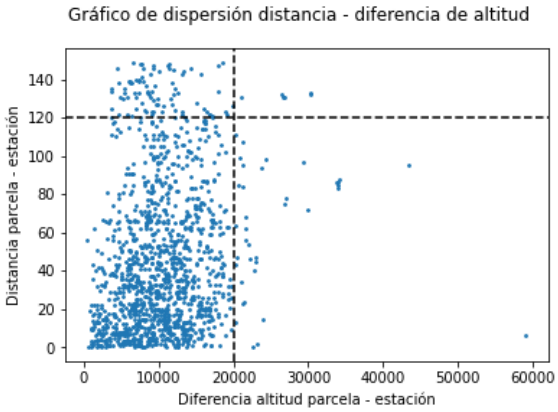


Figura 5.15: Dispersión entre diferencia de altitud y distancia parcela - estación

Primera solución para asociar parcelas y estaciones meteorológicas

La primera solución para la asignación de estaciones meteorológicas a las parcelas consiste en asignar a cada parcela una estaciones que se encuentre situada a una distancia máxima de cincuenta kilómetros y una diferencia de altitud máxima de ciento setenta metros.

Se calcula el número de estaciones a menos de 50 kilómetros y a una diferencia de altitud menor de 170 metros para cada parcela. En la gráfica 5.16 se representa en el eje x el número de estaciones y en el eje y el porcentaje de parcelas que tienen asignadas ese número de estaciones. Se puede ver en esta misma gráfica que la mayoría de las parcelas tienen asignadas más de cinco estaciones y se calcula que el 99,69 % de las parcelas tienen al menos tres estaciones meteorológicas.

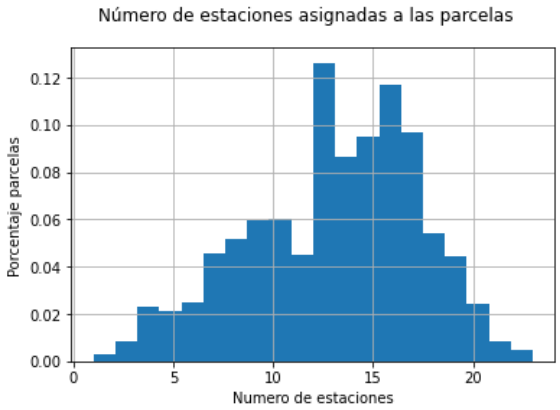


Figura 5.16: Porcentaje de parcelas frente a número de estaciones que tienen asignadas

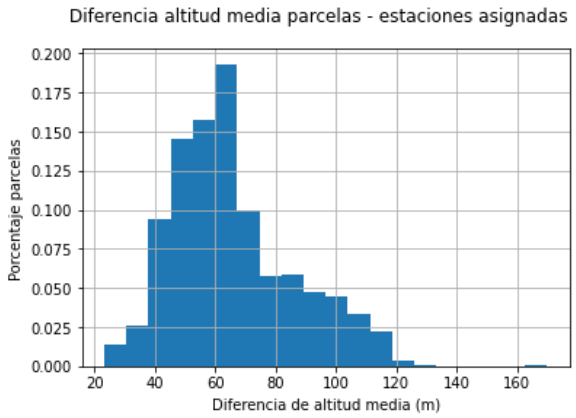


Figura 5.17: Diferencia de altitud media parcela - estaciones

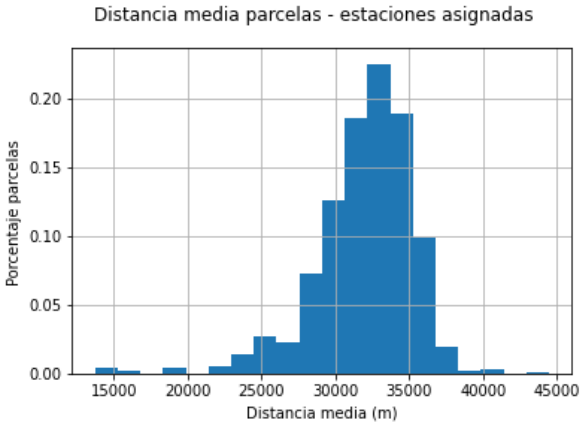


Figura 5.18: Distancia media parcela - estaciones asignadas

Se consigue disminuir el porcentaje de parcelas que tienen asignadas estaciones a una diferencia de altitud media mayor de 120 metros, tal como se puede ver en la figura 5.17. A pesar de ello, la distancia media a la que se encuentran las estaciones asociadas a las parcelas aumenta. Se puede ver este aumento comparando la figura 5.14 con la figura 5.18.

Segunda solución para asociar parcelas y estaciones meteorológica

Como solución alternativa se propone, en lugar de asignar estaciones meteorológicas a las parcelas, tratar de caracterizar a las estaciones con parcelas que se encuentren a menos de veinte kilómetros y a una diferencia de altitud menor de ciento cincuenta metros.

En la figura 5.19 se representa en el eje y el porcentaje de estaciones que tienen asignadas el número de parcelas indicadas en en eje x. Aproximadamente el 3 % de las estaciones no tienen parcela asignada y el resto tienen un número de parcelas mayor de diez.

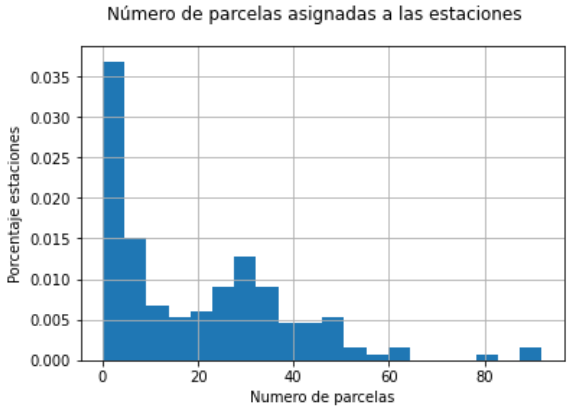


Figura 5.19: Porcentaje de estaciones frente a número de parcelas que tienen asignadas

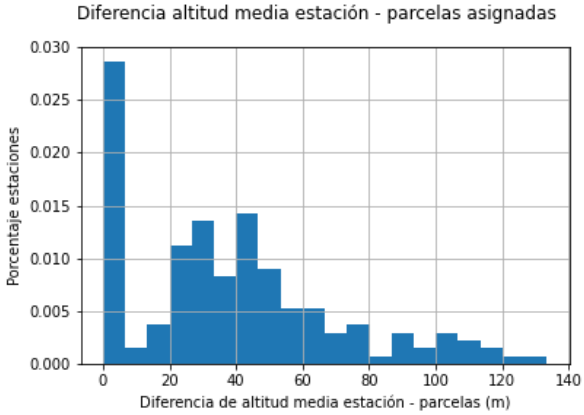


Figura 5.20: Diferencia de altitud media parcela - estaciones

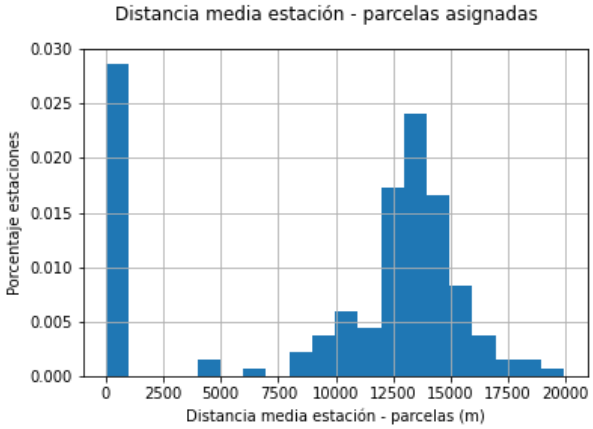


Figura 5.21: Distancia media parcela - estaciones asignadas

Aproximadamente en el 97 % de los casos, el número de parcelas asignadas a las estaciones es mayor de uno, por otro lado, tal como se puede ver en la figura 5.21, la distancia media entre la estación y las parcelas próximas no supera los 20 kilómetros y la diferencia de altitud media representada en la figura 5.20 es menor de 120 metros.

Con esta nueva asignación:

- Al tener más datos, se reduce el rango de distancias a veinte kilómetros y el rango de las altitudes a 160 metros.
- Se reduce el error en las variables meteorológicas, se utiliza una media de valores cercanos.
- Se aumenta el tamaño muestral y los patrones meteorológicos. Cada parcela puede aportar información a varias estaciones cercanas.
- Se consigue mayor robustez al evitar elegir una estación meteorológica poco representativa de la parcela.

Si tomamos las estaciones con al menos una parcela a menos de 20 kilómetros y a menos de 100 metros de diferencia de altitud se utilizan 161 estaciones. En estudios previos del ITACYL únicamente con las estaciones de inforiego y asignando estaciones a las parcelas, el número de estaciones utilizadas eran aproximadamente 30, en este Trabajo FIn de Grado se aumentan los datos incorporando un año más.

6. Creación del fichero para el modelado

En la sección 5.1 se hace referencia a la etapa de preparación de datos propuesta por la empresa *International Business Machines Corporation* o IBM. En esta etapa se incluye el proceso de ingeniería de características que conlleva la creación de variables explicativas adicionales, también conocidas como indicadores o características, a través de una combinación de conocimiento en el dominio y de variables existentes. Este procedimiento permite enriquecer el conjunto de indicadores y mejorar la precisión del modelo. Forma parte de las etapas de comprensión y preparación de los datos representadas en la figura 6.1 [8].

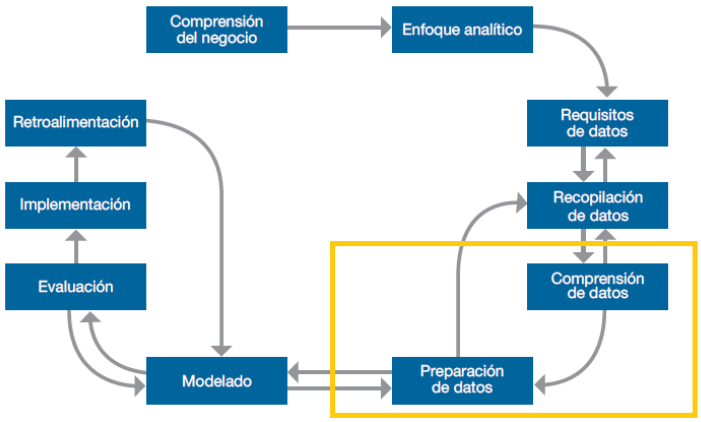


Figura 6.1: Metodología Fundamental para la Ciencia de Datos - Comprensión y preparación de los datos (extraída de [8])

En la sección 1.1 se describe el contexto en el que se desarrolla el proyecto y se mencionan dos papers tomados como referencia para la realización del mismo. En esta misma sección 1.1 se especifican los dos determinantes en la aparición de la plaga Roya. Siguiendo los estudios mencionados se deciden crear variables que describan estos dos determinantes:

- **Condiciones meteorológicas.** Las variables meteorológicas son directamente medibles pero se necesita saber su evolución en los días previos. Se utilizan resúmenes a diferentes rangos (10,30,60 días) con el fin de dar continuidad a las variables y considerar las condiciones meteorológicas de días previos. En los estudios mencionados se ha visto que estos rangos han resultado relevantes para medir evolución de diferentes plagas.
- **Evolución del cultivo.** La plaga solo se puede desarrollar en determinadas fases del cultivos. Cuanta más abundancia de cereal haya, es decir, más biomasa, existe una mayor facilidad de

desarrollo de la plaga. La evolución del cultivo depende de factores como la fecha de siembra, la variedad, las labores del agricultor... Estas variables no son recogidas durante la inspección de las parcelas, por ello se utilizan otras que podrían dar información sobre la evolución del cultivo. Entre las variables *proxy* utilizadas se encuentran:

- **Estado fenológico.** Ofrece información sobre el crecimiento del cultivo.
- **Mes.** La variable mes se encuentra relacionada positivamente con el crecimiento del cultivo, a medida que avanza la campaña, el cereal crece.
- **Abundancia de biomasa.** Esta variable puede ser medida mediante la abundancia de lluvias. La lluvia favorece el crecimiento del cultivo y por otro lado, lo limpia, limpiando así posibles plagas.
- **Zona agroclimática.** El clima define un ecosistema en el que puede convivir la plaga y el cultivo por lo tanto, se encuentra estrechamente relacionado con ambos.
- **Altitud.** La variable altitud afecta a la temperatura y al clima de una zona, esto condiciona tanto a los cultivos que se desarrollan como a las plagas.
- **Distancia.** La distancia de la parcela de la estación en la que se toman las medidas meteorológicas también afecta a las predicciones.

Seguidamente se describen las características creadas en los conjuntos de datos meteorológicos, las características creadas de las inspecciones para la aplicación de la segunda solución de la sección 5.6, el conjunto de datos definitivo utilizado para el modelado.

