eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Tesis de Máster

Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

# Desarrollo de modelos de aprendizaje automático para la estimación del riesgo de incendios forestales

*Julen Ercibengoa Calvo*

**Dirección**
Dra. Izaro Goienetxea Urkizu
Dra. Meritxell Gómez Omella

2 de enero de 2025

# Resumen

Los incendios forestales tienen impactos devastadores en los ecosistemas naturales, en la biodiversidad e incluso en la economía y vidas humanas. España es uno de los países más afectados de la zona mediterránea habiendo sufrido más de 7000 incendios desde 2008, los cuales han quemado alrededor de 1.300.000 hectáreas. Por todo ello, para que los cuerpos de bomberos y guardabosques puedan distribuir sus unidades de forma adecuada, es importante tener sistemas de predicción de riesgo de incendio forestal precisos.

Este artículo de investigación presenta varios modelos de Deep Learning que predicen el riesgo de incendio forestal de un día para otro en toda España con una resolución espacial de 1km, alcanzando una sensibilidad del 93 % y una especificidad del 60 %. Los modelos han obtenido los resultados siendo testados en conjuntos de datos que replican las distribuciones extremadamente desbalanceadas que ocurren en la práctica.

# Índice de contenidos

# Índice de figuras

# Índice de tablas

# Lista de Algoritmos

CAPÍTULO 1

# Introducción

Los incendios forestales tienen multiples consecuencias ambientales, sociales y económicas. Por un lado, destruyen vastas áreas de bosque, provocan la pérdida de hábitats naturales y emiten grandes cantidades de dióxido de carbono a la atmósfera. Además, erosionan el terreno y la vegetación necesita años en volver a aparecer [1].

Por otro lado, los incendios forestales pueden ser mortales tanto para los montañeros como para bomberos y personal de emergencia, ya que en algunas ocasiones la velocidad de propagación de un incendio puede superar los 22 km/h [2]. Además, los incendios causan la pérdida de hogares y medios de vida como terrenos agrícolas, lo cual se traduce en pérdidas económicas enormes tanto para la población como para el gobierno.

Esta problemática es particularmente importante en el caso de España, ya que es uno de los países más afectados en la Unión Europea [3, 4]. Según [5], cerca del 40 % de la superficie total quemada de toda Europa mediterránea entre 1980 y 2008 fue en España. Según datos del Sistema de Información Europeo sobre Incendios Forestales (EFFIS, European Forest Fire Information System [6]), desde 2008 en España se han producido más de 7000 incendios, que han causado decenas de muertos [7] y han quemado más de 1.300.000 hectáreas (similar a la superficie de Montenegro o Bahamas), ver figura 1.1. Como consecuencia, se invierten alrededor de 1.000 millones de euros anuales en extinción y gestión forestal. Se estima que solamente en 2022 los incendios ocasionaron pérdidas de más de 2.000 millones de euros [8].

Es por todo ello imprescindible tener herramientas de apoyo para combatir los incendios forestales. Para eso se estableció el proyecto GAIA [9] (del griego, *Tierra*), iniciativa cuyos objetivos son la prevención, detección de incendios forestales y la reforestación [10], de la cual es parte la empresa Tekniker. Este trabajo se ha realizado en dicha empresa con el objetivo de desarrollar un modelo de aprendizaje automático capaz de predecir el riesgo de incendio forestal con una temporalidad diaria.

Para poder seleccionar las variables adecuadas y entrenar un buen predictor de riesgo de incendio forestal, primero hace falta entender las causas que provocan los incendios forestales


**Figura 1.1:** Área de los incendios forestales en España coloreados en base a la cantidad de hectáreas quemadas, desde 2008 hasta 2024. Datos de EFFIS.

y qué factores hacen que un incendio pueda llegar a ser grande.

## 1.1. Causas principales de los incendios forestales

La literatura científica coincide en que la gran mayoría de los incendios forestales son causados por humanos. En un artículo del 2009, San Miguel Ayanz y Camiá dicen que en el sur de Europa más del 95 % de los incendios forestales se crean por causas humanas [11]. Un artículo del 2021 que analiza los incendios forestales de Australia también dice que la gran mayoría se generan por culpa de los humanos [12]. En cuanto a España, Helena Liz-López y col. en un artículo reciente del 2023 dicen que los humanos causan alrededor del 90 % de los incendios, provocados o accidentales [13].

Una de las causas más comunes de los incendios forestales son las quemas agrícolas y ganaderas. Tal y como se dice en un artículo del año 2000 [14], el uso del fuego para eliminar rastrojos, limpiar fincas o regenerar pastos es una práctica común que causa grandes incendios

al descontrolarse. Consultándolo con guardabosques, hoy en día en algunas zonas del norte de España estas prácticas siguen siendo muy comunes para quemar malas hierbas. Estos incendios 'controlados' pueden llegar a tener hasta 5 hectáreas, lo que hace que sea relativamente sencillo que se descontrolen.

Otro factor importante pueden ser las actividades recreativas como excursionistas y cazadores, que pueden generar incendios al no apagar correctamente fogatas o tirar cerillas [14].

La maquinaria agrícola y líneas eléctricas son también causantes de muchos incendios, ya que pueden generar chispas que, junto a las condiciones adecuadas, den lugar a un gran incendio. Es por eso que es fundamental realizar podas en las áreas donde atraviesan tendidos eléctricos, ya que la caída de un árbol sobre ellos podría provocar un incendio.

Por último, las causas naturales representan un porcentaje significativamente menor en comparación con las causas humanas, sin embargo, en algunos contextos pueden ser relevantes. Por ejemplo, en Aragón, Castilla-La Mancha y Cataluña los rayos provocan un número significativo de incendios por culpa de las tormentas secas [14].

## 1.2. Factores que afectan en la propagación del fuego

Para poder entender mejor los factores que afectan en la propagación de los incendios, conversamos con un guardabosques. Él nos habló sobre el bien conocido 'triángulo del fuego': temperatura, oxígeno y combustible. Aunque sean pautas simples, siguen siendo los factores más importantes.

Por un lado está la temperatura, que hace que la vegetación esté más seca. En esta sección podríamos introducir otros factores climáticos como la lluvia y la humedad: si ha llovido durante varios días o hay mucha humedad relativa, es mucho más difícil que un fuego se esparza. También podemos incluir el viento: a mayor viento más rápido se propaga el incendio.

