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

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

Índice de tablas

Capítulo 1

Capítulo 1: Introducción

1.1. Introducción

1.2. 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 incendioa forestalwa en Andalucía a partir de un estudio previo del problema.
2. Modelizar el riesgo de de incendio forestal usando distintos algoritmos de ML y comparar sus resultados
3. Analizar potenciales casos de interés.

1.3. 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"

1.4. 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 1.1: 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 2

Preliminares

2.1. Datos georreferenciados

Todos los datos empleados en este trabajo son georreferenciados, lo que significa que están asociados a ubicaciones geográficas específicas. Por ello, resulta esencial introducir, aunque sea de forma general, 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 rasters, al ser los tipos de datos fundamentales en este contexto, con características bien diferenciadas entre ellos.

2.1.1. Datos Vectoriales

El modelo de datos vectoriales geográficos 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.

2.1.1.1. Simple features

Las "Simple features" son un estándar abierto ampliamente usado para la representación de datos vectoriales, desarrollado y respaldado por el Open Geospatial Consortium (OGC, por sus siglas en inglés), 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 ingés) 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 geoprocesamiento (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 geometríca, dimensión, "Bounding Box" (límites o extensión geográfica) e información sobre el Sistema de referencia de coordenadas. Y además, presentan una columna de geometrías que tiene el nombre de "geom". Algunas ventajas del uso del modelo de "simple features" 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 and GEOMETRYCOLLECTION.

2.1.2. Datos Raster

El modelo de datos raster representa el espacio con una cuadrícula de celdas (también llamadas píxeles), que generalmente es regular, es decir, con todas las celdas de igual tamaño. Aunque no se tratarán en el presente trabajo, cabe mencionar que existen otros modelos de raster más complejos en los que se usan cuadrículas irregulares (rotadas, truncadas, rectilíneas o curvilíneas) y que pueden manipularse con el paquete de R (stars)[https://cran.r-project.org/web/packages/stars/index.html]. A cada una de estas celdas se le asocia uno (rasters de una sola capa) o varios (rasters multicapa).

Los datos en formato raster constan de una cabecera y una matriz cuyos elementos representan celdas equipespaciadas. 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:

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

La representación en forma de matriz evita tener que almacenar explícitamente las coordenadas de cada una de las cuatro esquinas de cada píxel, 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 raster 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 rásters regulares con una o varias capas a través de la clase de objetos `SpatRaster`. Sin embargo, existen otras alternativas, como el paquete (stars)[https://cran.r-project.org/web/packages/stars/index.html], que además de ser más potente, permite trabajar con rásters no regulares y ofrece una mejor integración con el paquete sf y el entorno tidyverse.

2.1.3. 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.

2.1.3.1. Sistemas de Coordenadas Geográficas

Los sistemas de coordenadas geográficas (GCS por sus siglas en inglés) 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 longitud 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 el 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 el el centro del elipsoide el que se alinea con el centro de la Tierra. Ejemplos de datums 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.

2.1.3.2. Sistemas de Coordenadas Proyectadas

Un Sistema de Coordenadas Proyectadas (PCS por sus siglas en inglés) 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, 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, 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. La mayoría de las proyecciones también se pueden clasificar en planas, cilíndricas o cónicas en función de cómo se realiza la proyección.

Un caso particular y ampliamente usado de PCS cilíndrico son los Universe Transverse Mercator (UTM), en el 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^{o} de longitud, para cada una de las cuales existe un PCS UTM correspondiente que está asociado al meridiano central. Se trata de proyecciones conformales, 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).

2.2. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA, por sus siglas en inglés), 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 buenos resultados con ellos.

El EDA hace referencia al conjunto de técnicas estadísticas 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.

2.2.1. 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 R4DS, Wickman) 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.

Entendemos por datos técnicamente correcto cuando cada valor pertenece a una variable y está almacenado en el tipo que 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 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.

2.2.2. PCA

2.3. 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.

