

FireGroundAI

Un modelo de inteligencia artificial para predecir la severidad de un incendio forestal

UOC

Patricia Luengo Carretero

Area 5: Data Science in
Complex Systems,
Sustainability and Ecology

Tutor de TFM

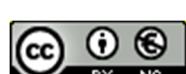
Bernat Bas Pujols

**Profesora responsable de
la asignatura**

Susana Acedo Nadal

16/01/2025

Universitat Oberta
de Catalunya



Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

Ficha del Trabajo Final

Título del trabajo:	FireGroundAI. Un modelo de inteligencia artificial para predecir la extensión de la superficie quemada en un incendio forestal.
Nombre del autor/a:	Patricia Luengo Carretero
Nombre del Tutor/a de TF:	Bernat Bas Pujols
Nombre del/de la PRA:	Silvia Acedo Nadal
Fecha de entrega:	01/2025
Titulación o programa:	Máster Ciencia de Datos
Área del Trabajo Final:	Area 5: <i>Data Science in Complex Systems, Sustainability and Ecology</i>
Idioma del trabajo:	Castellano
Palabras clave	Incendios forestales, <i>machine learning, gradient boosting</i>
Resumen del Trabajo	
El objetivo principal de este trabajo es desarrollar un modelo de inteligencia artificial que prediga la severidad de un incendio forestal a partir de unas condiciones de inicio del incendio. Para ello, se utilizó la base de datos del EGIF (Estadística General de Incendios Forestales), que incluye datos meteorológicos, topográficos, temporales, de vegetación, factores humanos y otros elementos relevantes. Además, se obtuvieron distintas variables derivadas, como la densidad de población o la densidad de incendios en la zona y se extrajo el FWI (<i>Fire Weather Index</i>) para complementar la base de datos y mejorar la capacidad predictiva del modelo. El análisis permitió identificar patrones relevantes para discernir entre las distintas clases de incendios: conato, incendio y gran incendio forestal (GIF).	
Utilizando técnicas de <i>machine learning</i> , se entrenaron diferentes modelos con la finalidad de encontrar el que mejor se adaptara al objetivo. Los modelos empleados fueron <i>Random Forest, CatBoost, XGBoost</i> y <i>LightGBM</i> . Los	

resultados obtenidos muestran que, en general, todos los modelos tienen dificultades para predecir la clase minoritaria, es decir, los grandes incendios forestales (GIFs), debido a su baja representación en la base de datos.

En cuanto a los resultados, *CatBoost* ha demostrado ser el modelo con el mejor rendimiento global para predecir los GIFs, logrando un equilibrio moderado entre *precision* (0,73) y *recall* (0,80) tras la validación cruzada. Este modelo podría ser el punto de partida para desarrollar herramientas más avanzadas que mejoren la predicción de los GIFs, lo que a su vez facilitaría la implementación de estrategias más efectivas en la prevención y gestión de incendios forestales.

Abstract

The main objective of this work is to develop an artificial intelligence model to predict the severity of a forest fire based on the initial conditions of the fire. For this, the EGIF (General Statistics of Forest Fires) database was used, which includes meteorological, topographic, temporal, vegetation, human factors, and other relevant elements. Additionally, various derived variables, such as population density and fire density in the area, were obtained, and the FWI (Fire Weather Index) was extracted to complement the database and improve the predictive capacity of the model. The analysis of these data allowed the identification of relevant patterns to distinguish between the different classes of fires: incipient fire, fire, and large forest fire (GIF).

Using machine learning techniques, different models were trained to find the one that best fits the objective. The models used were Random Forest, CatBoost, XGBoost, and LightGBM. The results show that, in general, all models have difficulty predicting the minority class of large forest fires (GIFs) due to their low representation in the database.

Regarding the results, CatBoost has proven to be the model with the best overall performance in predicting GIFs, achieving a well-balanced combination of precision (0.73) and recall (0.80) after cross-validation. This model could serve as a starting point for developing more advanced tools that improve GIF prediction, which in turn would facilitate the implementation of more effective strategies in forest fire prevention and management.