Por otro lado está el oxígeno. El fuego necesita oxígeno para mantenerse y a menor oxígeno, menos fuegos. Esto se traduce en que la elevación del terreno (altitud) es un factor muy importante. También se observa empíricamente en la mayoría de artículos que intentan predecir el riesgo de incendio [15, 16, 17]. Además de la elevación, la pendiente y orientación del terreno son otros factores a tener en cuenta: un fuego en pendiente y orientado hacia la dirección del viento, con rachas de vientos fuertes, es un fuego que se esparce muy rápido.

Por último, el que podría ser el factor más importante es el combustible: la vegetación que el fuego quema para esparcirse. Si hay mucha vegetación alta y seca, el fuego se esparcirá muy rápido, por lo que las interacciones entre la temperatura y el tipo de terreno o vegetación son importantes.

En conclusión, la propagación de un incendio es un fenómeno multifactorial y dinámico, donde no existe una fórmula única que explique todos los escenarios. Que un incendio sea grande es resultado de complejas interacciones entre estos factores, junto a mucha aleatoriedad.

## 1.3. Dificultad de la tarea de predicción

Una vez que conocemos los factores más importantes a tener en cuenta a la hora de predecir el riesgo de incendio forestal, analicemos la dificultad que supone dicha tarea.

Hemos visto que la mayoría de incendios son causados por factores no predecibles: que un excursionista tire un cigarrillo en el monte, chispas que ocurren por la actividad ganadera o agrícola... Esto hace que predecir que se produzca un incendio con éxito sea prácticamente imposible.

Por otro lado, el hecho de que no haya un incendio no significa que no haya peligro de incendio, por lo que en el conjunto de datos podemos observar varias instancias de 'no-fuego', pero que hayan tenido riesgo alto de incendio.

La mayoría de días en la mayoría de sitios no hay incendios, por lo que el problema al que nos enfrentamos sufre de desbalanceo de datos masivo. Más adelante entraremos más en detalle en cómo definimos una instancia positiva y negativa de 'fuego' pero por ejemplo, en [18] tienen una instancia de 'fuego' cada 100.000 'no-fuegos'. Esto hace también que la cantidad de datos sea masiva. En ese mismo artículo el conjunto de datos sin submuestrear tiene 830 millones de instancias.

Por último, también nos enfrentamos a un 'concept drift', ya que desde hace años se ha observado un aumento en las hectáreas quemadas por los Grandes Incendios Forestales (incendios de más de 500 hectáreas). Cada vez más comunidades autónomas se enfrentan a incendios de 5.000, 10.000 y 20.000 hectáreas y, tal y como dicen en [19], se puede concluir que ha habido un cambio de régimen de incendios en España. A esto se le suma el incremento de la biomasa y el abandono del medio rural [20] que afecta de forma negativa al estado de los bosques. Además, el cambio climático hace que las condiciones meteorológicas en las cuales un incendio se propaga mucho sean cada vez más frecuentes.

Todo lo anterior hace que la tarea de predecir el riesgo de incendio forestal sea muy compleja por lo que hay una gran variedad en cuanto a la metodología seguida en la literatura científica. En la próxima sección analizaremos diferentes puntos de vista utilizados hasta el momento y nos centraremos en los que sirvan para nuestra tarea: predecir el riesgo de incendio forestal con un día de antelación, asegurando la eficacia en un escenario desbalanceado realista.

CAPÍTULO 2

# Aplicación de la Inteligencia Artificial en tareas de prevención

Tradicionalmente para hacer predicciones de riesgo de incendio forestal, se han utilizado índices de clasificación de peligro de incendio como el índice Canadiense 'Fire Weather Index' [21]. De hecho, EFFIS [6] utiliza dicho índice para hacer sus predicciones. La primera aparición de un sistema inteligente aplicado en incendios forestales es de 1989 [22]. En años posteriores se han aplicado en muchas ocasiones algoritmos clásicos de Machine Learning para predecir el riesgo de incendio forestal. En la actualidad, con el auge del Deep Learning, se ha vuelto una práctica muy común el usar redes neuronales, ya que consiguen captar patrones no lineales muy complejos.

Como hemos dicho anteriormente, en el estado del arte hay mucha variedad en cuanto a la metodología utilizada para resolver el problema. Cada artículo define el problema de forma distinta y crea conjuntos de datos en base a esa definición, por lo que comparar los resultados puede llegar a ser difícil. En este capítulo analizaremos los diferentes enfoques del estado del arte para resolver el problema y decidiremos cuál es la mejor metodología en nuestro caso.

## 2.1. Metodología común

Empecemos analizando la metodología general que se sigue en el estado del arte. Aunque la mayoría de artículos difieran en muchos aspectos de la metodología, todos dividen el área a analizar en celdas espaciales, donde cada celda representa una instancia con variables tabulares como temperatura, humedad o el tipo de terreno. Por ejemplo, Oliveira y col. [4] dividieron toda la zona Mediterránea de la Unión Europea en 16.000 celdas de 10km $\times$ 10km. En [18] generan un total de 360.000 celdas de 500m $\times$ 500m que cubren toda Grecia. Hacen algo parecido también en Grecia en [23] y en [17] pero con celdas de 1km $\times$ 1km. En [2] generan un total de 63 celdas de 1km $\times$ 1km para representar una zona de California. En el caso de

España, Alonso-Betanzos y col. en un artículo del 2002 dividieron Galicia en 360 celdas de 10km $\times$ 10km [1]. Un artículo más reciente del 2014 divide los bosques de toda España en celdas de 10km $\times$ 10km [3].

En esta metodología común hay una diferencia importante entre algunos artículos: la temporalidad. Algunos artículos no dividen el área en celdas espaciales, sino en cubos temporales. Por ejemplo en [23] cada celda de 1km $\times$ 1km también tiene la dimensión temporal diaria, por lo que una instancia sería un cubo espacio-temporal que tiene asociadas varias variables tabulares. Esto se hace porque hay muchas variables meteorológicas que cambian diariamente y que afectan en los incendios forestales. En [2] también utilizan cubos espacio-temporales diarios durante 4 años y en [18] lo hacen durante 11 años. Otros artículos sin embargo, no tienen en cuenta el aspecto temporal.

Teniendo en cuenta el objetivo que tenemos, la metodología que vamos a emplear a la hora de crear instancias en nuestro conjunto de datos será la de generar cubos espacio-temporales: crearemos cubos de 1km $\times$ 1km $\times$ 1 día que cubran toda España durante varios años.