6.1. Obtención de características de los datos meteorológicos

De las variables meteorológicas descritas en el conjunto de datos del apartado 4.1.1, se seleccionaran para este estudio en un principio únicamente las variables *temperatura, humedad y precipitaciones*, por ser las más influyentes en la aparición de plaga [5]. Se calculan estadísticos resumen de las mismas y algunas variables relacionadas con las condiciones meteorológicas que influyen en la aparición de la plaga Roya. Se obtienen características a partir de los ficheros de datos obtenidos en la sección 5.1.

6.1.1. Resumen diario.

Los ficheros de datos diarios se obtienen a partir de de los datos horarios obtenidos en la sección 5.1. Para cada una de las variables temperatura, humedad y precipitación se obtienen los siguientes estadísticos:

- **Mínimo.** *temp_min, prec_min y hum_min.*
- **Máximo.** *temp_max, prec_max y hum_max.*
- **Media.** *temp_mean, prec_mean y hum_mean.*
- **Mediana.** *temp_median, prec_median y hum_median.*

- **Desviación estándar.** *temp_std, prec_std y hum_std.*
- **Suma.** *temp_sum, prec_sum y hum_sum.*
- **Quantil 10, 20, 30, 40, 60, 70, 80 y 90.** *temp_q10, prec_q10, hum_q10, temp_q20, prec_q20, hum_q20, temp_q30, prec_q30, hum_q30, temp_q40, prec_q40, hum_q40, temp_q60, prec_q60, hum_q60, temp_q70, prec_q70, hum_q70, temp_q80, prec_q80, hum_q80, temp_q90, prec_q90, hum_q90.*

A los anteriores estadísticos se añaden las siguientes variables de interés para la detección de plaga.

- **Integral térmica, grados día.** La variable integral térmica se encuentra desarrollada en el apartado 6.1.3. En este caso concreto se obtiene a partir de la temperatura media del día, de ahí su nombre grados día. Corresponde a los grados acumulados en los últimos 10 días, contabilizando solo grados que se encuentran entre los 5 y 25 grados. Se representa mediante *temp_int_gd.*
- **Lluvia acumulada.** Se calcula la lluvia acumulada durante los últimos 10, 30 y 60 días y se selecciona la correspondiente a las doce de la noche para caracterizar el día. Se representa mediante *prec_10d, prec_30d y prec_60d.*
- **Lluvia esperada.** Se obtiene la lluvia esperada en los últimos 30 y 60 días teniendo en consideración la lluvia de los últimos 30 años *perc_expected_30d, perc_expected_60d.*
- **Ratio precipitación acumulada frente esperada.** Para los 30 y 60 días. Se denominan *prec_exp_perc_30d* y *prec_exp_perc_60d* respectivamente y se obtiene de la siguiente forma:

$$ratio = \frac{acumulada - esperada}{esperada}$$

- **Ratio precipitación acumulada frente esperada sigmoide**. Se aplica la función sigmoide al ratio anterior, puesto que no se desea una progresión lineal. Se sabe que la lluvia influye en la cantidad de biomasa de manera positiva y se quiere que una vez alcanzado un valor la influencia sea alta. La variable obtenida se denomina *prec_exp_sgm_30d.*

$$\varphi(ratio) = \frac{1}{1 + e^{-ratio}}$$

Las variables descritas anteriormente se calculan para cada año y se guardan en ficheros distintos para cada año, tal como aparece en la Figura 6.2

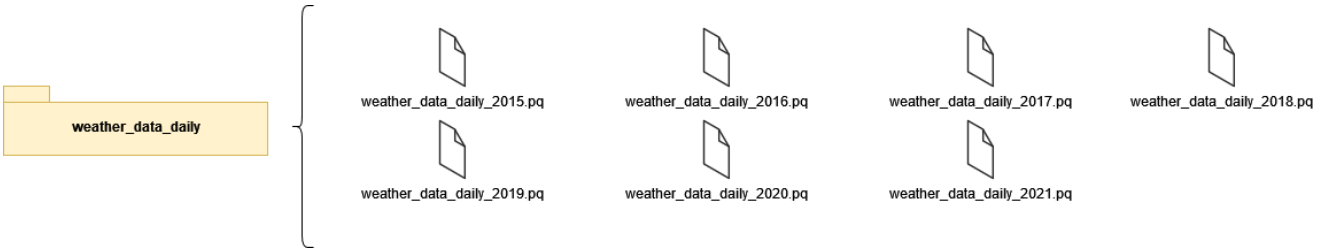


Figura 6.2: Ficheros de datos *weather_data_daily*

6.1.2. Resumen horario

Para cada una de las variables, temperatura, humedad y precipitaciones, se calculan para cada hora una ventana móvil de 15 días y una ventana móvil de 40 días con los siguientes estadísticos:

- **Mínmo.** *temp_min_15d_h, hum_min_15d_h, prec_min_15d_h, temp_min_40d_h, hum_min_40d_h, prec_min_40d_h.*
- **Máximo.** *temp_max_15d_h, hum_max_15d_h, prec_max_15d_h, temp_max_40d_h, hum_max_40d_h, prec_max_40d_h.*
- **Mediana.** *temp_median_15d_h, hum_median_15d_h, prec_median_15d_h, temp_median_40d_h, hum_median_40d_h, prec_median_40d_h.*
- **Media.** *temp_mean_15d_h, hum_mean_15d_h, prec_mean_15d_h, temp_mean_40d_h, hum_mean_40d_h, prec_mean_40d_h.*
- **Desviación estándar.** *temp_std_15d_h, hum_std_15d_h, prec_std_15d_h, temp_std_40d_h, hum_std_40d_h, prec_std_40d_h.*
- **Quantil 20, 40, 60 y 80.** *temp_q20_15d_h, hum_q20_15d_h, prec_q20_15d_h, temp_q20_40d_h, hum_q20_40d_h, prec_q20_40d_h, temp_q40_15d_h, hum_q40_15d_h,prec_q40_15d_h, temp_q40_40d_h, hum_q40_40d_h, prec_q40_40d_h, temp_q60_15d_h, hum_q60_15d_h, prec_q60_15d_h, temp_q60_40d_h, hum_q60_40d_h, prec_q60_40d_h, temp_q80_15d_h, hum_q80_15d_h, prec_q80_15d_h, temp_q80_40d_h, hum_q80_40d_h, prec_q80_40d_h*
- **Suma.** *temp_sum_15d_h, hum_sum_15d_h, prec_sum_15d_h, temp_sum_40d_h, hum_sum_40d_h, prec_sum_40d_h.*

A estos estadísticos se añaden las siguientes variables de interés para la detección de la plaga Roya:

- **Integral térmica, grados hora**. Como ya se ha mencionado anteriormente, la variable integral térmica se encuentra desarrollada en la sección 6.1.3. En este caso concreto se obtiene a partir de la temperatura horaria y se denomina grados hora. Esta variable, para cada hora, recoge los grados acumulados en los diez días anteriores.
- **Eventos de Roya.** Se sigue el modelo para la predicción plaga propuesto en el artículo *A Threshold-Based Weather Model for Predicting Stripe Rust Infection in Winter Wheat* [6] . Se registra plaga en una hora dada cuando durante 4 hora seguidas se han dado las siguientes condiciones:
  - La temperatura se encuentra ente 4ºC y 12ºC.
  - La humedad es mayor del 92 %.
  - Las precipitaciones son menores de 0.1ml.
- **Eventos de Roya con saltos.** Se modifica ligeramente el modelo descrito anteriormente y se permiten periodos de media hora en los que no se den las tres condiciones citadas relacionadas con la temperatura, humedad y precipitaciones.

Una vez calculadas las características horarias y diarias, se unen los ficheros y para ello, de las características horarias, se seleccionan las doce de la noche. Se dispone de un único fichero por año con todas las estaciones y todas las características siguiendo el esquema de la figura 6.3.

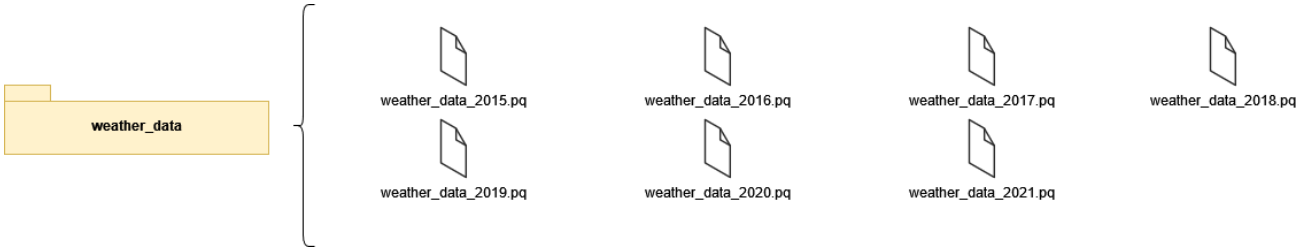


Figura 6.3: Ficheros de datos *weather_data*

6.1.3. Variable integral térmica

Las plantas responden a un ciclo vital marcado por los estados fenológicos descritos en el anexo A y estos estados se completan cuando la planta alcanza una temperatura más o menos concreta que puede variar en los distintos años. Aunque la velocidad de crecimiento esté influenciada por otros factores como podría ser la nutrición del suelo, la humedad, las lluvias, la radiación ... la variable temperatura es el factor más importante que induce el desarrollo de la planta a través de sus fases y se puede considerar que crece de forma lineal. Para poder calcular la integral térmica o el acumulado de la temperatura de la planta es necesario conocer dos valores [37]:

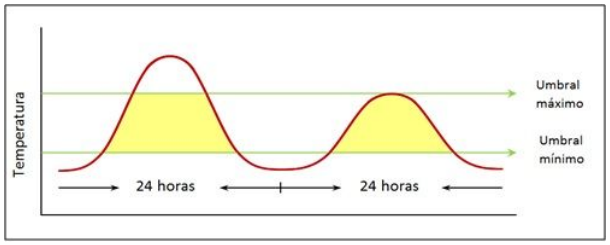


Figura 6.4: Integral térmica

- **Umbral inferior o temperatura base.** Se conoce también como umbral térmico inferior o temperatura cero de crecimiento a la temperatura por debajo de la cual la planta detiene su crecimiento por completo. Los valores que se encuentran por debajo de este valor no contabilizan para el desarrollo del cultivo o para el cálculo de la integral térmica.

- **Umbral superior o temperatura máxima de crecimiento.** El umbral térmico superior es el valor por encima del cual la planta detiene su desarrollo o este es muy lento. Los valores que se encuentran por encima de este valor tampoco contabilizan para el cálculo de la integral térmica.

Para cada planta, estos valores son distintos, así por ejemplo para el trigo, el umbral inferior se encuentra en 5ºC y el umbral superior o temperatura máxima de crecimiento en 25ºC, los grados comprendidos ente estos dos valores son los que se contabilizan para el cálculo de la integral térmica y son los que aparecen en la figura 6.4 en color amarillo [5].

6.2. Obtención de características de las inspecciones

Se parte del fichero de datos con las inspecciones que se han realizado en las distintas parcelas de trigo de Castilla y León descrito en la sección 4.2.1. Tras la evaluación de las distancias entre las parcelas y las estaciones meteorológicas de la sección 5.6 se decide seguir con la segunda solución propuesta y asignar a las distintas estaciones meteorológicas parcelas que se encuentran a menos de veinte kilómetros de distancia y a una diferencia de altitud menor de ciento cincuenta metros.

El objetivo es evaluar la afectación de las parcelas por la plaga Roya mediante variables que miden el desarrollo de la plaga y variables que miden el desarrollo del cultivo. Se calculan valores medios de las variables de las inspecciones asignadas en cada día a cada estación y se incluyen algunas variables que consideran la pervivencia de la plaga, es decir, si un días se han realizado una serie de inspecciones en las que se ha detectado la plaga Roya, se supone que esta pervive y si en los siguientes diez días si se realizase el mismo número de inspecciones, el número de detecciones sería el mismo. La pervivencia es un mecanismo que se utiliza para paliar el problema de no tener una toma de datos con un periodo concreto, las inspecciones se han realizado cada dos, tres semanas. Las variables que se calculan para cada estación son:

- **Día.** Día en el cuál se han realizado las inspecciones.
- **Identificador.** Identificador de la estación.
- **Número de inspecciones.** Suma del número de inspecciones que se han realizado en las parcelas cercanas a la estación en cada uno de los días.
- **Detectadas.** Suma del número de inspecciones en las que se ha detectado la plaga.
- **Porcentaje de plaga detectada.** Número de inspecciones con plaga detectada dividido entre el número total de inspecciones.
- **Número de inspecciones con pervivencia.** Número de inspecciones suponiendo que la plaga pervive diez días desde que fue detectada.
- **Detectadas con pervivencia.** Número de inspecciones con plaga detectada suponiendo pervivencia de la plaga en los diez días siguientes.
- **Porcentaje de plaga detectada con pervivencia.** Número de inspecciones con plaga detectada entre el número total de inspecciones realizados suponiendo la pervivencia de la plaga los diez días siguientes.
- **Altitud.** Media de la altitud de la parcelas que han contribuido al calculo de las inspecciones.
- **Distancia.** Media de la distancia de las parcelas a la estación meteorológica que contribuyen.
- **Estado fenológico.** El máximo estado fenológico que alcanzan las parcelas.
- **Clima.** Zona agroclimática de la estación.