Index

1. Introducción	1
1.1. Contexto y justificación del Trabajo	1
1.2. Motivación	2
1.3. Objetivos del Trabajo	3
1.4. Impacto en sostenibilidad, ético-social y de diversidad	3
1.5. Enfoque y método seguido	4
1.6. Planificación del trabajo	5
1. Estado del arte	10
2.1. Análisis de las principales variables que intervienen en los incendios forestales	10
2.2. Análisis de los principales modelos de predicción utilizados	12
2.3. Conclusiones del estado del arte	19
3. Materiales y métodos.....	22
3.1. Descripción de la base de datos del EGIF	22
3.2. Tratamiento de la base de datos del EGIF	23
3.3. Preprocesamiento de los datos	26
3.4. Análisis Exploratorio de Datos	31
3.4.1. Análisis Univariante	33
3.4.2. Análisis Multivariante	41
3.5. Conclusiones del EDA	46
4. Resultados.....	48
4.1. Selección de características	48
4.2. Pruebas preliminares con <i>Random Forest</i>	49
4.3. Submuestreo de las clases conato e incendio	50
4.4. Sobremuestreo de la clase minoritaria GIF	51
4.5. Estrategia híbrida: submuestreo clases conato e incendio y sobremuestreo clase GIF	52
4.6. Modelos	54
4.7. Desempeño de <i>CatBoost</i> con los incendios de 2017	58
5. Conclusiones y trabajos futuros.....	59
6. Referencias.....	61

Listado de Figuras

Ilustración 1. Superficie afectada en hectáreas y el número de GIFs Anual (autoría propia)	1
Ilustración 2. Planificación. Fases 1 y 2	8
Ilustración 3. Planificación. Fase 3.....	8
Ilustración 4. Planificación. Fases 4, 5 y 6	9
Ilustración 5. Esquema visual estado del arte	21
Ilustración 6. Relación de tablas en la base de datos del EGIF	22
Ilustración 7. Discrepancias en las coordenadas geográficas.	27
Ilustración 8. Registros de incendios forestales que carecen de coordenadas geográficas.....	27
Ilustración 9. Sistema de Hoja (izquierda) y Cuadrícula Militar (derecha).....	28
Ilustración 10. Porcentaje de datos faltantes en la base de datos de la EGIF	29
Ilustración 11. Porcentaje de datos faltantes después de la incorporación de los datos de AEMET.....	32
Ilustración 12. Porcentaje de datos faltantes en función de la clase de incendio.....	32
Ilustración 13. Visualización de los datos atípicos de las variables numéricas	34
Ilustración 14. Visualización de las distribuciones de las variables numéricas	35
Ilustración 15. Visualización de las distribuciones de las variables numéricas por clase de incendio.....	36
Ilustración 16. Visualización de las distribuciones de las variables numéricas en clase GIF	37
Ilustración 17. Visualización de las ocurrencias de las variables categóricas.....	38
Ilustración 18. Variables categóricas en función de la clase de incendio.....	39
Ilustración 19. Variables categóricas en la clase GIF	40
Ilustración 20. Distribución del número de ocurrencias por clase de incendio	41
Ilustración 21. Matriz de correlación de Spearman.....	42
Ilustración 22. Test Chi-cuadrado para las variables categóricas.....	44
Ilustración 23. Visualización de las variables categóricas en función de la clase de incendio.....	45
Ilustración 24. Selección de características con Random Forest.....	49
Ilustración 25. Gráficos comparativos del número de registros en el dataset desbalanceado, submuestreado y sobremuestreado	50
Ilustración 26. Gráfico con el número de registros por clase después de realizar el submuestreo con ENN y Tomek Links.....	51
Ilustración 27. Curvas de precisión en test y train para CatBoost y Curvas ROC para cada clase de incendio.....	57

Listas de Tablas

Tabla 1. Resumen de estudios sobre Modelos de Predicción de Incendios Forestales.....	17
Tabla 2. Relación de combinaciones de los tipos de combustibles	24
Tabla 3. Relación de combinaciones del lugar de inicio del incendio	25
Tabla 4. Relación de combinaciones de los tipos de fuego	26
Tabla 5. Descriptivos básicos de las variables numéricas.....	33
Tabla 6. Selección de características. Resultados de una Regresión Logística Multiclasa.....	48
Tabla 7. Selección de características. Resultados de Random Forest.....	48
Tabla 8. Resultados de Random Forest: desbalanceado, submuestreado y sobremuestreado	49
Tabla 9. Resultados de Random Forest con diferentes técnicas de submuestreo de las clases mayoritarias conato e incendio	50
Tabla 10. Resultados de Random Forest con diferentes técnicas de sobremuestreo de la clase GIF	52
Tabla 11. Submuestreo aleatorio de las clases conato e incendio	53
Tabla 12. Resultados de Random Forest con estrategia combinada de submuestreo de las clases conato e incendio y sobremuestreo de la clase GIF	53
Tabla 13. Comparativa de distintos modelos utilizando el dataset balanceado	54
Tabla 14. Resultados de Random Forest optimizado con Optuna.....	55
Tabla 15. Resultados de XGBoost optimizado con Optuna.....	55
Tabla 16. Resultados de LightGBM optimizado con Optuna	56
Tabla 17. Resultados de CatBoost optimizado con Optuna	56
Tabla 18. Resultados de CatBoost tras validación cruzada.....	57
Tabla 19. Resultados de CatBoost con los incendios de 2017	58