## 2.2. Tipo de variable dependiente

Una vez que conocemos qué representan las instancias de los conjuntos de datos, veremos qué tipo de problemas se resuelven en el estado del arte. De acuerdo con lo que conocemos, la mayoría de los artículos plantean el problema como un problema de aprendizaje supervisado. Es decir, para cada celda o cubo, hay un valor que representa el estado de la misma. En general, en la mayoría del estado del arte se define como un problema de clasificación: cada celda tiene valores discretos para indicar la existencia de un fuego o no (normalmente 0 para 'no-fuego' y 1 para 'fuego'). Sin embargo, hay algunos trabajos en los que lo plantean como un problema de regresión.

Algunos trabajos que plantean el problema como un problema de regresión son por ejemplo Oliveira y col. [4] que predicen la densidad de fuegos a nivel europeo utilizando regresión lineal multivariante; y Cortez y Morais que utilizan redes neuronales para predecir el área quemada por un incendio [24]. Estos obtienen resultados muy prometedores pero no son aplicables para nuestro caso, ya que el primero hace una predicción atemporal y el segundo predice el área quemada una vez que se ha iniciado el incendio.

Respecto a la clasificación, el estado del arte se predomina por clasificación dicotómica. Un ejemplo de ello sería tomar como una instancia de 'fuego' si ha habido un fuego en ese cubo espacio-temporal y una instancia de 'no-fuego' si no lo ha habido. Alonso-Betanzos y col. [1] entrenan una red neuronal con este método y luego dividen las predicciones en 4 clases de riesgo de incendio utilizando las probabilidades de pertenencia a las clases de 'fuego' y 'no-fuego'. [17, 18, 23] también emplean el mismo método de clasificación binaria para predecir el riesgo de incendio forestal de un día para otro en Grecia, y utilizan las probabilidades de pertenencia de las clases para definir la cantidad de riesgo de incendio.

Para el caso en el que no se toma en cuenta la temporalidad, en [3] toman como instancia de 'fuego' si ha habido 2 o más fuegos en esa celda durante 20 años. Es decir, introducen la

temporalidad en la variable predictora.

En nuestro caso, definiremos el problema como un problema de clasificación binaria, y utilizaremos las probabilidades de pertenencia a las clases 'fuego' y 'no-fuego' para obtener el riesgo de incendio forestal.

## 2.3. Definición del riesgo de incendio forestal

Aunque ya hayamos definido el problema como problema de clasificación binaria, es interesante analizar lo que hacen Ioannis Prapas, Spyros Kondylatos y col. en el trabajo [23]. Como hemos visto en la introducción, predecir el inicio de un fuego es prácticamente imposible, debido a la aleatoriedad intrínseca que tiene. Por ello, en el artículo solo toman como instancias de 'fuego' los fuegos que han quemado más de 30 hectáreas. Es decir, no intentan entrenar el algoritmo para que prediga la aparición de fuegos, sino que prediga el riesgo de que en caso de iniciarse un fuego, ese fuego sea grande (más de 30 hectáreas). Además, cuanto mayor sea un fuego, más instancias de 'fuego' generará, puesto que el fuego cubrirá más celdas espaciales (ver figura 2.1).

Creemos que esta metodología es la apropiada para nosotros puesto que no es lo mismo una instancia de 'fuego', a que el fuego afecte a solamente 2 hectáreas, o una instancia de 'fuego' que cubra la totalidad de la celda (100 hectáreas). El fuego más grande se habrá esparcido más por culpa de diferentes factores que puede que no aparezcan en el fuego pequeño.

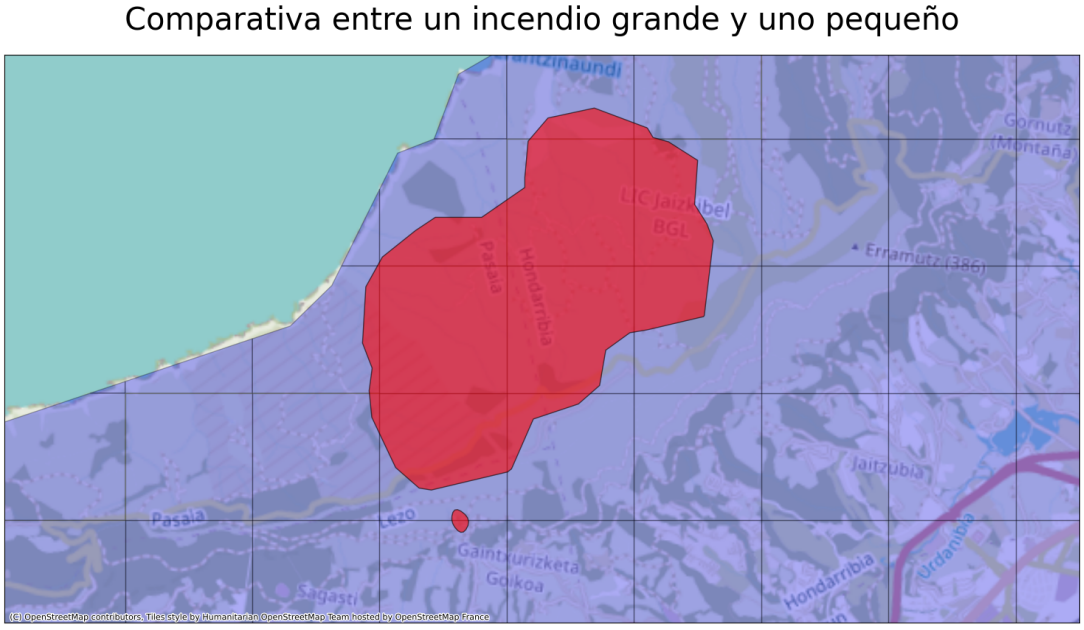


**Figura 2.1:** Visualización de dos incendios junto a celdas espaciales de 1km × 1km: un incendio de 2 hectáreas y otro de 501, datos de EFFIS.

Esta metodología es también intuitiva, ya que los fuegos grandes provocan más destrucción que los pequeños, por lo que asignar mayor riesgo a las condiciones necesarias para que haya incendios grandes cobra sentido.

## 2.4. Desbalanceo de datos

El utilizar como instancias cubos espacio-temporales trae consigo un problema: la mayoría de días y en la mayoría de sitios no hay fuegos. Por ejemplo, en la base de datos que generan en [18], hay alrededor de una instancia de 'fuego' cada 100.000 instancias de 'no-fuego'. En dicho trabajo utilizan 360.000 celdas diarias durante 7 meses consecutivos (en la época de incendios) en 11 años; con un total de 830 millones de instancias, por lo que la mayoría de esas instancias serán de 'no-fuego'.