- **Parcelas.** Una lista con las parcelas que han contribuido a la estación para comprobar cuántas parcelas han participado en las instancias creadas.

Una vez calculadas las anteriores características de las inspecciones aplicadas a las las estaciones se guardan en el fichero *parcel_features.pq*.

6.3. Creación del fichero final

Se realiza la unión de las características de las inspecciones con las características meteorológicas creadas en la sección 6.1, tal como se indica en la Figura 6.5 para la obtención de los datos finales que se utilizan para el modelado. Como variable respuesta se puede utilizar el porcentaje de plaga detectado y el porcentaje de plaga detectado con pervivencia.

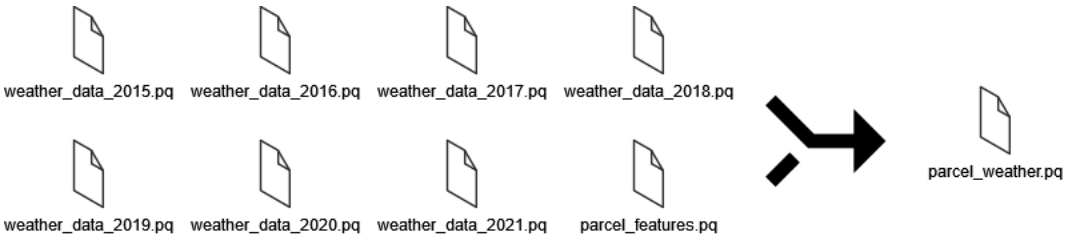


Figura 6.5: Fichero final para el modelado

El fichero inicial con las inspecciones de las parcelas del que se parte cuenta con 5.912 instancias, finalmente el fichero `parcel_weather.pq` dispone de 38.996 instancias. Cada instancia queda identificada por el identificador de la estación meteorológica y el día en el que se registran inspecciones en las parcelas cercanas. Las variables restantes son las variables calculadas para cada estación descritas en la lista inmediatamente anterior, las variables meteorológicas temperatura, humedad y precipitaciones del día y las 115 características creadas en la sección 6.1.

6.4. Comentario sobre la calidad de los datos

Los datos de las investigaciones en las que se basa este trabajo, *Disease–Weather Relationships for Powdery Mildew and Yellow Rust on Winter Wheat* [5] y *A Threshold-Based Weather Model for Predicting Stripe Rust Infection in Winter Wheat* [6], son datos obtenidos con un seguimiento exhaustivo de las parcelas. El error en las variables meteorológicas que caracterizan a las parcelas es mínimo debido a una distancia de la estación meteorológica respecto a las parcelas menor de cinco kilómetros y las inspecciones tienen un seguimiento diario.

Los datos que se utilizan en este trabajo, sin embargo, se encuentran condicionados por varios factores:

- **Condicionamiento meteorológico.** Las características meteorológicas considerabas para las parcelas tienen un error debido a las distancias considerada entre parcelas y estaciones y un error proveniente de la interpolación realizada.

- **Condicionamiento de las inspecciones.** Las inspecciones de seguimiento no se realizan con una periodicidad constantes, un alto porcentaje se visita cada 2 semanas pero el resto se visita cada 15-25 días dependiendo de la prioridad que se le de y la disponibilidad de recursos.
- **Condicionamiento del inspector.** Algunas de las características de la inspección, por ejemplo el estado fenológico pueden verse influidas por la subjetividad del inspector y no estamos modelando para modelar esa fuente de variabilidad.
- **Variedad del cultivo.** Hay variedades de trigo que son resistentes a plagas mientras que otras pueden presentar desarrollo de plaga más temprano. Es una variable que no ses tiene registrada, una variable latente que está afectando al modelo y que no se puede medir ni directa ni indirectamente.
- **Fecha de siembra.** El estado del cultivo depende de cuándo se haya plantado el mismo. No se tiene registrada esta variable y esto condiciona el desarrollo de la plaga en una zona y fecha.
- **Labores agrícolas.** Las labores del agricultor afectan a la evolución de la plaga, puede haber echado productos fitosanitarios para prevenir la plaga, puede haber dejado en barbecho la parcela y el germen de la plaga no se haya mantenido en el terreno, etc.
- **Cantidad de cereal en la zona**. En zonas de mucho cereal es más probable que se de y se expanda la roya que en zonas con poco cereal. Las parcelas de seguimiento se cogen con características más o menos homogéneas y en zonas dispersas para controlar las zonas de cultivo de Castilla y León, pero puede que una parcela en concreto no tenga ese año cereal alrededor y eso condicione que no se llegue a infectar. Es como si estás midiendo la gripe, y uno de los individuos de control que usas para ver la evolución en una zona está todo el día en casa encerrado, no se va a contagiar.

7. Modelado y validación

En esta sección se desarrollan las etapas de modelado y evaluación de la Metodología Fundamental para la Ciencia de Datos representadas en la figura 7.1. Consiste en la obtención de un modelo según el enfoque descrito en el capítulo 3 y en la validación de este mismo modelo.

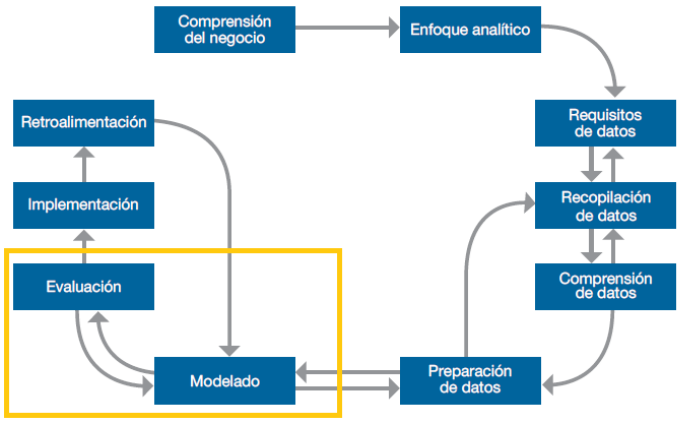


Figura 7.1: Metodología Fundamental para la Ciencia de Datos - Modelado y evaluación (extraída de [8])

7.1. Datos de entrenamiento validación y test

A partir del fichero de datos obtenido en la sección 6.3 se crea el conjunto de datos de entrenamiento para el ajuste de los modelos, el conjunto de datos de validación para la selección del modelo y el conjunto de datos de test para la evaluación del poder de generalización del modelo seleccionado. [38].

Se desea que los tres conjuntos datos contengan estaciones con distinto porcentaje de plaga detectada y que estas proporciones sean similares. Debido a ello, se obtiene el porcentaje de plaga total detectada para cada una de las estaciones, se discretiza este porcentaje en valores “baja”, “‘media” y “alta” y se seleccionan tres y siete estaciones de cada clase para el conjunto de test y validación respectivamente. El número de instancias y el porcentaje de datos del total que se disponen para cada conjunto se encuentra indicado en la tabla 7.1

	Número de instancias	**% total**
Entrenamiento	29.111	0,747
Validación	6.421	0,165
Test	3.464	0,088

Cuadro 7.1: Número y porcentaje de instancias conjuntos de datos

Como se pude ver en la Figura 7.2 las distribuciones de la variable respuesta con pervivencia es similar en los tres conjuntos de datos y existe un desequilibrio notable en esta variable, con predominio de estaciones cuyas parcelas no tienen plaga detectada o su porcentaje es bajo. Este desequilibrio se acentúa más si se toma como variable respuesta el porcentaje de plaga sin considerar la pervivencia de la misma, tal como se representa en la figura 7.3. Debido a ello y por el hecho de ser más razonable considerar la pervivencia en la plaga, se elige como variable respuesta el porcentaje de plaga con pervivencia. Si se encuentra plaga un día, es esperable que esta se encuentre en los próximos diez días aproximadamente.

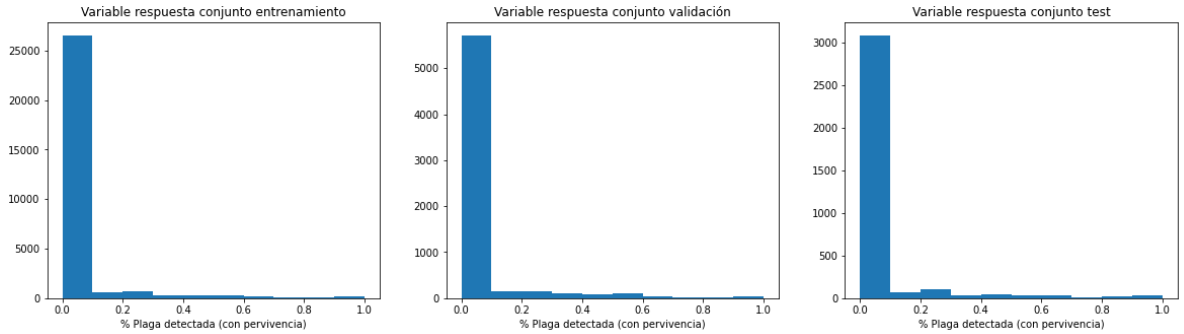


Figura 7.2: Distribución de la variable respuesta con pervivencia en el conjunto de datos de entrenamiento, validación y test datos de validación - Problema de regresión

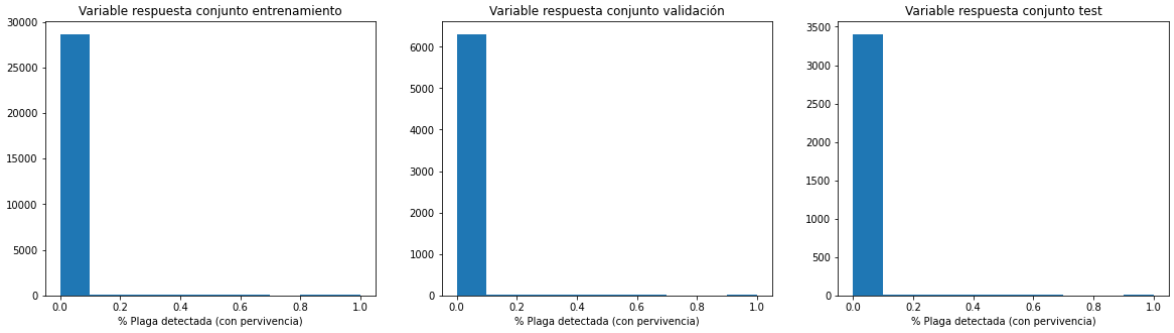


Figura 7.3: Distribución de la variable respuesta sin pervivencia en el conjunto de datos de entrenamiento, validación y test - Problema de regresión

El problema de predicción de plagas se puede tratar como un problema de clasificación en el que se predice la existencia de plaga o valor uno cuando existe un porcentaje de plaga detectada mayor que cierto valor y se predice un valor cero en caso contrario. En la Figura 7.4 se puede ver representado el desequilibrio de las clases en el problema de clasificación, con la variable respuesta detección de plaga considerando la pervivencia de la misma, si se escoge como variables respuesta el porcentaje de plaga sin

pervivencia este desequilibrio se acentúa más.

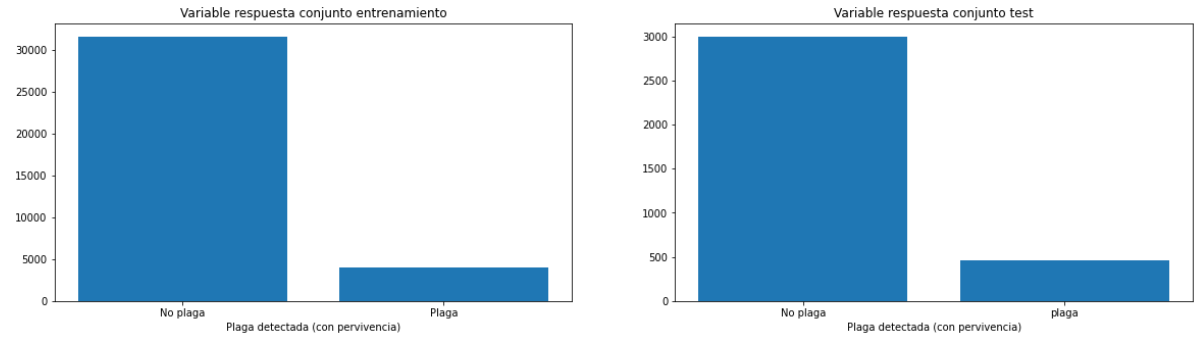


Figura 7.4: Distribución de la variable respuesta con pervivencia en el conjunto de datos de entrenamiento y en el conjunto de datos de test - Problema de clasificación

7.2. Modelo de regresión ExtraTreesRegessor

Se parte de un modelo base donde la para una estación y un año dado se predice la probabilidad de roya tomando como referencia la probabilidad de roya del año anterior. Se obtiene un MSE sobre el conjunto test de 0.0157 y un MAE de 0.0571. A continuación se trata de mejorar el ajuste del modelo. La librería TPOT de Python descrita en la sección 2.6 obtiene como modelos óptimos el modelo de regresión `ExtraTreesRegressor` y el modelo de clasificación `ExtraTreesClassifier`.