1. Introducción

1.1. Contexto y justificación del Trabajo

En los últimos años se ha vuelto a experimentar un incremento en la frecuencia y severidad de los incendios forestales en España. Si observamos la gráfica que muestra la superficie afectada en ha. y el número de GIFs (Grandes Incendios Forestales) anual, notamos que en el 2022 se experimenta un pico de 61 incendios con una superficie afectada de 267.940 ha. Esto es debido a distintas causas como el aumento de las temperaturas, la despoblación de las zonas rurales, el descenso de la actividad agrícola y ganadera, las políticas de prevención, o la regulación del empleo de bombero forestal entre otras [1].

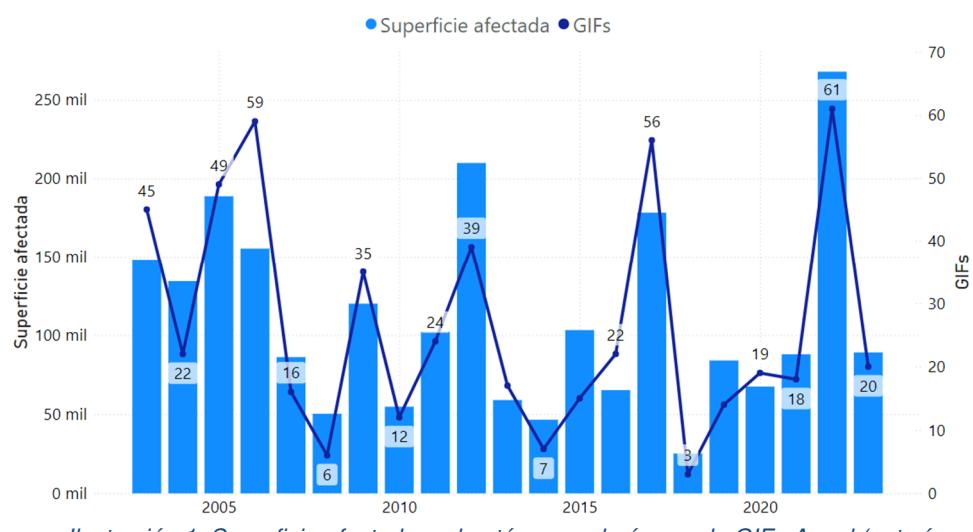


Ilustración 1. Superficie afectada en hectáreas y el número de GIFs Anual (autoría propia)

A su vez, el interés por incorporar el uso de la inteligencia artificial en el análisis y la predicción de los distintos factores que intervienen en un incendio forestal ha crecido significativamente [2].

El objetivo de esta propuesta de TFM es la de desarrollar un modelo de inteligencia artificial que mediante características meteorológicas, geográficas, temporales, tipología de vegetación, intervención humana obtenidas de la base de datos del EGIF (Estadística General de Incendios Forestales) [3] y de otras fuentes permita predecir la severidad de un incendio forestal.

Se entrenarán distintos modelos predictivos empleando diferentes técnicas de *machine learning*, desde soluciones más generales y conservadoras como los modelos de regresión lineal con datos tabulares al uso del *deep learning* con redes neuronales.

El resultado final será *FireGroundAI*. Un modelo de inteligencia artificial capaz de predecir la severidad que pueda alcanzar un incendio y así contribuir a las labores de contención y extinción de los incendios forestales reduciendo los daños económicos, ambientales y humanos.

1.2. Motivación

Imagina que de repente el aire se torna pesado, ya no escuchas los pájaros pero sí un crepitar. Caen destellos de luz desde tu copa mientras asistes desde un plano subjetivo a la destrucción de todo lo que hay a tu alrededor.

Imagina que ahora que ya no estás en peligro de extinción pero todavía eres una especie vulnerable tienes que huir con tus crías ante la presencia inminente de las llamas. Tu instinto de supervivencia se activa, pero sabes que no podrás salvarlas a todas.

