



DOBLE GRADO EN MATEMÁTICAS Y
ESTADÍSTICA

— TRABAJO FIN DE ESTUDIOS —

*Modelos de predicción de
incendios forestales en
Andalucía*

Juan Baeza Ruiz-Henestrosa

Sevilla, Mayo de 2024

Índice general

Prólogo	V
Resumen	VII
Abstract	VIII
Índice de Figuras	XI
Índice de Tablas	XIII
0.1. Introducción	1
0.2. Objetivos	1
0.3. Hipótesis	1
0.4. Revisión bibliográfica	1
1. Preliminares	3
1.1. Datos georreferenciados	3
1.1.1. Datos Vectoriales	3
1.1.1.1. Simple features	3
1.1.2. Datos Ráster	4
1.1.3. Sistemas de Referencia de Coordenadas	4
1.1.3.1. Sistemas de Coordenadas Geográficas	5
1.1.3.2. Sistemas de Coordenadas Proyectadas	5
1.2. Análisis exploratorio de datos	6
1.2.1. Depuración de los datos	6
1.2.2. Análisis de componentes principales	7
1.3. Modelos	7
1.3.1. Regresión logística (con penalización)	7
1.3.2. Support Vector Machine	9
1.3.2.1. SVM lineal	9
1.3.2.2. SVM no lineal	10
1.3.3. Decision Trees	10
1.3.4. Random Forest	12

1.3.5. K-Nearest Neighbors	13
1.4. Validación del ajuste	13
1.5. Evaluación de los modelos	13
1.5.1. Clasificación binaria	13
1.6. Herramientas	14
2. Construcción del conjunto de datos	17
2.1. Determinación del marco del estudio	18
2.1.1. Incendios forestales	18
2.1.2. Variables predictoras	19
2.2. Fuentes de datos	20
2.3. Procesamiento de los datos	22
2.3.1. Generación de una muestra balanceada de casos positivos y negativos.	22
2.3.2. Asignación de las variables descriptivas a cada observación	24
2.3.3. Depuración de la muestra	25
3. Cuerpo	29
3.1. Análisis exploratorio de datos	29
3.1.1. Distribución de la variable objetivo	30
3.1.2. Análisis univariantes variables numéricas	31
3.1.3. Análisis multivariantes de las variables numéricas	36
3.1.4. Análisis de las variables categóricas	38
3.2. Modelización	39
3.2.1. Regresión logística con penalización	40
3.2.2. Regresión logística con penalización + PCA	41
3.2.3. Árboles de decisión	41
3.2.4. Bosques aleatorios	42
3.2.5. KNN	43
3.2.6. SVM lineal	43
3.2.7. SVM radial	43
3.3. Comparación	44
4. Aplicación de los modelos	49
4.1. Visión general del desempeño de los modelos	49
4.2. Caso estudio	51

5. Conclusión	55
A. Apéndice: Gráficos espaciales EDA	57
A.1. Variables meteorológicas	57
A.2. Variables demográficas	61
A.3. Variable de vegetación	62
A.4. Variables topográficas	62
A.5. Variables antropológicas	64
B. Apéndice: Salidas de los modelos	69
B.1. Regresión logística con penalización	69
Bibliografía	71

Prólogo

Escrito colocado al comienzo de una obra en el que se hacen comentarios sobre la obra o su autor, o se introduce en su lectura; a menudo está realizado por una persona distinta del autor.

También se podrían incluir aquí los agradecimientos.

Resumen

Resumen...

Abstract

Abstract...

Índice de figuras

1. Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches, Omid Ghorbanzadeh, Khalil Valizadeh Kamran, Thomas Blaschke, Jagannath Aryal, Amin Naboureh, Jamshid Einali and Jinhua Bian.	2
1.1. Cadena de valor estadística. Fuente: [@van2018statistical]	6
2.1. Áreas recorridas por el fuego entre 2002 y 2024 en incendios mayores de 100 hectáreas en Andalucía.	19
2.2. Incendios durante el periodo de estudio	23
2.3. Observaciones para los que no está disponible alguna de las variables topográficas	26
3.1. Incendios durante el periodo de estudio	30
3.2. Distribución temporal de las observaciones en función de la variable objetivo.	31
3.3. Distribución espacial de las observaciones en función de la variable objetivo.	32
3.4. Boxplot de cada variable numérica en función de la variable objetivo.	32
3.5. Media mensual de la temperatura en función de fire.	33
3.6. Media mensual de la humedad relativa en función de fire.	34
3.7. Media mensual de la precipitación en función de fire.	34
3.8. Media mensual de la humedad del suelo en función de fire.	35
3.9. Media mensual de la humedad del suelo en función de fire.	35
3.10. Media mensual de la humedad del suelo en función de fire.	36
3.11. Correlaciones entre variables numéricas	37
3.12. Gráfico de coordenadas paralelas.	38
3.13. PCA sobre la matriz de correlaciones de las variables numéricas	38
3.14. Histogramas de las variables categóricas en función de fire.	39
3.15. Métricas de rendimiento de los modelos de regresión logística con penalización.	41
3.16. Métricas de rendimiento del árbol de decisión en función de el parámetro de costocomplejidad.	42

3.17. Métricas de rendimiento de Random Forest en función de los parámetros.	43
3.18. Métricas de rendimiento de KNN en función del número de vecinos.	44
3.19. Métricas de rendimiento de svm en función del coste.	44
3.20. Métricas obtenidas sobre el conjunto de validación por cada uno de los modelos seleccionados.	45
3.21. Curvas ROC sobre el conjunto de validación.	46
3.22. Métricas obtenidas sobre el conjunto test por cada uno de los modelos seleccionados.	47
3.23. Curvas ROC sobre test.	47
 4.1. Malla de puntos con una resolución de 10km por 10km.	49
4.2. Probabilidades de incendios estimadas el día 15 de cada mes de 2022 con el modelo de regresión logística con penalización. Los triángulos indican los incendios de más de 100ha registrados en ese mes.	50
4.3. Probabilidades de incendios estimadas el día 15 de cada mes de 2022 con el modelo de SVM lineal.Los triángulos indican los incendios de más de 100ha registrados en ese mes.	51
4.4. Probabilidades de incendios estimadas el día 15 de cada mes de 2022 con el modelo de random forest.Los triángulos indican los incendios de más de 100ha registrados en ese mes.	52
4.5. Área recorrida por el fuego en el incendio de Sierra Bermeja.	52
4.6. Mapa con las probabilidades de incendio estimadas en los días en torno al origen del incendio de Sierra Bermeja. El área total recorrida por el fuego se muestra en rojo.	53
 A.1. Distribución espacial de T2M por mes.	57
A.2. Distribución espacial de RH2M por mes.	58
A.3. Distribución espacial de GWETTOP por mes.	58
A.4. Distribución espacial de WS10M por mes.	59
A.5. Distribución espacial de PRECTOTCORR por mes.	59
A.6. Distribución espacial de WD10M por mes.	60
A.7. Distribución espacial de poblacion.	61
A.8. Distribución espacial de dens_poblacion.	61
A.9. Distribución espacial de NDVI por mes.	62
A.10.Distribución espacial de elevacion.	62
A.11.Distribución espacial de pendiente.	63
A.12.Distribución espacial de curvatura.	63
A.13.Distribución espacial de dist_carretera.	64

A.14.Distribución espacial de dist_sendero.	64
A.15.Distribución espacial de dist_camino.	65
A.16.Distribución espacial de dist_poblacion.	65
A.17.Distribución espacial de dist_electr.	66
A.18.Distribución espacial de dist_ferrocarril.	66
A.19.Distribución espacial de uso_suelo.	67
B.1. Coeficientes del modelos de regresión logística lasso seleccionado.	70

Índice de tablas

2.1.	Datos brutos	21
2.2.	Códigos de uso de suelo	25
2.3.	Conjunto de datos depurados	27
3.1.	Métricas de los modelos seleccionados sobre el conjunto de validación. . . .	45
3.2.	Métricas sobre el conjunto test.	46

0.1. Introducción

0.2. Objetivos

El objetivo de esta investigación será construir modelos que permitan predecir el riesgo de incendio forestal en la Comunidad Autónoma de Andalucía.

Subobjetivos:

1. Construir un conjunto de datos que permita la realización de análisis y la posterior construcción de modelos de Machine Learning para la predicción de incendios forestales en Andalucía a partir de un estudio previo del problema.
2. Modelizar el riesgo de incendio forestal usando distintos algoritmos de ML y comparar sus resultados
3. Analizar potenciales casos de interés.

0.3. Hipótesis

“Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches, Omid Ghorbanzadeh, Khalil Valizadeh Kamran, Thomas Blaschke, Jagannath Aryal, Amin Naboureh, Jamshid Einali and Jinhu Bian”

0.4. Revisión bibliográfica

Table 6. Cont.

No	Factors	Impacts	References
3	Altitude (m)	Altitude is an essential feature of fire danger distribution that should be considered. The wildfires that occur at higher altitudes are less severe because of the increase in moisture.	Koutsias et al. 2002, [30]; Ganteaume, et al. 2013, [31]; Jaafari et al. 2019, [26]
4	Annual temperature (°C)	There is a direct relationship between temperature increase and wildfires.	Baltar et al. 2015, [32]; Oulad Sayad et al. 2019, [10]
5	Annual rainfall (mm)	The annual rainfall parameter is one of the most significant variables of wildfires; rainfall moisture influences the speed of wildfires, which makes more extension of the burned area.	Vasilakos et al. 2009, [33]; Tanskanen et al. 2005, [34]
6	Wind effect	Wind can affect the extension and direction of the wildfires immediately after their ignition.	Darvishsefat et al. 2018, [11]; Sakellariou et al. 2016, [3]; Fovell and Gallagher et al. 2018, [35]
7	Plan curvature (100/m)	The positive curvature can be considered convex, such as the top of the hills, while negative curvature is concave, which refers to features like valleys. These criteria have different effects on the dynamics of wildfires.	Hilton et al. 2016, [36]; Pourtaghi et al. 2015, [4]
8	Topographic wetness index (TWI)	Fuel moisture is directly related to the required heat of ignition occurs. The actual relationship between the TWI and wildfires differs from other ground conditions and features.	Porensky et al. 2018, [37]; Ghorbanzadeh and Blaschke, 2018, [12]
9	Landform	Areas with steep slopes usually present the highest percentage of wildfires	Cantarello et al. 2011, [38];
10	Land use	Land use patterns based on shape and type have different impacts on wildfire risk.	Pourghasemi et al. 2016, [29]
11	NDVI	Reduction of the NDVI can cause an increase in water stress and the risk of fire.	Verbesselt et al. 2006, [39]; Pourtaghi et al. 2015, [4]
12	Distance to stream (m)	There is an indirect relationship between the distance from water sources and wildfire risk.	Razali and Sheriza 2010, [40]; Lee et al. 2010
13	Distance to road (m)	Roads provide access to forest areas; as a result, the risk of wildfire increases.	Syphard et al. 2008 Lee et al. 2010, [9]
14	Recreation area (m)	Recreation areas are places for human gatherings; humans, intentional or unintentional, can increase the risk of wildfire.	Stephens, 2005, [41]; Keeley and Fotheringham, 2003, [42]
15	Potential solar radiation	Increasing solar radiation can cause a reduction in the soil moisture and an increase in temperature and, consequently, wildfire risk.	Peters et al. 2013, [43]; Oulad Sayad et al. 2019, [10]
16	Distance to villages (m)	Expansion of residential area can increase the risk of wildfires, mostly because of human activities.	Canu et al. 2017, [44]; Lee et al. 2010, [9]

Figura 1: Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches, Omid Ghorbanzadeh, Khalil Valizadeh Kamran, Thomas Blaschke, Jagannath Aryal, Amin Naboureh, Jamshid Einali and Jinhu Bian.

Capítulo 1

Preliminares

1.1. Datos georreferenciados

Todos los datos empleados en este trabajo son datos georreferenciados, lo que significa que están asociados a ubicaciones geográficas específicas. Por ello, resulta esencial introducir los tipos de datos más utilizados para trabajar con esta información, sus características y las herramientas disponibles para manipularlos. Se tratarán los datos vectoriales y los datos ráster, al ser los tipos de datos fundamentales en este contexto, con características bien diferenciadas entre ellos.

1.1.1. Datos Vectoriales

El modelo de datos vectoriales se basa en puntos ubicados dentro de un sistema de referencia de coordenadas (CRS, por sus siglas en inglés). Estos puntos pueden representar características independientes o pueden estar conectados para formar geometrías más complejas como líneas y polígonos.

1.1.1.1. Simple features

Las *Simple features* son un estándar abierto ampliamente utilizado para la representación de datos vectoriales, desarrollado y respaldado por el *Open Geospatial Consortium* (OGC), una organización sin ánimo de lucro dedicada a la creación de estándares abiertos e interoperables a nivel global dentro del marco de los sistemas geográficos de información (GIS, por sus siglas en inglés) y de la *World Wide Web*.

El paquete *sf* proporciona clases para datos vectoriales geográficos y una interfaz de línea de comandos consistente para importantes bibliotecas de bajo nivel para geoprocесamiento (*GDAL*, *PROJ*, *GEOS*, *S2*,...).

Los objetos *sf* son fáciles de manipular ya que son dataframes o tibbles con dos características fundamentales. En primer lugar, contienen metadatos geográficos adicionales: tipo de geometría, dimensión, *Bounding Box* (límites o extensión geográfica) e información sobre el Sistema de referencia de coordenadas. Y además, presentan una columna de geometrías. Algunas ventajas del uso de este modelo de datos en R son que en la mayoría

de operaciones los objeto *sf* se pueden tratar como data frames, los nombres de las funciones son consistentes (todos empiezan por `st_`), las funciones se pueden combinar con el operador tubería y además funcionan bien con el ecosistema de paquetes *tidyverse*.

El paquete *sf* de R soporta 18 tipos de geometrías para las *simple features*, de las cuales las más utilizadas son: *POINT*, *LINESTRING*, *POLYGON*, *MULTIPOINT*, *MULTILINESTRING*, *MULTIPOLYGON* y *GEOMETRYCOLLECTION*.

1.1.2. Datos Ráster

El modelo de datos ráster representa el espacio con una cuadrícula de celdas (o píxeles) a cada una de las cuales se le asocia un valor o varios, tratándose así de ráster de una o varias capas, respectivamente. Lo más común es trabajar con cuadrículas regulares, es decir, formadas por celdas rectangulares de igual tamaño. Sin embargo, existen otros modelos de ráster más complejos en los que se usan cuadrículas irregulares (rotadas, truncadas, rectilíneas o curvilíneas).

Los datos en formato ráster constan de una cabecera y una matriz cuyos elementos representan celdas equipespacadas. En la cabecera del raster se definen el Sistema de referencia de coordenadas, la extensión (o límites espaciales del área cubierta por el ráster), la resolución y el origen. El origen son las coordenadas de uno de los píxeles del ráster, que sirve de referencia para los demás, siendo generalmente utilizado el de la esquina inferior izquierda (aunque el paquete *TERRA*, usado en este trabajo, usa por defecto el de la esquina superior izquierda). La resolución se calcula como:

$$\text{resolution} = \frac{x_{\max} - x_{\min}}{ncol}, \frac{y_{\max} - y_{\min}}{nrow}$$

La representación en forma de matriz evita tener que almacenar explícitamente las coordenadas de cada una de las cuatro esquinas de cada pixel, debiendo almacenar solamente las coordenadas de un punto (el origen). Esto, unido a las operaciones del álgebra de mapas hacen que el procesamiento de datos ráster sea mucho más eficiente que el de datos vectoriales.

Se usará el paquete *TERRA* para tratar los datos en formato ráster. Este paquete permite tratar el modelo de rásteres regulares con una o varias capas a través de la clase de objetos **SpatRaster**. Sin embargo, existen otras alternativas, como el paquete *stars*, que además de ser más potente, permite trabajar con rásteres no regulares y ofrece una mejor integración con el paquete *sf* y el entorno *tidyverse*.

1.1.3. Sistemas de Referencia de Coordenadas

Intrínseco a cualquier modelo de datos espaciales está el concepto de Sistema de referencia de coordenadas (CRS), que establece cómo la geometría de los datos se relaciona con la superficie terrestre. Es decir, es el nexo de unión entre el modelo de datos y la realidad, por lo que juega un papel fundamental. Los CRS pueden ser de dos tipos: geográficos o proyectados.

1.1.3.1. Sistemas de Coordenadas Geográficas

Los sistemas de coordenadas geográficas (GCS) identifican cada punto de la superficie terrestre utilizando la longitud y la latitud. La longitud es la distancia angular al Meridiano de Greenwich medida en la dirección Este-Oeste. La latitud es la distancia angular al Ecuador medida en la dirección Sur-Norte.

Cualquier sistema de coordenadas geográficas se compone de tres elementos: el elipsoide, el geoide y el *datum*. El primero es el elipsoide (o esfera) utilizado para representar de forma simplificada la superficie terrestre, sobre el que se supone que se encuentran los datos y que permitirá realizar mediciones. El segundo, el geoide, es el modelo matemático que representa la verdadera forma de la Tierra, que no es suave sino que presenta ondulaciones debidas a las fluctuaciones del campo gravitatorio a lo largo de la superficie terrestre, que además cambian a una amplia escala temporal. Y el tercero, el *datum*, indica cómo se alinean el elipsoide y el geoide, es decir, cómo el modelo matemático se ajusta a la realidad. Este puede ser local o geocéntrico, en función de si el elipsoide se ajusta al geoide en un punto concreto de la superficie terrestre o de si es el centro del elipsoide el que se alinea con el centro de la Tierra. Ejemplos de *datum* geocéntricos usados en este trabajo son:

- *European Terrestrial Reference System 1989* (ETRS89), usado ampliamente en la Europa Occidental.
- *World Geodetic System 1984* (WGS84), usado a nivel global.

1.1.3.2. Sistemas de Coordenadas Proyectadas

Un Sistema de Coordenadas Proyectadas (PCS) es un sistema de referencia que permite identificar localizaciones terrestres y realizar mediciones en una superficie plana, es decir, en un mapa. Estos sistemas de coordenadas se basan en las coordenadas cartesianas, por lo que tienen un origen, un eje X y un eje Y y usan una unidad lineal de medida (en este trabajo se usará el metro). Pasar de una superficie elíptica (GCR) a una superficie plana (PCS) requiere de transformaciones matemáticas apropiadas y siempre induce deformaciones en los datos.

Al proyectar la superficie terrestre en una superficie plana siempre se modifican algunas propiedades de los objetos, como el área, la dirección, la distancia o la forma. Un PCS solo puede conservar alguna de estas propiedades pero no todas, por lo que es habitual clasificar los PCS en función de la propiedad que mantienen: las proyecciones de igual área preservan el área, las azimutales preservan la dirección, las equidistantes preservan la distancia y las conformales preservan la forma local. En función de como se realice la proyección, estas también se pueden clasificar en planas, cilíndricas o cónicas.

Un caso particular y ampliamente usado de PCS cilíndrico son los *Universal Transverse Mercator* (UTM), en los que se proyecta el elipsoide sobre un cilindro tangente a este por las líneas de longitud (los meridianos). De esta forma, se divide el globo en 60 zonas de 6° de longitud, para cada una de las cuales existe un PCS UTM correspondiente que está asociado al meridiano central. Se trata de proyecciones conformes, por lo que preservan ángulos y formas en pequeñas regiones, pero distorsionan distancias y áreas.

A lo largo de este trabajo se utilizará ampliamente el Sistema de coordenadas proyectadas UTM30N (es habitual especificar el hemisferio para evitar confusión en los valores del eje Y, ya que miden distancia al ecuador, de ahí la N de hemisferio norte).

1.2. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA), es una parte fundamental de todo proyecto de *Machine Learning* y en general de cualquier proyecto en el que se deba trabajar con datos de cualquier procedencia para extraer de ellos conclusiones. Antes del procesamiento de los datos es siempre necesario explorar, entender y evaluar la calidad de estos, pues como indica la expresión inglesa *garbage in, garbage out*, si trabajamos con datos pobres, no podemos esperar obtener de ellos buenos resultados.

El EDA hace referencia al conjunto de técnicas estadísticas (tanto numéricas como gráficas) con las que se pretende explorar, describir y resumir la naturaleza de los datos, comprender las relaciones existentes entre las distintas variables presentes, identificar posibles errores o revelar posibles valores atípicos. Todo esto con el objetivo de maximizar nuestra compresión sobre el conjunto de datos.

1.2.1. Depuración de los datos

La depuración de los datos o *data cleaning* es el proceso de detectar y corregir o eliminar datos incorrectos, corruptos, con formato incorrecto, duplicados o incompletos dentro de un conjunto de datos. Puede considerarse una fase dentro del EDA (como se sugiere en Wickham y Grolemund [2]) o una fase previa a este.

Puede entenderse que el *data cleaning* es el proceso de pasar de *raw data* o datos en bruto a datos técnicamente correctos y finalmente a datos consistentes.

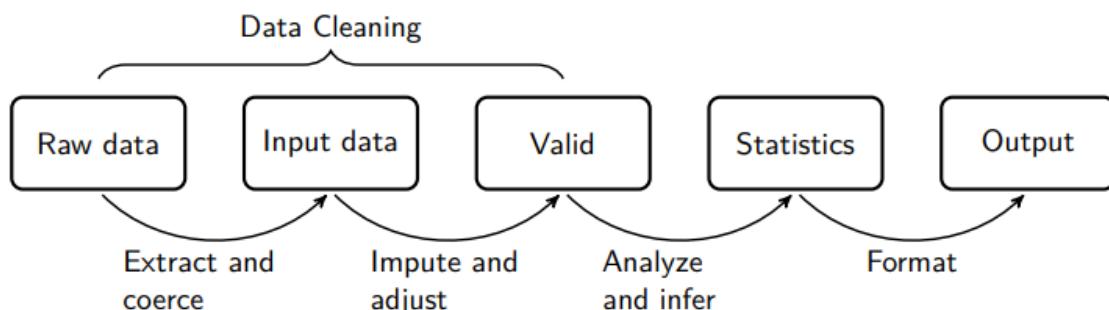


Figura 1.1: Cadena de valor estadística. Fuente: [@van2018statistical]

Entendemos que un conjunto de datos es técnicamente correcto cuando cada valor pertenece a una variable y está almacenado en el tipo que le corresponde en base al conocimiento del dominio del problema. Para ello se debe reajustar el tipo de cada variable al que le corresponda en base al conocimiento que se tenga sobre esta, codificando los valores en las clases adecuadas si fuese necesario.

Decimos que un conjunto de datos es consistente cuando es técnicamente correcto y, además, adecuado para el análisis estadístico. Se trata, por tanto, de datos que han eliminado, corregido o imputado los valores faltantes, los valores especiales, los valores atípicos y los errores.

1.2.2. Análisis de componentes principales

El Análisis de Componentes Principales (PCA) es una técnica de reducción de la dimensionalidad ampliamente usada en el análisis de datos multivariante. Al emplear técnicas de reducción de dimensionalidad como PCA, se persiguen diversos objetivos: eliminar correlaciones redundantes, reducir el ruido presente en los datos y facilitar el uso de algoritmos cuya eficiencia computacional está fuertemente influenciada por la dimensionalidad de los datos.

Definición 1.2.1 Dadas $\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^k$ realizaciones de un vector aleatorio \underline{X} en \mathbb{R}^k , se dice que los vectores $\underline{c}_1, \underline{c}_2, \dots, \underline{c}_p \in \mathbb{R}^k$ son las $*k*$ componentes principales (muestrales) del vector aleatorio \underline{X} si forman una base ortonormal del espacio $V \subset \mathbb{R}^k$ de dimensión $*p*$ que minimiza la media del cuadrado de la distancia euclídea entre los \underline{x}_i y su proyección $\pi_V(\underline{x}_i)$ sobre V .

De la propia definición se desprende que las componentes principales no son únicas.

Proposición 1.2.1 Las $*p*$ primeras componentes principales se corresponden con los autovectores unitarios asociados a los p mayores autovalores λ_j de la matriz de covarianzas muestrales.

Definición 1.2.2 Se define la fracción de varianza explicada por las p primeras componentes principales como $\frac{\sum_{j=1}^p \lambda_j}{\sum_{j=1}^k \lambda_j}$.

El número de componentes principales necesarias para explicar un porcentaje dado de la varianza de los datos puede servir para caracterizar la magnitud del problema.

1.3. Modelos

El problema que se aborda en este trabajo se engloba dentro de lo que se conoce como aprendizaje supervisado, ya que para cada observación del conjunto de entrenamiento se conoce el valor de la variable objetivo (en este caso si ha habido incendio o no). Más concretamente, se trata de un problema de clasificación binaria, ya que el objetivo es asignar cada observación a una de las dos clases posibles (incendio o no incendio). Existen numerosas técnicas de clasificación binaria supervisada, en este trabajo se explorarán algunas de las de uso más común en problemas similares.

1.3.1. Regresión logística (con penalización)

La regresión logística es un caso particular de modelo lineal generalizado basado en las siguientes hipótesis:

- Hipótesis distribucional. Dadas las variables explicativas, \underline{X}_i con $i = 1, 2, \dots, n$, se verifica que las variables $Y|_{\underline{X}=\underline{x}_i}$ y su distribución pertenece a la familia Bernoulli, es decir,

$$Y|_{\underline{X}=\underline{x}_i} \sim Be(\pi(x_i))$$

- Hipótesis estructural. La esperanza $E(Y|_{\underline{X}=\underline{x}_i}) = \pi_i$ está relacionada con un predictor lineal ($\eta_i = \beta^t z_i$) a través de la función *logit* (con $\underline{z}_i = (1, \underline{x}_i)$). Es decir, dado que

$$\eta_i = \underline{\beta}^t \underline{z}_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right)$$

O equivalentemente,

$$\pi_i = \frac{\exp(\underline{\beta}^t \underline{z}_i)}{1 + \exp(\underline{\beta}^t \underline{z}_i)}$$

Bajo estas hipótesis, la función de log-verosimilitud dada una muestra $\{(\underline{x}_i, y_i)\}_{i=1,\dots,n}$ es:

$$l(\underline{\beta}) = \sum_{i=1}^n \left[y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) + \ln(1 - \pi_i) \right]$$

En la regresión logística clásica se estima el vector de parámetros $\underline{\beta}$ maximizando la función de log-verosimilud, o lo que es equivalente, minimizando su opuesta. Por tanto, el problema de optimización a resolver será

$$\min_{\underline{\beta}} -l(\underline{\beta})$$

Sin embargo, con el objetivo de evitar el sobreajuste y construir modelos con mayor capacidad de generalización existen variaciones de la regresión logística que incluyen un término de penalización en la función objetivo. Las dos variantes de uso más extendido son la regresión *ridge* y *lasso*.