`ExtraTreesRegressor` es una clase de la librería `sklearn.ensembles` que implementa un ensemble basado en la combinación de las estimaciones de varios árboles de decisión aleatorios y de regresión con una elección aleatoria de las variables, al igual que los *Random Forests*. A diferencia de estos últimos, `ExtraTreesRegressor` elige de forma aleatoria el valor o umbral que se fija en los nodos para separar las instancias y con ello reduce algo más la varianza a cambio de aumentar el sesgo. Se calcula la media de las predicciones de los estimadores aleatorios con el objetivo de mejorar la robustez frente a un único estimador.

7.2.1. Selección de variables

El número total de variables del que se dispone finalmente es de 118 variables meteorológicas, dos variables que indican la distancia y altitud media de la estación a las parcelas, la variable estado fenológico y la variable tipo de clima codificada con `one_hot_encoding` para poder ser utilizada en los árboles aleatorios. Todas las variables iniciales se enumeran en la tabla 7.2.

Tanto la obtención de las variables meteorológicas como el ajuste de un modelo que las considere en su totalidad supone un coste en tiempo elevado y en ocasiones innecesario si el modelo predictivo puede obtener resultados similares con una pequeña parte de las mismas.

	Variables
100 % de variables	"mean_palt", "mean_dist", "fenol_state", "temp", "hum", "prec", "temp_min_15d_h", "temp_max_15d_h", "temp_mean_15d_h", "temp_std_15d_h", "temp_q20_15d_h", "temp_q40_15d_h", "temp_median_15d_h", ,temp_q60_15d_h", "temp_q80_15d_h" , "temp_sum_15d_h", "prec_min_15d_h","prec_max_15d_h", "prec_mean_15d_h", "prec_std_15d_h", "prec_q20_15d_h", "prec_q40_15d_h", "prec_median_15d_h", "prec_q60_15d_h", "prec_q80_15d_h", "prec_sum_15d_h", "hum_min_15d_h", "hum_max_15d_h", "hum_mean_15d_h", "hum_std_15d_h", "hum_q20_15d_h", "hum_q40_15d_h", "hum_median_15d_h", "hum_q60_15d_h", "hum_q80_15d_h", "hum_sum_15d_h", "temp_min_40d_h", "temp_max_40d_h", "temp_mean_40d_h", "temp_std_40d_h", "temp_q20_40d_h", "temp_q40_40d_h", "temp_median_40d_h", "temp_q60_40d_h", "temp_q80_40d_h","temp_sum_40d_h", "prec_min_40d_h", "prec_max_40d_h", "prec_mean_40d_h", "prec_std_40d_h", "prec_q20_40d_h", "prec_q40_40d_h", "prec_median_40d_h", "prec_q60_40d_h", "prec_q80_40d_h", "prec_sum_40d_h", "hum_min_40d_h","hum_max_40d_h", "hum_mean_40d_h", "hum_std_40d_h", "hum_q20_40d_h", "hum_q40_40d_h"," hum_median_40d_h", "hum_q60_40d_h", "hum_q80_40d_h", "hum_sum_40d_h", "temp_int_gh", "hum_max", "hum_mean", "hum_median", "hum_min", "hum_q10", "hum_q20", "hum_q30", "hum_q40", "hum_q60", "hum_q70", "hum_q80", "hum_q90", "hum_std", "hum_sum", "prec_max", "prec_mean", "prec_median", "prec_min", "prec_q10", "prec_q20", "prec_q30", "prec_q40", "prec_q60", "prec_q70", "prec_q80", "prec_q90", "prec_std", "prec_sum", "temp_max","temp_mean", "temp_median", "temp_min", "temp_q10", "temp_q20", "temp_q30", "temp_q40", "temp_q60", "temp_q70", "temp_q80", "temp_q90", "temp_std", "temp_sum", "temp_int_gd", "prec_10d", "prec_30d", "prec_60d", "prec_expected_30d", "prec_expected_60d", "prec_exp_perc_30d", "prec_exp_perc_60d", "prec_exp_sgm_30d", "prec_exp_sgm_60d", "roya_event_10d", "roya_event_skip_10d", "mes", "climType_ATLAN", "climType_ESTEF", "climType_MEDIT", "climType_OCEAN", "clim_Type_SUBTH"

Cuadro 7.2: Conjunto de variables iniciales

Para la selección de variables se utiliza la eliminación recursiva de variables basada en su importancia. La importancia de cada variable se obtiene con el propio ensemble de árboles de regresión aleatorios y se mide en función de la reducción en la métrica o error cuadrático medio, que produce dicha variable. De forma recursiva se entrena el árbol de regresión, se calcula la importancia de cada variable y se elimina o poda la variable menos relevante para la solución. Este procedimiento se repite hasta alcanzar el número de variables mínimo indicado. En la Figura 7.5 se representa la media de la disminución en impureza del árbol que produce cada variable. Algunas variables apenas disminuyen esta impureza y podrían ser eliminadas [39].

El método ideal para la selección de variables con árboles aleatorios en Python es `RFECV` de la librería `sklearn.feature_selection`. Este método implementa la eliminación recursiva con validación cruzada de variables y determina el número de variables y el conjunto de variables en concreto que, para todos los folds, obtiene un error cuadrático medio menor. Sin embargo, debido al gran número de variables disponibles y debido al gran desequilibrio presente en la variable respuesta, resulta computacionalmente costoso e inapropiado aplicar este procedimiento.

En la práctica se aplica la eliminación recursiva de variables mediante el método `RFE` sin validación cruzada de la librería mencionada anteriormente. Este método elimina las variables menos importantes del modelo hasta llegar a un número de variables indicado. Como inconveniente, no obtiene el número de variables óptimo. Para elegir este número óptimo se utiliza el método `RFE` fijando distinto porcentaje de variables, se ajusta el ensemble con estas variables y se obtienen las métricas sobre el conjunto de entrenamiento y el conjunto de validación.

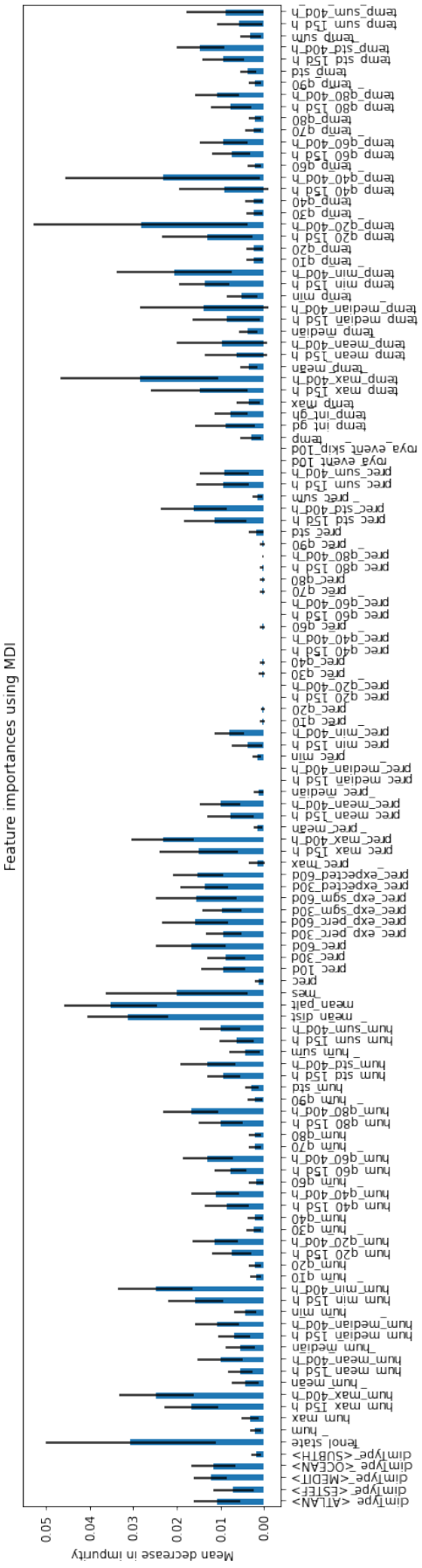


Figura 7.5: Importancia de las variables explicativas utilizando el criterio MDI con el modelo con el 100 % de las variables

En la tabla 7.3 se incluye el 50 % de las variables seleccionadas, en total 64 variables. Entre las variables que no se seleccionan se encuentran las variables meteorológicas "temp", "hum", "prec" y "hum", parece razonable descartarlas y mantener las variables que recogen las condiciones meteorológicas de días anteriores. Sin embargo, resulta extraño que no se hayan seleccionado las variables que recogen los eventos de Roya "roya_event_10d", "roya_event_skip_10d".

	Variables
50 % de variables	"mean_palt", "mean_dist", "fenol_state", "temp_min_15d_h","temp_max_15d_h", "temp_std_15d_h", "temp_q20_15d_h", "temp_q40_15d_h", "temp_median_15d_h", "temp_q60_15d_h", "temp_q80_15d_h", "prec_max_15d_h", "prec_mean_15d_h", "prec_std_15d_h", "prec_sum_15d_h", "hum_min_15d_h", "hum_max_15d_h", "hum_std_15d_h", "hum_q20_15d_h", "hum_q40_15d_h", "hum_median_15d_h", "hum_q60_15d_h", "hum_q80_15d_h", "temp_min_40d_h", "temp_max_40d_h", "temp_mean_40d_h", "temp_std_40d_h", "temp_q20_40d_h", "temp_q40_40d_h", "temp_median_40d_h", "temp_q60_40d_h", "temp_q80_40d_h", "temp_sum_40d_h", "prec_min_40d_h", "prec_max_40d_h", "prec_mean_40d_h", "prec_std_40d_h", "prec_sum_40d_h", "hum_min_40d_h", "hum_max_40d_h", "hum_mean_40d_h", "hum_std_40d_h", "hum_q20_40d_h", "hum_q40_40d_h", "hum_median_40d_h", "hum_q60_40d_h", "hum_q80_40d_h", "hum_sum_40d_h", "temp_int_gh", "hum_median", "temp_int_gd", "prec_10d", "prec_30d", "prec_60d", "prec_expected_30d", "prec_expected_60d", "prec_exp_perc_30d", "prec_exp_perc_60d", "prec_exp_sgm_30d", "prec_exp_sgm_60d", "mes", "climType_ATLAN", "climType_MEDIT", "climType_OCEAN"

Cuadro 7.3: Selección del 50 % de las variables (64 variables) mediante *Recursive feature elemination*

A continuación, se selecciona el 25 % de las variables totales. En la tabla 7.4 se recogen las 32 variables seleccionadas. Se han descartado las variables integral térmica grados día y grados hora, posiblemente las características meteorológicas son suficientes para la descripción y por ello no se seleccionan.

	Variables
25 % de variables	"mean_palt", "mean_dist", "fenol_state", "temp_min_15d_h", "temp_max_15d_h", "temp_q20_15d_h", "prec_max_15d_h", "prec_std_15d_h", "hum_min_15d_h", "hum_max_15d_h", "temp_min_40d_h", "temp_max_40d_h", "temp_std_40d_h", "temp_q20_40d_h", "temp_q40_40d_h", "temp_median_40d_h", "temp_q80_40d_h", "prec_max_40d_h", "prec_std_40d_h", "hum_min_40d_h", "hum_max_40d_h", "hum_std_40d_h", "hum_q40_40d_h", "hum_q60_40d_h", "hum_q80_40d_h", "prec_60d", "prec_expected_30d","prec_expected_60d", "prec_exp_perc_60d", "prec_exp_sgm_60d", "mes", "climType_OCEAN""

Cuadro 7.4: Selección del 25 % de las variables (32 variables) mediante *Recursive feature elemination*

Si se selecciona el 20 % y el 10 % de las variables, 25 y 13 variables en total, se descartan además de resúmenes de variables meteorológicas, las variables relacionadas con el clima, no quedando presente ningún clima. Se incluyen las variables seleccionadas en las tablas 7.5 y 7.6.