Imagina que llevas toda tu vida viviendo y trabajando en un entorno natural que de la noche a la mañana se ve amenazado por un incendio forestal. A pesar de las altas temperaturas un escalofrío recorre tu cuerpo ante la incertidumbre de no saber qué va a suceder durante las próximas horas, días, semanas.

Imagina que es tu primer día en extinción de incendios. No sabes exactamente por qué estás aquí, pero después de asegurar unas casas con tus compañeros ya sabes que aún a pesar de la crudeza no querrás dejar este trabajo jamás.

Imagina que ya casi no recuerdas cuando empezaste en esto, otro incendio más y todas las miradas puestas en ti y en tu vasta experiencia. Lo que nadie sabe es que aun así necesitas mentalizarte para la devastación y la pérdida.

Imagina que un día amaneces con que el lugar más bello del mundo para ti está siendo pasto de las llamas y te encuentras a cientos de kilómetros de allí, impotente, sin poder hacer nada.

Pero ¿y si no tuviera que ser así? ¿Qué sucedería si tuvieras los datos, las herramientas y la motivación para poder hacer algo?

Este trabajo tiene la pretensión de transformar todas esas pérdidas en aprendizaje para contribuir a la extinción de incendios forestales a través del uso de la ciencia de datos; porque para que algo renazca primero debemos haberlo dejado arder.

1.3. Objetivos del Trabajo

Objetivo principal. Predecir la severidad que podría alcanzar un incendio forestal, conato, incendio o gran incendio forestal, a partir de unos indicadores dados en el inicio del incendio basados en condiciones meteorológicas, características geográficas, topográficas, poblacionales y de tipo de vegetación.

Objetivos secundarios:

- **Enriquecer y consolidar la base de datos**, con datos provenientes de otras fuentes adicionales (meteorológicas, geográficas, poblacionales) o mediante transformaciones de las variables que ya contiene que permitan mejorar la aplicabilidad al modelo.
- **Análisis de los principales factores** que influyen en la severidad alcanzada en un incendio forestal.
- **Generación y comparación** de distintos modelos.

1.4. Impacto en sostenibilidad, ético-social y de diversidad

Los impactos que el siguiente TFM tendrá en cuanto a las dimensiones de la competencia transversal UOC “Compromiso ético y global” serán los siguientes:

Impacto en sostenibilidad. El proyecto tiene un impacto positivo con respecto a la protección del medio ambiente ya que una extinción temprana de un incendio forestal permite optimizar los recursos y reducir los daños ecológicos y materiales colaborando con el **ODS 15: vida de ecosistemas terrestres**.

Mejorará la seguridad de las zonas rurales afectadas por los incendios forestales minimizando la respuesta de los equipos de emergencias, lo que se alinea con el **ODS 11: ciudades y comunidades sostenibles**.

Además, el trabajo también contribuye con el **ODS 13: acción por el clima** ya que la severidad y recurrencia de los incendios forestales se han visto condicionados en parte por el cambio climático.

Impacto ético y responsabilidad social. El modelo predictivo que se pretende desarrollar en este trabajo tiene un profundo componente de responsabilidad social contribuyendo al bien común como se menciona en el **ODS 16: Paz, justicia e instituciones sólidas** ya que será diseñado para ser utilizado de forma equitativa por los diferentes actores relacionados con la prevención y extinción de incendios forestales sin favorecer el interés privado.

Diversidad y derechos humanos. FireGroundAI es una solución que se desarrolla para todos los usuarios y comunidades de forma ecuánime, independientemente de su situación geográfica lo que se vincula con el **ODS 10: reducción de las desigualdades** ya que se asegura que tanto recursos como tecnología estén disponibles en aquellas zonas rurales donde más afectadas se encuentran por los incendios forestales.

1.5. Enfoque y método seguido

La metodología para realizar el proyecto se divide en las siguientes fases:

1. Investigación previa y revisión del estado del arte

En primer lugar se llevará a cabo una investigación en torno a los últimos estudios en cuanto a predicción de incendios forestales utilizando inteligencia artificial, revisando artículos académicos, revistas especializadas y proyectos relevantes.