2.3.1. 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$i = 1, 2, ..., n$, se verifica que las variables $Y|_{\underline{X}=\underline{x}_i}$ y su distribución pertenece a la famila Bernouilli, es decir,

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

- Hipótesis estructural. La esperanzara $E(Y|_{\underline{X}=\underline{x}_i}) = \pi_i$ 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

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

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 $\{(\underline{x}_i, y_i)\}_{i=1,...,n}$ 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 $\underline{\beta} = \left(\beta_0, \underline{\beta}_1\right)$, 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*.

El paquete glmnet implementa una combinación de ambos métodos (llamada *elastic net*), en la que se añade un parámetro de mezcla $\alpha \in [0,1]$ que combina ambos enfoques. El problema de optimización resultante es:

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

2.3.2. Support Vector Machine

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

2.3.2.1. SVM lineal

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

$$h(x) = w_1 x_1 + w_2 x_2 + ... + w_d x_d + b = \underline{w}^t \underline{x} = 0$$

que mejor separe a la muestra.

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

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

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

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.

Dado un punto $\underline{x}_i$ y un hiperplano $h(x) = w_1 x_1 + w_2 x_2 + ... + w_d x_d + b = \underline{w}^t \underline{x} = 0$, la distancia entre ambos viene dada por:

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

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

Dada una muestra linealmente separable $\{(\underline{x}_i, y_i)\}_{i=1,...,n}$ 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

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

O equivalentemente,

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

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

De está ultima expresión se deduce que maximizar el margen τ es equivalente a minimizar la norma euclídea de 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:

$$\begin{array}{ll} \min\limits_{w,b} & \frac{1}{2}w^t w \\ \text{s.t.} & \underline{w}^t \underline{x}_i + b \geq 1, \qquad \forall i \in \{1, ..., n\} \\ & w \in \mathbb{R}^d, b \in \mathbb{R} \end{array} \tag{2.1}$$

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 $\xi_i, \quad i = 1, ..., n$. Se habla entonces de hiperplano de separación *soft margin*. De esta forma, se llega al problema de optimización siguiente:

$$\begin{array}{ll} \min\limits_{w,b,\xi} & \frac{1}{2}w^t w + C\sum\limits_{i=1}^{n} \xi_i \\ \text{s.t.} & \underline{w}^t \underline{x}_i + b \geq 1, \qquad \forall i \in \{1, ..., n\} \\ & \xi \geq 0, \qquad \forall i \in \{1, ..., n\} \\ & w \in \mathbb{R}^d, b \in \mathbb{R} \end{array} \tag{2.2}$$

donde $C > 0$ es un parámetro de regularización que permite controlar los errores de clasificación permitidos por el modelo. Recibe el nombre de coste (*cost* en inglés).

2.3.2.2. 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 (*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 (o gaussiano): $k(x, z) = \exp(-\gamma\|x - z\|^2)$

2.3.3. Decision Trees

Un árbol de decisión (DT por sus siglas en inglés) 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 raiz, ramas, nodos internos y nodos hojas. Cada nodo representa una test sobre

una variable, y cada rama que nace de ese nodo uno de 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), la que aquí se plantea es una de las más usadas y recibe el nombre de CART (Clasification And Regression Trees). Se explica para el caso de árboles de clasificación binarios, i.e. en los que de cada nodo salen dos ramas.

Dada una muestra $\{(\underline{x}_i, y_i)\}$ con $\underline{x}_i = (x_{i1}, ..., xid)$, 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 $\{R_j\}_{j=1,...,J}$ es una partición del espacio predictivo y c_j es la clase asignada en R_j para todo $j \in 1, ..., J$.

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

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 separa 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 es 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,división) que conduce a una mayor reducción de la impureza. Es decir, dada una medida de impureza Φ se busca la variable $j \in 1, .., d$ y la división $s \in \mathbb{R}$ solución de

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

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. La idea fundamental es que 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 son modelos con una varianza elevada, por lo que tienden a ser muy inestables y a producir sobreajuste. Para evitar esto, se recurre al uso de técnicas de *Bagging* y *Boosting*. El ejemplo más extendido con árboles de decisión son los Bosques Aleatorios (*Random Forest* en inglés).