Este problema hace que al entrenar un clasificador en la base de datos extremadamente desbalanceada, el clasificador tienda a ser un clasificador trivial (que siempre predice 'no-fuego'). Por ello se suele submuestrear la clase mayoritaria ('no-fuego'), aunque en la práctica, como las instancias de 'no-fuegos' se generan de forma manual, se generan la cantidad de instancias 'no-fuego' deseadas. Esto se debe a que la información disponible que tenemos es la de los incendios que han ocurrido (ver figura 1.1), por lo que hay que generar instancias de 'no-fuego' artificialmente (en días y celdas que no hubo fuegos).

La parte negativa de submuestrear la clase mayoritaria es que la mayoría de trabajos del estado del arte también lo hacen para el conjunto de test. Es decir, los algoritmos se testan en conjuntos de datos que no cumplen la distribución extremadamente desbalanceada que ocurre en la realidad.

En [1] dicen que alrededor del 5 % de su conjunto de datos real son instancias de 'fuego', por lo que deciden generar aleatoriamente en el espacio y en el tiempo hasta tener la misma cantidad de instancias de 'fuegos' y 'no-fuegos'. Esta distribución también la mantienen en el conjunto de test.

Sin embargo, en [2] en vez de submuestrear la clase mayoritaria lo que hacen es sobre-muestrear la clase minoritaria utilizando el algoritmo 'Synthetic Minority Oversampling Technique' (SMOTE) [25]. Sin embargo, no mencionan la proporción final utilizada.

En [23] generan dos instancias de 'no-fuego' por cada instancia de 'fuego' manteniendo la distribución del tipo de terreno que siguen las instancias de 'fuego'. En este trabajo también mantienen esa distribución para el conjunto de test.

En nuestro caso, no mantener la distribución desbalanceada en el conjunto de test haría que se sobreestimasen las capacidades reales del modelo predictivo a la hora de aplicarlo en la práctica. Por ello, conviene seguir la metodología que usan en [18]: utilizan un conjunto de datos balanceado para entrenar, uno con solamente 10 veces menos instancias de 'no-fuego' que la proporción real para validar, y otro con la proporción real para testar.

## 2.5. Bases de datos utilizadas

Otro aspecto muy importante (si no el más importante) son los datos tabulares utilizados para entrenar al modelo. En general todos los artículos del estado del arte utilizan datos meteorológicos, del estado de la vegetación, del uso del terreno y de la elevación del terreno. Otros artículos también introducen el factor humano como distancia a carreteras o densidad poblacional.

En cuanto a los datos históricos de meteorología, las variables utilizadas suelen ser la temperatura, la humedad relativa y el viento entre otros. Cada artículo obtiene los datos de distintas fuentes, aunque hay varios artículos que usan una fuente común. [17, 18, 23] utilizan datos de ERA5-Land [26]: una base de datos con varias variables meteorológicas estandarizadas en una resolución espacial de 9km $\times$ 9km en todo el mundo y una resolución temporal horaria desde 1950 hasta hoy en día. Otros artículos utilizan bases de datos específicos de cada territorio geográfico. Por ejemplo, en [15] utilizan una base de datos China y en [12] una base de datos del gobierno Australiano.

Los datos del estado de la vegetación consisten en los índices NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index) y LST (Land Surface Temperature). Por ejemplo, [18] obtiene dichos índices mediante las mediciones MODIS (Moderate Resolution Imaging Spectroradiometer) de la NASA, los cuales están disponibles cada 8 días. En [17] y [23] obtienen los índices de la misma fuente y además utilizan otros dos índices: LAI (Leaf Area Index) y Fpar (Fraction of Photosynthetically Active Radiation). En [2] también obtienen los datos de la NASA pero solamente utilizan el NDVI, EVI y añaden otro índice llamado NDWI (Normalized Difference Water Index). Como antes, [12, 15] utilizan bases de datos de sus territorios geográficos.

Sobre el uso del terreno, es muy común en trabajos realizados en Europa utilizar la base de datos 'Corine Land Cover' [27]: una base de datos de Copernicus que divide toda Europa en una resolución espacial de 100m $\times$ 100m en 44 clases distintas (ver figura 2.2). Por ejemplo, los trabajos [4, 16, 17, 18, 23] utilizan dicha base de datos.

Las mediciones de elevación del terreno se llaman DEM (Digital Elevation Model). Mediante los DEM se pueden obtener variables como la elevación del terreno, la pendiente, la orientación y la rugosidad a diferentes resoluciones espaciales. Los trabajos [18, 17, 23] utilizan la base de datos Europea Copernicus EU-DEM [28] a 30m $\times$ 30m para obtener dichas variables. En [16] utilizan la base de datos 'United States Geological Survey' y en [15] una base de datos de China.

Por último, la mayor diferencia en cuanto a datos utilizados entre los artículos se encuentra en el factor humano. Algunos artículos no introducen ninguna variable que tenga que ver con el factor humano como por ejemplo [18]. Otros artículos en cambio sí que las introducen. Por ejemplo, [23, 17] utilizan datos de OpenStreetMap [29] para calcular la densidad de carreteras y datos de WorldPop [30] para calcular la densidad poblacional. En [15] utilizan tanto la distancia a carreteras, como la distancia a área residencial. También utilizan variables socioeconómicas como PIB per cápita y si es día de festividad China (en dichos días se suelen