Sea $\underline{\beta} = (\beta_0, \underline{\beta}_1)$, donde $\underline{\beta}_1$ contiene los coeficientes de las covariables. En la regresión *ridge* el término de penalización es de la forma $\|\underline{\beta}_1\|_2^2$ mientras que en la regresión *lasso* el penalización es de la forma $\|\underline{\beta}_1\|_1$. Por tanto, el problema de optimización será

$$\min_{\underline{\beta}} -l(\underline{\beta}) + \lambda \sum \beta_i^2$$

en el caso de la regresión logística *ridge* y

$$\min_{\underline{\beta}} -l(\underline{\beta}) + \lambda \sum |\beta_i|$$

en el caso de la regresión logística *lasso*.

En este trabajo se usará el paquete *glmnet*, que implementa una combinación de ambos métodos (llamada *elastic net*), en la que se añade un parámetro de mixtura $\alpha \in [0, 1]$ que combina ambos enfoques. El problema de optimización resultante en este caso será:

$$\min_{\underline{\beta}} -l(\underline{\beta}) + \lambda \left[(1 - \alpha) \sum \beta_i^2 + \alpha \sum |\beta_i| \right]$$

1.3.2. Support Vector Machine

Las Máquinas de Vector Soporte (SVM) son una familia de modelos principalmente usados en problemas de clasificación binaria (si bien se pueden extender a problemas de clasificación multiclas o de regresión) que parten de la idea de encontrar el hiperplano que “mejor” separa al conjunto de puntos.

1.3.2.1. SVM lineal

Dada una muestra $\{(\underline{x}_i, y_i)\}_{i=1,\dots,n}$ con $\underline{x}_i \in \mathbb{R}^d$ y $y_i \in \{-1, 1\}$ para todo $i \in \{1, \dots, n\}$, el objetivo es encontrar al hiperplano de la forma

$$h(x) = w_1x_1 + w_2x_2 + \dots + w_dx_d + b = \underline{w}^t\underline{x} = 0$$

que mejor separe a la muestra.

Definición 1.3.1 Se dice que la muestra es linealmente separable si existe un hiperplano, denominado hiperplano de separación, que cumple, para todo $i \in \{1, \dots, n\}$:

$$\underline{w}^t\underline{x}_i + b \geq 0 \text{ si } y_i = +1$$

$$\underline{w}^t\underline{x}_i + b \leq 0 \text{ si } y_i = -1$$

Definición 1.3.2 Dado un hiperplano de separación de una muestra linealmente separable, se define el margen como la menor de las distancias del hiperplano a cualquier elemento de la muestra. Se denotará por τ .

Proposición 1.3.1 Dado un punto \underline{x}_i y un hiperplano $h(x) = w_1x_1 + w_2x_2 + \dots + w_dx_d + b = \underline{w}^t\underline{x} = 0$, la distancia entre ambos viene dada por:

$$d(h, \underline{x}_i) = \frac{|h(\underline{x}_i)|}{\|\underline{w}\|} = \frac{y_i(\underline{w}^t\underline{x}_i + b)}{\|\underline{w}\|}$$

Donde $\|\cdot\|$ hace referencia a la norma euclídea.

Proposición 1.3.2 Dada una muestra linealmente separable $\{(\underline{x}_i, y_i)\}_{i=1,\dots,n}$ con $\underline{x}_i \in \mathbb{R}^d$ y $y_i \in \{-1, 1\}$ y un hiperplano de separación $h(x) = \underline{w}^t\underline{x} = 0$ con margen τ , se verifica que

$$\frac{y_i(\underline{w}^t\underline{x}_i + b)}{\|\underline{w}\|} \geq \tau \quad \forall i \in \{1, \dots, n\}$$

O equivalentemente,

$$y_i(\underline{w}^t\underline{x}_i + b) \geq \tau \|\underline{w}\| \quad \forall i \in \{1, \dots, n\}$$

Y, además, es posible reescribir el mismo hiperplano h de forma que $\tau \|\underline{w}\| = 1$.

De esta última expresión se deduce que maximizar el margen τ es equivalente a minimizar la norma euclídea de \underline{w} . Por tanto, para encontrar el hiperplano de separación óptimo para una muestra en las condiciones de la proposición anterior basta resolver el problema de optimización siguiente:

$$\begin{aligned}
 & \min_{w,b} \quad \frac{1}{2} w^t w \\
 \text{s.t.} \quad & \underline{w}^t \underline{x}_i + b \geq 1, \quad \forall i \in \{1, \dots, n\} \\
 & w \in \mathbb{R}^d, b \in \mathbb{R}
 \end{aligned} \tag{1.1}$$

En general, las muestras no son separables, por lo que es necesario permitir que pueda haber casos mal clasificados, y penalizarlos proporcionalmente a la distancia a la que se encuentren del subespacio correcto (holgura). Para ello, se introducen en la formulación del modelo las variables artificiales ξ_i , $i = 1, \dots, n$. Se habla entonces de hiperplano de separación *soft margin*. De esta forma, se llega al problema de optimización siguiente:

$$\begin{aligned}
 & \min_{w,b,\xi} \quad \frac{1}{2} w^t w + C \sum_{i=1}^n \xi_i \\
 \text{s.t.} \quad & \underline{w}^t \underline{x}_i + b \geq 1, \quad \forall i \in \{1, \dots, n\} \\
 & \xi \geq 0, \quad \forall i \in \{1, \dots, n\} \\
 & w \in \mathbb{R}^d, b \in \mathbb{R}
 \end{aligned} \tag{1.2}$$

donde $C > 0$ es un parámetro de regularización que permite controlar los errores de clasificación permitidos por el modelo, controlando así el sobreajuste. Este parámetro recibe el nombre de coste (*cost*).

1.3.2.2. SVM no lineal

Existen muchos casos en los que el SVM no es capaz de obtener buenos resultados, debido a la estructura de la distribución de las clases en la muestra. En estos casos, es común recurrir a una técnica llamada *kernel trick*. Esta técnica consiste en realizar una inmersión del conjunto de los vectores de la muestra en un espacio de dimensión superior (llamado *feature space*) en el que los casos sí sean separables (o al menos mejore la separabilidad de estos). Esta inmersión en un espacio de dimensión superior se hace indirectamente a través de funciones *kernel*, que calculan los productos escalares entre los vectores de la muestra en el espacio de inmersión. Existen distintos tipos de funciones *kernel* que se corresponden con distintas inmersiones en espacios de dimensión superior:

- Kernel polinomial: $k(x, z) = (\gamma(x^t z + c_0))^p$
- Kernel RDF (radial o gaussiano): $k(x, z) = \exp(-\gamma \|x - z\|^2)$

1.3.3. Decision Trees

Un árbol de decisión (DT) es un algoritmo de aprendizaje supervisado no paramétrico, que puede aplicarse tanto a problemas de clasificación como de regresión. La idea de este método es segmentar el espacio predictor en rectángulos, de forma que para predecir una observación se usa la moda (o la media) de la región a la que pertenece. Se trata de un modelo jerárquico con estructura de árbol, que consta de un nodo raíz, ramas, nodos internos y nodos hojas. Cada nodo representa un test sobre una variable, y cada las ramas que nacen de ese nodo representan los posibles valores que puede tomar esa

variable. De esta forma, para clasificar una nueva instancia basta comenzar en el nodo raíz e ir descendiendo por el árbol hasta llegar al nodo hoja correspondiente, que indicará la clasificación asignada a dicha instancia. La simplicidad del método muestra su principal ventaja, su fácil comprensión dada su estructura de árbol.

Existen diversas técnicas para construir árboles de clasificación (y regresión), aquí se ilustra es una de las más usadas que recibe el nombre de CART (*Classification And Regression Trees*). Se explica para el caso de árboles de clasificación binarios, es decir, en los que de cada nodo salen dos ramas.

Dada una muestra $\{(\underline{x}_i, y_i)\}$ con $\underline{x}_i = (x_{i1}, \dots, x_{id})$, un árbol de clasificación con J hojas se puede expresar como

$$f(\underline{x}) = \sum_{j=1}^J c_j I(\underline{x} \in R_j)$$

donde $\{R_j\}_{j=1, \dots, J}$ es una partición del espacio predictivo y c_j es la clase asignada en R_j para todo $j \in 1, \dots, J$.

En la práctica, c_j se estima asignando la clase mayoritaria en el recinto R_j . Es decir, $\hat{c}_j = \text{moda}(\{y_i | \underline{x}_i \in R_j\})$.

Para construir un árbol de clasificación, el algoritmo necesita decidir las variables tests y los puntos de corte en cada nodo, así como la topología del árbol. Para realizar esto, se vale de un método *greedy*, que en cada nodo elige la variable y el punto de corte que mejor separan los datos en base a una medida de impureza. Es decir, la construcción de un árbol de clasificación no se hace mediante la resolución de un solo problema de optimización global, si no a partir de la resolución de muchos problemas de optimización locales, con las implicaciones que esto pueda tener.

Las medidas de impureza más comúnmente usadas son:

- Error de clasificación: $1 - \max(p, 1 - p)$
- Índice de Gini: $2p(1 - p)$
- Entropía: $-p \log p - (1 - p) \log(1 - p)$

donde p denota la proporción de casos positivos en la muestra.

Así, el algoritmo de construcción de un árbol de clasificación es:

1. Comenzar con el nodo raíz, que incluye todos los casos.
2. Determinar el par (variable,corte) que conduce a una mayor reducción de la impureza. Es decir, dada una medida de impureza Φ se busca la variable $j \in 1, \dots, d$ y el corte $s \in \mathbb{R}$ solución de

$$\min_{j,s} \left[\frac{|R_1|}{|R_1| + |R_2|} \min_{c_i} \Phi(\{y_i | \underline{x}_i \in R_1(j, s)\}) + \frac{|R_2|}{|R_1| + |R_2|} \min_{c_i} \Phi(\{y_i | \underline{x}_i \in R_2(j, s)\}) \right]$$

donde $R_1(j, s) = \{X | X_j \leq s\}$ y $R_2(j, s) = \{X | X_j > s\}$.

3. Aplicar iterativamente el proceso anterior a cada nuevo nodo, hasta que se verifiquen las condiciones de finalización. En este caso, el criterio será finalizar el proceso de división en el nodo una vez que el número de casos en este sea igual o inferior a una cantidad n_{min} fijada de antemano. En los nodos hoja se asigna la clase mayoritaria en el nodo.
4. Podar o recortar el árbol obtenido en base a un criterio de coste-complejidad. Dado un árbol completo T y un valor del parámetro de coste-complejidad α , se elije el subárbol $T_0 \subset T$ obtenido a partir de T mediante poda, es decir, colapsando nodos no terminales, que minimice el criterio de coste complejidad definido como:

$$C_\alpha(T) = \Phi(T) + \alpha|T_0|$$

El parámetro α permite controlar la capacidad de generalización del modelo (*Bias-Variance tradeoff*) y se estima mediante Validación Cruzada.

El gran inconveniente de los árboles de decisión es que en general son modelos con una varianza elevada, por lo que tienden a ser inestables y a producir sobreajuste. Para evitar esto, se recurre al uso de técnicas de *Bagging* y *Boosting*. Una de las técnicas más extendida con árboles de decisión son los Bosques Aleatorios (*Random Forest* en inglés).

1.3.4. Random Forest

La idea detrás del modelo de bosques aleatorios es reducir la varianza de los árboles de decisión sin aumentar el sesgo. Para intentar conseguir este objetivo, la idea es aplicar *Bagging* (*Bootstrap Aggregating*) al modelo de árbol de decisión. Sin embargo, ya que al aplicar *Bagging* la reducción de la varianza es mayor cuanto más incorrelados sean los predictores individuales, en cada nuevo nodo de cada árbol construido se selecciona la variable que más disminuya la impureza de entre un conjunto aleatorio de $m_{try} < d$ predictores.

El algoritmo para construir un bosque aleatorio es el siguiente:

1. Para $b = 1, \dots, B$:
 - a) Seleccionar una muestra bootstrap Z^* de tamaño n del conjunto de entrenamiento.
 - b) Construir un árbol de decisión T_b a partir de la muestra bootstrap b , aplicando recursivamente los siguiente pasos para cada nodo terminan del árbol, hasta que se alcance el tamaño mínimo de nodo n_{min} :
 - I. Seleccionar aleatoriamente m_{try} variables de entre las d variables predictoras.
 - II. Elegir el mejor par variable/división de entre las m_{try} variables seleccionadas en función de la reducción del criterio de impureza.
 - III. Dividir el nodo en dos nodos hijos.
2. De esta forma se obtiene el conjunto de árboles de decisión bootstrap $\{T_b\}_{b=1}^B$.

Para predecir la clase de un nuevo punto \underline{x} se aplica la regla de la clase más votada al conjunto de clases predichas por los B árboles de decisión bootstrap para \underline{x} .

1.3.5. K-Nearest Neighbors

El método de k vecinos más cercanos (KNN) clasifica una nueva observación \underline{x} en base a las clases de las k observaciones del conjunto de entrenamiento más cercanas a estas en el espacio muestral aplicando la regla de la clase más votada. Es decir, dado un espacio muestral Θ con una distancia d definida sobre él, dado un conjunto de entrenamiento $T \subset Y$ y dado $k \in \mathbb{N}^{+*}$ la función calculada por el algoritmo para estimar la clase de $\underline{x} \in \Theta$ es:

$$f(\underline{x}) = \text{majority vote } \{y_i \mid \underline{x}_i \in N_k(\underline{x})\}$$

donde $N_k(\underline{x})$ es el conjunto de los k puntos $\underline{x}_i \in \Theta$ más próximos a \underline{x} en Θ en base a la distancia d .

El parámetro k permite controlar el sobreajuste del modelo.

1.4. Validación del ajuste

Para validar el ajuste de los modelos comentados en los datos, se utilizará una partición temporal en entrenamiento/ validación/ test. Es decir, se asignará el primer 60 % de los datos (de acuerdo al día de la observación) a entrenamiento, el 20 % siguiente a validación y el último 20 % a test. Este enfoque permite evitar el sesgo positivo debido al efecto *look-ahead* en la estimación de la capacidad de generalización de los modelos.

1.5. Evaluación de los modelos

Una vez construido un modelo predictivo es necesario conocer el rendimiento de este sobre nuevos datos, con el objetivo de estimar su capacidad de generalización. Esto es fundamental de cara a determinar si el modelo es adecuado para el propósito previsto o si necesita ajustes o mejoras. La evaluación del rendimiento permite comparar entre diferentes modelos y seleccionar el que mejor se adapte a las necesidades específicas del problema en cuestión. Para ello, se recurre al uso de distintas métricas, en función de las características propias de cada problema.

1.5.1. Clasificación binaria

En el presente trabajo el problema que se aborda es un problema de clasificación binaria, pues tenemos solo dos clases que son la clase positiva y la clase negativa. A la hora de clasificar una nueva instancia pueden darse 4 situaciones:

- Que se clasifique como positiva siendo realmente positiva, en cuyo caso se dirá que forma parte de las *True Positives (TP)*
- Que se clasifique como negativa siendo realmente negativa, en cuyo caso se dirá que forma parte de las *True Negatives (TN)*

- Que se clasifique como positiva siendo realmente negativa, en cuyo caso se dirá que forma parte de las *False Positives (FP)*
- Que se clasifique como negativa siendo realmente positiva, en cuyo caso se dirá que forma parte de las *False Negatives (FN)*

Se considerarán las siguientes métricas de rendimiento para problemas de clasificación binaria:

Tasa de acierto o exactitud. Mide la proporción de casos que han sido correctamente clasificados.

$$\text{Exactitud} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precisión. Mide la proporción de casos clasificados como positivos que realmente lo son.

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Especificidad. Mide la proporción de casos negativos que han sido correctamente clasificados por el modelo.

$$\text{Especificidad} = \frac{TN}{TN + FP}$$

Sensibilidad o recall. Mide la proporción de casos positivos que han sido correctamente clasificados por el modelo.

$$\text{Recall} = \frac{TP}{TP + FN}$$

AUC-ROC. Mide el área bajo la curva ROC (*Receiver Operating Characteristic* o Característica Operativa del Receptor en castellano). Esta curva es una representación gráfica del rendimiento de un modelo de clasificación binaria para todos los umbrales de clasificación. Representa la sensibilidad frente a la proporción de falsos positivos para cada posible umbral de clasificación. El AUC está comprendido entre 0 y 1, entendiéndose que el rendimiento del es mejor cuanto mayor sea su valor. En general, se suelen considerar aceptables modelos con un valor del AUC superior a 0.75.

1.6. Herramientas

Toda la parte práctica del presente trabajo se ha llevado a cabo empleado el lenguaje de programación R a través del entorno de desarrollo integrado que ofrece RStudio. R es un lenguaje y entorno de programación de código abierto desarrollado dentro del proyecto GNU y orientado a la computación estadística. Este lenguaje puede extender sus funcionalidades fácilmente a través de la gran cantidad de paquetes disponibles dentro del repositorio de paquetes de CRAN (The Comprehensive R Archive Network), siendo este uno de sus puntos fuertes, dada la gran comunidad de usuarios y desarrolladores con la que cuenta.

Los paquetes que se han utilizado han sido:

- tidyverse:
 - ggplot2, para la visualización.
 - dplyr, para la manipulación.
 - tidyr, para la ordenación.
 - readr, para la importación.
 - purrr, para la programación funcional
- tidymodels
 - parsnip -...
- sf
 - GDAL
 - ...
- terra
- nasapower: obtención de información climática satelital
- mapSpain:

...

Capítulo 2

Construcción del conjunto de datos

El primer paso a la hora de construir cualquier modelo de predicción es disponer de datos adecuados que permitan explicar correctamente el fenómeno en estudio, en este caso los incendios forestales en Andalucía. Con este fin, se ha llevado a cabo un extenso estudio previo del dominio del problema para conocer qué variables son relevantes de cara a la predicción de incendios forestales, analizando estudios similares realizados anteriormente así como otras fuentes relativas a la ecología del fuego, que nos permitiesen conocer el efecto que cabría esperar de estas variables.

Se ha querido adoptar un enfoque dinámico, es decir, el objetivo no es construir un modelo estacionario que nos indique si una determinada zona se verá afectada por un incendio forestal a lo largo de un amplio periodo temporal, si no que se pretende ser capaz de predecir si un determinado punto del territorio se verá afectado por un incendio forestal en un momento concreto, en base a las covariables correspondientes a ese lugar en ese momento. Es decir, se considera no solo la dimensión espacial de los datos si no también la temporal, al mayor nivel de desagregación disponible. Este es un enfoque mucho menos explorado, debido fundamentalmente a dos factores:

1. La dificultad de disponer de información fiable y de calidad desagregada espacio-temporalmente
2. La dificultad de trabajar con datos de estas características de cara al análisis y principalmente a la modelización, ya que son datos correlados en el tiempo y en el espacio.

Queda claro, por tanto, que se trata de un problema complejo que requiere de simplificaciones para poder ser abordado, más aun dadas las limitaciones en los recursos computacionales disponibles y la enorme cantidad de datos que se están considerando y que requieren de un procesamiento sumamente costoso desde un punto de vista computacional.

Por todo ello, esta sección es probablemente la de mayor importancia y dificultad de todo el trabajo, ya que implica la toma de decisiones que serán determinantes de cara al correcto desempeño de los modelos que se construirán más adelante, requiere de un vasto conocimiento del problema que permita un enfoque adecuado que haga posible la consecución de los objetivos que se esperan conseguir, necesita del uso de técnicas específicas de procesamiento de datos espaciales que no han sido tratadas durante el grado y se ve fuertemente limitada por los escasos recursos computacionales disponibles.

2.1. Determinación del marco del estudio

El primer paso ha sido limitar el área y la franja temporal que abarcará el estudio. Para ello, ha sido necesario basarse principalmente en la disponibilidad y consistencia de la información requerida para el proyecto y en las limitaciones computacionales impuestas por el equipo disponible.

En cuanto a la disponibilidad de información, hay que diferenciar entre la información de incendios forestales y la información de variables que permitan explicar este fenómeno considerando la mayor desagregación espacial y temporal posible.

2.1.1. Incendios forestales

En lo referente a los datos sobre incendios forestales cabe mencionar que España cuenta con una de las mayores y más completas bases de datos sobre incendios forestales a nivel europeo. Se trata de la Estadística General de Incendios Forestales (EGIF), que en su versión definitiva actualmente contiene toda la información que se recoge en cada parte de incendio forestal que ha tenido lugar en España desde 1983 hasta 2015, incluyendo su información espacial con sus coordenadas de origen. Se ha explorado extensamente el uso de esta base de datos para el proyecto, dada su exhaustividad y completitud. Sin embargo, lamentablemente no ha sido posible en este caso incorporarla al trabajo por diversas razones.

La principal de ellas fue que hasta marzo de 2024 la base de datos de la EGIF solo se encontraba disponible en el Catálogo de Datos del Gobierno de España en formato TURTLE¹ y esto conllevó numerosas dificultades. Se exploraron distintas librerías de R (y alguna de Python) para el manejo de datos en este formato como *RDFlib*. Sin embargo, al tratarse de una base de datos de un tamaño considerable (aproximadamente 1GB y con más de una decena de millones de tripletas), esta librería no era suficientemente eficiente para poder realizar consultas en un tiempo razonable al conjunto de datos. Tras explorar otras alternativas, se valoró la posibilidad de usar un *triplestore*, es decir, una base de datos especialmente diseñada para el almacenamiento y recuperación de tripletas a través de consultas semánticas. En este caso se usó *Apache Jena Fuseki*, ya que cuenta con una interfaz que facilita su uso. Sin embargo, aunque esto supuso una mejora considerable en la eficiencia y permitió realizar consultas sencillas a la base de datos, en este caso fue la complejidad del gráfico de datos (ontología) y la escasa documentación disponible sobre esta, la que impidió que se pudiesen realizar las consultas más complejas requeridas para llevar a cabo el proyecto. Además, se debe tener en cuenta que se trata de una base de datos muy heterogénea y con numerosos datos faltantes debida su naturaleza, por lo que requiere de un preprocesamiento que probablemente será complicado y costoso en tiempo y en recursos computacionales. Al no disponer de ninguno de estos, finalmente se optó por buscar una alternativa más abordable dada las limitaciones con las que cuenta un Trabajo de Fin de Estudios, aunque queda abierta la posibilidad de explorar esta base de datos en futuros estudios, la cual aportar nuevas dimensiones al estudio de los incendios forestales en España gracias a la enorme cantidad de información que ofrece.

¹TURTLE es una sintaxis para RDF con compatible con SPARQL. RDF (*Resource Description Framework*) es un estándar de semántica web utilizado para el intercambiar de datos en la Web.

Ante esta situación, la solución planteada fue limitar el área en estudio a la Comunidad Autónoma de Andalucía, aprovechando la enorme disponibilidad de información medioambiental que ofrece la Red de Información Ambiental de Andalucía (REDIAM). En particular, se emplea la cartografía generada por la REDIAM sobre las áreas recorridas por los incendios forestales entre 1975 y 2022. Esta contiene los perímetros de incendios forestales mayores de 100 ha en Andalucía obtenidos a partir de imágenes de satélite y datos de campo. Se trata por tanto de una información que no es exhaustiva, pues los incendios con una extensión inferior a 100ha no han sido considerados. Sin embargo, frente a no disponer de otra información operativa de mayor calidad, se utilizará esta teniendo en cuenta que tendrá un efecto sobre las conclusiones que se puedan sacar de los modelos que se construyan.

De esta forma, se han recopilado los polígonos de 1090 incendios forestales ocurridos en Andalucía entre 2002 y 2022, junto con la fecha de inicio de cada uno de ellos (Figura 2.1).

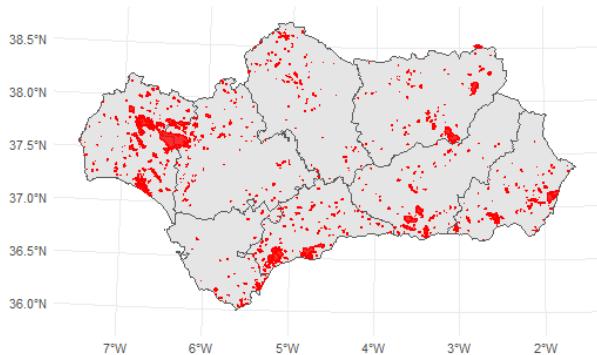


Figura 2.1: Áreas recorridas por el fuego entre 2002 y 2024 en incendios mayores de 100 hectáreas en Andalucía.

2.1.2. Variables predictoras

Una vez limitada la extensión territorial del estudio el siguiente paso era acotar la franja temporal que abarcaría el estudio en base a la disponibilidad de datos adecuados para explicar el fenómeno en cuestión desagregando espacial y temporalmente.

Los incendios forestales son un proceso sumamente complejo, en el que actúan numerosos factores de muy distinta índole (citar estudio...). Además, dentro de un incendio forestal se pueden distinguir distintas fases que presentan características muy diversas y sobre las que actúan distintos agentes: ignición, propagación y extinción. Dada la información sobre incendios forestales disponible, se está obligado a adoptar un enfoque

global, pues no se dispone de los puntos de ignición u origen de los incendios forestales. El enfoque será, por tanto, intentar predecir si una determinada localización se verá afectada por un incendio forestal (de más de 100 ha) en un momento concreto.