	Variables
20 % de variables	"mean_palt", "mean_dist", "fenol_state", "temp_min_15d_h", "temp_max_15d_h", "prec_max_15d_h", "hum_min_15d_h", "hum_max_15d_h", "temp_min_40d_h", "temp_max_40d_h", "temp_std_40d_h", "temp_q20_40d_h", "temp_q40_40d_h", "temp_median_40d_h", "prec_max_40d_h", "hum_min_40d_h", "hum_max_40d_h", "hum_std_40d_h", "hum_q60_40d_h", "hum_q80_40d_h", "prec_60d", "prec_expected_60d", "prec_exp_perc_60d", "prec_exp_sgm_60d", "mes"

Cuadro 7.5: Selección del 20 % de las variables (25 variables) mediante *Recursive feature elemination*

	Variables
10 % de variables	"mean_palt", "mean_dist", "fenol_state", "temp_min_15d_h", "temp_min_40d_h", "temp_max_40d_h", "temp_q20_40d_h", "prec_max_40d_h", "hum_min_40d_h", "hum_max_40d_h", "hum_q80_40d_h", "prec_60d", "prec_exp_sgm_60d"

Cuadro 7.6: Selección del 10 % de las variables (13 variables) mediante *Recursive feature elemination*

Tras la evaluación de las variables seleccionadas por la eliminación recursiva y con las métricas obtenidas sobre el conjunto de validación en la tabla 7.7, se decide seguir en la etapa de modelado con el 20 % de las variables, 25 variables en total. Entre las variables se encuentran de mayor a menor proporción, variables de temperatura, humedad y precipitación. Además de estas, las variables relacionadas con las parcelas seleccionadas son la altitud media y la distancia media de las parcelas que caracterizan a la estación meteorológica, el estado fenológico máximo de las parcelas y la variable mes que se había incluido como variable *proxy* de la evolución del cultivo.

	MSE / MEA entrenamiento	**MSE / MEA validación**
100 %	0.0 / 0.0	0.019 / 0.059
50 %	0.0 / 0.0	0.019 / 0.060
25 %	0.0 / 0.0	0.019 / 0.061
20 %	0.0 / 0.0	0.019 / 0.062
10 %	0.0 / 0.0	0.020 / 0.064

Cuadro 7.7: Métricas sobre conjunto de entrenamiento y validación tras la selección de variables

7.2.2. Poda del árbol

La obtención de un error cero en el conjunto de entrenamiento en la tabla 7.7 es un claro indicio de sobreajuste. En la sección 3.2.1 se describe el método *cost complexity pruning*, un método que permite reducir este sobreajuste mediante la incorporación de un valor α que penaliza el tamaño del árbol. Los árboles de clasificación `DecisionTreeClassifier` ofrecen el método `cost_complexity_pruning_path` para la obtención de la evolución de la impureza del árbol con distintos valores de α, sin embargo, el ensemble `ExtraTreesRegressor` no.

El método `ExtraTreesRegressor` de la librería `sklearn.ensemble` de Python permite ajustar el ensemble con un valor de penalización α indicado en el parámetro `ccp_alpha`. Es necesario manualmente comprobar con distintos valores de α el ajuste sobre el conjunto de datos de validación.

En el gráfico de la figura 7.6 se observa que con un valor α en torno al 0.01 es suficiente para hacer frente al sobreajuste presente. Se procede a ajustar el árbol con 25 variables, un α de 0.01 y se examinan las inferencias sobre el conjunto de validación.

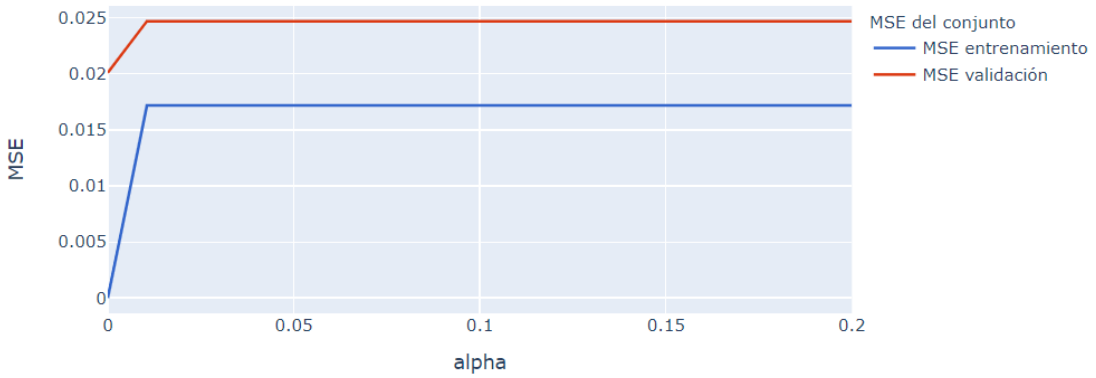


Figura 7.6: MSE frente al parámetro alpha de penalización

En la tabla 7.8 se observa que con el parámetro de penalización se incrementa ligeramente el MSE y el MEA sobre el conjunto de validación. En la figura 7.7 se muestra en cada fila una estación del conjunto de datos de validación y en cada columna el año de las inspecciones. Para cada gráfico se representa el porcentaje de plaga real en azul y el porcentaje de plaga predicha en rojo frente al día del año inspeccionada. Se puede ver que las inferencias con el modelo con parámetro de penalización son siempre nulas. Por ello se descarta seguir con un modelo con penalización.

	MSE / MEA entrenamiento	**MSE / MEA validación**
Con penalización	0.018 / 0.066	0.012 / 0.061
Sin penalización	0.0 / 0.0	0.009 / 0.053

Cuadro 7.8: Modelos `ExtraTreesRegressor` con 25 variables

Figura 7.7: Porcentaje de plaga real y predicho con 25 variables y parámetro α=0.01

7.2.3. Análisis del poder de generalización

Con el modelo `ExtraTreesRegressor` ajustado con el conjunto de entrenamiento y validación con 25 variables y sin parámetro de penalización se examinan las inferencias sobre el conjunto test, en la tabla 7.9 se muestran las métricas obtenidas.

	MSE / MEA entrenamiento + validación	**MSE / MEA test**
Sin penalización	0.0 / 0.0	0.009 / 0.053

Cuadro 7.9: Métricas `ExtraTreesRegressor` sobre conjunto entrenamiento + validación y test

En las figuras 7.8, 7.9, 7.10 y 7.11 se representan los valores reales y los predichos con el modelo de 25 variables y sin parámetro de penalización para cada una de las estaciones meteorológicas del conjunto test. Se representan los años comprendidos ente 2015 y 2021, a excepción del año 2020, donde no se inspeccionaron las parcelas por motivos de la pandemia COVID-19. Los huecos en blanco se debe a que la estación de la fila no contiene inspecciones para el año de la columna.

Los resultados obtenidos son bastante aceptables pues se consigue un modelo con un MSE de 0.009, un modelo adecuado al objetivo en el que prima la exhaustividad frente a la precisión. Se aceptan las inferencias positivas con detección de plaga, aunque esta no esté presente, con el fin de detectar todas las plagas existentes. Las inferencias con detección de plaga errónea en el año 2018 para las tres primeras filas de la primera columna de la figura 7.9, son más aceptadas por el ITACYL (como se indica en el capítulo 1) que las del gráfico de la esquina inferior derecha de esta misma figura, donde no se detecta la plaga.

En las inferencias positivas de la plaga suele haber una tendencia bastante razonable en la que las detecciones de plaga se producen los días finales de la campaña de inspección, en mayo y junio, precisamente cuando la cantidad de biomasa está en su máximo desarrollo. No importa que la cantidad detectada sea menor o mayor, lo importante es la tendencia mencionada de la plaga detectada, como se puede ver en la fila dos y columna uno de la figura 7.8.

El modelo `ExtraTreesRegressor` en ocasiones obtiene resultados más razonables que los datos reales cuya calidad está condicionada por los determinantes expuestos en la sección 6.4. En la figura 7.9, en el gráfico de la segunda columna y tercera fila en azul se representa la detección de plaga en un momento concreto, las inferencias son más acertadas porque es de esperar que la plaga perviva en los siguientes días.

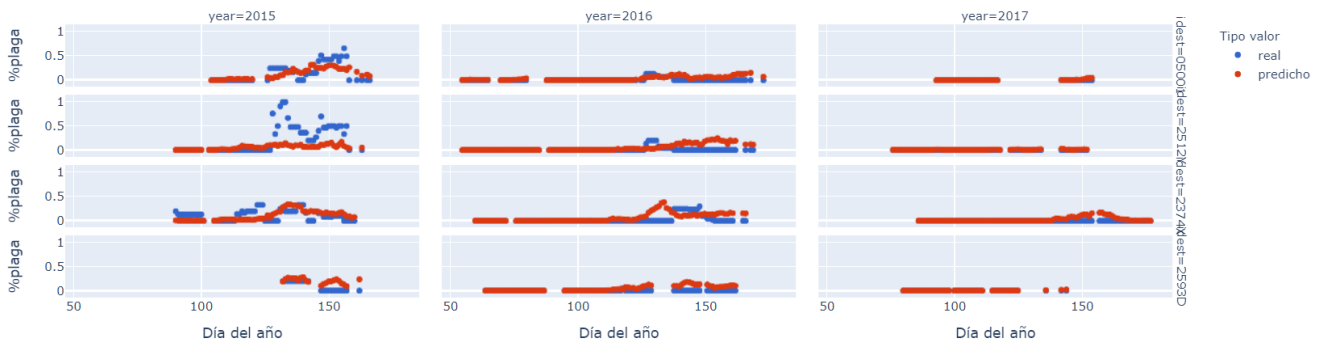


Figura 7.8: Porcentaje de plaga real y predicho con 25 variables y sin parámetro de penalización

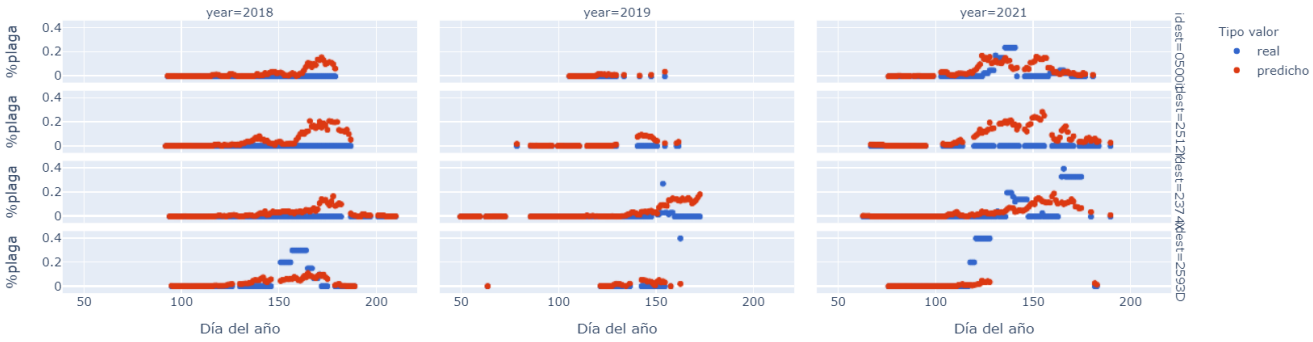


Figura 7.9: Porcentaje de plaga real y predicho con 24 variables y sin parámetro de penalización

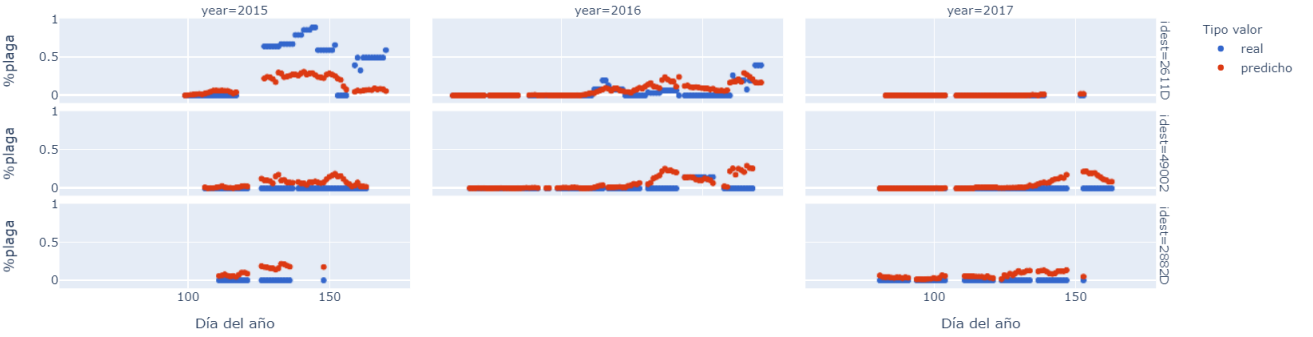


Figura 7.10: Porcentaje de plaga real y predicho con 25 variables y sin parámetro de penalización

Figura 7.11: Porcentaje de plaga real y predicho con 25 variables y sin parámetro de penalización

7.2.4. Representación de los resultados

Para las estaciones del conjunto test se muestra para cada año, el porcentaje de plaga real y el porcentaje de plaga predicho en los mapas mediante la librería `GeoPandas`. Para ello se utilizan las coordenadas x e y de las estaciones y un archivo `.shp` que contiene el mapa de Castilla y León. En las figuras 7.12, 7.13, 7.14, 7.15, 7.16 y 7.17 se observa que las inferencias por lo general predicen plagas de más.