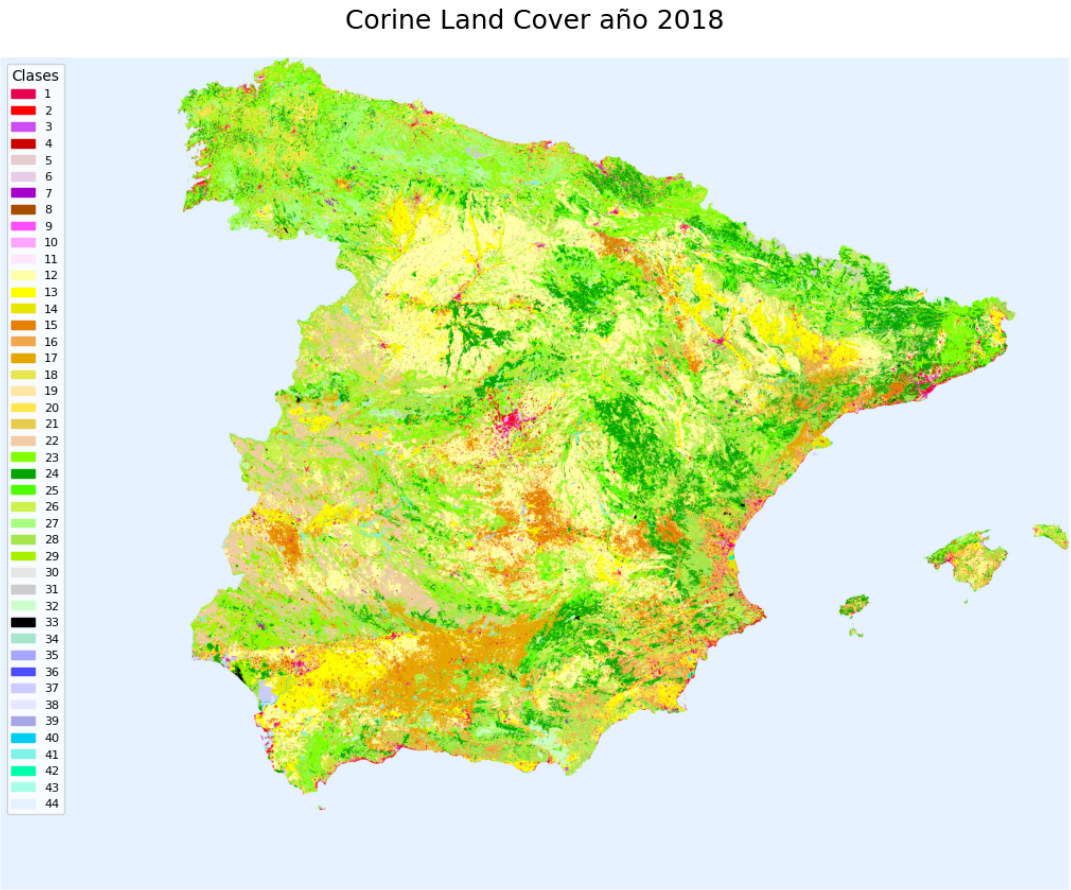


**Figura 2.2:** Base de datos Corine Land Cover del año 2018.

lanzar farolillos al aire con velas dentro). En [2] también utilizan información de líneas de alta tensión, una variable que podría afectar en el riesgo de incendio tal y como hemos explicado en la introducción. [16] utiliza muchas variables relacionadas con los humanos: además de las que se utilizan en los trabajos ya comentados, también utiliza la distancia a terreno cultivable y distancia a terreno agrícola, generados mediante el conjunto de datos 'Corine Land Cover'. Por último, [4] entre otras muchas variables relacionadas con los humanos utiliza la tasa de desempleo y dice que, aunque no se sepa realmente por qué, tiene relación con el riesgo de incendio.

## 2.6. Métodos de validación

Ya hemos visto cómo se plantea el problema en el estado del arte y qué datos se utilizan: se divide el área en cubos espacio-temporales y cada cubo representa una instancia en el conjunto de datos. Ahora vamos a analizar los métodos de validación adecuados para esta forma de plantear el problema.

Si se utiliza el método clásico de dividir el conjunto de datos de forma aleatoria en un subconjunto de entrenamiento y un subconjunto de test, puede que un mismo incendio tuviese varias instancias de 'fuego' en el conjunto de entrenamiento y de test al mismo tiempo. Por ejemplo, el fuego de 501 hectáreas de la figura 2.1 genera 12 instancias espaciales de 'fuego', por lo que si tomásemos aleatoriamente el 20 % del conjunto de datos como datos de test, seguramente algunas de esas 12 instancias estarían en el conjunto de entrenamiento y otras en el conjunto de test. Esto es un problema ya que testaríamos el modelo en instancias muy parecidas en las que han sido entrenadas (generadas por el mismo fuego), sobrestimando las capacidades reales del modelo.

Para corregirlo, algunos trabajos del estado del arte han propuesto esquemas de validación propias de series temporales para evitar fugas de datos ('data leakage'). Por ejemplo, en [17] y [23] entrenan los algoritmos en datos de 2009-2018, validan en datos de 2019 y testan en 2020. Además, enfatizan en que ésta debería de ser la forma correcta de validar, ya que a la hora de utilizar el modelo en la práctica, se entrena con datos del pasado con el objetivo de predecir en el futuro.

En [18] proponen dos métodos de validación. Por un lado, proponen un método clásico adaptado a este problema: hacer una validación cruzada con k particiones pero asegurándose que un mismo día no se repite en distintas particiones. Esto hace que un mismo incendio no pueda generar instancias de 'fuego' que estén en diferentes particiones. Por otro lado, proponen otro esquema de validación cruzada usada en series temporales: se mantiene el orden temporal en las particiones utilizadas para validar, mientras se va aumentando el tamaño del conjunto de entrenamiento (ver figura 2.3).

Por ejemplo, si se tuvieran datos desde 2008 hasta 2024, se podría entrenar y validar en datos de 2008-2022 y testar en datos de 2023-2024. A la hora de hacer la validación se tomaría

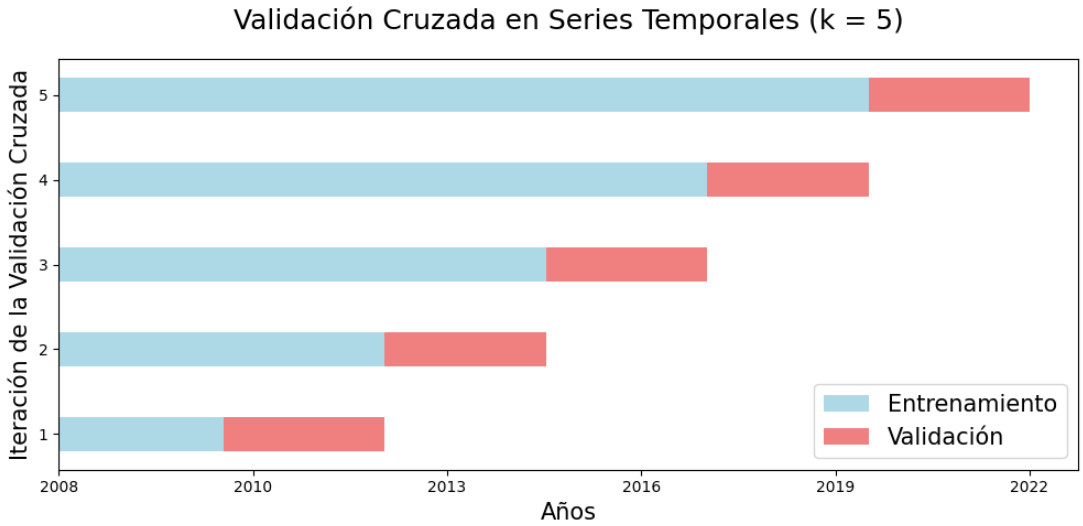


**Figura 2.3:** Esquema de validación cruzada para series temporales.

como primera partición 2008 para entrenar y 2009 para validar, luego la siguiente partición sería 2008-2009 para entrenar y 2010 para validar, la siguiente 2008-2010 para entrenar y 2011 para validar, y así sucesivamente hasta entrenar en 2008-2021 y validar en 2022. Luego se testaría en 2023-2024.