Además, es importante tener en cuenta que existen factores estructurales que tienen una influencia directa sobre los regímenes de incendios forestales como son las tendencias de uso y explotación de los bosques, la presencia de interfaz urbano forestal, los tipos y técnicas de agricultura que se llevan a cabo, la presencia e intensidad del pastoreo, los cambios en los usos de suelo e incluso conductas sociales y tendencias demográficas diversas. Se trata de variables que cambian a lo largo de periodos relativamente largos de tiempo y que muy difícilmente pueden ser incluidos en los modelos, dada la falta de datos sobre ellas, así como su carácter transversal. Por ello, se ha considerado conveniente no extender en exceso el periodo de estudio, reconocida la imposibilidad de incluir en el modelo todas las variables que tienen un impacto relevante en la aparición de incendios y que son cambiantes en el tiempo.

Todo ello hace necesario que el conjunto de datos utilizado contenga información sobre todas las dimensiones (o al menos las principales) que influyen en cualquiera de las fases de un incendio forestal. Es decir, se deben incluir la dimensión antropogénica, la demográfica, la hidrográficas, la topográfica, la meteorológica y la vegetación. Es importante recalcar que siempre se hace referencia a datos geoespaciales pues debe ser la información relativa al lugar (y al momento) del incendio, con la dificultad posterior que esto supondrá.

Por último, es importante diferenciar entre características que se considerarán estructurales (y por tanto invariantes a lo largo del periodo de estudio) y aquellas que se considerarán variables en el tiempo. Dentro de las primeras se encuentran todas las características relacionadas con la topografía del terreno, las infraestructuras y los usos del suelo, como por ejemplo el modelo de elevaciones, la distribución de asentamientos de población, la red de carreteras y el uso de suelo. Todas las demás variables de carácter demográfico, meteorológico o de vegetación se considerarán, por tanto, desagregadas temporalmente.

En base a todo lo mencionado y a la disponibilidad de información de calidad de las categorías comentadas, se ha decidido limitar la franja temporal del estudio a los 20 años que van de 2002 a 2022, ambos inclusive.

2.2. Fuentes de datos

Como se ha comentado en la sección anterior, los datos sobre los incendios forestales se han obtenido de los perímetros de incendios forestales mayores de 100 ha en Andalucía entre 1975 y 2020 disponibles la REDIAM. De cada incendio registrado se dispone de su fecha de inicio y del polígono del área recorrida por el fuego, así como de otras variables que dependen del año de la campaña y que no serán relevantes de cara al estudio.

Tomando como base estudios similares (...) y partiendo de las 6 categorías ya mencionadas se han recopilado 23 conjuntos de datos de distinto tipo que se usarán para explicar y predecir los incendios forestales en Andalucía. Estos conjuntos se recogen en la Tabla 2.1, donde también se indica la fuente de la que ha sido obtenido cada uno de ellos, el tipo de datos que contiene (indicando su resolución en el caso de los datos ráster) y la frecuencia de las observaciones (o resolución temporal) en el caso de las variables temporales. En realidad, el número de archivos de datos que se manejan es mucho mayor,

Categoría	Datos	Fuente	Tipo de dato	Frecuencia
Topográficas	Altitud	DERA ^a	TIFF (100m)	-
	Orientación	REDIAM ^b	TIFF (100m)	-
	Pendiente	REDIAM	TIFF (100m)	-
	Curvatura	REDIAM	TIFF (100m)	-
Vegetación	NDVI	REDIAM	TIFF (250m)	Mensual
Antropogénicas	Uso de suelo	DERA	Shapefile	-
	Red de carreteras	DERA	Shapefile	-
	Red de ferrocarril	DERA	Shapefile	-
	Línea eléctrica	DERA	Shapefile	-
	Espacio protegido	DERA	Shapefile	-
	Senderos / Vías Verde / Carriles Bici	DERA	Shapefile	-
	Caminos / Vías Pecuarias	DERA	Shapefile	-
Demográficas	Número de habitantes por municipio y año	IECA ^c	csv	Anual
	Extensión municipal	IECA ^c	csv	-
Hidrográficas	Principales Ríos	MAGRAMA ^d	Shapefile	-
Meteorológicas	Precipitación (mm/day)	NASA POWER ^e	df (0.5° x 0.625°)	Diaria
	Temperatura a 2m sobre la superficie (°)	NASA POWER	df (0.5° x 0.625°)	Diaria
	Humedad del suelo (%)	NASA POWER	df (0.5° x 0.625°)	Diaria
	Dirección del viento a 10 metros sobre la superficie terrestre(°)	NASA POWER	df (0.5° x 0.625°)	Diaria
	Humedad relativa a 2m sobre la superficie (%)	NASA POWER	df (0.5° x 0.625°)	Diaria
	Cantidad de precipitaciones (mm/day)	NASA POWER	dfdf (0.5° x 0.625°)	Diaria

Fuente: Elaboración propia

^a Datos Espaciales de Referencia de Andalucía (DERA)

^b Descargas Rediam

^c Instituto de Estadística y Cartografía de Andalucía (IECA)

^d Ministerio de Agricultura, Alimentación y Medio Ambiente (MAGRAMA)

^e NASA Prediction Of Worldwide Energy Resources (NASA POWER)

Tabla 2.1: Datos brutos

ya que, por ejemplo, para la variable *NDVI* se dispone de un archivo *tiff* para cada mes del periodo de estudio, resultando en un total de unos 240 archivos ráster diferentes solo para esta variable. Inevitablemente, esto añade cierta complejidad al manejo de los datos.

Es relevante la heterogeneidad de los datos recopilados, pues se dispone tanto de datos tabulares como de datos espaciales y dentro de estos últimos de datos vectoriales y datos ráster, con distintas resoluciones, distintas frecuencias y distintos sistemas de referencia de coordenadas. Esto hará que el procesamiento de estos datos hasta obtener datos adecuados para el análisis estadístico sea costoso y que deban utilizarse técnicas específicas de geocomputación.

Cabe también mencionar que se ha optado por el uso de datos meteorológicos basados en modelos y en observaciones satelitares, en lugar del uso de datos provenientes de estaciones meteorológicas. Si bien la información de estaciones meteorológica puede ser más precisa, la dificultad de disponer de datos consistentes y continuos en el tiempo a lo largo del periodo de estudio de las variables meteorológicas seleccionadas ha hecho que este enfoque no sea viable. En esta dirección se ha explorado la API de la AEMET y algunos paquetes de R como *climate*, sin llegar a resultados satisfactorios. Por otro lado, el paquete *nasapower* permite la descarga de una gran cantidad de variables meteorológicas con frecuencia diaria y con una resolución de aproximadamente 0.5×0.625 grados de latitud y longitud (unos 50km). Si bien es cierto que no es lo ideal, es la única opción que se ha considerado viable y, de cara a la construcción de unos primeros modelos approximativos podría ser suficiente. Si quisiese extenderse el estudio, sería conveniente profundizar en la búsqueda de alternativas que permitan obtener información meteorológica de una mayor calidad y detalle.

2.3. Procesamiento de los datos

Una vez se dispone de todos los conjuntos de datos que se usarán en el estudio, el siguiente paso será combinarlos de manera adecuada y transformarlos a un formato tabular apto para el análisis estadístico y la construcción de modelos predictivos. Dado que el objetivo que se persigue es predecir si, dada unas condiciones meteorológicas concretas en un momento dado, un punto del territorio andaluz se verá afectado o no por un incendio forestal, será necesario disponer de una cantidad suficiente de muestras negativas y positivas distribuidas espacial y temporalmente que tengan asociadas las variables explicativas correspondientes.

Intuitivamente, las muestras positivas serán aquellas observaciones (puntos definidos en el tiempo y en el espacio) dentro del marco espacio-temporal del estudio en las que se ha detectado un incendio forestal en el día de la observación. Es decir, son observaciones dentro de los polígonos de incendios el día que estos se han producido. Por tanto, las muestras negativas serán observaciones dentro del marco espacio-temporal definido en las que no se ha detectado un incendio forestal. Es importante tener en cuenta que dado que solo se dispone de los incendios con una extensión mayor a 100 ha, la muestra cuenta con un importante sesgo, ya que los casos positivos están infrarepresentados. Por ello, no podremos hacer inferencia a todos los incendios forestales, si no solo a los de una extensión superior a 100ha.

A continuación se detalla el proceso seguido para generar el conjunto de datos depurado sobre el que desarrollar el estudio a partir de los distintos conjuntos de datos en bruto:

1. Generación de una muestra balanceada de casos positivos y negativos.
2. Asignación de las variables descriptivas a cada observación.
3. Depuración de la muestra.

2.3.1. Generación de una muestra balanceada de casos positivos y negativos.

Para poder construir cualquier modelo de clasificación binaria se necesita disponer de una muestra que cuente con un número suficiente de casos positivos y negativo. Además, es aconsejable trabajar con conjuntos de datos balanceados para evitar sesgos en los modelos de clasificación [ARTICULO].

Como ya se ha comentado, se considerarán observaciones positivas aquellas que se hayan visto afectadas por un incendio forestal en el día y lugar de la observación. En cambio, serán observaciones negativas aquellas que no se hayan visto afectadas por un incendio forestal en el día y lugar de la observación. Estas observaciones deberán generarse a partir de los polígonos de incendios disponibles. Para ello, se usará un enfoque similar al utilizado en [https://www.researchgate.net/publication/228527438_Learning_to_predict_forest_fires_with_different_data_mining_techniques], con algunas diferencias importantes. En [cita al paper] usan los puntos de ignición como muestras positivas y su objetivo es predecir los puntos de origen de los incendios forestales. En cambio, en el presente trabajo no se disponen de los puntos de ignición de los incendios, por lo que el enfoque adoptado es ligeramente diferente; el objetivo es predecir las zonas que pueden

verse afectadas por un incendio forestal (superior a 100ha) bajo unas circunstancias concretas. De esta forma, para construir la muestra de casos positivos se han generado 10 puntos aleatorios dentro de cada polígono de incendio y se les ha asignado la fecha del día de inicio del incendio. Los casos negativos se generarán igual que en [cita paper]: se toman fechas aleatorias dentro del periodo de estudio y a cada una de ellas se le asocia una localización aleatoria dentro del área de estudio satisfaciendo que deben estar a al menos 15km de de cualquier incendio detectado en un margen de ± 3 días. Esta forma de tomar los casos negativos asegura que estén lo suficientemente alejados de los incendios forestales para representar condiciones no influidas por estos, dando prioridad así a las áreas con una menor prioridad de ocurrencia de incendio en un período definido. Sin embargo, sería conveniente en estudios futuros plantearse si esos parámetros (franja de ± 3 días y distancia superior a 15km) son adecuados o tal vez sería más adecuado tomar otros valores, basados, por ejemplo, en la duración media de los incendios en Andalucía y otras características propias de los incendios en la región.

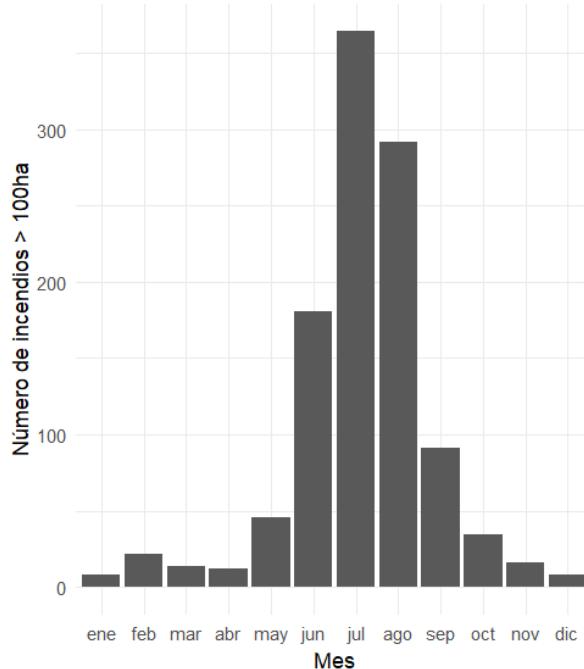


Figura 2.2: Incendios durante el periodo de estudio

Tanto en (cita paper anterior) como en otros estudios consultados se generan los casos negativos tomando fechas completamente aleatorias dentro de la franja temporal del estudio. Sin embargo, esto induce un claro sesgo en los datos, ya que las observaciones positivas no se distribuyen uniformemente entre los 12 meses, si no que se concentran marcadamente en los meses de verano, como puede observarse en la Figura 2.2. Generar el conjunto de datos sin tener en cuenta este hecho hace que las muestras positivas y negativas tengan características meteorológicas muy diferenciadas que no responden al verdadero proceso latente de aparición de incendios forestales sino al proceso de selección de la muestra, pudiendo provocar un marcado sesgo positivo en las medidas de evaluación de los modelos que en realidad no estarían reflejando la realidad si no los sesgos introducidos en los conjuntos de datos. Por ello, en el presente trabajo, para generar aleatoriamente los días de las muestras negativas se seguirá una distribución de probabilidad proporcional a la cantidad de incendios observados a lo largo del periodo de estudio en cada mes.

Mediante este enfoque se espera obtener una muestra balanceada y estratificada por mes que permita construir modelos de clasificación capaces de captar los patrones latentes de aparición de incendios forestales. Además, este enfoque permite centrar el esfuerzo computacional en los períodos en los que más incendios se producen.

2.3.2. Asignación de las variables descriptivas a cada observación

Una vez generada una muestra balanceada de 22170 observaciones dentro de Andalucía entre el 1 de enero de 2002 y el 31 de diciembre de 2022 siguiendo el enfoque apenas descrito, el siguiente paso es asignar a cada una de ellas los valores correspondientes a ese día y a esa localización concreta de todas las variables predictoras a partir de los conjuntos de datos que se han recopilado (recogidos en la Tabla 2.1). Dado que será necesario calcular distancias, se usa un Sistema de Referencia de Coordenadas proyectado al generar la muestra. En particular se usa una variante del UTM30N por ser el que traen los archivos *raster* de las variables topográficas (es conveniente, siempre que sea posible, no transformar el crs de los datos *raster* pues al hacerlo se deben interpolar los valores de los píxeles, produciendo pérdida de información).

A continuación se detalla el proceso seguido para manipular construir cada variable:

- Variables topográficas: Simplemente se extrae el dato del píxel en el que cae cada observación para cada uno de los conjuntos de datos.
- Variables antropológicas:
 - Se calculan las distancias de cada observación a la red de carreteras, a la red de ferrocarril, a las poblaciones y a la línea eléctrica, creando así las variables *dist_carreteras*, *dist_ferrocarril*, *dist_poblaciones* y *dist_electr*, respectivamente. Las geometrías de los senderos, de las vías verdes y de los carriles bici se unen y se calcula la distancia de cada observación a este conjunto de geometrías, generando así la variable *dist_sendero*. De la misma forma se procede con los caminos y las vías pecuarias, que dan origen a la variable *dist_camino*. Estas uniones se han llevado a cabo por considerar que sus elementos tienen características similares. Siempre se hace referencia a la distancia euclídea.
 - Para construir una variable dicotómica *enp* que indique si la observación se encuentra o no dentro de un Espacio Natural Protegido primero se ha rasterizado el conjunto de polígonos de Espacios Naturales Protegidos de Andalucía (de forma que en cada píxel se indica 1 si el centro de este está dentro del polígono de un ENP o 0 si no) y posteriormente se ha extraído el valor del píxel que contiene a cada observación. En la rasterización se ha usado como modelo el ráster de elevaciones con una resolución de $100m \times 100m$. Proceder de este modo hace que se pierda algo de resolución pero resulta muchísimo más eficiente computacionalmente que comprobar para cada observación la relación espacial “estar dentro del polígono de algún ENP” (este enfoque resultaba computacionalmente inviable dado el equipo disponible).
 - Para construir la variable *uso_suelo* se ha procedido de manera similar. Primero se ha rasterizado, usando como modelo el mapa de elevaciones con una

Nivel 1	Nivel 2	Código
Superficies artificiales	Zonas urbanas	11
	Zonas industriales, comerciales y de transporte	12
	Zonas de extracción minera, vertederos y de construcción	13
	Zonas verdes artificiales, no agrícolas	14
Zonas agrícolas	Tierras de labor	21
	Cultivos permanentes	22
	Prados y praderas	23
	Zonas agrícolas heterogéneas	24
Zonas forestales con vegetación natural y espacios abiertos	Bosque	31
	Espacios de vegetación arbustiva y/o herbácea	32
	Espacios abiertos con poca o sin vegetación	33
Zonas húmedas	Zonas húmedas continentales	41
	Zonas húmedas litorales	42
Superficies de agua	Aguas continentales	51
	Aguas marinas	52

Tabla 2.2: Códigos de uso de suelo

resolución de $100m \times 100m$, asignando a cada píxel la categoría de uso de suelo del polígono que cubriese el centro del píxel. Y posteriormente, para cada observación se ha extraído el valor del píxel sobre el que estuviese. De nuevo, se ha procedido así por cuestiones de eficiencia. Es necesario hacer algunos comentarios más sobre esta variable. La información de uso de suelo proviene del mapa de Ocupación de Uso de Suelo CORINE Land Cover 2018, donde se establece 3 niveles de clasificación con 5, 15 y 44 clases, respectivamente. La primera clase se corresponde con el primer dígito del código, las segunda con el segundo y la tercera con el tercero. En este trabajo se ha decidido trabajar con el segundo nivel de clasificación, por lo que se consideran solo los dos primeros dígitos del código de cada observación. En la Tabla 2.2 se recoge la clase correspondiente a cada código.

- Para construir la variable *poblacion* se ha asignado a cada observación el código del municipio en el que está y se ha hecho un *left_join* con el código del municipio y el año de la observación. Para la variable *dens_población* se ha procedido de la misma manera pero se ha dividido por la extensión del municipio en km^2 .
- La distancia a los ríos *dist_rios* se ha obtenido simplemente calculando la distancia de cada observación al conjunto de geometrías de los principales ríos de España.
- El NDVI viene en archivos *raster* mensuales (lo que supone un total de unos 240 archivos en formato *.tiff*). Para cada observación se ha extraído el valor del píxel correspondiente (en función de las coordenadas del punto) del archivo correspondiente (que depende del mes y año de la observación).

2.3.3. Depuración de la muestra

Una vez construido el conjunto de datos “en bruto”, se tratan los valores perdidos y se ajustan adecuadamente los tipos de las variables. En primer lugar se convierten en factores las variables *fire*, *enp* y *uso_suelo*. A continuación, se codifican las variables numéricas *WD10M* y *orientacion* en los 4 puntos cardinales y sus bisectrices, generando así 8 clases

(“N”, “NW”, “W”, “SW”, “S”, “SE”, “E”, “NE”). En el caso de la variable orientación se añade también la clase “plano”, si la pendiente en ese punto es 0.

El conjunto de datos construido tiene 200 registros incompletos, lo cual supone un 0.1 % del total de registros. De estos, el 68 % son casos negativos y el 32 % son casos positivos. Los valores perdidos se encuentran en las variables demográficas (85), en *uso_suelo* (8), en NDVI (85) y en las variables topográficas (53). Las causas de los datos faltantes son:

- El dato no está disponible para esa observación. Esto sucede con las variables demográficas (hay años para los que no está disponible el número de habitantes de algunos municipios) y el NDVI (para algunos meses no se dispone del archivo correspondiente).
- En el caso de las variables topográficas los valores perdidos se encuentran todos en los límites de la comunidad (Figura 2.3). Al proceder de datos en formato *raster*, los píxeles con información no se ajustan exactamente a los límites de Andalucía (ya que son cuadrados). Esto provoca que para algunos puntos situados en los bordes del polígono no esté disponible la información de las variables topográficas.

Tras explorar otras alternativas y teniendo en cuenta tanto el reducido número de registros incompletos como la naturaleza de los valores desconocidos, se opta simplemente por eliminar estos registros.

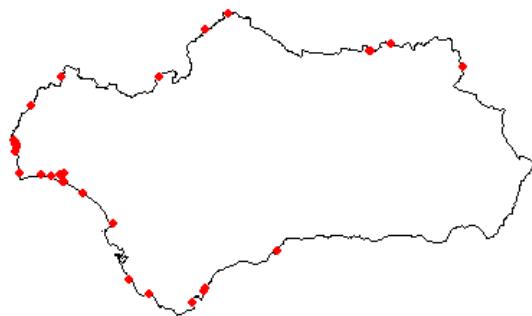


Figura 2.3: Observaciones para los que no está disponible alguna de las variables topográficas

El resultado de todo este proceso es un conjunto de datos con 21546 registros y 27 variables, las cuales se detallan en la Tabla 2.3.

Categoría	Nombre	Descripción	Tipo
Topográficas	elevacion	Elevación sobre el nivel del mar (m)	numérica
	orientacion	Orientación de la pendiente descendiente	categórica
	pendiente	Pendiente del terreno (°)	numérica
	curvatura	Curvatura de la superficie	numérica
Vegetación	NDVI	Índice de vegetación de diferencia normalizada	numérica
Antropogénicas	uso_suelo	Clasificación del uso del suelo	categórica
	dist_carretera	Distancia a la carretera más cercana (m)	numérica
	dist_ferrocarril	Distancia a la vía de ferrocarril más cercana (m)	numérica
	dist_electr	Distancia a la línea electrica más cercana (m)	numérica
	enp	Espacio Natural Protegido	categórica
	dist_sendero	Distancia a la vía verde, al carril bici o al sendero más cercano (m)	numérica
	dist_camino	Distancia al camino o a la vía pecuaria más cercano (m)	numérica
Demográficas	poblacion	Número de habitantes del municipio	numérica
Hidrográficas	dens_poblacion	Densidad de población del municipio	numérica
Meteorológicas	dist_rios	Distancia al río más próximo (m)	numérica
	PRECTRORCORR	Promedio corregido del total de precipitaciones en la superficie	numérica
	T2M	Temperatura promedio del aire a 2 metros sobre la superficie de la tierra (°C)	numérica
	GWETTOP	Porcentaje de humedad del suelo	numérica
	WD10M	Promedio de la dirección del viento a 10 metros sobre la superficie de la tierra	categórica
	WS10M	Promedio de la velocidad del viento a 10 metros sobre la superficie de la tierra (m/s)	numérica
Variable Objetivo	RH2M	Humedad relativa a 2 metros sobre la superficie de la tierra	numérica
	fire	Incendio forestal	categórica
Identificadoras	date	Fecha de la observación	fecha
	municipio	Nombre del municipio	texto
	cod_municipio	Código del municipio	texto
	geometry	Geometría de los puntos	sfc

Tabla 2.3: Conjunto de datos depurados

Capítulo 3

Cuerpo

3.1. Análisis exploratorio de datos

En esta sección se aplicarán distintos métodos numéricos y gráficos de análisis de datos a la muestra generada siguiendo el procedimiento detallado en el capítulo anterior. Se usarán principalmente técnicas de estadística descriptiva para comprender las características del conjunto de datos y extraer conocimiento útil para el problema que se intenta abordar, predecir incendios forestales. Es importante tener presente que se trata de datos correlados espacial y temporalmente, lo que hace necesario el uso de métodos específicos para este tipo de datos. Los objetivos de esta etapa son:

1. Generar conocimiento sobre el conjunto de datos que nos permita evaluar la calidad de este, sin olvidar las limitaciones que ya se han comentado en la sección anterior.
2. Conocer, al menos de forma descriptiva, el impacto de cada variable en la variable objetivo. Este conocimiento será necesario para evaluar e interpretar los modelos que se construirán en la próxima sección.
3. Analizar las características de las distintas variables, de cara a usar posteriormente técnicas de preprocesamiento adecuadas para cada modelo.

Antes de abordar el estudio detallado de cada una de las variables y las relaciones entre estas, en la Figura 3.1 se recoge un resumen de todo el conjunto de datos, sin incluir la columna de geometría. En este resumen se puede observar que en el conjunto de datos hay 4 tipos de variables (además de la variable *geometry* que es de tipo *simple feature column POINT*, abreviado como *sfc_POINT*): cadenas de caracteres, fechas, factores y variables numéricas.

Se puede observar que hay registros en 749 municipios diferentes (de los 785 municipios de que hay en Andalucía). Probablemente el hecho de que en algunos municipios no haya habido observaciones sea debido a los datos faltantes. Las variables *municipio* y *cod_municipio* no se incorporarán a los modelos. De la misma forma, se puede ver que hay observaciones en 3691 días diferentes. El conjunto cuenta con 5 variables de tipo factor: *fire* (la variable objetivo), *WD10M*, *orientacion*, *enp* y *uso_suelo*; y con 18 variables numéricas. Aunque cada una de ellas se analizará a continuación con detalle, ya cabe hacer algunos comentarios:

- El 38 % de las observaciones se encuentran en espacios de vegetación arbustiva y/o herbácea (código 32).
- Como era de esperar, por la forma en la que se ha tomado la muestra, el conjunto está balanceado.
- El 81 % de las observaciones se encuentran fuera de Espacios Naturales Protegidos.
- Todas las variables, salvo *T2M* y *curvatura*, son positivas y la mayoría de ellas presentan una marcada distribución asimétrica hacia la derecha.
- Las variables muestran escalas muy diversas entre ellas, siendo *GWETTOP* la que presenta menor desviación típica (0.145) y *poblacion* la que tiene una desviación típica mayor (64453). Se evidencia la necesidad de incluir algún método de normalización de las variables en el preprocesamiento de los datos.