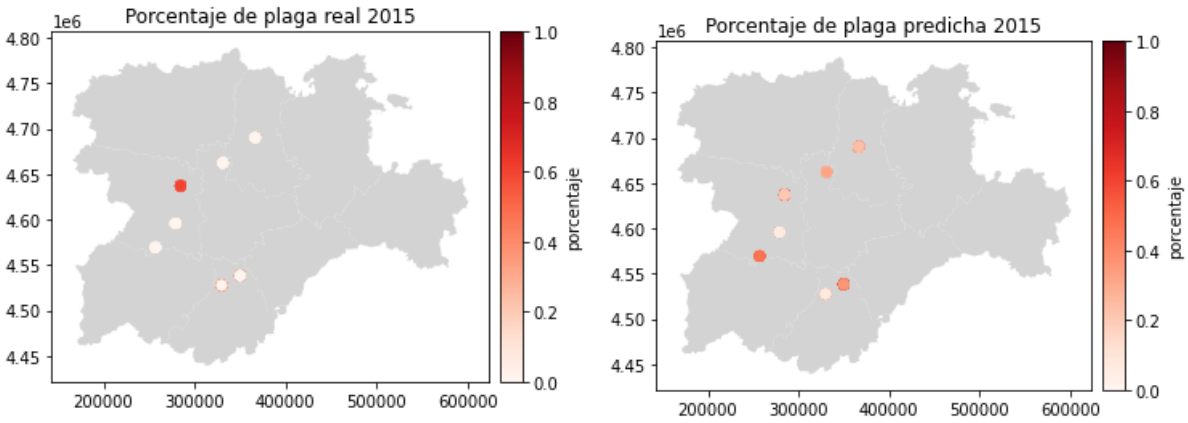


Figura 7.12: Porcentaje de plaga real y predicha conjunto test 2015

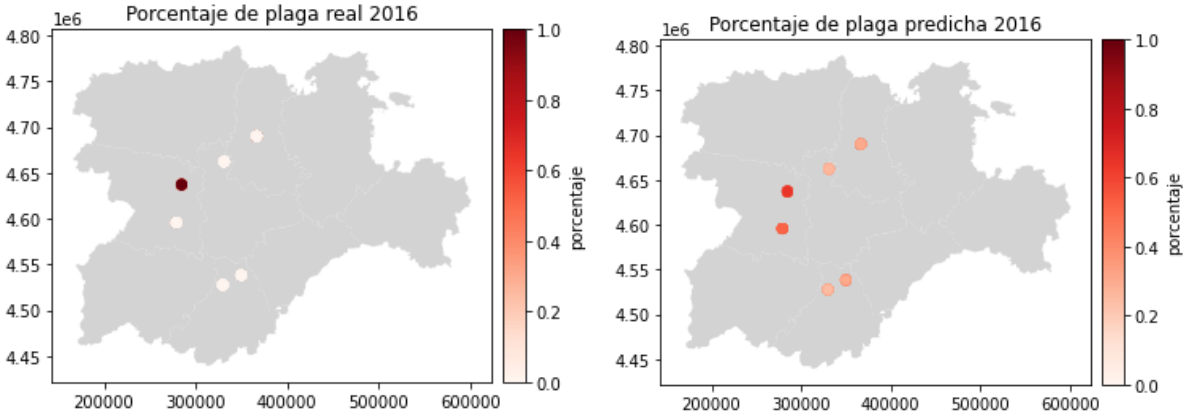


Figura 7.13: Porcentaje de plaga real y predicha conjunto test 2016

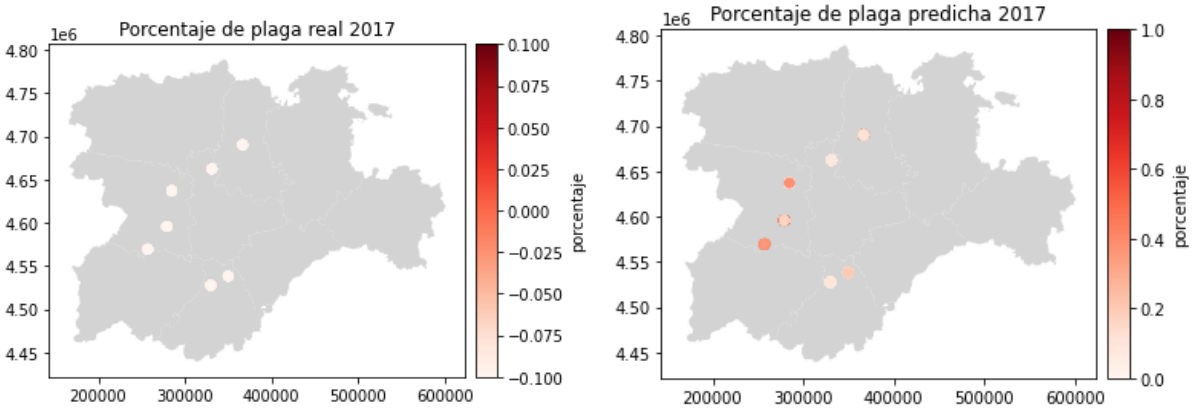


Figura 7.14: Porcentaje de plaga real y predicha conjunto test 2017

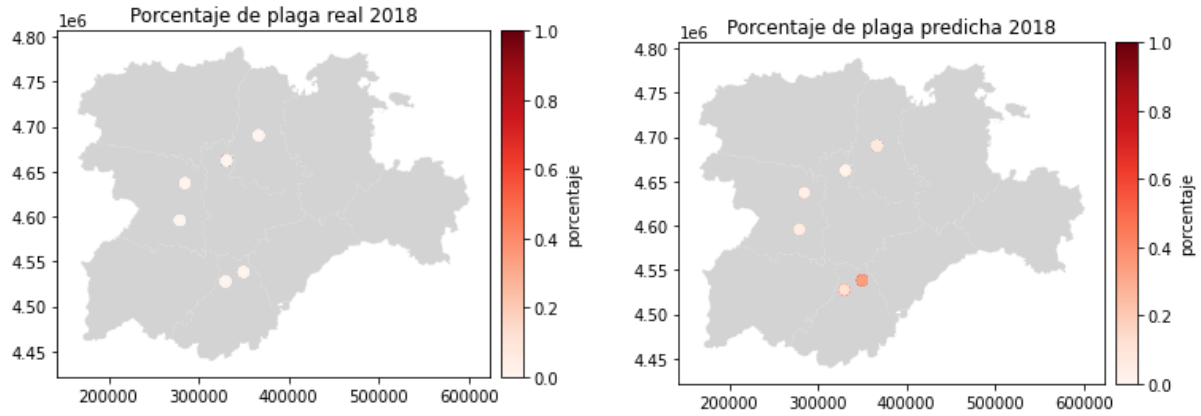


Figura 7.15: Porcentaje de plaga real y predicha conjunto tes 2018

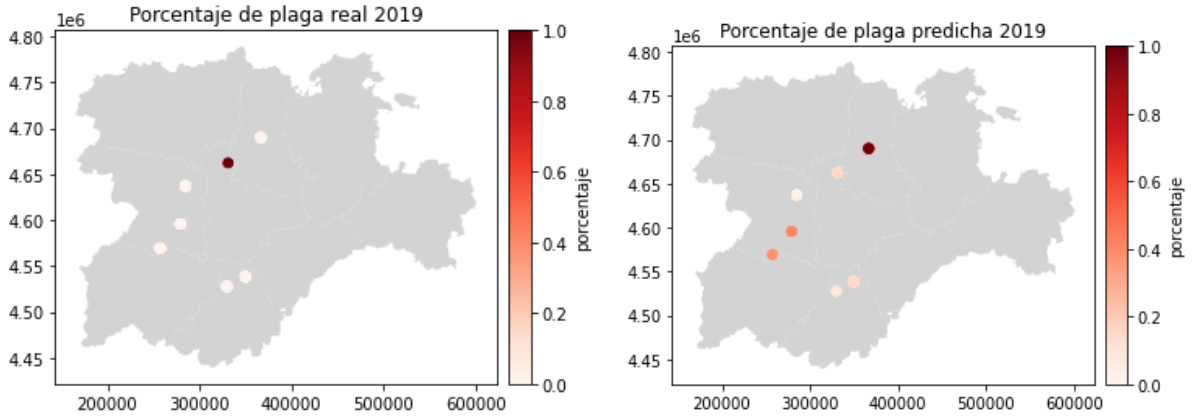


Figura 7.16: Porcentaje de plaga real y predicha conjunto test 2019

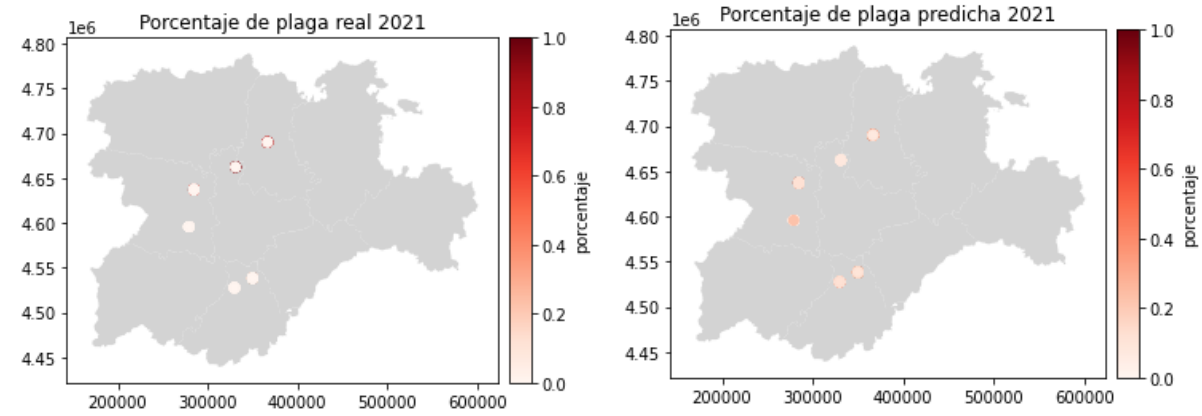


Figura 7.17: Porcentaje de plaga real y predicha conjunto test 2021

7.2.5. Reformulación problema de regresión en problema de clasificación

Los patrones de diseño en el desarrollo de software son técnicas que resuelven problemas comunes. En *Machine Learning* o aprendizaje automático existen también una serie de patrones que permiten resolver problemas concretos y se encuentran explicados en el libro *Machine Learning Design Patterns: Solutions to Common Challenges in*

Data Preparation, Model Building, and MLOps [25].

El patrón *reframing* o reformulación surge en problemas en los que un mismo conjunto de variables explicativas, obtienen una variable respuesta diferente. Relacionado con la predicción de la plaga Roya un ejemplo sería, ante unas mismas condiciones meteorológicas obtener distinto porcentaje de plaga predicha. El patrón de diseño propone como solución la reformulación de la salida del problema de regresión en salida propia de un problema de clasificación. Esta última aproximación al problema permite capturar probabilidades de distribución de la variable respuesta en lugar de calcular la media de la distribución y permite ponderar las clases imbalanceadas.