En nuestro caso, creemos que ambos esquemas de validación de series temporales tanto la de [23] como la segunda de [18] son las adecuadas para aplicar en nuestro problema.

## 2.7. Modelos utilizados

Ahora que hemos analizado las diferentes metodologías que se han utilizado en el estado del arte para plantear el problema y estructurar los datos, vamos a comparar los modelos predictivos utilizados en la literatura.

En la revisión de la literatura que hacen en [31] los algoritmos más típicos a la hora de predecir el riesgo de incendio forestal son las redes neuronales [32, 33, 34] (NN), regresión logística [35] y los árboles de decisión [36] y sus derivados como 'Random Forest' [37] (RF). También hablan de que los algoritmos de Deep Learning [38] dan buenos resultados y de que las Redes Neuronales Convolucionales [39] (CNN) se han utilizado mucho. Otro algoritmo que también se ha utilizado es el 'Support Vector Machine' [40] (SVM).

En otra revisión del estado del arte [41] también dicen que las CNN se han utilizado mucho y que otros algoritmos utilizados son la regresión logística, los perceptrones multi-capa [42] (MLP), los 'Bayesian Neural Networks' [43] (BNN) y el algoritmo Long Short-Term Memory [44] (LSTM).

Según nuestra búsqueda bibliográfica, el algoritmo RF es el más popular para esta tarea y además, obtiene muy buenos resultados siendo el mejor algoritmo en muchos artículos. Por ejemplo, en [15] compara los algoritmos NN, SVM y RF y obtiene que el mejor es el último. En [2] también compara RF con MLP y LSTM y obtienen que el primero es el mejor. En [3] también obtienen que el algoritmo RF es el mejor comparándolo con SVM entre otros.

Sin embargo, si comparamos el algoritmo RF con algoritmos más complejos de Deep Learning éste se queda atrás. Se debe a que mediante el aprendizaje profundo se puede introducir contexto espacial, temporal y contexto espacio-temporal. Por ejemplo, en [23] comparan los RF con CNN, LSTM y LSTM convolucionales [45] (ConvLSTM).

Los LSTM se aplican introduciendo datos históricos de las variables del pixel, durante 10 días. Las CNN se aplican introduciendo como número de filtros inicial la cantidad de variables que hay, y las dimensiones de esos filtros siendo $25 \times 25$ (25 celdas espaciales $\times$ 25 celdas espaciales). Es decir, se les introduce contexto espacial. Por último, las ConvCNN se introducen añadiendo contexto espacial como con las CNN y añadiendo contexto temporal como con las LSTM. En la figura 2.4 que proporcionan en [23] se puede ver cómo se introducen los datos en cada algoritmo.

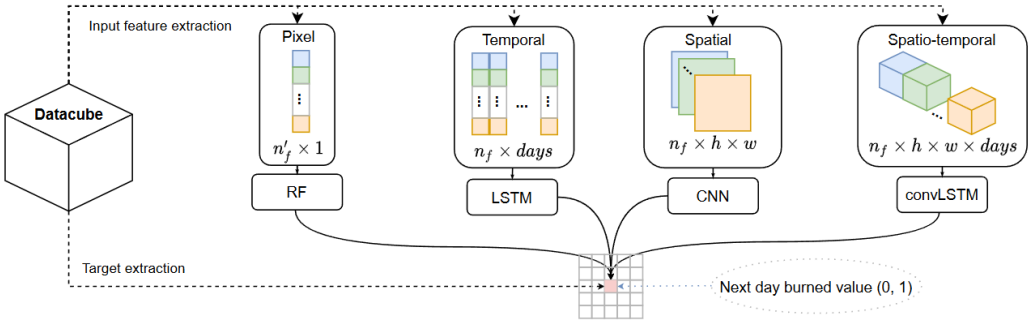


**Figura 2.4:** Estructura de los datos usada para cada algoritmo.

## 2.8. Medidas para calcular la efectividad del modelo

Para concluir con la revisión del estado del arte vamos a analizar qué medidas de efectividad se han utilizado. Al tratar de un problema de clasificación binaria extremadamente desbalanceada, el utilizar la precisión como medida de efectividad sería un error, ya que con un ratio de 'fuego' 'no-fuego' de 1:100.000 (que tienen por ejemplo en [18]), un clasificador trivial tendría una precisión del más del 99.99 %. Por ello se deberían de utilizar otras medidas de efectividad, siempre que estemos testando con una distribución desbalanceada. Como hemos mencionado anteriormente, la mayoría de trabajos testan en conjuntos de datos que no siguen la distribución real, por lo que en algunos de ellos sí que utilizan la precisión como medida de efectividad.

La mayoría de trabajos del estado del arte utilizan las mismas medidas:

- Precisión (para la clase 1):

$$\frac{\text{Número de instancias de clase 1 clasificadas correctamente como 1}}{\text{Número total de instancias clasificadas como clase 1}} \tag{2.1}$$

- Recall (para la clase 1):

$$\frac{\text{Número de instancias de clase 1 clasificadas correctamente como 1}}{\text{Número total de instancias de clase 1}} \tag{2.2}$$

- F1-score: Media harmónica de la precisión y el recall.

$$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

- Sensibilidad: Recall de la clase 1 ('fuego'). De todas las instancias de 'fuego', cuántas ha clasificado correctamente como 'fuego'.
- Especificidad: Recall de la clase 0 ('no-fuego'). De todas las instancias de 'no-fuego', cuántas ha clasificado correctamente como 'no-fuego'.