— Data Summary —																				
	Values																			
Name	datos																			
Number of rows	21546																			
Number of columns	26																			
Column type frequency:																				
character	2																			
Date	1																			
factor	5																			
numeric	18																			
Group variables																				
None																				
— Variable type: character —																				
skim_variable n_missing complete_rate min max empty n_unique whitespace																				
1 cod_municipio	0	1	5	5	0	749	0													
2 municipio	0	1	3	32	0	749	0													
— Variable type: Date —																				
skim_variable n_missing complete_rate min max median n_unique																				
1 date	0	1	2002-01-02	2022-11-29	2012-08-04		3691													
— Variable type: factor —																				
skim_variable n_missing complete_rate ordered n_unique top_counts																				
1 fire	0	1	FALSE	2	1: 10794, 0: 10752															
2 WD10M	0	1	FALSE	8	SW: 4965, W: 4867, S: 3786, SE: 3316															
3 orientacion	0	1	FALSE	9	S: 3267, SW: 3090, SE: 2956, W: 2544															
4 emp	0	1	FALSE	2	0: 17393, 1: 4153															
5 uso_suelo	0	1	FALSE	15	32: 8068, 21: 3128, 24: 2798, 22: 2786															
— Variable type: numeric —																				
skim_variable n_missing complete_rate mean sd p0 p25 p50 p75 p100 hist																				
1 T2M	0	1	24.6	5.20	-0.8	22.7	25.6	28.0	36.0											
2 GWETTOP	0	1	0.318	0.145	0.09	0.2	0.29	0.42	0.91											
3 RH2M	0	1	46.4	16.1	9.23	33.8	44.6	57.5	96.8											
4 WS10M	0	1	3.64	1.42	0.99	2.67	3.34	4.25	13.3											
5 PRECTOTCORR	0	1	0.229	1.38	0	0	0	0	46.1											
6 elevacion	0	1	494.	400.	0	169.	425.	711.	3351.											
7 pendiente	0	1	12.3	11.8	0	3.46	8.24	17.9	98.0											
8 curvatura	0	1	0.00549	0.0563	-0.294	-0.0193	-0.000100	0.0255	0.420											
9 dist_carretera	0	1	1799.	1907.	0.0706	507.	1200.	2425.	17662.											
10 dist_poblacion	0	1	902.	800.	0	379.	713.	1170.	8215.											
11 dist_electr	0	1	5253.	5640.	0.220	1216.	3462.	7276.	48069.											
12 dist_ferrocarril	0	1	15311.	13521.	0.824	5086.	11855.	21467.	85242.											
13 dist_camino	0	1	762.	741.	0.0238	240.	539.	1060.	7090.											
14 dist_sendero	0	1	5619.	4805.	0.0465	1781.	4361.	8276.	25103.											
15 poblacion	0	1	22095.	64453.	114	2252	4860	16759	704414.											
16 dens_poblacion	0	1	105.	320.	2.30	12.2	30.8	71.1	4974.											
17 dist_rios	0	1	6781.	5836.	1.94	2162.	5237.	10123.	37074.											
18 NDVI	0	1	0.412	0.136	0	0.314	0.393	0.495	0.944											

Figura 3.1: Incendios durante el periodo de estudio

3.1.1. Distribución de la variable objetivo

En primer lugar, se estudiará la distribución de la *fire* espacial y temporalmente.

En la Figura 3.2 se muestran los histogramas de la variable objetivo en función del día de la semana, del mes y del año, respectivamente. En primero de ellos se observa que mientras que la distribución de los casos negativos es uniforme entre los días de la semana, en los casos positivos se aprecia un ligero aumento en el fin de semana, especialmente en

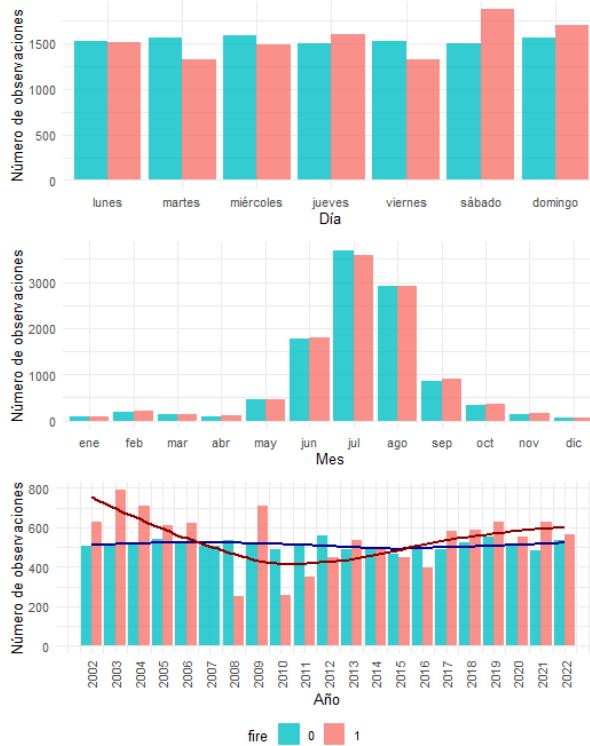


Figura 3.2: Distribución temporal de las observaciones en función de la variable objetivo.

el sábado. En el segundo histograma, se observa como las observaciones se concentran en los meses de verano y en cada mes hay una cantidad balancedada de muestras de ambas clases (esto es fruto del proceso de muestreo de las observaciones negativas, que como se ha explicado en la sección anterior, se ha llevado acabo asegurando que la proporción de casos negativos en cada mes sea igual a la de los casos positivos). En el tercer histograma es remarcable que, mientras las observaciones negativas están uniformemente distribuidas entre los 20 años del estudio, las positivas muestran una disminución importante en los años 2008 y 2010. En el año 2007 no hay observaciones positivas, debido a que los 4 polígonos de incendios mayores de 100ha que había registrados ese año no disponían de la fecha de inicio del incendio, por lo que no pudieron usarse para el estudio. Se desconoce la causa del reducido número de incendios (mayores de 100ha) en 2007, 2008 y 2010.

Dada la clara influencia el mes y la aparente influencia del día de la semana en la aparición de incendios, estas variables serán incluidas en los modelos a través del procesamiento de la variable *date*.

En la Figura 3.3 se observa claramente como las 10752 muestras negativas están uniformemente distribuidas dentro de los límites de la Comunidad Autónoma de Andalucía, mientras que las 10794 muestras positivas se concentran a ambos lados de la cuenca del río Guadalquivir, con una mayor densidad de observaciones en la provincia de Huelva y en algunas zonas de la costa mediterránea (como ya se apreciaba en la Figura 2.1).

3.1.2. Análisis univariantes variables numéricas

El análisis univariante de las variables numéricas se lleva a cabo desde 3 enfoques complementarios:

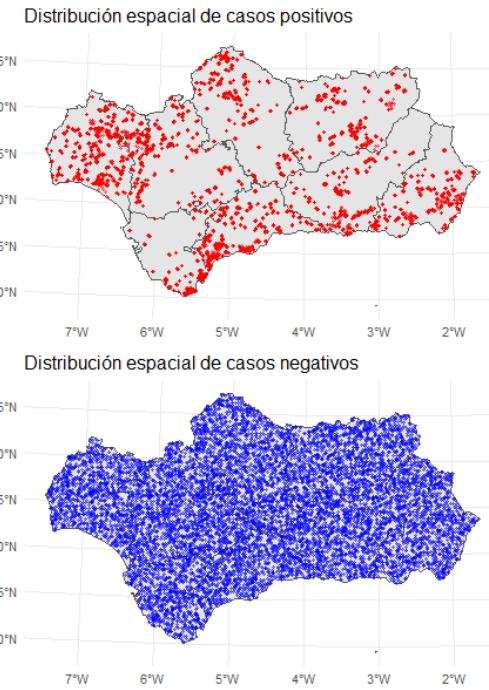


Figura 3.3: Distribución espacial de las observaciones en función de la variable objetivo.

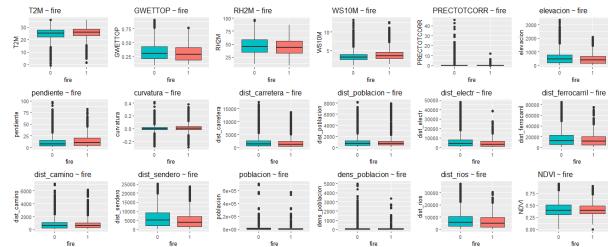


Figura 3.4: Boxplot de cada variable numérica en función de la variable objetivo.

1. A través de los resúmenes numéricos recogidos en la Figura 3.1 y del análisis gráfico de los diagramas de caja y bigote (3.4).
2. Estudiando la media mensual de cada variable en función de la variable *fire*.
3. Analizando la distribución espacial de cada variable separando por mes si corresponde.

En los *boxplots* de las variables numéricas en función de la variable *fire* (3.4) destacan varios aspectos. Por un lado, como ya se había comentado anteriormente, que las variables presentan escalas muy diferentes y que la mayoría de las variables tienen una marcada asimetría hacia la derecha. Por otro lado, es evidente la gran cantidad de valores *outliers* que se observan en los datos, lo que tendrá implicaciones en los modelos que se construyan con ellos. Sin embargo, es importante destacar que no se trata de observaciones erróneas, si no que son inherentes a la naturaleza de los datos. Por ejemplo, en el caso de la variable *PRECTOTCORR* el valor máximo observado es 46.06mm en un día, un valor elevado que sin duda es atípico en esta región de clima seco, pero sin embargo, posible. Es también remarcable que todas las variables presentan una variabilidad similar en ambos niveles del factor *fire*, lo que indica que no será un problema de clasificación trivial. A

priori, solo con los diagramas de caja y bigotes y los resúmenes numéricos es difícil llegar sacar más conclusiones, sin embargo, sí pueden observarse sutiles diferencias entre las distribuciones de algunas variables para ambos niveles del factor *fire*.

Dada la naturaleza temporal de algunas variables, el análisis gráfico de los *boxplots* resulta insuficiente. Con el fin de considerar la componente estacional de las variables climáticas y de vegetación, a continuación, se estudiará la media mensual de cada una de estas variables función de la variable objetivo.

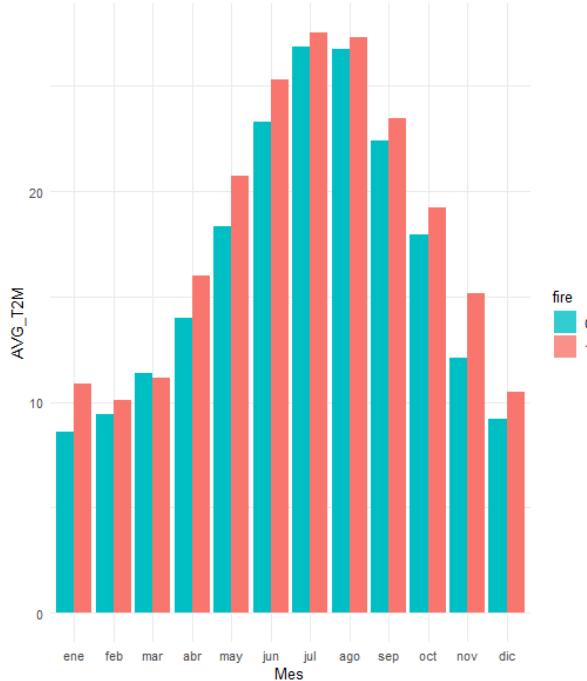


Figura 3.5: Media mensual de la temperatura en función de fire.

En la Figura 3.5 se puede observar como en casi todos los meses, la temperatura media mensual es superior en las observaciones en las que se ha registrado un incendio forestal.

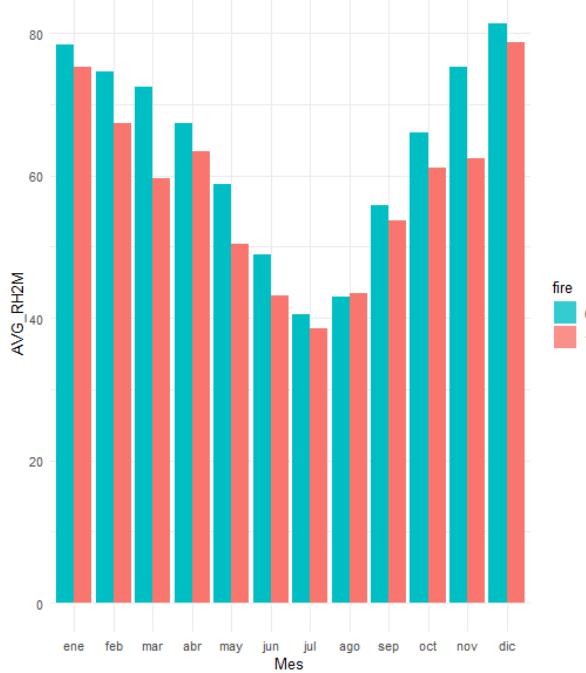


Figura 3.6: Media mensual de la humedad relativa en función de fire.

En la Figura 3.6 se puede observar que en todos los meses la media mensual de la humedad relativa del aire a 2m sobre la superficie es menor en las observaciones en las que se ha registrado un incendio forestal. Sin embargo, las diferencias se reducen durante los meses de verano, en los que la humedad presenta valores bajos en ambas clases.

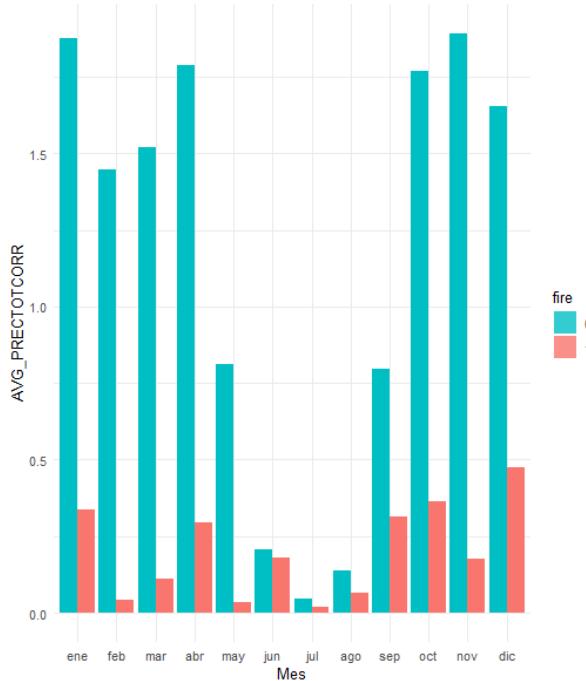


Figura 3.7: Media mensual de la precipitaciones en función de fire.

En la Figura 3.7 se observa una clara diferencia en la media mensual de las precipitaciones diarias en función del de si se ha registrado o no un incendio forestal en la observación,

siendo significativamente mayor en este último caso.

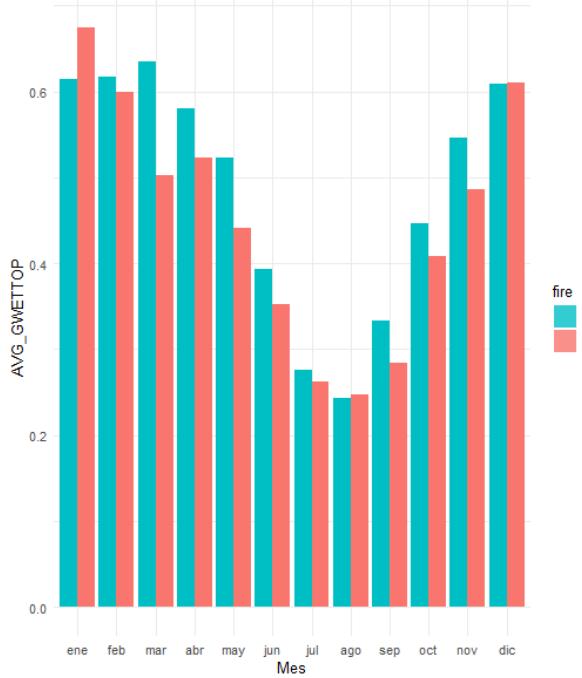


Figura 3.8: Media mensual de la humedad del suelo en función de fire.

En la Figura 3.8 se observa un gráfico similar al de la humedad relativa del aire, con valores medios más elevados en las observaciones en las que no se han registrado incendio forestal. Sin embargo, también parece que las diferencias son más reducidas durante la estación estival.

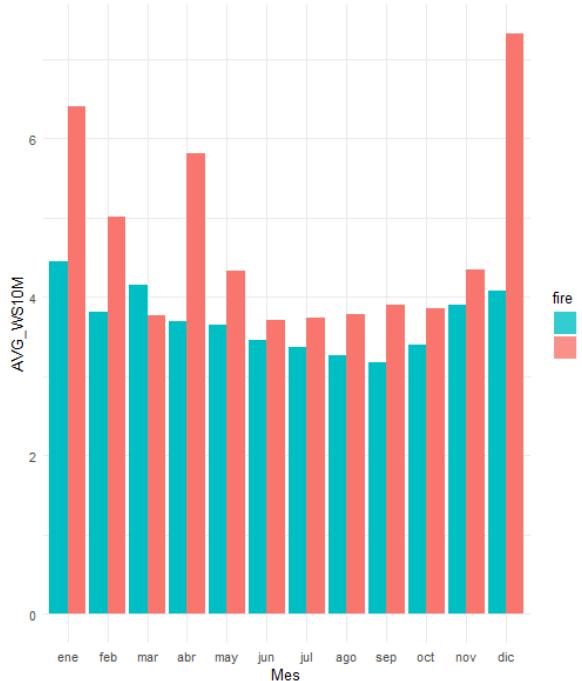


Figura 3.9: Media mensual de la humedad del suelo en función de fire.

En la Figura 3.9 se observa como durante todos los meses, la media mensual de la velocidad del viento a 10 metros sobre la superficie es mayor en los registros en los que ha habido un incendio forestal.

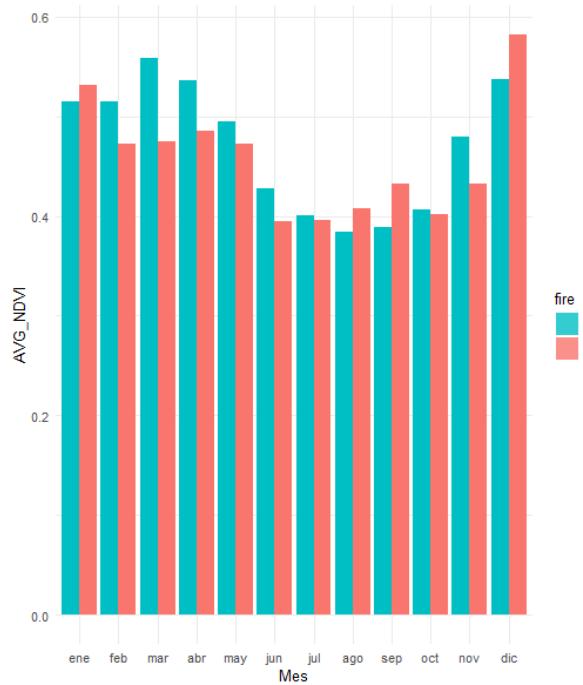


Figura 3.10: Media mensual de la humedad del suelo en función de fire.

Como se observa en la Figura 3.10, las diferencias entre los casos en los que se ha registrado incendio y los que no en términos del *NDVI* no están claras.

En el Apéndice: Gráficos espaciales EDA se recogen los gráficos espaciales y espacio_temporales de todas las variables numéricas. En ellos se refleja como la valores de las variables en estudio son coherentes con lo que cabría esperar de la realidad. Además, permiten una comprensión mayor de la distribución espacial (y temporal) de las variables en el área de estudio, lo que será útil de cara a interpretar los modelos que se construyan.

3.1.3. Análisis multivariantes de las variables numéricas

En la Figura 3.11 se muestra un gráfico con las correlaciones entre las variables. La interpretación es sencilla, cuanto más intenso sea el color y cuanto mayor sea la excentricidad de la elipse, mayor será la correlación (en valor absoluto) para ese par de variables. El color de la elipse indica el signo del coeficiente de correlación. De esta forma, se observa que las variables más correlacionadas en la muestra son:

- *T2M* con *RH2M* (negativamente, -0.71)
- *T2M* con *GWETTOP* (negativamente, -0.69)
- *GWETTOP* con *R2HM* (positivamente, 0.68)
- *poblacion* con *dens_poblacion* (positivamente, 0.63)

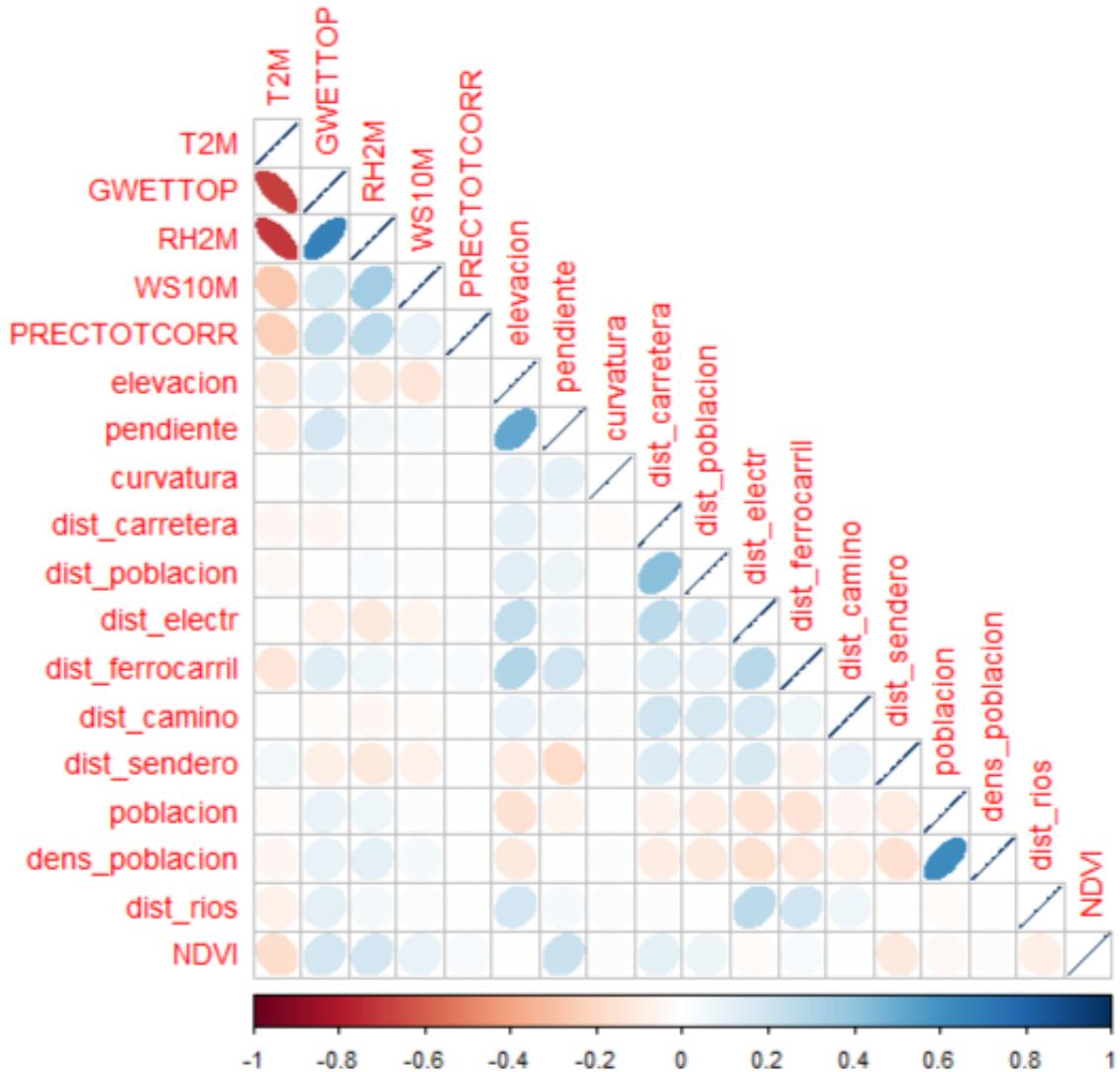


Figura 3.11: Correlaciones entre variables numéricas

En la Figura 3.12 se muestra el gráfico de coordenadas paralelas de las variables tipificadas a una normal estándar, es decir, restándole la media y dividiendo por la desviación típica. Este gráfico complementa la información de los *boxplots*, pues refleja también las relaciones entre las variables. Si bien es cierto que al tener un número bastante elevado de observaciones el gráfico no es tan claro, pueden hacerse algunas observaciones importantes.

En primer lugar, se observa que la variable con mayor variabilidad (una vez tipificada) es PRECTOTCORR, que presenta bastantes valores atípicos, todos ellos en observaciones en las que no se ha registrado incendio. También destacan en este sentido *dens_poblacion* y *poblacion*, entre las que además puede observarse que no hay una relación lineal clara (hay municipios con un elevado número de habitantes pero con una densidad de población reducida y viceversa). Además, puede verse que todas las variables tienen una marcada asimetría positiva (salvo *curvatura*, *T2M* y *NDVI*). Este gráfico es útil también pues permite ver a qué clase de *fire* corresponden los valores más atípicos de cada variable. Por ejemplo: la mayor parte de los valores más elevados de *WS10M*, *dist_poblacion*, *curvatura* y *dist_camino* se dan en observaciones positivas, mientras que en *PRECTOTCORR*,

elevacion, GWTTOP, dist_Carretera, dist_electr y dist_rios sucede lo contrario.