2. Consolidación de la base de datos:

La base de datos principal tiene su origen en la Estadística General de Incendios Forestales (EGIF) que se complementará con otras fuentes adicionales. Para consolidar la base de datos se realizarán las siguientes acciones:

- **Localización.** La base de datos presenta distintas formas de localización del punto de inicio del incendio, mediante coordenadas UTM, cuadrícula sobre mapa militar o mediante el nombre de las entidades de población proporcionado por el IGN (Instituto Geográfico Nacional). Se unificarán en latitud y longitud.
- **Altitud, Superficie, Población.** Se obtendrán a partir de las coordenadas de la localización del incendio y el núcleo de población más cercano con datos proporcionados por el IGN.
- **Índice FWI.** A partir de la extracción de datos de *Climate Data Store*.
- **Densidad de incendios** en la zona en función de una ventana de cinco años mediante cálculos propios.
- **Densidad de población histórico** mediante cálculos propios.
- **Diversas variables meteorológicas.** Mediante la API de AEMET.

3. Análisis exploratorio de datos

Una vez consolidada la base de datos, se procederá a realizar un análisis exploratorio donde se profundizará en la estructura de los datos y en la detección de valores atípicos, valores faltantes o nulos mediante análisis gráfico y medidas estadísticas, además de realizar una evaluación de la distribución de las principales variables.

4. Selección de características

Se llevará a cabo una Regresión Logística Multiclasa y se implementará un modelo de *Random Forest* para evaluar la importancia de las características. Con base en estos resultados, se seleccionarán las variables más relevantes para reducir la dimensionalidad y mejorar la eficiencia del modelo.

5. Modelos predictivos

Se emplearán distintos enfoques predictivos, comenzando con soluciones más sencillas, como la Regresión Logística Multiclasa para datos tabulares, y avanzando hacia técnicas más sofisticadas, incluyendo árboles de decisión, *Random Forest*, modelos de *Gradient Boosting*, así como redes neuronales para capturar patrones más complejos en los datos.

6. Validación del modelo

Validar el modelo predictivo generado con datos de incendios forestales procedentes de distintas ubicaciones y contextos para evaluar el desempeño del modelo.

El modelo será validado utilizando datos de incendios forestales correspondientes al año 2017, que serán excluidos del proceso de entrenamiento. Esto permitirá evaluar su desempeño en un escenario real evaluando cómo el modelo generaliza a datos no vistos.

1.6. Planificación del trabajo

Para realizar el proyecto serán necesarios diferentes recursos de software:

- Access y SQL para trabajar directamente son la base de datos de la EGIF.
- *Python*, será el lenguaje principal para realizar el procesamiento, análisis y modelado de los datos. Se utilizarán diferentes librerías como *pandas*, *numpy*, *tensorflow*, *keras*, *seaborn*, *scikit-learn*, etc.
- Entorno en *Colaboratory* con el uso de *Jupyter Notebooks*
- Se usarán APIs externas AEMET para extraer datos meteorológicos históricos
- *Google Drive* para la gestión de datos y almacenamiento.
- *ChatGPT* como asistente para mejorar las distintas partes del proyecto.
- *GitHub*, como repositorio de los *notebooks*, modelos y *dataset*.

La planificación se llevará a cabo en las siguientes fases:

1. Definición del TFM (del 25 de septiembre al 13 de octubre de 2024)

- **Día 1-3:** Definir el problema a resolver. Objetivo principal y secundarios.
- **Día 4-5:** Especificar las fuentes de datos y APIs a utilizar.
- **Día 6-7:** Estructurar la metodología.
- **Día 8-10:** Redactar el borrador inicial.
- **Día 11-14:** Desarrollar la propuesta y las motivaciones personales.
- **Día 15-16:** Realizar la planificación.
- **Día 17-18:** Revisión y ajustes según *feedback* del tutor.
- **Día 19-20:** Redactar el documento final y enviar la propuesta del TFM.

2. Estado del arte (del 14 de octubre al 3 de noviembre de 2024)

Semana 1 (del 14 al 20 de octubre):

- **Día 1-2:** Búsqueda de artículos científicos y referencias relevantes
- **Día 3-4:** Selección de las fuentes clave que están relacionadas con los incendios forestales y el uso generalizado de la IA en la prevención y extinción de incendios forestales.
- **Día 5-7:** Revisión de trabajos previos realizados sobre el uso de *machine learning* y redes neuronales para la predicción de la superficie quemada en incendios forestales.

Semana 2 (del 21 de octubre al 27 de octubre):

- **Día 1-2:** Perfilar el software que se utilizará.
- **Día 3-4:** Seleccionar las fuentes de datos.
- **Día 5-6:** Revisar el contexto, la justificación de la propuesta y la metodología en función de la investigación realizada.
- **Día 7:** Redacción del estado del arte preliminar.