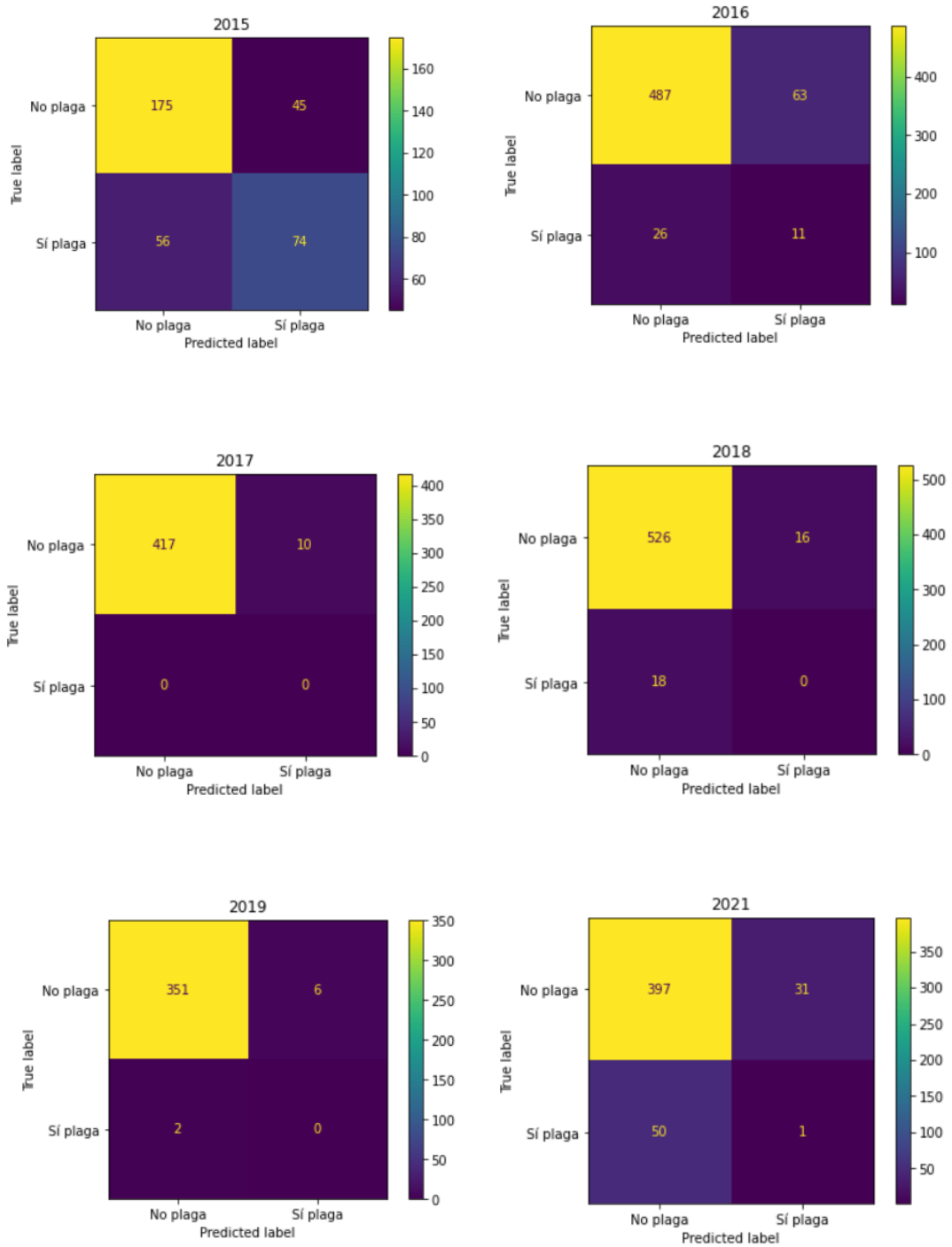


Figura 7.18: Matrices de confusión tras reformular las salidas de regresión a salidas de clasificación

Para obtener las matrices de confusión de la Figura 7.18 se han reformulado las salidas del ensemble de regresión a salidas binarias. Para aquellas instancias que superen el 0.15 % de plaga detectada se considera que hay plaga con el valor 1, para aquellas que no lo alcancen, cero. Con la salida se puede ver un alto número de falsos negativos, es decir, instancias con plaga que no han sido detectadas. Existen varias soluciones al problema:

- **Tratar los datos.** Se puede utilizar la técnica de *boosting* para la generación de instancias con la clase minoritaria.
- **Modificar la métrica.** En los árboles de clasificación se pueden utilizar métricas como la *F-score* descrita en la sección 3.4.2 que guardan un compromiso entre la sensibilidad del modelo y la precisión.
- **Modificación de los pesos.** Durante el ajuste del modelo se puede dar mayor peso o importancia a la clase minoritaria y que esta influya más en la clasificación.

7.3. ExtraTreesClassifier

`ExtraTreesClassifier` es una clase de la librería `sklearn.ensembles` que al igual que la clase `ExtraTreesRegressor` descrita en la sección 7.2, implementa un ensemble que realiza una elección aleatoria de las variables y del umbral que se fija en los nodos para separar las instancias. A diferencia de `ExtraTreesRegressor`, `ExtraTreesClassifier` realiza una clasificación.

7.3.1. Generación de instancias

Se generan instancias para el conjunto de datos de entrenamiento con el fin de solventar el problema de desequilibrio de la variable respuesta. Se utiliza la librería `imblearn` de Python, en concreto, el método `over_sampling`. Este método genera, mediante la elección aleatoria de las muestras ya existentes, nuevas muestras de la clase cuya proporción es inferior. En la figura 7.19 se representa la proporción de instancias con plaga e instancias sin plaga, ambas equilibradas, tras aplicar el método `over_sampling` con los datos obtenidos en la sección 7.1.

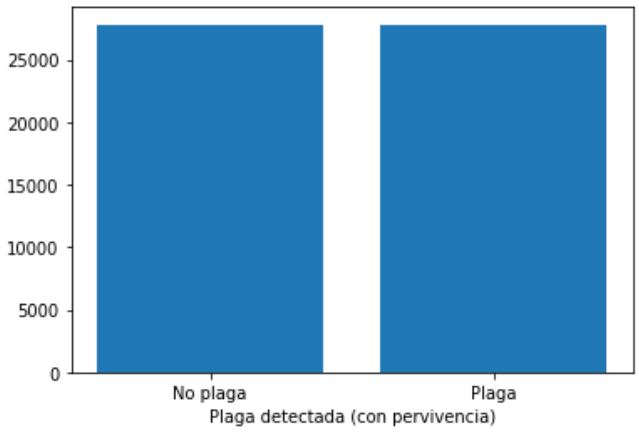


Figura 7.19: Proporción de clases después de la generación de instancias

7.3.2. Selección de variables

Con el conjunto de datos de entrenamiento equilibrado se utiliza el método `RFECV` de la librería `sklearn` de Python que implementa la eliminación recursiva de variables mediante validación cruzada para la selección de variables. La métrica que se trata de maximizar es la F_β descrita en la sección 3.4.2 con un valor β igual a dos que

da mayor importancia a la sensibilidad del modelo. La sensibilidad es considerada dos veces más importante que la precisión.

$$F_2 = (1+\beta^2)\frac{precision \cdot recall}{(\beta \cdot precision) + recall} = \frac{5 \cdot precision \cdot recall}{(2 \cdot precision) + recall} \tag{7.1}$$

En la figura 7.20 se muestra la evolución de la métrica F_2 de cinco folds en la validación cruzada de la selección de variables mediante eliminación recursiva. El método obtiene un valor óptimo con 123 variables, sin embargo, en la misma figura se observa que con 20 variables se obtiene una métrica bastante elevada.

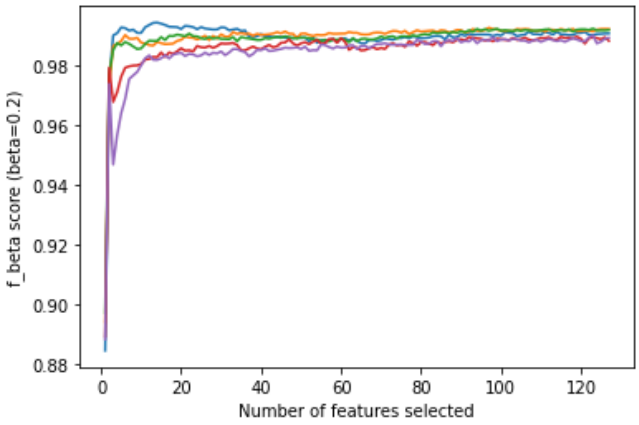


Figura 7.20: Selección de variables mediante *ExtraTreesClassifier* y eliminación recursiva de variables

En la tabla 7.10 se enumeran las veinte variables seleccionadas por el método *Recursive feature elemination*. Se observa que las variables *"mean_palt"*, *"mean_dist"*, *"fenol_state"* y *"mes"* al igual que en la selección de variables en el modelo de regresión de la sección 7.2.1, son seleccionadas. Entre el resto de variables seleccionadas se encuentra en mayor proporción las variables que describen la temperatura y en igual proporción variables de la humedad y de la precipitación.

	Variables
20 variables	"mean_palt", "mean_dist", "fenol_state", "temp_max_15d_h", "temp_q20_15d_h", "temp_q40_15d_h", "temp_median_15d_h", "temp_min_40d_h", "temp_max_40d_h", "temp_std_40d_h", "temp_q20_40d_h", "temp_q40_40d_h", "temp_median_40d_h", "prec_max_40d_h", "hum_q40_40d_h", "hum_median_40d_h", "hum_q80_40d_h", "prec_expected_30d", "prec_expected_60d", "mes"

Cuadro 7.10: Selección de 20 variables mediante *Recursive feature elemination* en problema de clasificación

8. Conclusiones y líneas futuras

En este capítulo se exponen las conclusiones obtenidas tras la realización de este Trabajo Fin de Grado, así como las etapas pendientes por completar para cumplir con la metodología de ciencias de datos propuesta por el IBM, posibles mejoras y lineas de trabajo futuras.

8.1. Conclusiones

Este Trabajo Fin de Grado se ha realizado siguiendo la metodología fundamental para la ciencia de datos propuesta por *International Business Machines Corporatio*n o IBM. Esta metodología permite de manera ordenada obtener respuestas o resultados, es aplicable a cualquier problema en el que se disponga de datos y no depende de tecnologías ni herramientas específicas.

La realización de este trabajo ha supuesto el afianzamiento del conocimiento sobre árboles de decisión adquirido a lo largo de la carrera así como el aprendizaje de dos programas imprescindibles para la ejecución de un proyecto en colaboración, Git y Enterprise Architect. Además de estos programas, se ha aprendido sobre el uso del sistema de información geográfica QGIS, el uso del entorno de desarrollo PyCharm, el uso de librerías fundamentales para la ciencia de datos en Python y el uso del sistema gestor de bases de datos SQLite.

El objetivo del trabajo: "desarrollo de un modelo *machine learning* que permita predecir de forma temprana la aparición de roya amarilla" ha sido alcanzado con éxito. Se ha obtenido un modelo exhaustivo con un MSE de 0.009 y un un MEA de 0.053 sobre el conjunto de datos test. Este modelo captura la tendencia de desarrollo de la plaga y en algunas ocasiones ofrece resultados más acertados que los datos disponibles teniendo en consideración la calidad de los mismos que se encuentra condicionada por variables que no se encuentran registradas como la variedad del cultivo, la fecha de siembra del cultivo, las labores agrícolas o la cantidad de cereal en la zona, entre otra.

Fruto de este trabajo se obtienen una serie de *scripts* de gran utilidad para el Instituto Tecnológico Agrario. Estos *scripts* han permitido automatizar el proceso de integración de la información meteorológica procedente de la Agencia Estatal de Meteorología de España y de la red de estaciones meteorológicas Inforiego del Instituto Tecnológico Agrario. Los *scripts* incluyen los procedimientos de limpieza e imputación de los datos y son utilizable por el Observatorio de Vigilancia y Control de Plagas de Castilla y León completo pues trata datos meteorológicos influyentes en en el desarrollo de cualquier tipo de plaga.

Los *scripts* más específicos para la obtención de las características de interés para las plaga roya también se encuentran listos para ser automatizados mediante Apache Airflow, una plataforma de gestión de flujo de trabajo de código abierto escrita en Python.

8.2. Líneas futuras

Queda pendiente por finalizar las dos últimas etapas de la metodología fundamental para la ciencia de datos propuesta por IBM, concretamente la implementación de los modelos y la retroalimentación representadas en la figura 8.1. Para la ejecución automática de los modelos es necesaria la creación de *scripts*, pues en este trabajo estos *scripts* han sido realizados en Notebooks, entornos similares a RMarkdown.

Los *scripts* para la ejecución de los modelo junto con los *scripts* implementados para la limpieza de los datos y la creación de característica permitirán automatizar el proceso de predicción de plaga. Una vez obtenidas las inferencias sobre el año actual 2022, sería posible realizar una retroalimentación y mejorar el modelo.