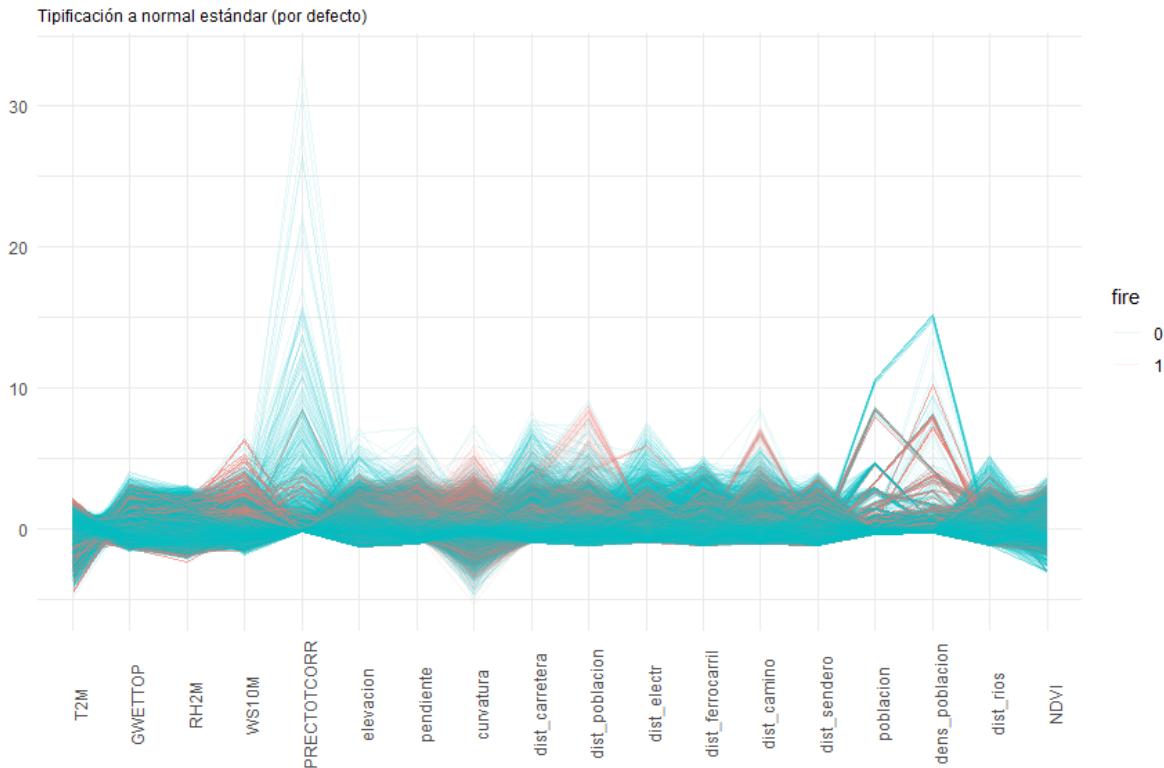


Figura 3.12: Gráfico de coordenadas paralelas.

Los resultados de aplicar análisis de componentes principales sobre la matriz de correlaciones de las 18 variables numéricas se muestran en la Figura 3.13. Como se puede observar, se necesitan al menos 11 componentes principales para lograr explicar el 80 % de la varianza de la muestra, y 14 para alcanzar el 90 % de la varianza de los datos. Estos resultados se aplicarán más adelante en los modelos, pero a nivel meramente explicativo ya indican que se trata de un conjunto de datos complejo en cuanto a la dimensión real de estos.

Importance of components:	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	1.6830901	1.5509109	1.26704790	1.18114361	1.13860921	1.00151777	0.98498153	0.92948032	0.92853341
Proportion of variance	0.1573773	0.1336291	0.08918946	0.07750557	0.07202394	0.05572433	0.05389937	0.04799631	0.04789857
Cumulative Proportion	0.1573773	0.2910065	0.38019595	0.45770152	0.52972546	0.58544979	0.63934916	0.68734547	0.73524404
Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	
Standard deviation	0.91036573	0.86805852	0.85739620	0.78965012	0.72570700	0.65041713	0.60872397	0.52343451	0.48001944
Proportion of variance	0.04604254	0.04186253	0.04084046	0.03464152	0.02925837	0.02350236	0.02058583	0.01522132	0.01280104
Cumulative Proportion	0.78128658	0.82314912	0.86398958	0.89863109	0.92788946	0.95139182	0.97197765	0.98719896	1.00000000

Figura 3.13: PCA sobre la matriz de correlaciones de las variables numéricas

3.1.4. Análisis de las variables categóricas

Las variables categóricas se analizarán a través de los histogramas de cada variable en función de la variable *fire* (Figura 3.14).

En la variable *WD10M* cabe destacar la escasez de observaciones con dirección del viento norte. En el histograma no se observa una clara relación de esta variable con la

variable objetivo, aunque entre las observaciones con viento con dirección sur o suroeste hay más observaciones negativas y entre las que tienen dirección noroeste o este hay una mayor presencia de observaciones positivas.

En el caso de la variable *orientación*, la relación tampoco está clara, aunque puede verse una mayor proporción de observaciones positivas en las superficies con orientación sur (sureste, sur y suroeste).

En términos de la variable *enp* por si sola no se observan diferencias significativas entre ambas clases.

La variable *uso_suelo* sí que muestra una distribución marcadamente diferenciada entre ambas clases. La mayoría de las observaciones positivas se dan en espacios de vegetación arbustiva y/o herbácea, clase en la que hay casi el doble de observaciones positivas que negativas. En tierras de labor y cultivos permanentes la proporción de observaciones negativas es mucho mayor, mientras que en zonas agrícolas heterogéneas y en espacios abiertos con poca o sin vegetación hay una mayor presencia de observaciones positivas. También es relevante el hecho de que casi la totalidad de las observaciones se encuentran en zonas agrícolas y en zonas forestales, mientras que en las demás clases la proporción de observaciones es mucho menor (3.5 de total). Es por ello que antes de construir los modelos, todas las categorías categorías de uso de suelo que no se corresponden con zonas agrícolas o forestales (es decir, todas cuyo código no comience por 2 o 3) se agruparán en el nivel *Otro*. En ?? puede observarse la distribución espacial de esta variable.

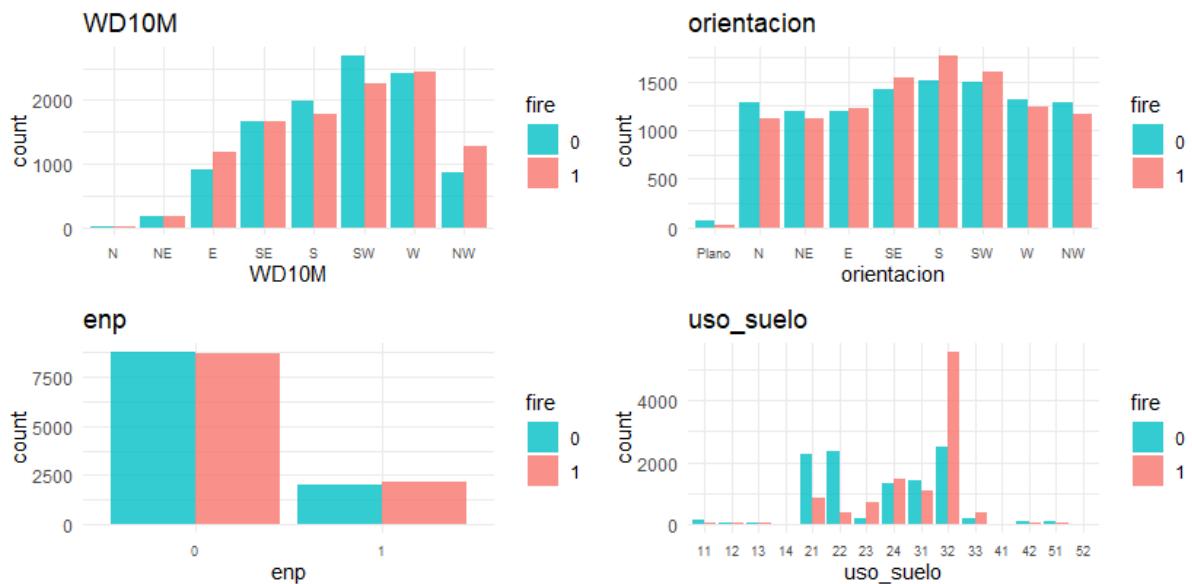


Figura 3.14: Histogramas de las variables categóricas en función de fire.

3.2. Modelización

A continuación se va a utilizar el conjunto de datos construido en el capítulo Construcción del conjunto de datos para entrenar los modelos de clasificación binaria explicados en la sección Modelos. Es evidente que el rendimiento de los modelos debe evaluarse en observaciones futuras, por lo que las técnicas habituales de validación cruzada o partición

aleatoria en entrenamiento/test no son adecuadas para este problema, ya que sufrirían el llamado efecto *look-ahead*. Por tanto, el enfoque que se seguirá en este trabajo será trabajar con una partición en entrenamiento/validación/test construida a partir de la ordenación temporal de las observaciones.

Se compararán los resultados obtenidos en 7 modelos diferentes: Regresión logística con penalización, Regresión logística con penalización + PCA, Árboles de decisión, Bosques aleatorios, KNN, SVM lineal y SVM radial.

Se ha seguido el flujo de trabajo habitual de *tidymodels*:

1º. Crear una partición temporal en entrenamiento (60 %), validación (20 %) y test (20 %), que será utilizada en todos los modelos.

2º. Definir cada uno de los modelos, indicando los parámetros del modelo deberán ajustarse.

3º. Crear la receta (*recipe*) con el preprocesamiento que se usará en cada modelo. Como se observó en la sección Análisis exploratorio de datos, se incluirán en todos los modelos variables categóricas que indiquen el día de la semana y el mes de cada observación, haciendo uso de la función `step_date`. Igualmente, como también se indicó en el EDA, se modificará la variable `uso_suelo` para unificar todos los niveles que no sean agrícolas o forestales en un solo nivel que se llamará *Otro*. Esto último se hará fuera del *workflow*, antes de realizar la partición del conjunto de datos, haciendo uso de la función `fct_lump` del paquete *forcats*. Los detalles del preprocesamiento que se ha llevado a cabo en cada modelo pueden consultarse en el código, que se adjunta en el Apéndice 2, donde aparecen debidamente comentados. 4º. Crear el *workflow* con el modelo y la receta.

5º. Crear la rejilla (*grid*) con los posibles valores de los parámetros que se deben ajustar.

6º. Entrenar el modelo para cada combinación de los valores de los parámetros a ajustar sobre los datos de entrenamiento.

7º. Evaluar el rendimiento de cada modelo sobre los datos de validación y seleccionar el mejor en base a las medidas de rendimiento ya mencionadas. El objetivo con estos modelos es predecir incendios forestales, por lo que es de vital importancia que los modelos funcionen especialmente bien en la clase positiva, es decir, que si un incendio se va a producir, que el modelo lo detecte. Sin embargo, es fundamental que el modelo tenga un buen desempeño general (un modelo que todo lo clasifique como incendio no serviría de nada, poniendo un ejemplo extremo). Por tanto, cada modelo se valorará de forma individual, considerando todas las métricas de rendimiento mencionadas y priorizando la sensitividad (o *recall*). Sin embargo, en la mayoría de los casos maximizar la tasa de acierto maximiza también la sensitividad, garantizando además un buen desempeño general. Por ello, en la mayoría de modelos se maximizará la tasa de acierto, pero porque analizando las salidas individualmente se ha considerado que es la mejor opción ya que produce la mayor sensitividad sin bajar demasiado las otras medias.

3.2.1. Regresión logística con penalización

Antes de aplicar este modelo, se han transformado las variables categóricas a variables *dummy* y se han tipificado todas las variables (media 0 y varianza 1). Los parámetros a ajustar son λ (parámetro de penalización o *penalty*) y α (parámetro de mixtura o *mixture*).

Se consideran 10 valores equiespaciados para cada parámetro (en el caso de λ entre 10^{-4} y 10^{-1} y el caso de α entre 0 y 1) y se construye el grid tomando todas las combinaciones de estos valores.

Las métricas obtenidas por cada combinación de parámetro sobre los datos de validación se representan en la Figura 3.15

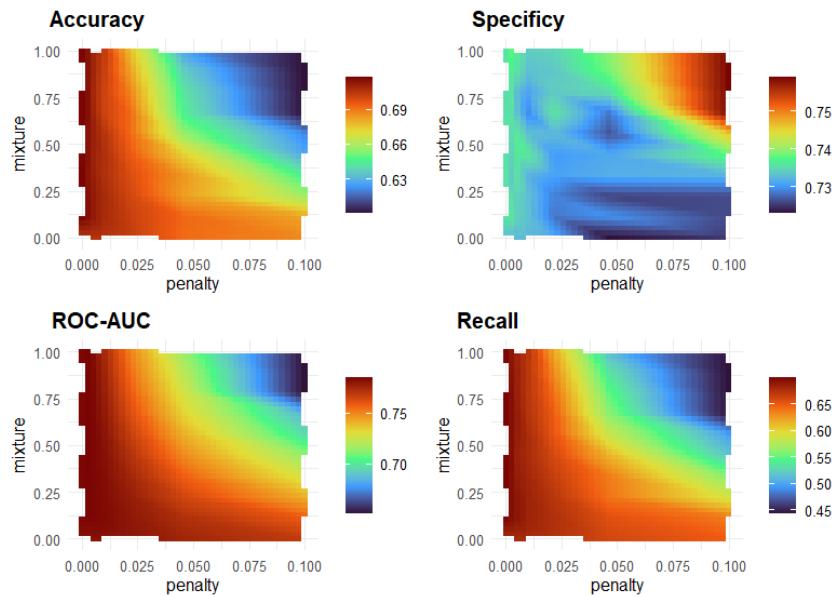


Figura 3.15: Métricas de rendimiento de los modelos de regresión logística con penalización.

Finalmente, se elige el modelo que maximiza la tasa de acierto, cuyos parámetros son: $\alpha = 1$ y $\lambda = 0.000464$. Es decir, un modelo de regresión logística *lasso* puro. Los coeficientes de este modelo se muestran en la Figura B.1.

3.2.2. Regresión logística con penalización + PCA

A continuación se considera el mismo modelo de regresión logística con penalización que en la sección anterior pero en lugar de trabajar con los datos directamente, se aplica análisis de componentes principales sobre los datos normalizados en el preprocesamiento, ajustando el número de componentes principales utilizadas. Para construir el *grid* de parámetros se consideran los mismos valores que en el modelo sin PCA para los parámetros de penalización y mixtura pero ahora se consideran también 7 posibles valores para el número de componentes principales ($\{20, 25, 30, \dots, 50\}$). Finalmente, el modelo que maximiza la tasa de acierto es el que tiene 40 componentes principales, $\alpha = 0.333$ y $\lambda = 0.00464$.

3.2.3. Árboles de decisión

Se construirán los árboles de decisión usando el índice de Gini como función de impureza y se elegirá el parámetro de coste-complejidad (α) que maximice la tasa de acierto. Se considera un *grid* con 10 valores del parámetro de coste-complejidad que oscilan entre

$1.28e - 10$ y $3.02e - 2$. La mejor tasa de acierto en el conjunto de validación se obtiene con $\alpha = 0.00182$. En la Figura 3.16 se muestran las distintas métricas de rendimiento sobre los datos de validación para cada uno de los valores del parámetro a ajustar.

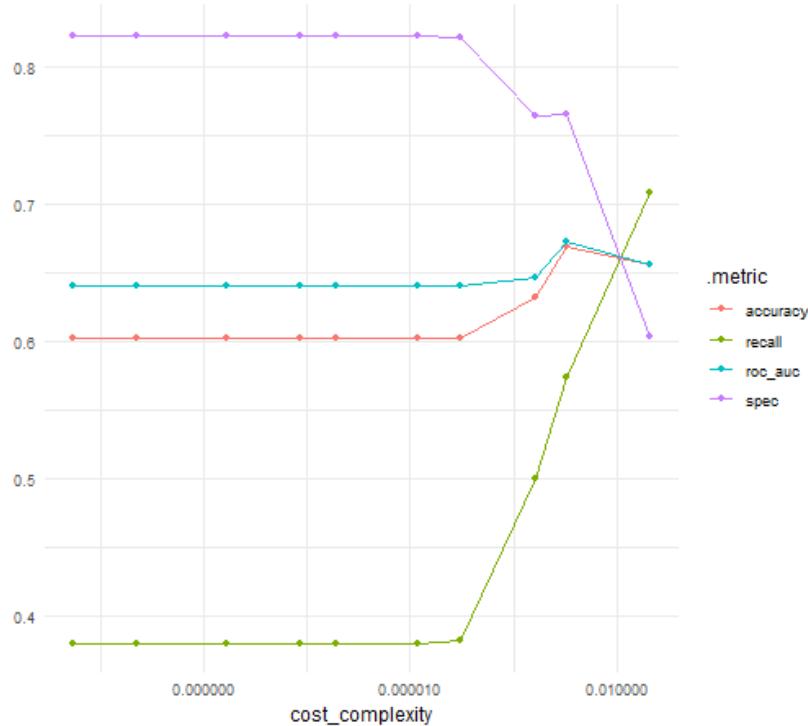


Figura 3.16: Métricas de rendimiento del árbol de decisión en función de el parámetro de costocomplejidad.

3.2.4. Bosques aleatorios

En este modelo se ha fijado el número de árboles a 1000 y se han ajustado los parámetros *mtry* (el número de variables que se seleccionarán aleatoriamente en cada nodo) y *min_n* (el número de observaciones en un nodo a partir del cual no se sigue dividiendo y se convierte en nodo hoja). En este caso se ha optado por un enfoque diferente, motivado por el amplio rango de valores que puede tomar el parámetro *min_n* y por las limitaciones computacionales del equipo disponible.

De esta forma, la estimación de parámetros se ha hecho en dos etapas. En una primera etapa se ha fijado el parámetro *mtry* = 4 y se ha estimado el parámetro *min_n* considerando para ello un *grid* equiespaciado de 1000 a 2500 tomando valores de 100 en 100 ($\{1000, 1100, \dots, 2500\}$). Para elegir entre los distintos modelos esta vez se ha usado como criterio la sensitividad, obteniendo el valor más elevado para *min_n* = 2100. En la segunda etapa, una vez *min_n*, este se ha considerado fijo y se ha estimado *m_try*, considerando una rejilla de 10 valores equiespaciados tomados del 1 al 10 ($\{1, 2, \dots, 10\}$). De nuevo se ha utilizado la sensitividad para elegir el modelo final, eligiendo así *min_n* = 7. En la Figura ?? se recogen los resultados de las dos etapas de *tuning*. El modelo final elegido tiene *min_n* = 2100 y *min_n* = 7.

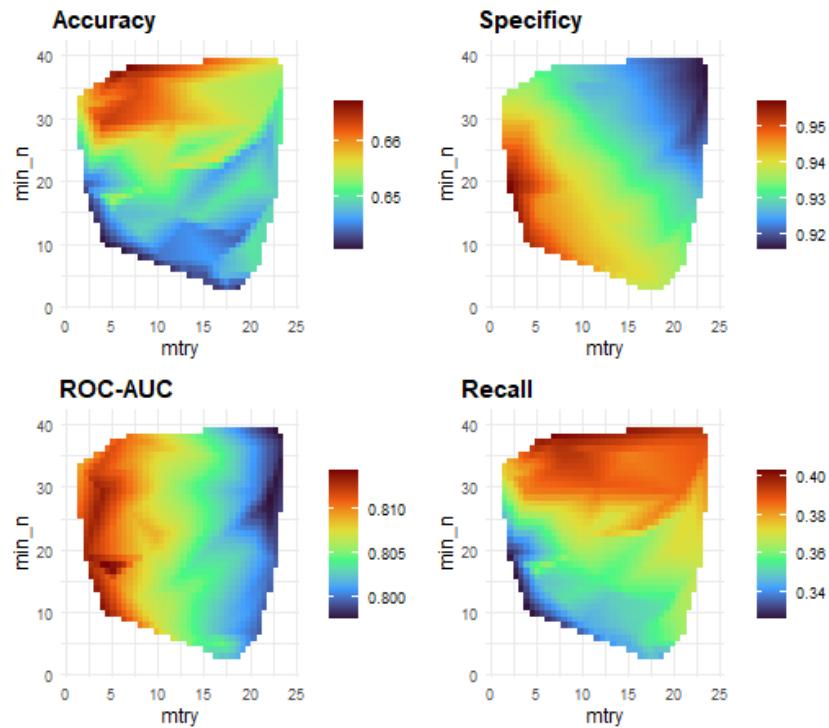


Figura 3.17: Métricas de rendimiento de Random Forest en función de los parámetros.

3.2.5. KNN

Para aplicar el modelo, primero se han transformado las variables categóricas en variables *dummy* y, posteriormente, se han tipificado las todas las variables. Se ha usado la distancia euclídea entre los vectores transformados. Para ajustar el parámetro *k* del modelo se han tomado valores entre 1 y 400. La mayor tasa de acierto sobre los datos de validación se ha obtenido con *k* = 275. Los resultados del *tuning* se muestran en la Figura 3.18.

3.2.6. SVM lineal

Antes de construir el modelo, se han transformado las variables categóricas usando variables *dummy* y se han tipificado todas las variables. Se ha probado con 15 valores del parámetro *coste* entre 0.001949 y 24.666648. La mayor tasa de acierto y el mayor *recall* se han obtenido para *C* = 0.0437. Los resultados del *tuning* se muestran en la Figura 3.19

3.2.7. SVM radial

Por último, se ha construido el modelo de SVM usando un kernel gaussiano. El preprocesamiento ha sido el mismo que en el caso del kernel lineal. Dado el elevado tiempo de entrenamiento de este modelo solo se ha probado con 8 combinaciones de valores para los parámetros *C* y γ , que oscilan entre 0.005 y 31.7 y entre 0 y 0.05. La mayor tasa de acierto sobre los datos de validación se ha conseguido para *C* = 31.7 y γ = 0.0000496.

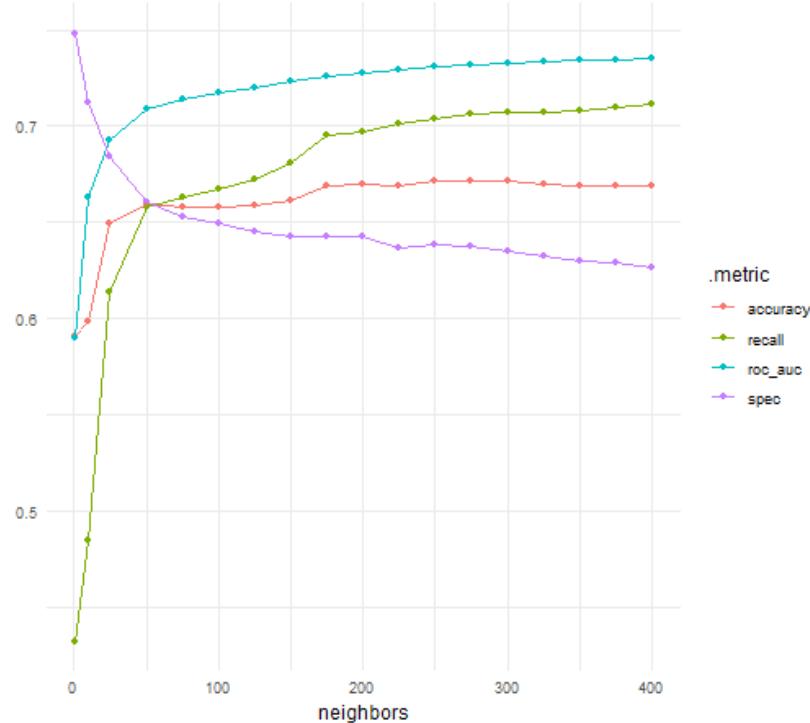


Figura 3.18: Métricas de rendimiento de KNN en función del número de vecinos.

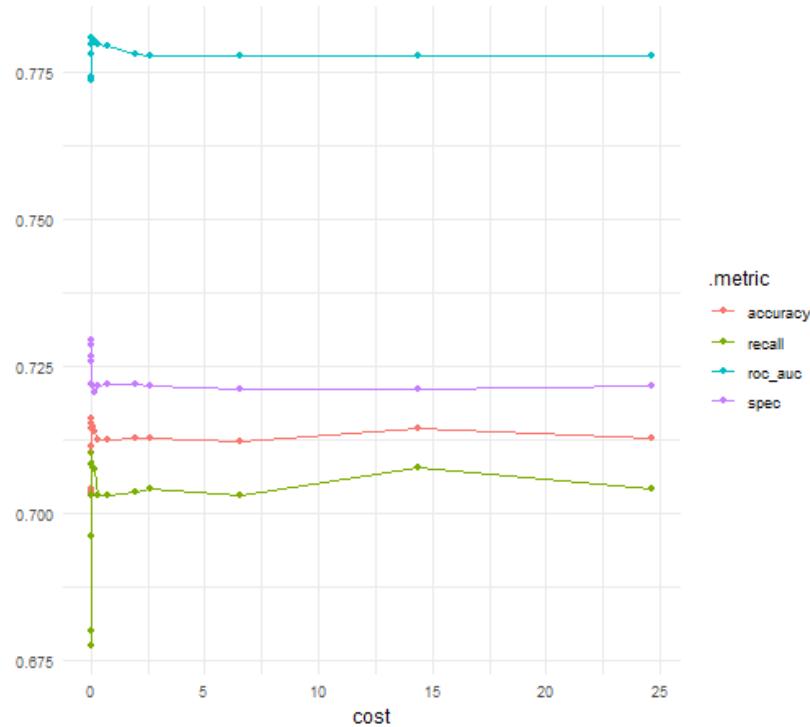


Figura 3.19: Métricas de rendimiento de svm en función del coste.

3.3. Comparación

A continuación se muestran las métricas de cada uno de los modelos seleccionados en los datos de validación en la Tabla 3.1 y en la Figura 3.20. Las curvas ROC de todos los

model_name	roc_auc	accuracy	recall	specificity	precision
lr	0.785	0.719	0.700	0.738	0.727
lr_pca	0.727	0.659	0.624	0.694	0.670
dt	0.656	0.656	0.709	0.604	0.641
rf	0.781	0.725	0.758	0.693	0.711
svm_linear	0.781	0.716	0.710	0.722	0.718
svm_rbf	0.777	0.710	0.692	0.728	0.717
knn	0.732	0.672	0.706	0.637	0.660

Tabla 3.1: Métricas de los modelos seleccionados sobre el conjunto de validación.

modelos se muestran en la Figura 3.21.