2.3.4. 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 redución 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.

Algoritmo:

1. Para $b = 1, ..., B$:

   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 alcanze 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 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 hacer una predicción en un nuevo punto $\underline{x}$ (clasificación): Sea $\hat{C}_b(\underline{x})$ la clase predicha por el b-ésimo árbol de decisión bootstrap. Entonces, $\hat{C}^B_{rf}(\underline{x}) = majorityvote\{\hat{C}_b(\underline{x})\}$. Es decir, se asigna la clase más votada.

2.3.5. Redes Neuronales

2.4. 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.

2.5. Evaluación 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. Además, 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 a distintas métricas, en función de las características propias de cada problema.

2.5.1. 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 definen las siguientes métricas de rendimiento de un modelo de clasificación binaria:

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

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

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

$$Precisión = \frac{TP}{TP + FP}$$

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

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

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

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

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.

2.6. 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:

. . .

library(tidyverse) library(skimr) library(sf) library(corrplot) library(GGally) # Coordenadas paralelas library(ggpubr) library(tidyverse) # Manipulación de datos library(sf) # Vector data library(terra) # Raster data library(mapSpain) # Polígonos de regiones de España library(magrittr) # Operador %<>% library(tidymodels) library(sf) library(ggplot2) library(akima) # interp library(magrittr) library(ggpubr) library(forcats)

Capítulo 3

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 pueden ser 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 andaluz 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 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.

3.1. 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.

3.1.1. 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 TURTLE[1] 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 que requería 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 abarcable 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.

[1] TURTLE 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.

3.1.2. 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 desagregados espacial y temporalmente.

Los incendios forestales son un proceso sumamente complejo, en el que actúan numerosos factores de muy distinta índole (...). 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 considera-

rán 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 20 años que van de 2002 a 2022, ambos inclusive.

3.2. 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, del área recorrida por el fuego y del municipio en el que originó, así como de otras variables que dependen del año de la campaña y que no son relevantes de cara a nuestro 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 3.1, 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, ya que por ejemplo para la variable NDVI se dispone de un archivo tiff para cada mes del periodo de estudio, lo que añade cierta complejidad al manejo de la información.

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 geocumputación.

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 aproximativos 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.

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	Población del municipio	IECA[c]	csv	Anual
Hidrográficas	Principales Ríos	MAGRAMA[d]	Shapefile	-
Meteorológicas	Precipitación (mm/day)	NASA POWER[e]	df (0.5º x 0.625º)	Diaria
	Temperatura a 2m sobre la superficie (º)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Humedad del suelo (%)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Dirección del viento a 10 metros sobre la superficie terrestre(º)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Humedad relativa a 2m sobre la superficie (%)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Cantidad de precipitaciones (mm/day)	NASA POWER	dfdf (0.5º x 0.625º)	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 Worlwide Energy Resources (NASA POWER)

Tabla 3.1: Datos brutos

3.3. 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 apto para el análisis estadístico y la construcción de modelos predictivos, es decir, a un data frame. 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 imporante 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.

3.3.1. 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 incio 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 $\pm$ 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 $\pm$ 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.

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 3.1. 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 de generar 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