- AUROC (Area Under the Receiver Operating Curve) [46]: El área que hay debajo de la curva ROC (Receiver Operating Characteristic). La curva ROC es una gráfica que se crea moviendo el umbral de discriminación de un clasificador binario probabilista dibujando la sensibilidad en el eje **Y** y (1 - especificidad) en el eje **X**. El valor del AUROC va desde 0.5 (para un clasificador aleatorio) a 1 (para un clasificador perfecto).

Algunos de los trabajos que utilizan estas medidas de efectividad son por ejemplo [15, 17, 23]. Estos trabajos utilizan la precisión, recall y el F1-score de la clase 'fuegos'. Es decir, miden de todas las instancias clasificadas como 'fuegos', cuantas son realmente 'fuegos'; y de todos los 'fuegos', cuántos se han clasificado como 'fuegos'. Y luego hacen la media harmónica de ambos. Además, en [23] dicen que de las cuatro medidas la más adecuada es la AUROC.

En [18] además de utilizar las medidas anteriormente comentadas, proponen dos medidas específicas para la tarea de predicción de riesgo de incendio forestal. Dichas medidas son combinaciones de la sensibilidad y la especificidad:

$$rhybrid_k = \frac{\text{sensibilidad} \cdot \text{especificidad}}{\text{sensibilidad} + k \cdot \text{especificidad}} \tag{2.4}$$

$$shybrid_k = k \cdot \text{sensibilidad} + \text{especificidad} \tag{2.5}$$

En dicho trabajo tenían como objetivo obtener una sensibilidad de alrededor del 90 % y una especificidad mínima del 50 % en el conjunto de test con proporción real. Es decir, tenían que predecir el 90 % de todos los 'fuegos' correctamente, y tenían que predecir al menos el 50 % de todos los 'no-fuegos' correctamente. Para ello utilizan ambas medidas con diferentes valores de $k$ para seleccionar los mejores hiperparámetros en el proceso de validación y obtienen que utilizando el $\text{shybrid}_5$ consiguen cumplir con la sensibilidad y especificidad deseadas en el conjunto de test.

En nuestro caso, como queremos tener en cuenta el desbalanceo de los datos en el conjunto de test, tal y como lo hacen en [18], utilizar las medidas de efectividad propuestas por ellos como método de selección de hiperparámetros puede ser adecuado.

CAPÍTULO 3

# Metodología y propuesta de investigación

La metodología que seguiermos será la que consideramos más adecuada para obtener resultados fiables a la hora de aplicar en la práctica. Para ello, dividiremos España en cubos espacio-temporales de 1km $\times$ 1km $\times$ 1 día durante 16 años (desde 2008 hasta 2024), y definiremos el problema como un problema de clasificación binaria.

Para definir las instancias de 'fuego', solamente tomaremos los fuegos mayores a 30 hectáreas. Para obtener las instancias de 'no-fuego', generaremos aleatoriamente en el tiempo y espacio instancias que no interfieran en instancias de 'fuego' hasta obtener un conjunto de datos balanceado. Esto lo haremos para el conjunto de entrenamiento, mientras que para la validación utilizaremos un conjunto con 10 veces menos instancias de 'no-fuegos' que la distribución real, y para el conjunto de test seguiremos la distribución real.

En cuanto a las bases de datos, utilizaremos ERA5-Land, Corine Land Cover, datos de la NASA para los índices de vegetación, Copernicus DEM, y todas las variables relacionadas con los humanos. Luego haremos una selección de variables para entrenar los modelos.

Para validar los modelos, utilizaremos el método de validación cruzada de series temporales (ver figura 2.3); y los modelos que utilizaremos serán RF, XGBoost, CNN, LSTM y ConvLSTM.

Para terminar, las medidas de efectividad que utilizaremos serán la sensibilidad, la especificidad, el AUROC y los $\text{rhybrid}_k$ y $\text{shybrid}_k$.

Lo que nosotros proponemos es combinar las diferentes metodologías y modelos utilizados en el estado del arte para generar un predictor fiable entrenado en una cantidad de datos (500.000 celdas espaciales durante 16 años) con los que, hasta donde sabemos, no se ha trabajado. Además, el análisis se realizará en España, donde no hay muchos trabajos recientes que apliquen el Deep Learning para esta tarea.

CAPÍTULO 4

# Resultados experimentales

CAPÍTULO 5

# Conclusiones y trabajo futuro

En este trabajo hemos visto cómo se puede crear un predictor diario de riesgo de incendio forestal que asegure su efectividad en la práctica.

Como trabajo futuro, ya que los incendios forestales son eventos muy poco comunes, se podría plantear como un problema de 'outlier detection'.

# Bibliografía


[1] Amparo Alonso-Betanzos, Oscar Fontenla-Romero, Bertha Guijarro-Berdiñas, Elena Hernández-Pereira, Juan Canda, Ezequiel Jiménez, José Luis Legido Soto, Susana Muñiz, Cristina Paz-Andrade, and Maria Paz-Andrade. A Neural Network Approach for Forestal Fire Risk Estimation. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI 2002)*, Lyon, France, January 2002. Ver páginas 1, 6, and 8.

[2] Ashima Malik, Megha Rajam Rao, Nandini Puppala, Prathusha Koouri, Venkata Anil Kumar Thota, Qiao Liu, Sen Chiao, and Jerry Gao. Data-driven wildfire risk prediction in northern california. *Atmosphere*, 12(1), January 2021. Ver páginas 1, 5, 6, 8, 9, 10, and 12.

[3] Marcos Rodrigues and Juan de la Riva. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, 57:192–201, July 2014. Ver páginas 1, 6, and 12.

[4] Sandra Oliveira, Friderike Oehler, Jesus San-Miguel-Ayanz, Andrea Camia, and Jose M. C. Pereira. Modeling spatial patterns of fire occurrence in mediterranean europe using multiple regression and random forest. *Forest Ecology and Management*, 275:117–129, July 1 2012. Ver páginas 1, 5, 6, 9, and 10.

[5] Francisco Javier de Vicente y López. *Diseño de un modelo de riesgo integral de incendios forestales mediante técnicas multicriterio y su automatización en sistemas de información geográfica. Una aplicación en la comunidad valenciana.* PhD thesis, Universidad Politécnica de Madrid, 2012. Ver página 1.

[6] European Commission. European forest fire information system (effis). Link, 2024. Consultado el 25 de diciembre de 2024. Ver páginas 1, 5.

[7] Europa Press. Los incendios forestales dejan 127 muertos en lo que va de siglo en españa. Link, 2022. Consultado el 25 de diciembre de 2024. Ver página 1.