Puede observarse que los resultados obtenidos por todos los modelos son bastante similares. Destacan el modelo de bosque aleatorio y el de regresión logística con penalización, el primero por ser el que tiene la tasa de acierto y la sensitividad más elevadas, y el segundo por dar los mejores resultado en cuanto a precisión y especificidad. Las curvas ROC de los modelos de bosque aleatorio, regresión logística con penalización, SVM lineal y SVM radial son prácticamente iguales. Los modelos más pobres son la regresión logística aplicando PCA, KNN y el árbol de decisión.

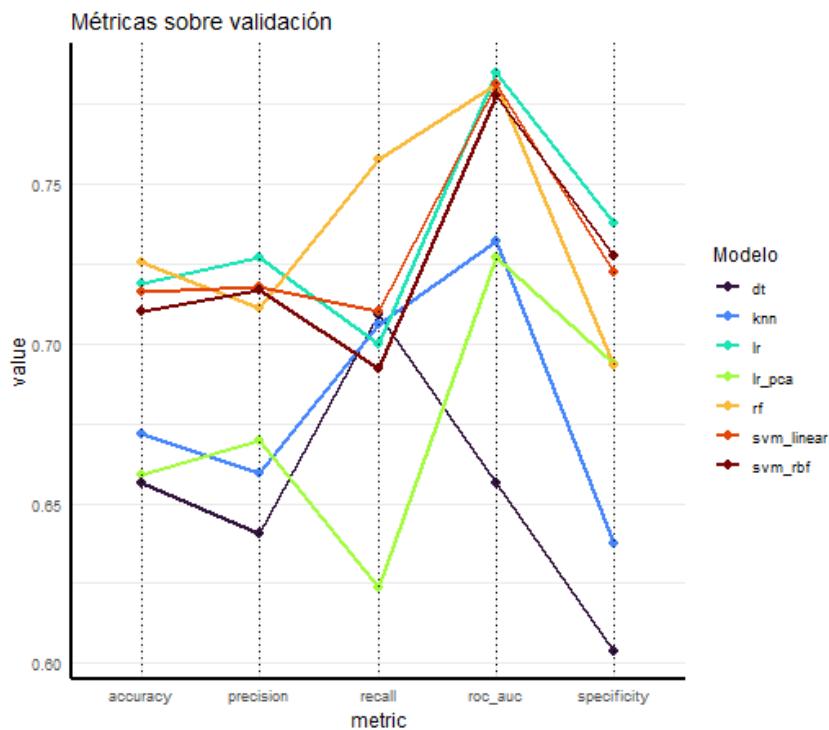


Figura 3.20: Métricas obtenidas sobre el conjunto de validación por cada uno de los modelos seleccionados.

Por último, para conocer la capacidad de generalización de los modelos construidos, estos se evaluarán sobre nuevas observaciones, el conjunto de datos test. Recuérdese que el entrenamiento de los modelos se ha realizado con el conjunto de entrenamiento, formado por 12927 observaciones tomadas entre el 2002 y mediados de 2014 y para ajustar los parámetros de cada modelo, se han utilizado 4309 observaciones tomadas entre mediados de 2014 y mediados de 2019. Finalmente, se evaluará la capacidad de predicción de los modelos sobre 4310 nuevas observaciones tomadas entre mediados de 2019 y 2022. Para

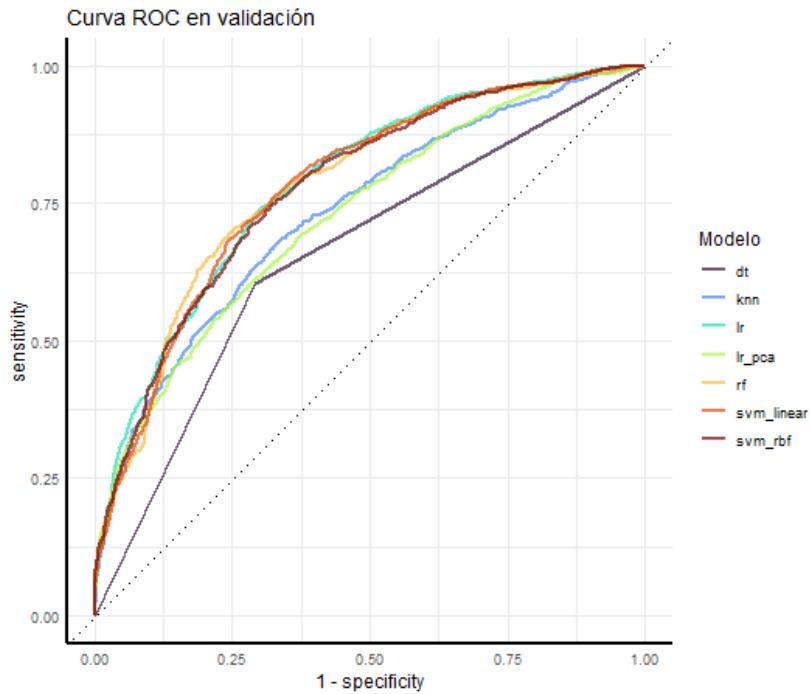


Figura 3.21: Curvas ROC sobre el conjunto de validación.

model_name	roc_auc	accuracy	recall	specificity	precision
lr	0.795	0.710	0.726	0.693	0.718
lr_pca	0.715	0.645	0.631	0.661	0.667
dt	0.667	0.668	0.712	0.621	0.670
rf	0.762	0.699	0.727	0.669	0.703
svm_linear	0.790	0.705	0.729	0.680	0.710
svm_rbf	0.789	0.706	0.727	0.683	0.712
knn	0.744	0.673	0.719	0.624	0.673

Tabla 3.2: Métricas sobre el conjunto test.

ello, primero se juntarán los conjuntos de entrenamiento y validación para reentrenar los modelos con la configuración de parámetros seleccionada en cada caso, y posteriormente se compararán los valores predichos por los modelos con los valores reales. Los resultados obtenidos se muestran en la Tabla 3.2 y en la Figura 3.22. Las curvas ROC de los distintos modelos sobre el conjunto de datos test se muestran en la Figura 3.23.

En este caso, los mejores resultados en todas las medidas los da el modelo de regresión logística con penalización. Los modelos de SVM muestran resultados bastante similares entre ellos y prácticamente iguales al modelo de regresión logística. Sobre los datos test, el modelo de bosque aleatorio ha dado un rendimiento peor que el obtenido en validación, quedando por detrás de los tres modelos ya comentados, aunque la sensibilidad de todos estos modelos es prácticamente igual. De nuevo, los peores resultados los dan los modelos de regresión logística aplicando PCA y el árbol de decisión, seguidos del KNN.

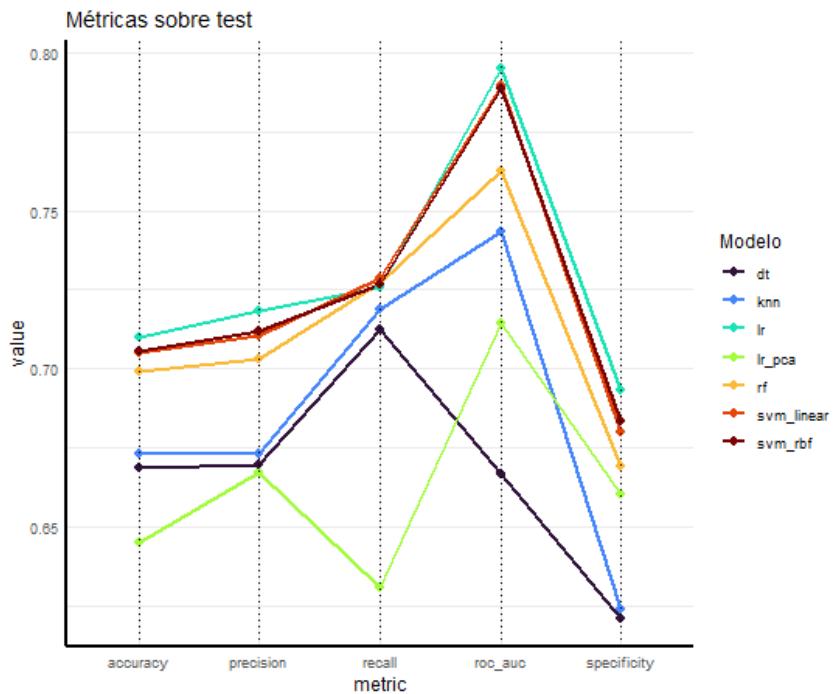


Figura 3.22: Métricas obtenidas sobre el conjunto test por cada uno de los modelos seleccionados.

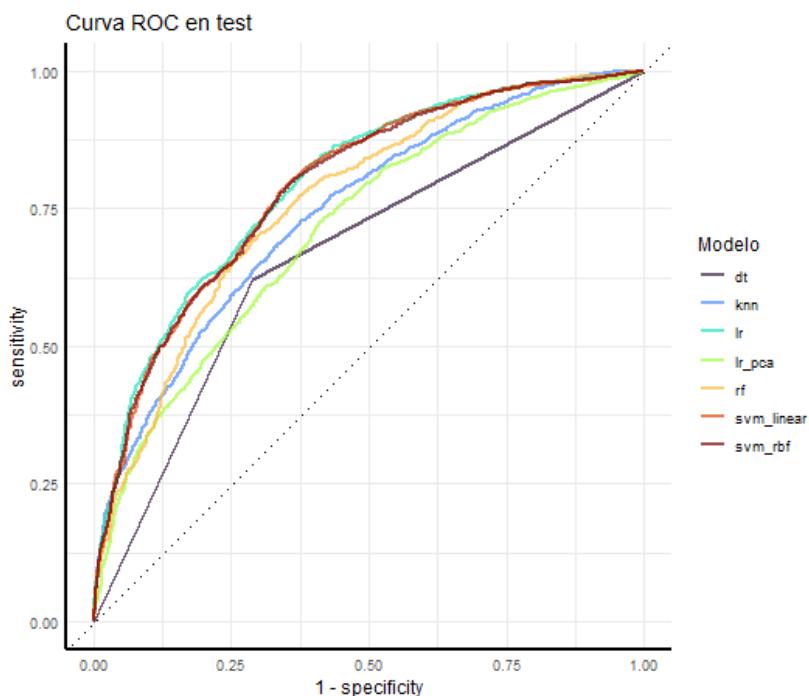


Figura 3.23: Curvas ROC sobre test.

Capítulo 4

Aplicación de los modelos

En esta sección se pretende ilustrar el funcionamiento de los modelos construidos en la sección anterior, valorar su desempeño al ser aplicados en la realidad y conocer sus limitaciones.

4.1. Visión general del desempeño de los modelos

Dado que los modelos se han construido y evaluado sobre un conjunto de datos que es inevitablemente limitado y que se ve fuertemente influenciado por decisiones metodológicas, surge la incógnita de si las métricas de rendimiento obtenidas están realmente reflejando la capacidad de generalización de los modelos al ser aplicados en la realidad o de si, en cambio, tan solo están reflejando sesgos introducidos en el conjunto de datos a través del proceso de selección. Para entender mejor el funcionamiento de los modelos y conocer si realmente esto sirven para predecir incendios forestales en la vida real, se ha decidido adoptar el siguiente enfoque.

En primer lugar, se ha construido una malla de puntos con una resolución de 10km por 10km cubriendo toda la extensión de Andalucía (en este caso, se entiende como resolución la distancia entre los puntos en la dirección Este-Oeste y Norte-Sur) (Figura 4.1).

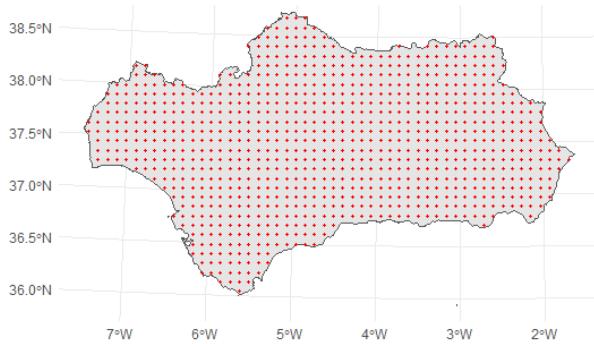


Figura 4.1: Malla de puntos con una resolución de 10km por 10km.

A continuación, se ha asociado a cada uno de los puntos de la malla el valor de todas las variables predictoras el día 15 de cada mes del año 2022 en esa localización, usando los métodos de preprocesamiento y depuración ya descritos. Estos datos se han utilizado

para predecir el riesgo de incendio forestal en cada uno de los puntos de la malla el día 15 de cada mes, utilizando para ello los modelos finales de la sección anterior que mejor rendimiento mostraron sobre los datos test. Los resultados se muestran en las Figuras 4.2 (Regresión logística con penalización) y 4.3 (SVM lineal) y 4.4 (Random Forest).

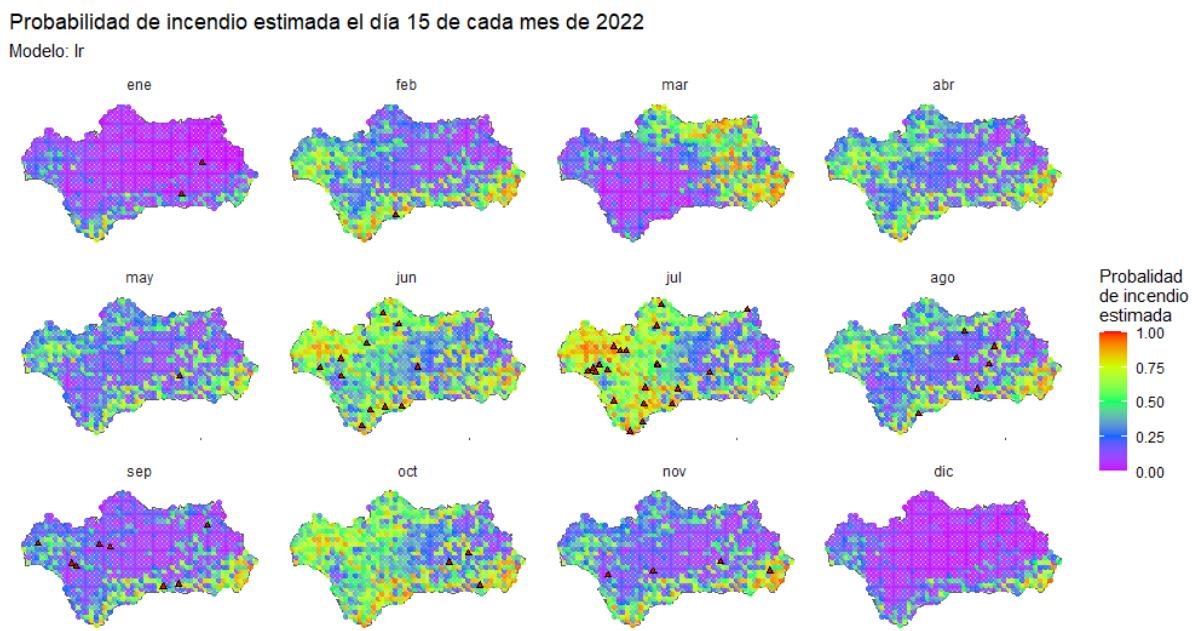


Figura 4.2: Probabilidades de incendios estimadas el día 15 de cada mes de 2022 con el modelo de regresión logística con penalización. Los triángulos indican los incendios de más de 100ha registrados en ese mes.

Se puede observar que las predicciones del modelo de regresión logística con penalización y del SVM lineal son muy parecidos. A diferencia de las predicciones del modelo de Bosque aleatorio, que son bastante similares todos los meses, estos dos modelos muestran bastante variación mensual. Seguramente esta sea la causa de que, si bien el modelo de bosque aleatorio mostraba un buen rendimiento en validación, al evaluar su rendimiento sobre los datos test, este bajó significativamente. Es por ello que, a falta de la opinión de un experto en ecología del fuego, se opta por descartar el modelo de bosque aleatorio ya que no parece reflejar correctamente la variación estacional que se observa en la aparición de incendios forestales.

Se analizan por tanto las predicciones de los otros dos modelos (regresión logística y SVM). Se puede observar una clara componente estacional en las observaciones. En los meses de diciembre y enero se observan los niveles de riesgo más bajos a nivel global, mientras que los niveles de riesgo más elevados se encuentran en los meses de junio y julio, aunque por algún motivo también se observan niveles de riesgo elevado en los meses de marzo y octubre. Se puede observar también como las zonas con una probabilidad alta de incendio forestal varían en función del mes. Es curioso que en marzo ambos modelos den probabilidades de incendio tan elevadas en la zona oriental de la comunidad. Al margen del estudio específico y detallado de los mapas presentados, lo cual correspondería a los expertos en la materia y escapa de los objetivos de este trabajo, se puede observar que prácticamente todos los incendios se producen en zonas con una probabilidad de incendio elevada y que los modelos son capaces de ir más allá del mero estudio de las variables meteorológicas, ya que se observan zonas más o menos aisladas con una mayor

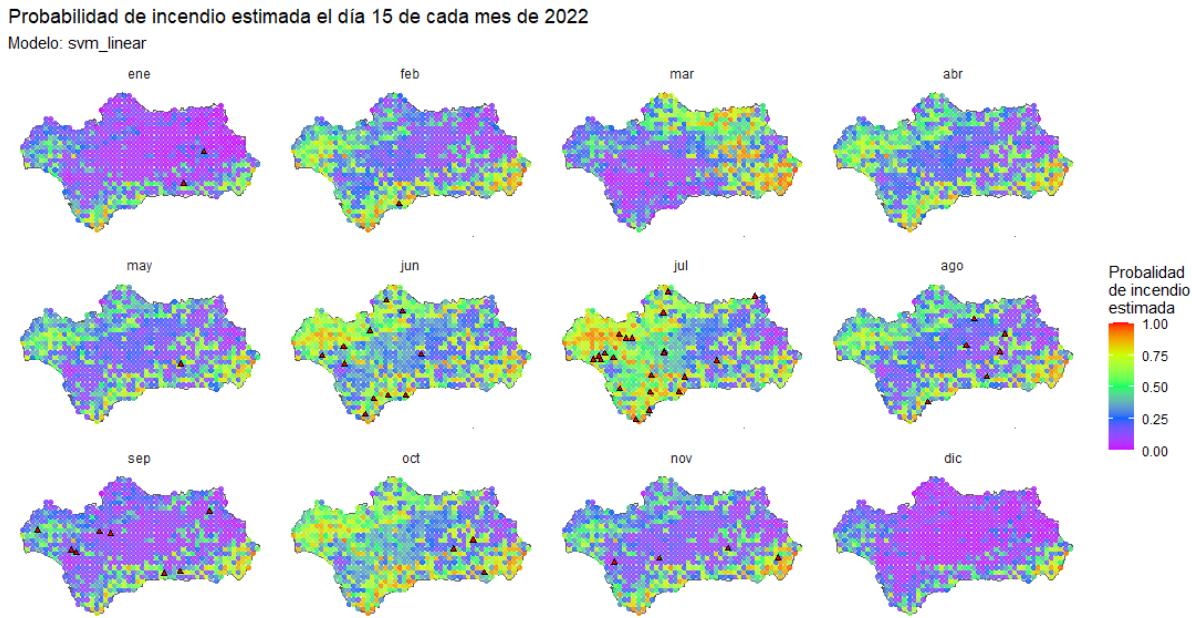


Figura 4.3: Probabilidades de incendios estimadas el día 15 de cada mes de 2022 con el modelo de SVM lineal. Los triángulos indican los incendios de más de 100ha registrados en ese mes.

probabilidad de incendio que se corresponden con las zonas en las que se ha observado un incendio. Esto indica que son otros factores los que el modelo está considerando para indicar riesgo de incendio, ya que la resolución espacial de las variables meteorológicas es bastante baja (50km), por lo que las variaciones a un mayor detalle son debidas a otros factores.

Este es, sin embargo, un enfoque bastante pobre, pues solo se está considerando el día 15 de cada mes, lo que podría llevar a conclusiones erróneas a la hora de evaluar los modelos (debidas, por ejemplo, a valores atípicos en ese día concreto). Esto es debido a las limitaciones computacionales del equipo disponible. Pese a ello, se ha podido ilustrar, aunque sin entrar en detalle, el desempeño de los modelos al aplicarlos para evaluar el riesgo de incendio en la realidad.

4.2. Caso estudio

A continuación, se pondrá a prueba el modelo de regresión logística construido con un caso real, el incendio de Sierra Bermeja, que se originó el 8 de septiembre de 2021 en el municipio de Jubrique en la provincia de Málaga (Figura 4.5). Se ha elegido este incendio por dos motivos. En primer lugar, porque fue el mayor incendio que hubo en España en el año 2021, con una superficie total afectada de 8607ha y una duración de 46 días hasta su extinción. Y en segundo lugar, porque fue un incendio intencionado, por lo que permitirá reflejar el comportamiento del modelo en incendios causados por el hombre.

Para analizar la capacidad de predicción del modelo para este incendio, se ha llevado a cabo el siguiente enfoque. Primero, se ha construido una malla de puntos con una resolución de 1km por 1km, cubriendo todo el *bounding box* de un *buffer* de 10km alrededor

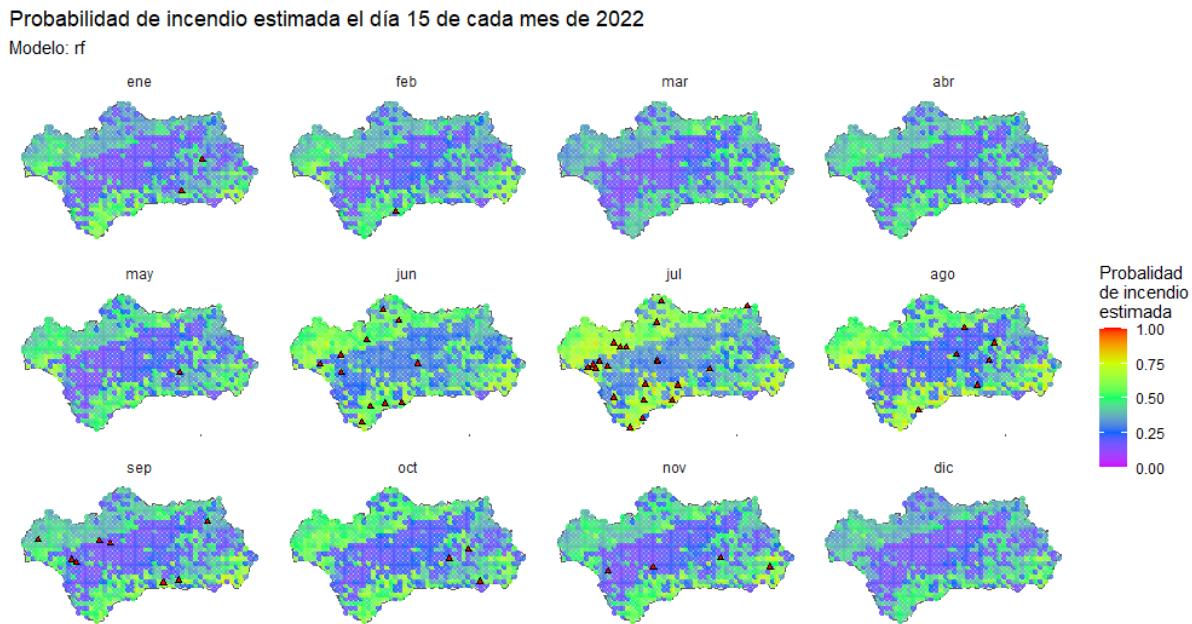


Figura 4.4: Probabilidades de incendios estimadas el día 15 de cada mes de 2022 con el modelo de random forest. Los triángulos indican los incendios de más de 100ha registrados en ese mes.

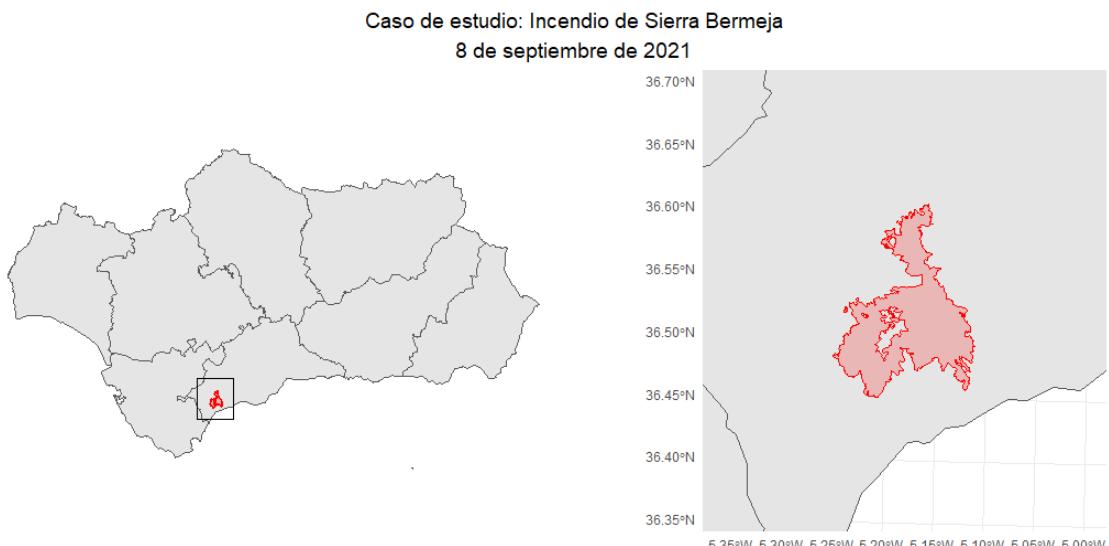


Figura 4.5: Área recorrida por el fuego en el incendio de Sierra Bermeja.

del perímetro del incendio. A continuación, en cada uno de estos puntos se han tomado todas las variables predictoras el día de origen del incendio, 15 y 30 días antes y 15, 30 y 45 días después. Con estos datos se ha utilizado el modelo de regresión logística con penalización para predecir la probabilidad de incendio forestal en cada uno de los días considerados en toda la malla de puntos. Los resultados se muestran en la Figura 4.6.