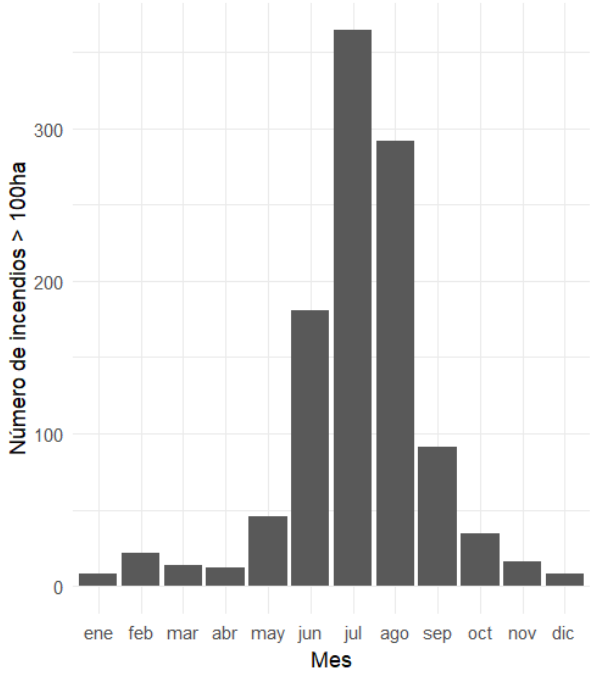

Figura 3.1: Incendios durante el periodo de estudio

de incendios forestales. Además, este enfoque permite centrar el esfuerzo en los periodos que más incendios se producen.

3.3.2. 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, 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 3.1). 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 electrica, 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 bicis

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 3.2: Códigos de uso de suelo

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 eleviaciones 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 inabarcable para 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 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 establecen con 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. En la Tabla 3.2 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 divido 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).

3.3.3. 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 *orientación* en los 4 puntos cárdinales 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 perdidos 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 3.2). 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.

El resultado de todo este proceso es un conjunto de datos con 21546 registros y 27 variables, las cuales se detallan en la Tabla 3.3.

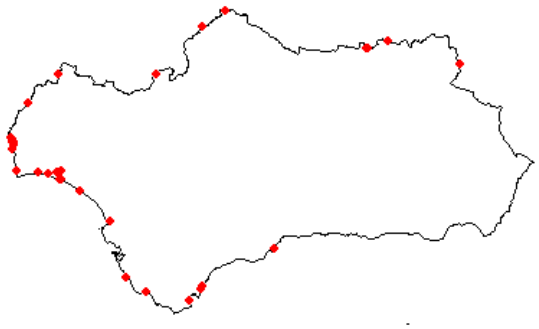

Figura 3.2: Observaciones para los que no está disponible alguna de las variables topográficas

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 (º)	numérica
	curvatura	Curvatura de la superfie	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_ferocarril	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	dist_rios	Distancia al río más próximo (m)	numérica
Meteorológicas	PRECTRORCORR	Promedio corregido del total de precipitaciones en la superficie de la tierra en masa de agua (incluye el contenido de agua en la nieve) (mm/día)	numérica
	T2M	Temperatura promedio del aire a 2 metros sobre la superficie de la tierra (ºC)	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
	RH2M	Humedad relativa a 2 metros sobre la superficie de la tierra	numérica
Variable Objetivo	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 3.3: Conjunto de datos depurados

Capítulo 4

Cuerpo

4.1. Análisis exploratorio de datos

3.1 3.3

4.2. Estudio de las variables

4.3. Modelización

4.3.1. Modelo 1

4.3.2. Modelo 2

4.3.3. ...

4.4. Comparación

Capítulo 5

Estudios de casos de interés

Capítulo 6

Conclusión

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

Bibliografía

[1] (Página web). «Universidad de Sevilla». Disponible en https://www.us.es.

[2] 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/.

[3] Facultad de Matemáticas (Univ. Sevilla) (s.f.).
https://www.matematicas.us.es.

[4] Lopez, Juan Fernando; Fernández Henao, Sergio y Morales, Marcela María (2007). «Aplicación de la programación por metas en la distribución de servicios entre empresas operadoras del sistema de transporte masivo». *Scientia et technica*, **13(37)**, pp. 339–343.

[5] Luque Calvo, Pedro L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown.*
http://destio.us.es/calvo.

[6] —— (2019). *Cómo crear Tablas de información en R Markdown.*
http://destio.us.es/calvo.

[7] Luque Calvo, Pedro L. (2021). «Página personal de Pedro L. Luque».
http://destio.us.es/calvo.

[8] R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org/.

[9] RStudio Team (2015). *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA.
http://www.rstudio.com/.

[10] Xie, Yihui (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R.*
https://yihui.org/knitr/. R package version 1.45.