[8] La Sexta. El coste económico directo de los incendios en españa supera los 2.000 millones de euros. Link, 2022. Consultado el 25 de diciembre de 2024. Ver página 1.

[9] Proyecto GAIA. Iniciativa para la gestión integral de incendios forestales. Link, 2024. Consultado el 25 de diciembre de 2024. Ver página 1.

[10] Cyber Human Systems. Gaia: Progresando en la gestión completa de incendios forestales. Link, 2024. Consultado el 25 de diciembre de 2024. Ver página 1.

[11] Jesus San-Miguel-Ayanz and Andrea Camia. Forest Fires at a Glance: Facts, Figures and Trends in the EU, 2009. Ver página 2.

[12] Mohsen Naderpour, Hossein Mojaddadi Rizeei, and Fahimeh Ramezani. Forest fire risk prediction: A spatial deep neural network-based framework. *Remote Sensing*, 13(13), July 2021. Ver páginas 2, 9.

[13] Helena Liz-Lopez, Javier Huertas-Tato, Jorge Perez-Aracil, Carlos Casanova-Mateo, Julia Sanz-Justo, and David Camacho. Spain on fire: A novel wildfire risk assessment model based on image satellite processing and atmospheric information. *Knowledge-Based Systems*, 283, January 11 2024. Ver página 2.

[14] R. Vélez Muñoz. Las quemas incontroladas como causa de incendios forestales. *Cuadernos de la Sociedad Española de Ciencias Forestales*, (9), June 2000. Ver páginas 2, 3.

[15] Yongqi Pang, Yudong Li, Zhongke Feng, Zemin Feng, Ziyu Zhao, Shilin Chen, and Hanyue Zhang. Forest fire occurrence prediction in china based on machine learning methods. *Remote Sensing*, 14(21), November 2022. Ver páginas 3, 9, 12, and 14.

[16] Slobodan Milanovic, Nenad Markovic, Dragan Pamucar, Ljubomir Gigovic, Pavle Kostic, and Sladjan D. Milanovic. Forest fire probability mapping in eastern serbia: Logistic regression versus random forest method. *Forests*, 12(1), January 2021. Ver páginas 3, 9, and 10.

[17] Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, Maria Piles, Miguel-Angel Fernandez-Torres, and Nuno Carvalhais. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17), September 16 2022. Ver páginas 3, 5, 6, 9, 11, and 14.

[18] Alexis Apostolakis, Stella Girtsou, Giorgos Giannopoulos, Nikolaos S. Bartsotas, and Charalampos Kontoes. Estimating next day's forest fire risk via a complete machine learning methodology. *Remote Sensing*, 14(5), March 2022. Ver páginas 4, 5, 6, 8, 9, 11, 12, 13, and 14.

[19] Xavier Úbeda, Joaquim Farguell Pérez, Marcos Francos, and Jorge Mataix Solera. Los grandes incendios forestales y sus consecuencias en el suelo. In *Geografía: cambios, retos y adaptación: libro de actas. XVIII Congreso de la Asociación Española de Geografía, Logroño, 12 al 14 de septiembre de 2023, 2023, ISBN 978-84-09-53925-3, págs. 87-96*, pages 87–96. Asociación Española de Geografía, 2023. Ver página 4.

[20] Raúl Quílez Moraga. *Prevención de megaincendios forestales mediante el diseño de planes de operaciones de extinción basados en nodos de propagación*. PhD thesis, Universidad de León, 2016. Consultado en Dialnet. Ver página 4.

[21] C. E. Van Wagner. Structure of the canadian forest fire weather index, 1974. Ver página 5.

[22] Paul R. Cohen, Michael L. Greenberg, David M. Hart, and Adele E. Howe. Trial by Fire: Understanding the Design Requirements for Agents in Complex Environments. *AI Magazine*, 10(3):32–48, September 1989. Ver página 5.

[23] Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, María Piles, Miguel-Ángel Fernández-Torres, and Nuno Carvalhais. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17):e2022GL099368, 2022. Ver páginas 5, 6, 7, 8, 9, 11, 12, and 14.

[24] Paulo Cortez and Anıbal Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In *New Trends in Artificial Intelligence: Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007)*, pages 512–523, Guimarães, Portugal, January 2007. Associação Portuguesa para a Inteligência Artificial (APPIA). Ver página 6.

[25] Jason Brownlee. Smote oversampling for imbalanced classification with python. Machine Learning Mastery, January 2020. Ver página 8.

[26] J. Muñoz Sabater. ERA5-Land hourly data from 1950 to present. Link, 2019. Consultado el 29 de diciembre de 2024. Ver página 9.

[27] Copernicus Land Monitoring Service. CORINE Land Cover. Link. Consultado el 29 de diciembre de 2024. Ver página 9.

[28] European Space Agency. Copernicus global digital elevation model. Link, 2024. Distribuido por OpenTopography, consultado el 29 de diciembre de 2024. Ver página 9.

[29] OpenStreetMap contributors. Openstreetmap. Link, 2024. Consultado el 29 de diciembre de 2024. Ver página 9.

[30] WorldPop. Worldpop project. Link, 2024. Consultado el 29 de diciembre de 2024. Ver página 9.

[31] Faroudja Abid. A survey of machine learning algorithms based forest fires prediction and detection systems. *Fire Technology*, 57(2):559–590, 2021. Ver página 12.

[32] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. Ver página 12.

[33] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. Ver página 12.

[34] Gérard Dreyfus. *Neural Networks: Methodology and Applications*. Springer, Berlin, Heidelberg, 2005. Ver página 12.

[35] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958. Ver página 12.

[36] Leo Breiman, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, New York, 1984. Ver página 12.

[37] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. Ver página 12.

[38] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017. Ver página 12.

[39] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324. IEEE, 1998. Ver página 12.

[40] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. Ver página 12.

[41] Ramez Alkhatib, Wahib Sahwan, Anas Alkhatieb, and Brigitta Schütt. A brief review of machine learning algorithms in forest fires science. *Applied Sciences*, 13(14):8275, 2023. Ver página 12.

[42] Frank Rosenblatt. *The perceptron: a probabilistic model for information storage and organization in the brain*, volume 65. American Psychological Association, 1958. Ver página 12.

[43] Radford M Neal. Bayesian learning for neural networks. *Lecture Notes in Statistics*, 118, 2012. Ver página 12.

[44] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. Ver página 12.

[45] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 802–810, 2015. Ver página 12.

[46] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. Ver página 14.