De este gráfico pueden extraerse varias conclusiones. Por un lado puede observarse que el día del origen del incendio se produce un aumento drástico de las probabilidades estimadas de incendio, las cuales continúan siendo muy altas 15 días después y, aunque disminuyen de forma general, se mantienen elevadas hasta 45 días después del inicio del fuego. Sin

Incendio de Sierra Bermeja

Model: Ir

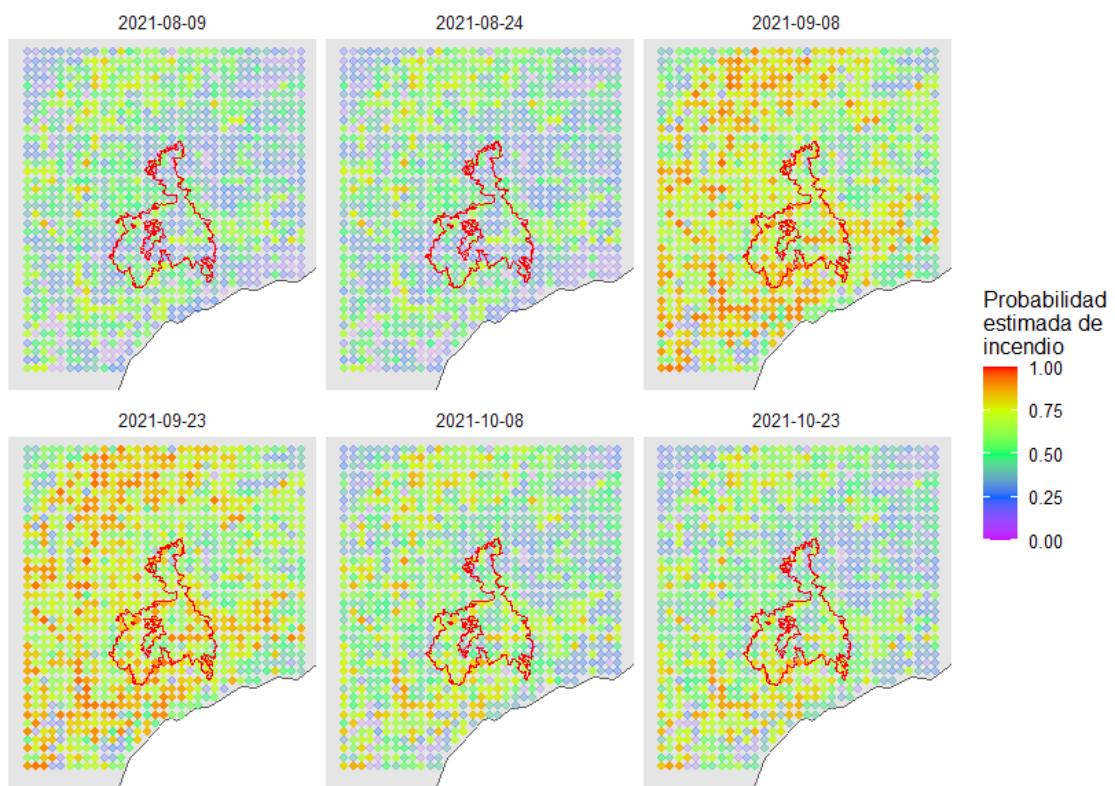


Figura 4.6: Mapa con las probabilidades de incendio estimadas en los días en torno al origen del incendio de Sierra Bermeja. El área total recorrida por el fuego se muestra en rojo.

embargo, si bien es cierto que a nivel global el modelo sí parece aportar información estimando un riesgo muy alto de incendio en la región, al aumentar el nivel de detalle puede observarse que la capacidad discriminatoria del modelo disminuye significativamente. Esto es coherente, ya que la resolución de las variables climáticas es de aproximadamente 50km por 50km, por lo que no son adecuadas para trabajar con un nivel de detalle tan reducido.

Cabe mencionar que la variación observada en el riesgo de incendio estimado en las distintas fechas es debida, únicamente, a los cambios en las variables meteorológicas y en el NDVI. Esto es debido a que en el modelo de regresión logística construido no se han considerado las posibles interacciones entre las variables, lo cual podría ser de gran interés dadas las características del problema.

Capítulo 5

Conclusión

Apéndice A

Apéndice: Gráficos espaciales EDA

A.1. Variables meteorológicas

Distribución espacial de T2M por mes

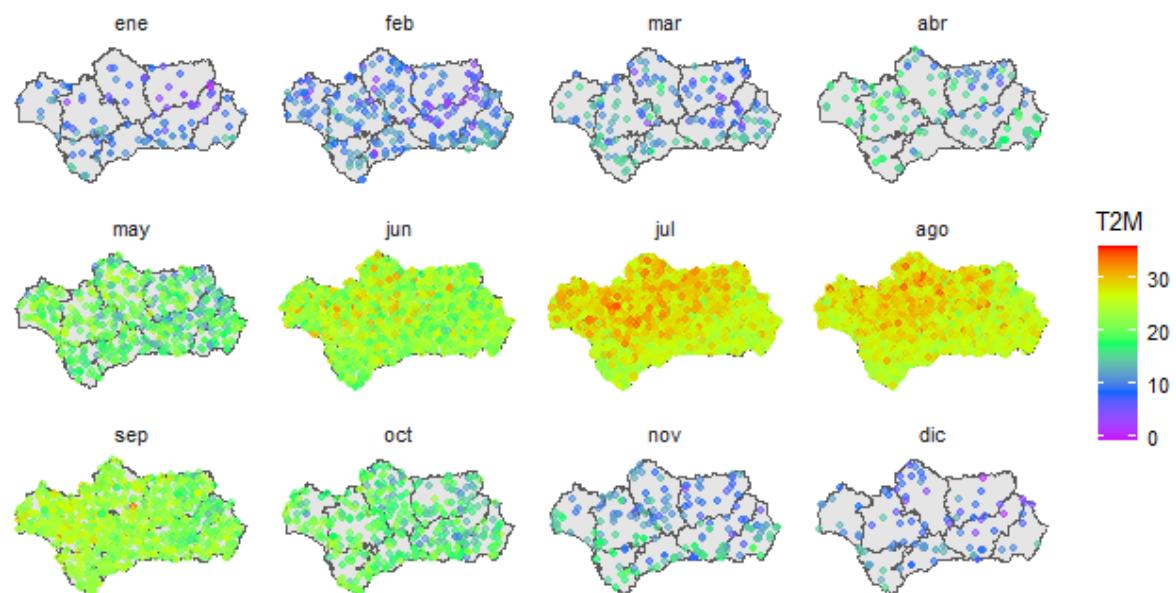


Figura A.1: Distribución espacial de T2M por mes.

Distribución espacial de RH2M por mes

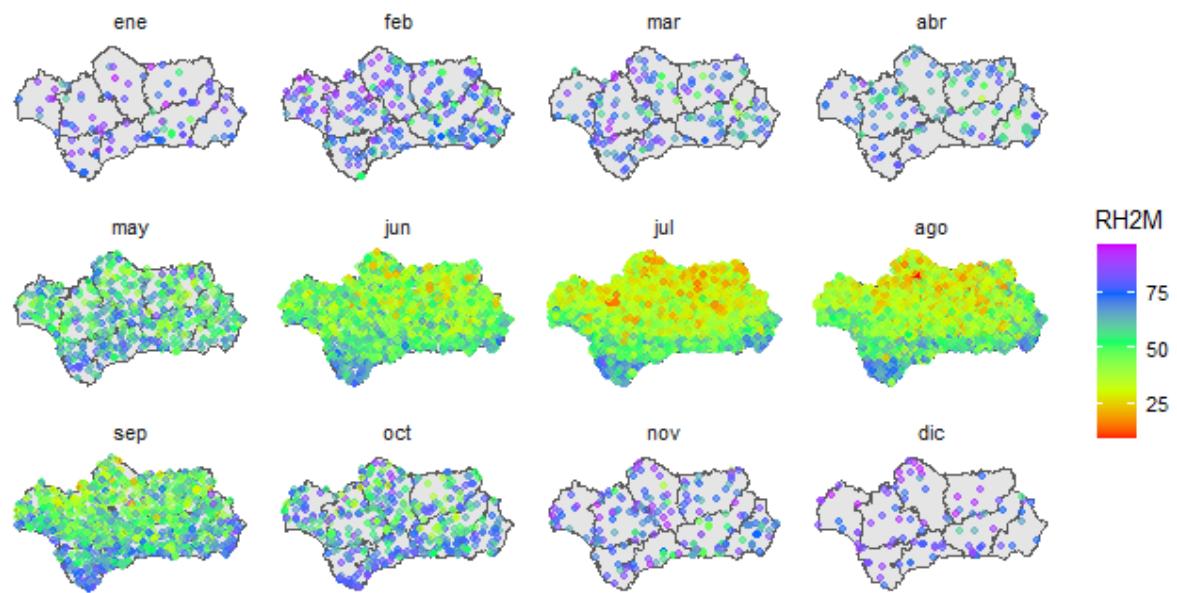


Figura A.2: Distribución espacial de RH2M por mes.

Distribución espacial de GWETTOP por mes

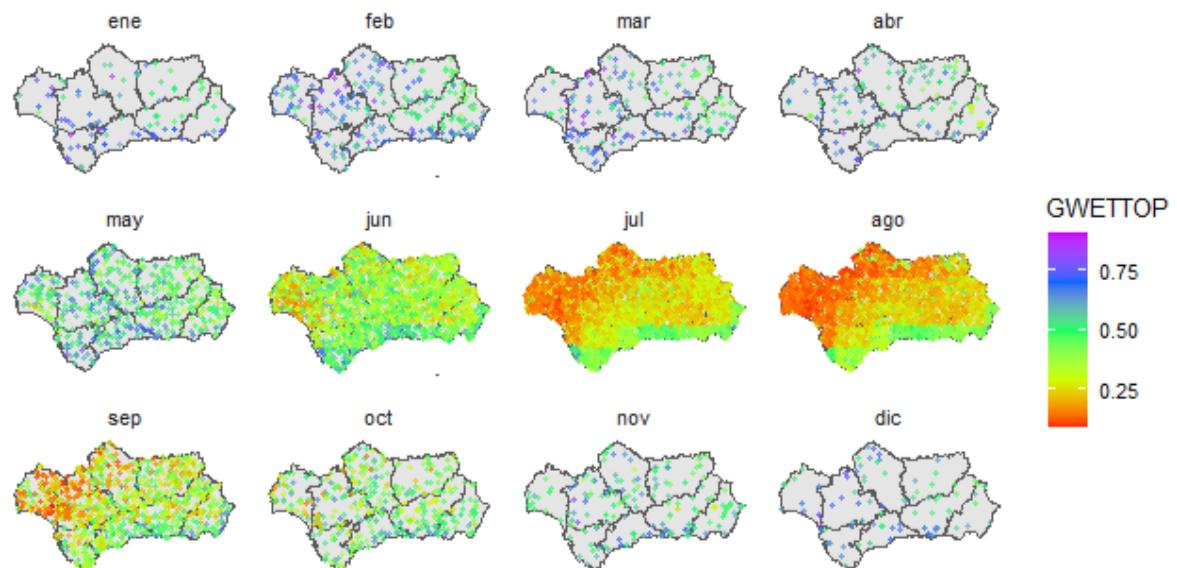


Figura A.3: Distribución espacial de GWETTOP por mes.

Distribución espacial de WS10M por mes

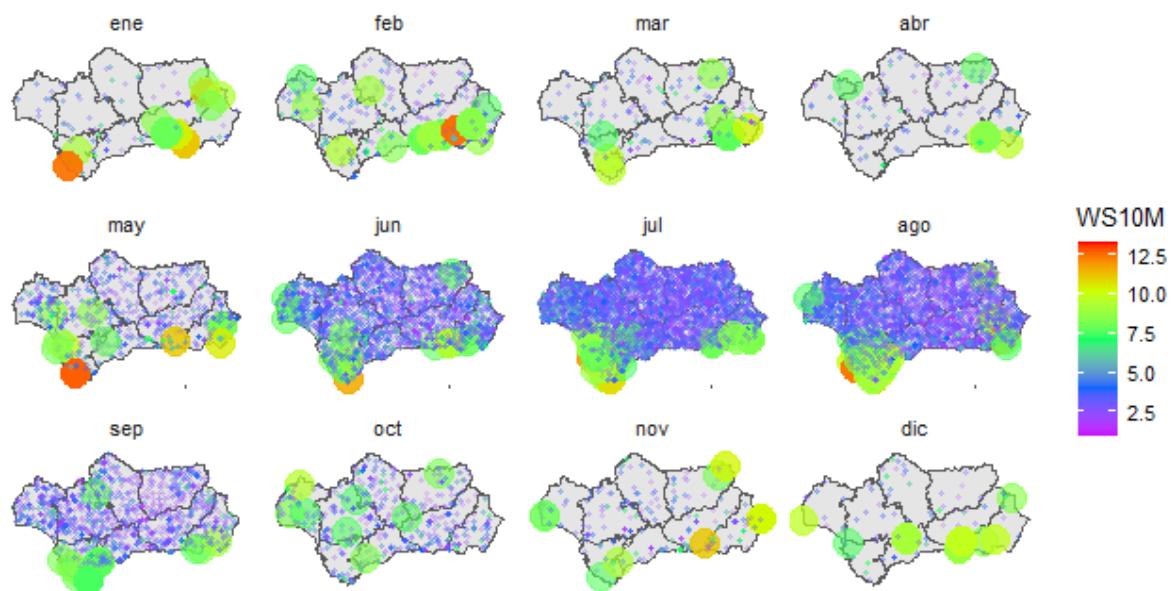


Figura A.4: Distribución espacial de WS10M por mes.

Distribución espacial de PRECTOTCORR por mes

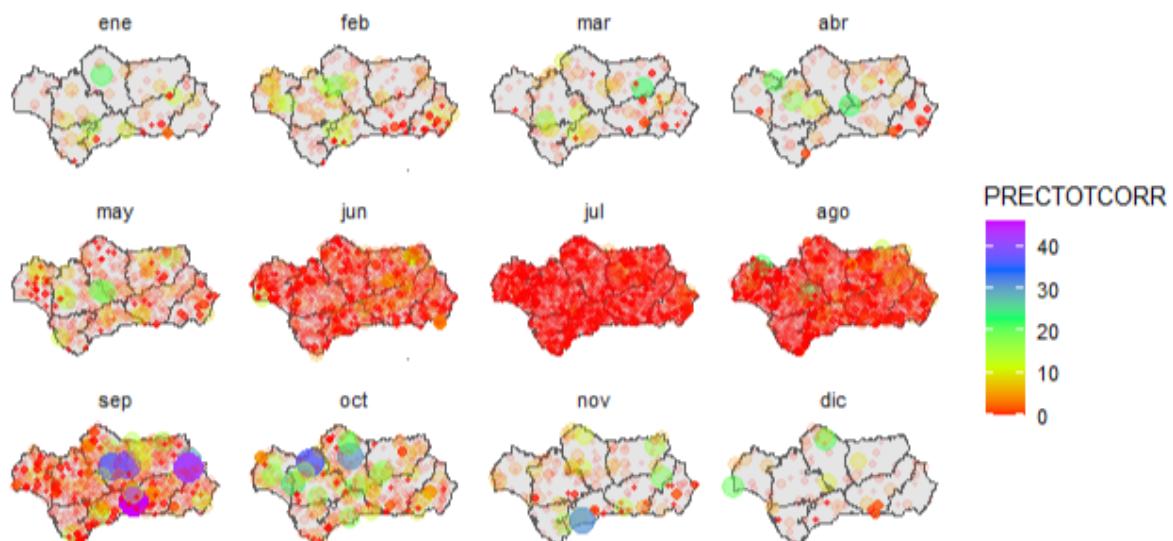


Figura A.5: Distribución espacial de PRECTOTCORR por mes.

Distribución espacial de WD10M por mes

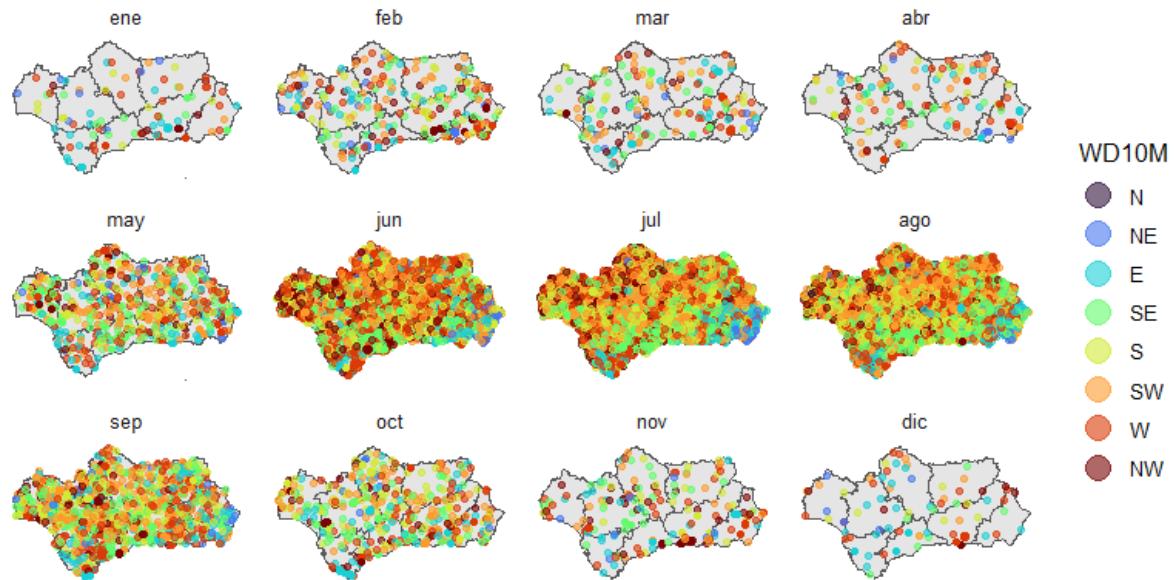


Figura A.6: Distribución espacial de WD10M por mes.

A.2. Variables demográficas

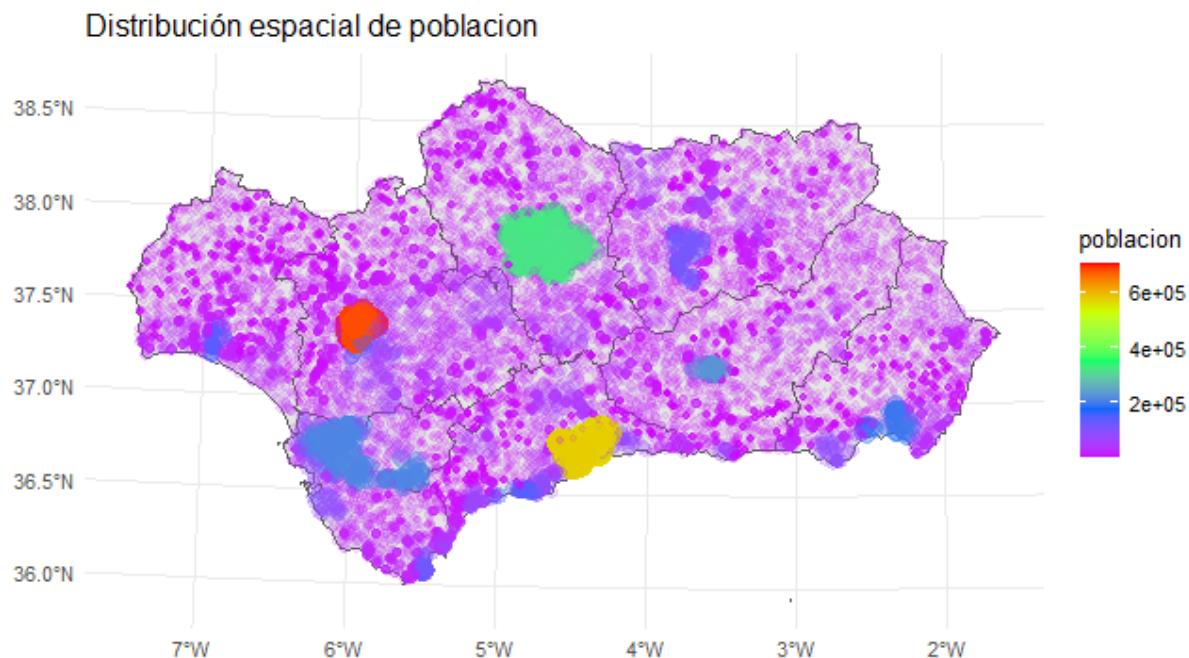


Figura A.7: Distribución espacial de población.

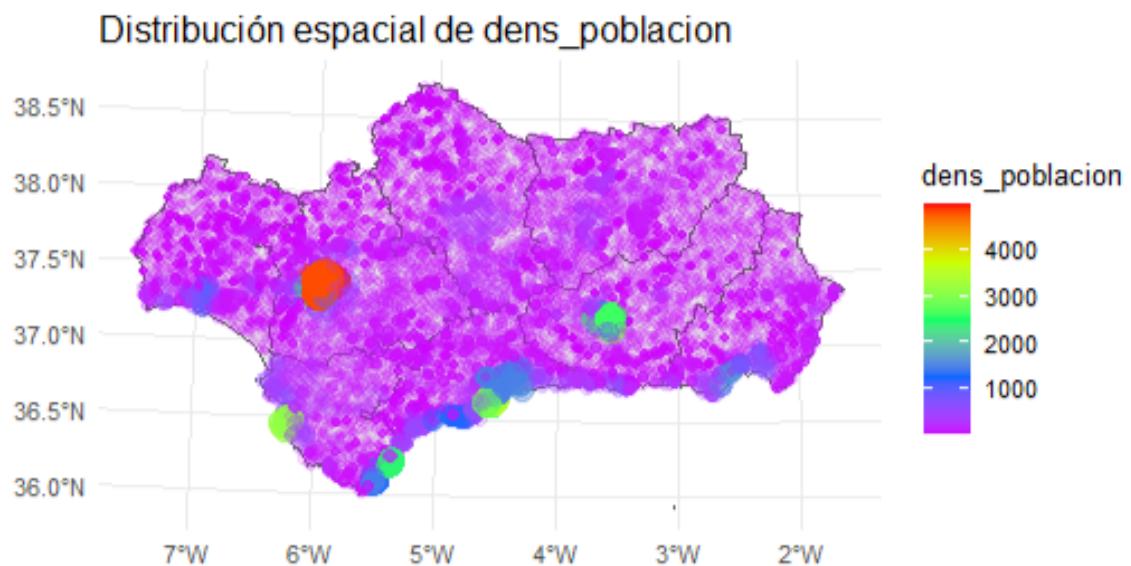


Figura A.8: Distribución espacial de dens_poblacion.

A.3. Variable de vegetación

Distribución espacial de NDVI por mes

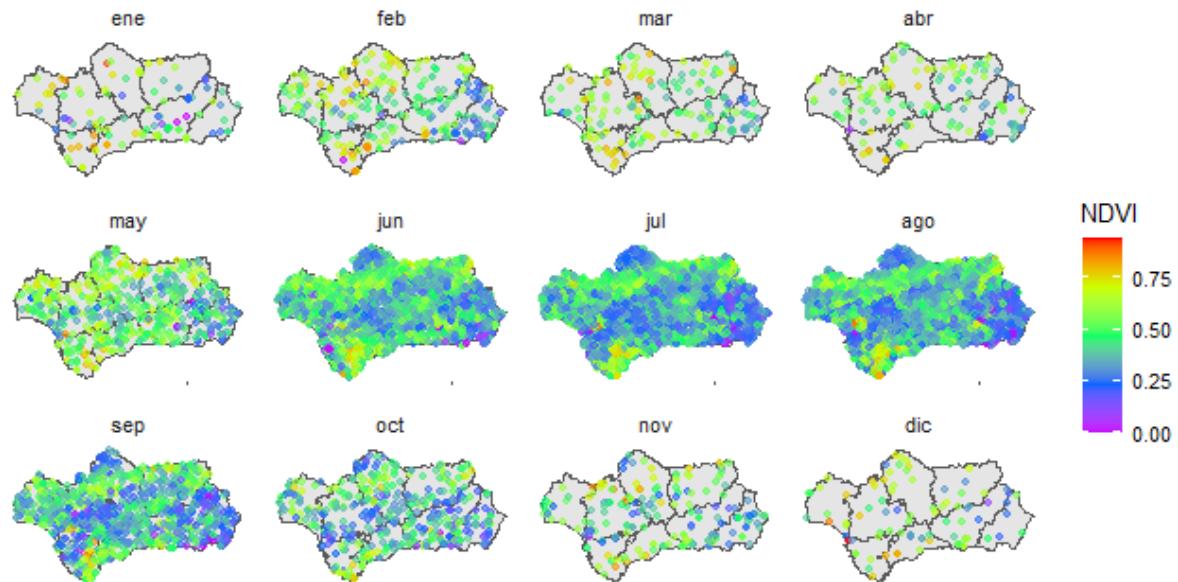


Figura A.9: Distribución espacial de NDVI por mes.