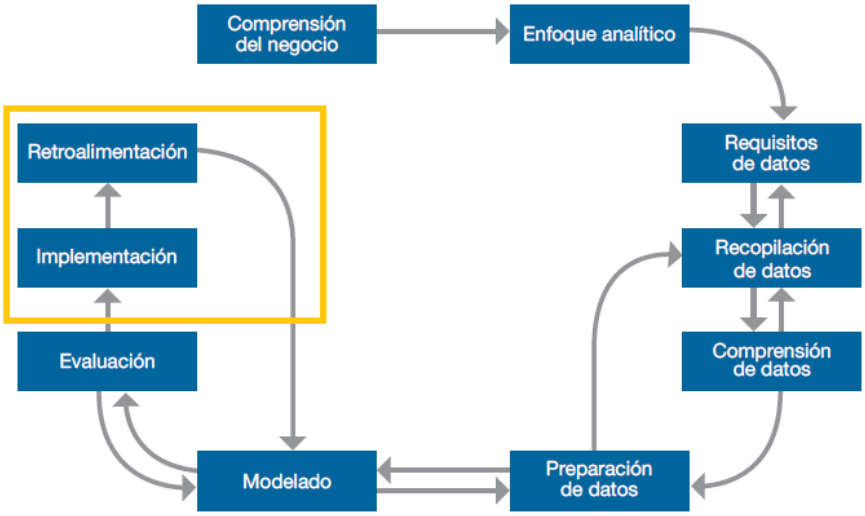


Figura 8.1: Metodología Fundamental para la ciencia de datos - Implementación y retroalimentación (extraída de [8])

Como futuro trabajo, se propone la implementación del modelo de clasificación presentado en la sección 7.3 mediante una selección de los hiperparámetros óptima. Además se propone la implementación de modelos con capacidad de modelar secuencias temporales como pueden ser las redes convolucionales en una dimensión o las redes LSTM.

A. Datos

A.1. Estados fenológicos

1	SEMILLA SECA	0
2	COMIENZA LA IMBIBICION DE LA SEMILLA	0
3	IMBIBICION COMPLETA DE LA SEMILLA	0
4	RADICULA (RAIZ EMBRIONAL) EMERGIDA DE LA SEMILLA	0
5	RADICULA ALARGADA, FORMANDO PELOS RADICULARES Y RA	0
6	COLEOPTILO, EMERGIDO DE LA SEMILLA	0
7	EMERGENCIA: EL COLEOPTILO TRASPASA LA SUPERFICIE D	0
8	1ª HOJA, ATRAVIESA EL COLEOPTILO	1
9	1ª HOJA, DESPLEGADA	1
10	2 HOJAS, DESPLEGADAS	1
11	3 HOJAS, DESPLEGADAS	1
12	4 O MAS HOJAS, DESPLEGADAS	1
13	5 O MAS HOJAS, DESPLEGADAS	1
14	6 O MAS HOJAS, DESPLEGADAS	1
15	7 O MAS HOJAS, DESPLEGADAS	1
16	8 O MAS HOJAS, DESPLEGADAS	1
17	9 O MAS HOJAS, DESPLEGADAS	1
18	NO HIJUELO VISIBLE	2
19	COMIENZO DEL MACOLLAMIENTO; 1 HIJUELO VISIBLE	2
20	2 HIJUELOS O MACOLLAS VISIBLES	2
21	3 HIJUELOS O MACOLLAS VISIBLES	2
22	4 HIJUELOS O MACOLLAS VISIBLES	2
23	5 HIJUELOS O MACOLLAS VISIBLES	2
24	6 HIJUELOS O MACOLLAS VISIBLES	2
25	7 HIJUELOS O MACOLLAS VISIBLES	2
26	8 HIJUELOS O MACOLLAS VISIBLES	2
27	FIN DEL MACOLLAMIENTO; EL MAXIMO DE HIJUELOS O MAC	2
28	COMIENZO DEL ENCAÑADO: PSEUDOTALLO E HIJUELOS, ERE	3
29	1ER NUDO, POR LO MENOS A 1 CM POR ENCIMA DEL NUDO	3
30	2º NUDO: PERCEPTIBLE, A 2 CM DEL 1ER NUDO	3
31	3º NUDO: PERCEPTIBLE, A 2 CM DEL 2º NUDO	3
32	4º NUDO: PERCEPTIBLE, A 2 CM DEL 3º NUDO	3
33	5º NUDO: PERCEPTIBLE, A 2 CM DEL 4º NUDO	3
34	6º NUDO: PERCEPTIBLE, A 2 CM DEL 5º NUDO	3

35	APARECE LA ULTIMA HOJA (HOJA BANDERA), AUN ENROLLA	3
36	ESTADIO HOJA BANDERA: HOJA BANDERA COMPLETAMENTE D	3
37	ESTADIO HINCHADO TEMPRANO: SE ALARGA LA VAINA DE L	4
38	ESTADIO HINCHADO MEDIO: SE EMPIEZA A VER LA VAINA	4
39	ESTADIO HINCHADO TARDIO: LA VAINA DE LA HOJA BANDE	4
40	SE EMPIEZA A ABRIR LA VAINA DE LA HOJA BANDERA	4
41	PRIMERAS ARISTAS (BARBAS), VISIBLES (SOLO EN VARIE	4
42	COMIENZO DEL ESPIGADO: LA PUNTA DE LA ESPIGA O DE	5
43	20 % DE LA ESPIGA EMERGIDA	5
44	30 % DE LA ESPIGA EMERGIDA	5
45	40 % DE LA ESPIGA EMERGIDA	5
46	MITAD DEL ESPIGADO: EMERGIDA LA MITAD DE LA ESPIGA	5
47	60 % DE LA ESPIGA EMERGIDA	5
48	70 % DE LA ESPIGA O PANÍCULA EMERGIDA	5
49	80 % DE LA ESPIGA EMERGIDA	5
50	FIN DEL ESPIGADO: LA ESPIGA O PANICULA COMPLETAMEN	5
51	COMIENZO DE LA FLORACION: PRIMERAS ANTERAS VISIBLE	6
52	PLENA FLORACION: 50 % DE LAS ANTERAS MADURAS	6
53	FIN DE LA FLORACION: TODAS LAS ESPIGUILLAS HAN TER	6
54	ESTADIO DE MADUREZ ACUOSA: PRIMEROS GRANOS HAN ALC	7
55	GRANO LECHOSO TEMPRANO	7
56	GRANO LECHOSO MEDIO: CONTENIDO DEL GRANO LECHOSO,	7
57	GRANO LECHOSO TARDIO	7
58	PASTOSO TEMPRANO	8
59	PASTOSO BLANDO: CONTENIDO DEL GRANO, BLANDO, PERO	8
60	PASTOSO DURO: CONTENIDO DEL GRANO, SOLIDO; SE MANT	8
61	MADUREZ COMPLETA: GRANO DURO, DIFICIL DE DIVIDIR C	8
62	SOBRE-MADUREZ: GRANOS, MUY DUROS, NO PUEDEN SER ME	9
63	GRANOS, DESPRENDIENDOSE DURANTE EL DÍA	9
64	PLANTA MUERTA, TALLOS SE QUIEBRAN	9
65	PRODUCTO COSECHADO	9

Cuadro A.1: Estados fenológicos

Bibliografía

[1] M. M. Roa, • *Gráfico: El Big Bang del Big Data — Statista*. dirección: https://es.statista.com/grafico/26031/volumen-estimado-de-datos-digitales-creados-o-replicados-en-todo-el-mundo/.

[2] *¿Qué es el big data? — Oracle España*. dirección: https://www.oracle.com/es/big-data/what-is-big-data/.

[3] *Agricultura 4.0. Nuevas tecnologías en la agricultura — ATRIA Innovation*. dirección: https://www.atriainnovation.com/agricultura-4-0-nuevas-tecnologias-en-la-agricultura/.

[4] *Solicitud Única de Ayudas PAC (2022) — Sede Electrónica — Junta de Castilla y León*. dirección: https://www.tramitacastillayleon.jcyl.es/web/jcyl/AdministracionElectronica/es/Plantilla100Detalle/1251181050732/Ayuda012/1285132235798/Propuesta.

[5] D. E. T. Beest, N. D. Paveley, M. W. Shaw y F. V. D. Bosch, "Disease-weather relationships for powdery mildew and yellow rust on winter wheat," *Phytopathology*, vol. 98, págs. 609-617, 5 mayo de 2008, ISSN: 0031949X. DOI: 10.1094/PHYTO-98-5-0609.

[6] M. E. Jarroudi y C. H. Bock, "A Threshold-Based Weather Model for Predicting Stripe Rust Infection in Winter Wheat," 2017. DOI: 10.1094/PDIS-12-16-1766-RE. dirección: http://dx.doi.org/10.1094/PDIS-12-16-1766-RE.

[7] *Home - Sativum*. dirección: https://www.sativum.es/.

[8] J. B. Rollins, "Metodología Fundamental para la Ciencia de Datos," 2015. dirección: https://www.ibm.com/downloads/cas/6RZMKDN8.

[9] *Python or R for Data Analysis: Which Should I Learn?* Dirección: https://www.coursera.org/articles/python-or-r-for-data-analysis.

[10] *NumPy*. dirección: https://numpy.org/.

[11] *pandas - Python Data Analysis Library*. dirección: https://pandas.pydata.org/.

[12] *GeoPandas 0.10.2+0.g04d377f.dirty — GeoPandas 0.10.2+0.g04d377f.dirty documentation*. dirección: https://geopandas.org/en/stable/.

[13] *Matplotlib — Visualization with Python*. dirección: https://matplotlib.org/.

[14] *SciPy*. dirección: https://scipy.org/.

[15] *scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation*. dirección: https://scikit-learn.org/stable/.

[16] *PyCharm: el IDE de Python para desarrolladores profesionales, por JetBrains*. dirección: https://www.jetbrains.com/es-es/pycharm/.

[17] *Sistema de información geográfica - Wikipedia, la enciclopedia libre*. dirección: https://es.wikipedia.org/wiki/Sistema_de_informaci%C3%B3n_geogr%C3%A1fica.

[18] *Datos Vectoriales*. dirección: https://docs.qgis.org/2.14/es/docs/gentle_gis_introduction/vector_data.html.

[19] *Datos Raster*. dirección: https://docs.qgis.org/2.14/es/docs/gentle_gis_introduction/raster_data.html.

[20] *SQLite Home Page*. dirección: https://www.sqlite.org/index.html.

[21] *AutoML - AutoML*. dirección: http://automl.info/automl/.

[22] *UML modeling tools for Business, Software, Systems and Architecture*. dirección: https://sparxsystems.com/.

[23] S. J. (J. Russell, P. Norvig, J. M. C. Rodríguez y L. J. Aguilar, "Inteligencia artificial : un enfoque moderno," pág. 1212, 2004.

[24] A. Munoz, "Machine Learning and Optimization,"

[25] V. Lakshmanan, S. Robinson y M. Munn, "Machine Learning Design Patterns Solutions to Common Challenges in Data Preparation, Model Building, and MLOps."

[26] M. E. Jarroudi, R. Lahlali, H. E. Jarroudi, B. Tychon, A. Belleflamme, J. Junk, A. Denis, M. E. Jarroudi y L. Kouadio, "Employing weather-based disease and machine learning techniques for optimal control of septoria leaf blotch and stripe rust in wheat," *Advances in Intelligent Systems and Computing*, vol. 1103 AISC, págs. 157-165, 2020, ISSN: 21945365. DOI: 10.1007/978-3-030-36664-3_18.

[27] M. Alejandro y L. Ánge, "Arboles de clasificación y bosques aleatorios (Random Forests) Análisis Multivariante," Universidad de Valladolid, 2020.

[28] G. James, D. Witten, T. Hastie y R. Tibshirani, *An Introduction to Statistical Learning with Application in R*, Second Edition. dirección: http://www.springer.com/series/417.

[29] E. del Barrio Tellado, "Métodos de regularización en regresión," Universidad de Valladolid.

[30] R. L. Lawrence y A. Wrlght, "Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis,"

[31] *1.11. Ensemble methods — scikit-learn 1.1.1 documentation*. dirección: https://scikit-learn.org/stable/modules/ensemble.html#forest.

[32] *Inicio - Atlas agroclimático - ITACyL Portal Web*. dirección: http://atlas.itacyl.es/.

[33] *Datos Raster*. dirección: https://docs.qgis.org/2.14/es/docs/gentle_gis_introduction/raster_data.html.

[34] *Determinación de zonas agroclimáticas para la producción de mango (Mangifera indica L. "Manila") en Veracruz, México*. dirección: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-46112007000200003.

[35] D. Nualart, "Skorokhod topology," *Encyclopaedia of Mathematics*, págs. 18-19, 2011. dirección: https://encyclopediaofmath.org/index.php?title=Linear_interpolation.

[36] *Inverse Distance Weighting • SOGA • Department of Earth Sciences*. dirección: https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/geostatistics/Inverse-Distance-Weighting/index.html.

[37] *Integral térmica. ¿Qué tiene que ver con la agricultura? - Agromática*. dirección: https://www.agromatica.es/integral-termica/.

[38] T. Hastie, R. Tibshirani y J. Friedman, "Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction."

[39] $sklearn.feature_selection.RFE|scikit-learn1{,}1{,}1documentation$. dirección: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html#sklearn.feature_selection.RFE.score.