A.4. Variables topográficas

Distribución espacial de elevacion

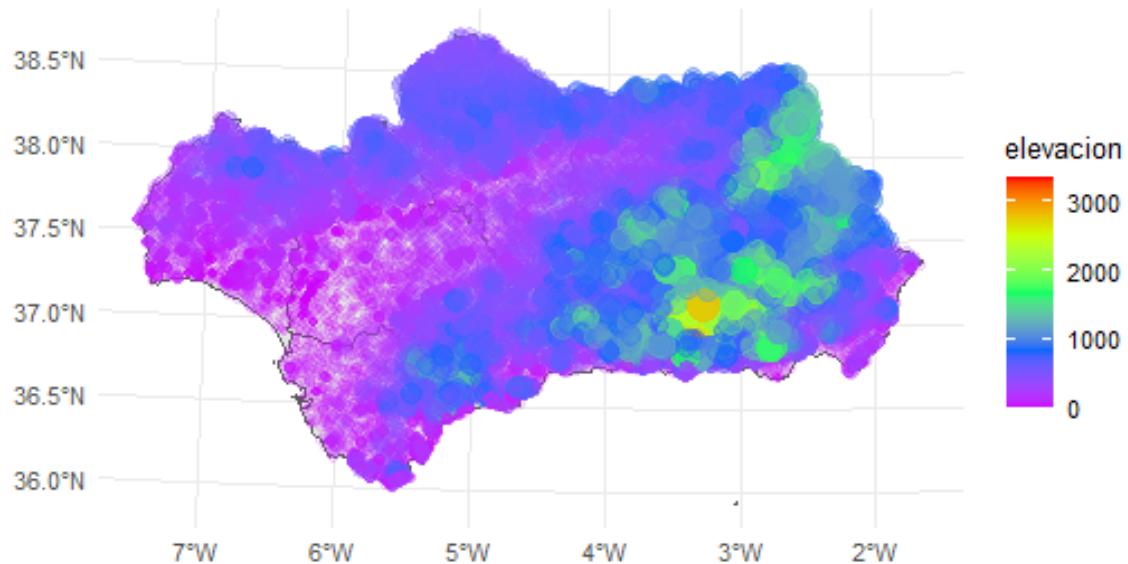


Figura A.10: Distribución espacial de elevacion.

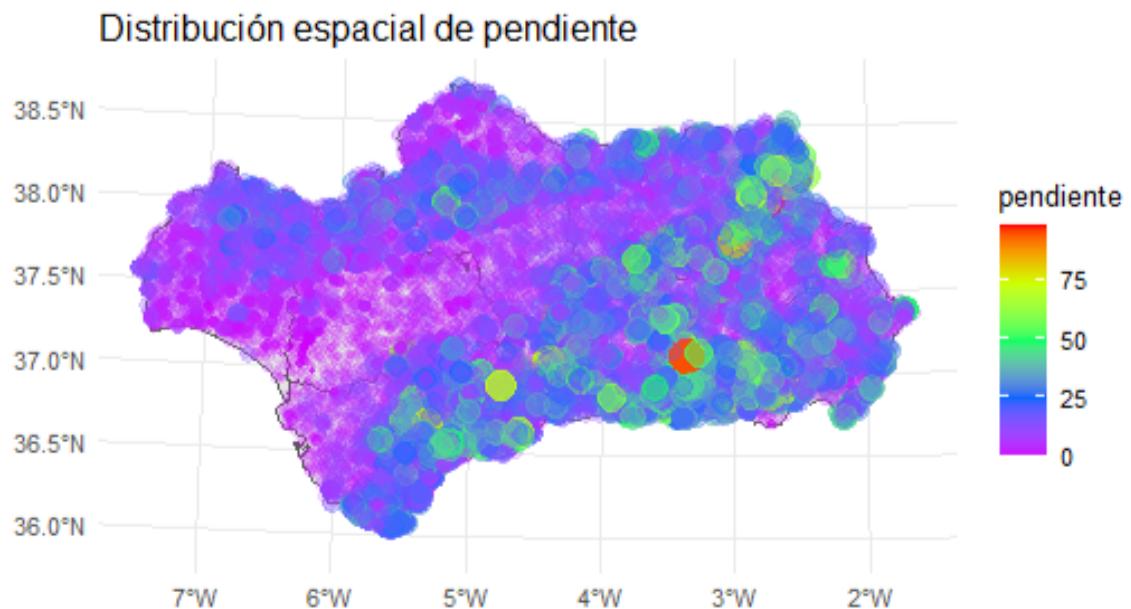


Figura A.11: Distribución espacial de pendiente.

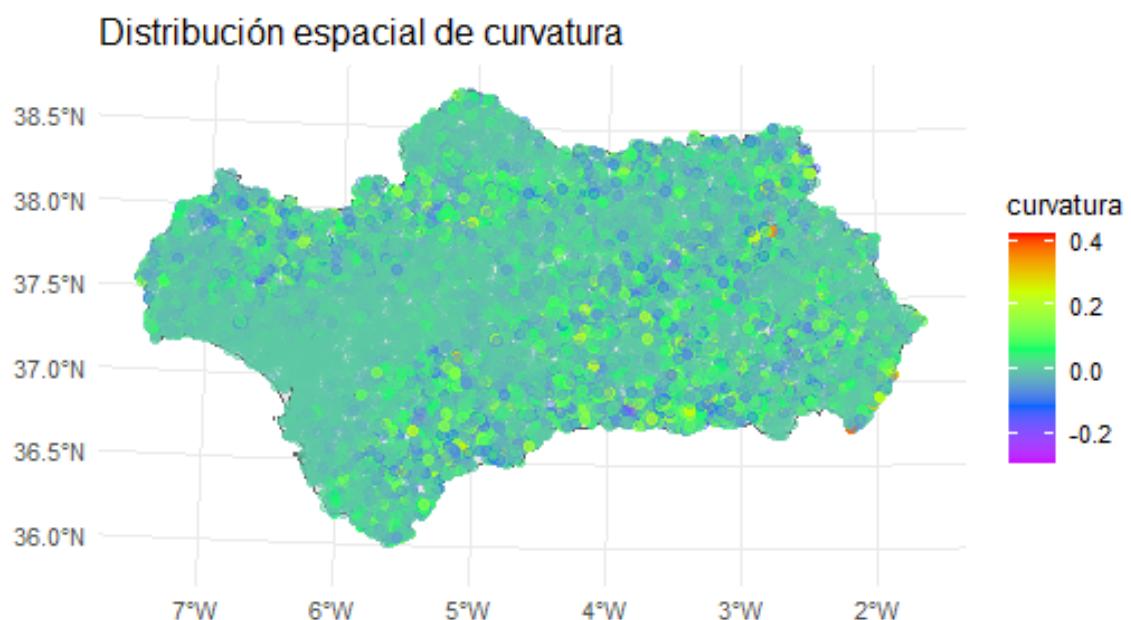


Figura A.12: Distribución espacial de curvatura.

A.5. Variables antropológicas

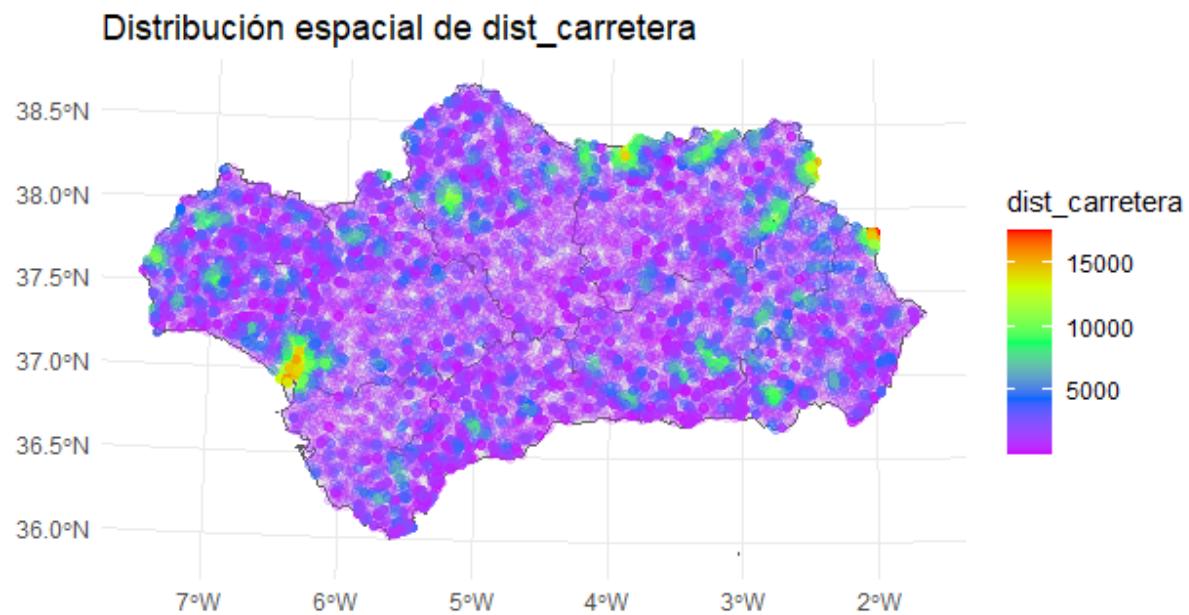


Figura A.13: Distribución espacial de dist_carretera.

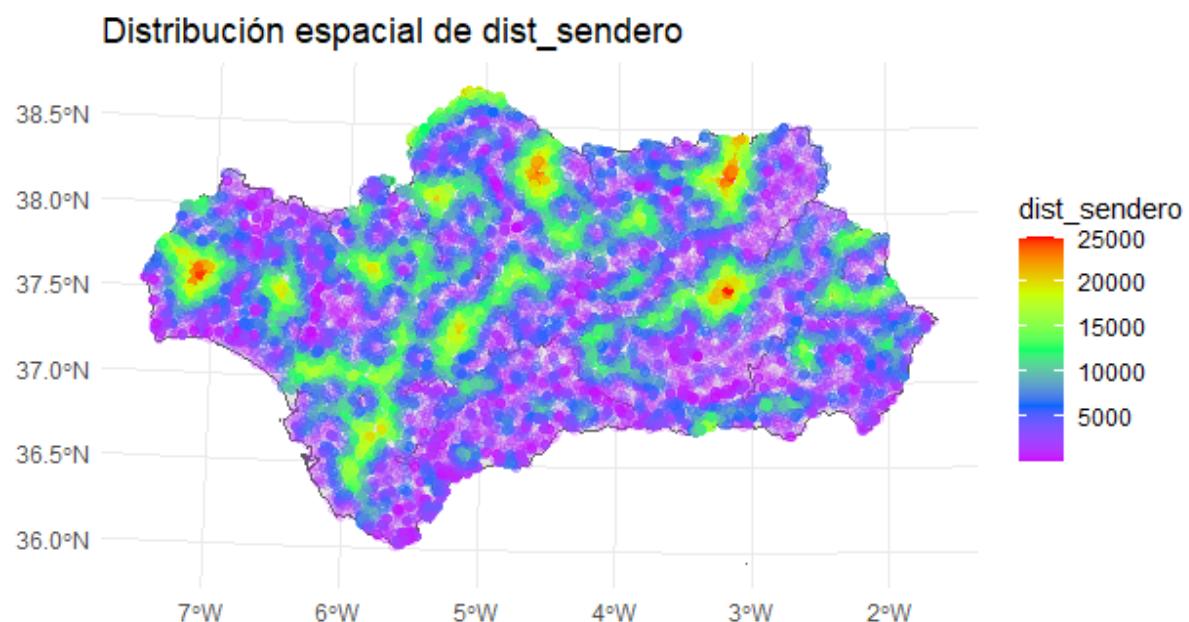


Figura A.14: Distribución espacial de dist_sendero.

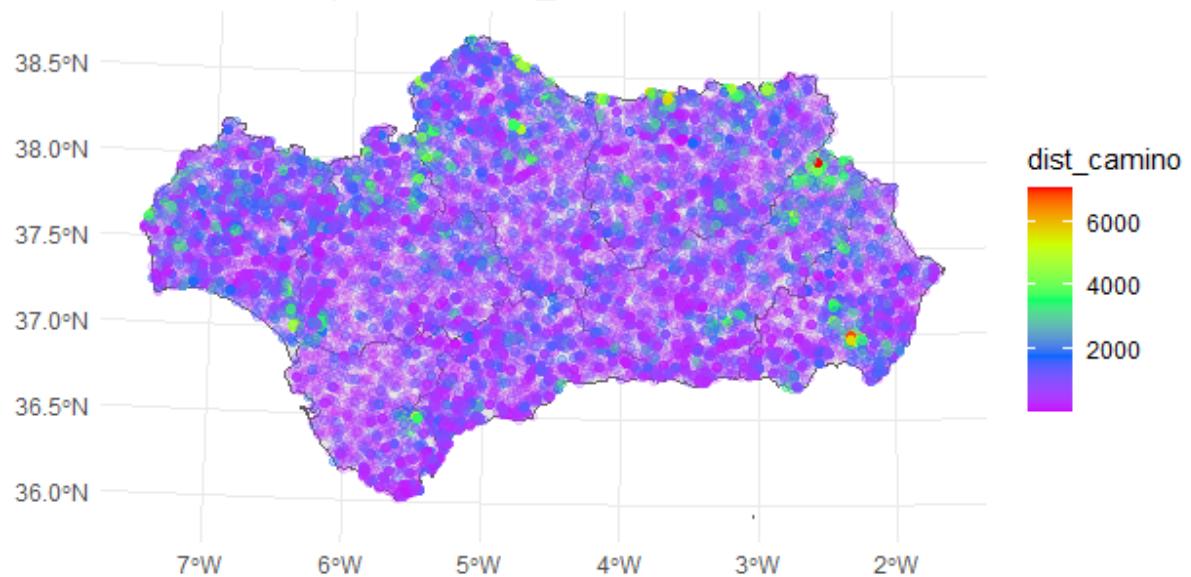
Distribución espacial de dist_camino

Figura A.15: Distribución espacial de dist_camino.

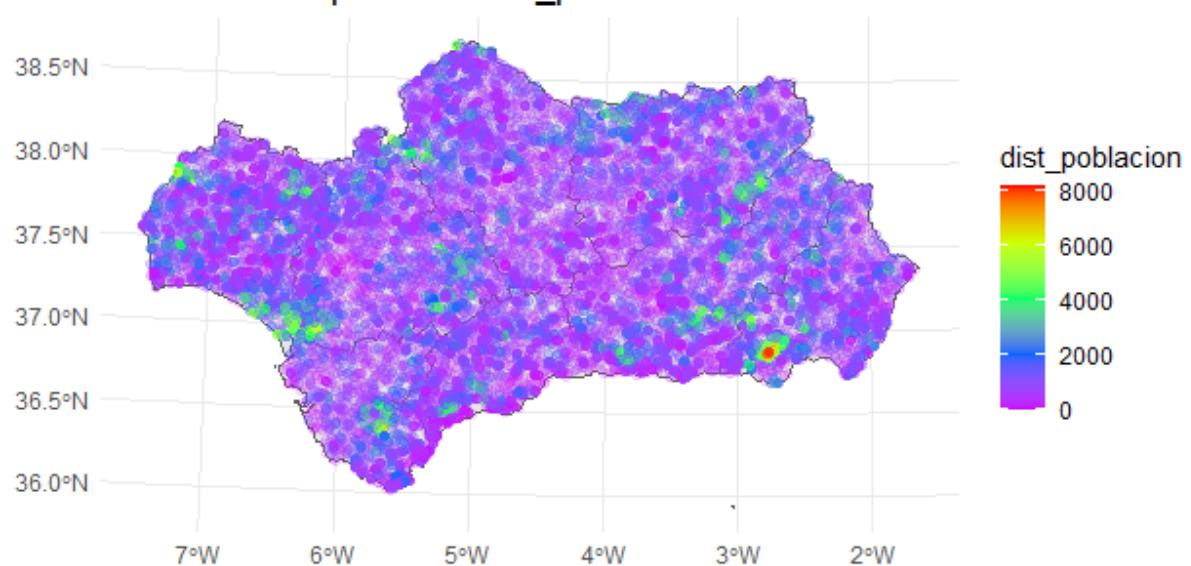
Distribución espacial de dist_poblacion

Figura A.16: Distribución espacial de dist_poblacion.

Distribución espacial de dist_electr

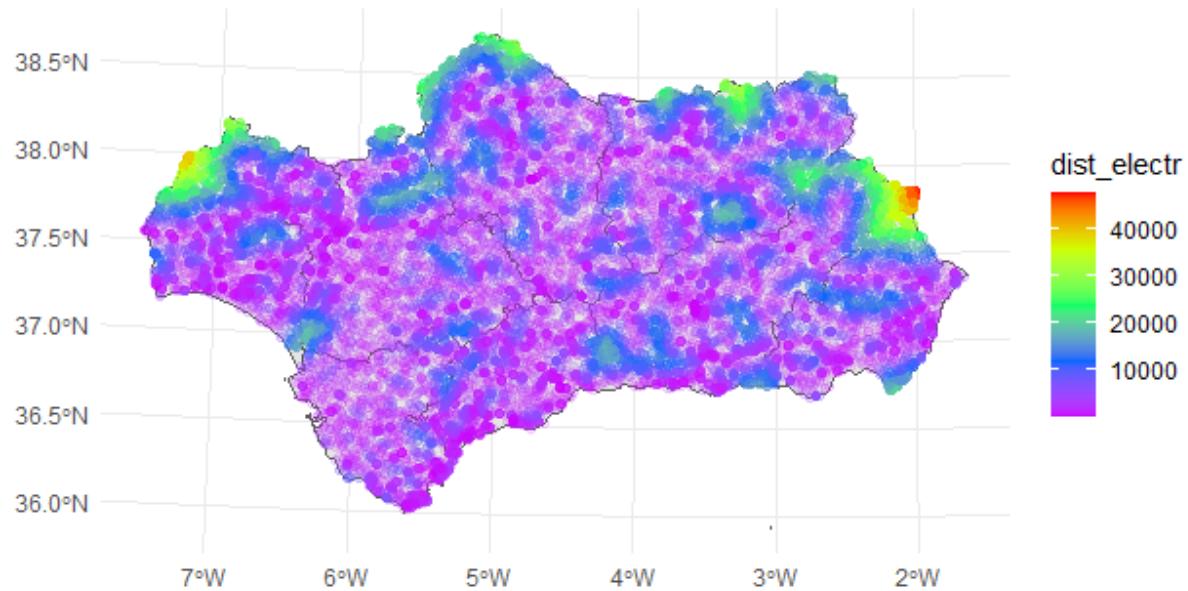


Figura A.17: Distribución espacial de dist_electr.

Distribución espacial de dist_ferrocarril

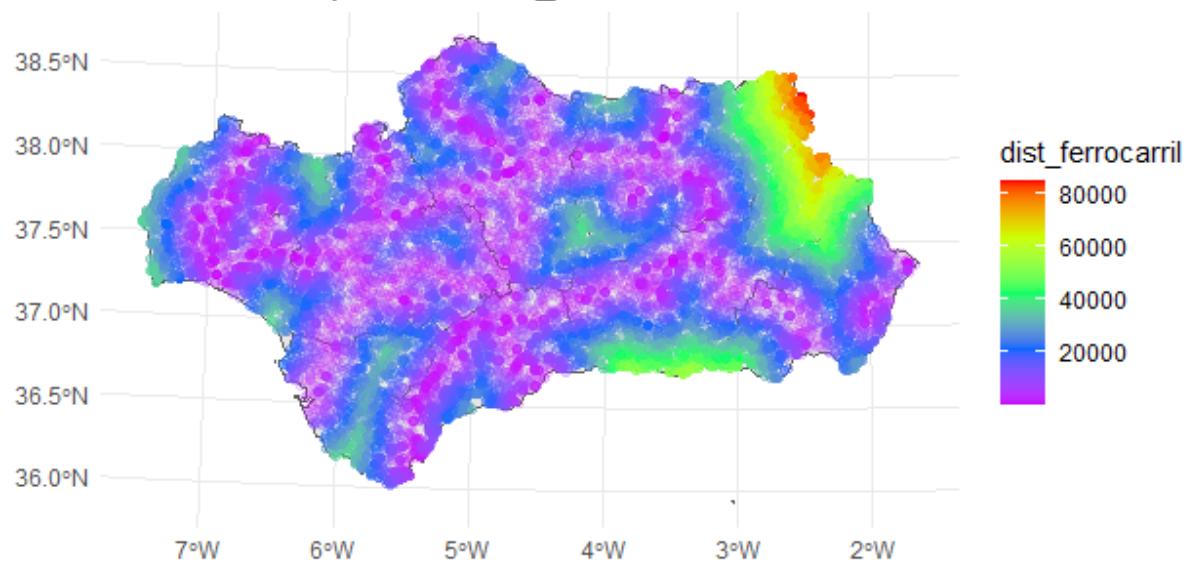


Figura A.18: Distribución espacial de dist_ferrocarril.

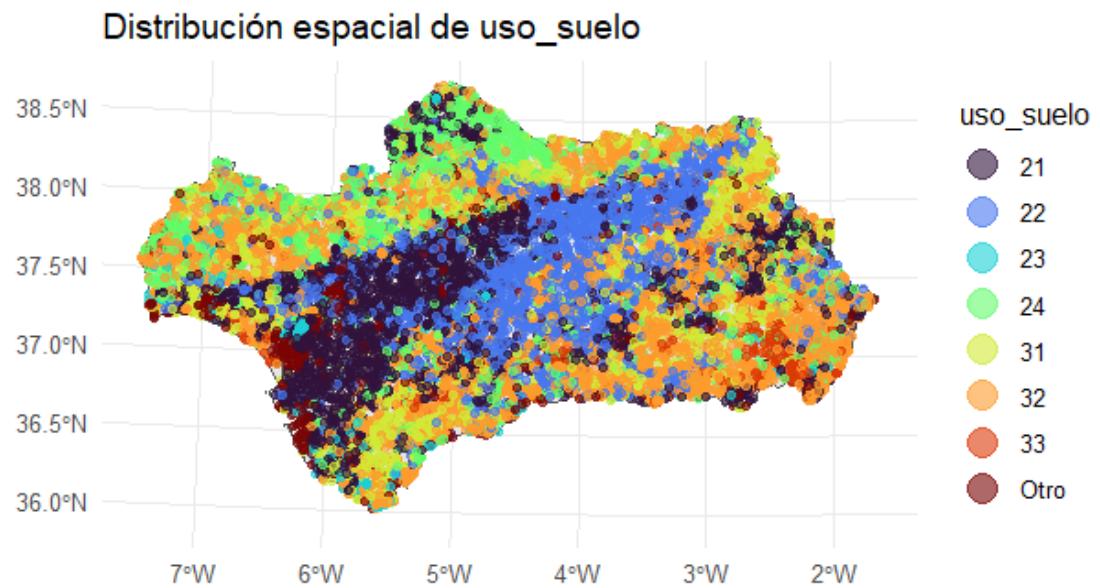


Figura A.19: Distribución espacial de uso_suelo.

Apéndice B

Apéndice: Salidas de los modelos

B.1. Regresión logística con penalización

```
# A tibble: 59 × 3
  term      estimate
  <chr>     <dbl>
1 (Intercept) -0.0672
2 T2M          0.585 
3 GWETTOP     -0.0235
4 RH2M         -0.417 
5 WS10M        0.516 
6 PRECTOTCORR -0.221 
7 elevacion    -0.296 
8 pendiente    0.320 
9 curvatura   0.149 
10 dist_carretera -0.186
11 dist_poblacion -0.0631
12 dist_electr  -0.167 
13 dist_ferrocarril -0.182
14 dist_camino   -0.0636
15 dist_sendero  -0.244 
16 poblacion    -0.134 
17 dens_poblacion 0.0674
18 dist_rios     -0.0431
19 NDVI          -0.127 
20 WD10M_NE     -0.106 
21 WD10M_E       0.0431
22 WD10M_SE     0.0740
23 WD10M_S       0.0711
24 WD10M_SW     -0.0302
25 WD10M_W       0.000845
26 WD10M_NW     -0.0260
27 orientacion_N -0.0127
28 orientacion_NE -0.000391
29 orientacion_E  0.0114
30 orientacion_SE 0.0228
31 orientacion_S  0.0454
32 orientacion_SW -0.00710
33 orientacion_W  -0.0686
34 orientacion_NW 0
35 enp_X1       -0.183 
36 uso_suelo_X22 -0.191 
37 uso_suelo_X23  0.567 
38 uso_suelo_X24  0.436 
39 uso_suelo_X31  0.368 
40 uso_suelo_X32  1.08 
41 uso_suelo_X33  0.356 
42 uso_suelo_Otro 0.100 
43 date_dow_Mon  -0.0554
44 date_dow_Tue  -0.0652
45 date_dow_Wed  0.00146
46 date_dow_Thu  0.0128
47 date_dow_Fri  -0.0116
48 date_dow_Sat  0.0545
49 date_month_Feb 0.209 
50 date_month_Mar 0.123 
51 date_month_Apr 0.0534
52 date_month_May -0.0752
53 date_month_Jun -0.366 
54 date_month_Jul -0.735 
55 date_month_Aug -0.600 
56 date_month_Sep -0.157 
57 date_month_Oct  0.0359
58 date_month_Nov  0.107 
59 date_month_Dec  0.00111
```

Figura B.1: Coeficientes del modelos de regresión logística lasso seleccionado.

Bibliografía

- [1] ALLAIRE, JJ; XIE, YIHUI; DERVIEUX, CHRISTOPHE; MCPHERSON, JONATHAN; LURASCHI, JAVIER; USHEY, KEVIN; ATKINS, ARON; WICKHAM, HADLEY; CHENG, JOE; CHANG, WINSTON y IANNONE, RICHARD (2024). *rmarkdown: Dynamic Documents for R*.
<https://github.com/rstudio/rmarkdown>. R package version 2.26,
<https://pkgs.rstudio.com/rmarkdown/>.
- [2] WICKHAM, H. y GROLEMUND, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. ISBN 9781491910368.
<https://r4ds.hadley.nz/>.
- [3] XIE, YIHUI (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*.
<https://yihui.org/knitr/>. R package version 1.45.