Semana 3 (del 21 de octubre al 3 de noviembre):

- **Día 1-2:** Consolidación de las fuentes bibliográficas.
- **Día 3-5:** Revisión y corrección del documento.
- **Día 6-7:** Entrega del estado del arte.

3. Implementación (del 4 de noviembre al 15 de diciembre de 2024)

Semana 1 (4 al 10 de noviembre):

- **Día 1-2:** Configuración del entorno de desarrollo y herramientas necesarias (*Python*, bibliotecas de *machine learning*, APIs de datos).
- **Día 3-5:** Importación y exploración inicial de los datos (datos del EGIF y otras fuentes). Análisis preliminar para identificar inconsistencias y patrones básicos.
- **Día 6-7:** Preparación y limpieza de los datos (detección de valores nulos, *outliers*, duplicados).

Semana 2 (11 al 17 de noviembre):

- **Día 1-3:** Transformación de los datos tabulares, consolidación de la base de datos y unificación de coordenadas (latitud y longitud).
- **Día 4-5:** Obtención de características adicionales (altitud, índice de recurrencia de incendios, densidad de población).
- **Día 6-7:** Extracción de datos meteorológicos de AEMET.

Semana 3 (18 al 24 de noviembre):

- **Día 1-3:** Extracción del índice FWI.
- **Día 4-5:** Análisis exploratorio de datos (gráficos, medidas estadísticas, correlaciones).
- **Día 6-7:** Selección de características con una Regresión Logística Multiclasa y un Random Forest.

Semana 4 (25 de noviembre al 1 de diciembre):

- **Día 1-3:** Implementación de modelos básicos (árboles de decisión y redes neuronales). Pruebas iniciales con datos tabulares.

- **Día 4-5:** Comparación de los mejores modelos utilizando métricas clave como *accuracy*, *recall*, *f1-score* y matriz de confusión.
- **Día 6-7:** Elección del mejor modelo y ajuste inicial de hiperparámetros con *Optuna*.

Semana 5 (2 al 8 de diciembre):

- **Día 1-3:** Implementación de técnicas de validación cruzada para evaluar la robustez del modelo seleccionado.
- **Día 4-5:** Evaluación preliminar del desempeño de los modelos con datos reales no vistos por el modelo con anterioridad.
- **Día 6-7:** Ajustes finales y optimización del mejor modelo seleccionado.

Semana 6 (9 al 15 de diciembre):

- **Día 1-2:** Consolidación de resultados y preparación de informes.
- **Día 3-6:** Subida de todos los documentos, notebooks y otros archivos relevantes al repositorio de GitHub.
- **Día 7:** Entrega de la implementación y documentación final.

4. Redacción de la memoria (del 16 de diciembre al 29 de diciembre de 2024)

Semana 1 (del 16 al 22 de diciembre) - Entrega preliminar: 22 de diciembre de 2024:

- **Día 1-2:** Estructurar el documento (Índice, Introducción, Estado del arte).
- **Día 3-4:** Redactar la metodología (Describir el proceso de consolidación de la base de datos, análisis exploratorio, selección de características y modelos).
- **Día 5-6:** Redactar los resultados obtenidos (Evaluación de los modelos).
- **Día 7:** Revisión preliminar del documento.

Semana 2 (del 23 al 29 de diciembre) - Entrega final: 29 de diciembre de 2024:

- **Día 1-2:** Redacción de conclusiones y trabajo futuro.
- **Día 3:** Formato final del documento (revisión de estilo, ortografía, coherencia).
- **Día 4-5:** Ajustes a partir del *feedback* final del tutor.
- **Día 6-7:** Edición final del documento.

5. Presentación audiovisual del trabajo (del 30 de diciembre al 7 de enero de 2025)

- **Día 1-2:** Redacción del guion de la presentación.
- **Día 3-5:** Creación de diapositivas.
- **Día 6-7:** Ensayo y grabación de la presentación en video.
- **Día 8-9:** Entrega final de la presentación audiovisual.

6. Entrega de la documentación al tribunal y defensa pública del trabajo (del 8 de enero al 16 de enero de 2025)

- **Día 1-2:** Revisión final del TFM y formato de la documentación.
- **Día 3:** Entrega de la documentación.
- **Día 5-8:** Preparación de la defensa.
- **Día 9 (16 de enero):** Defensa.

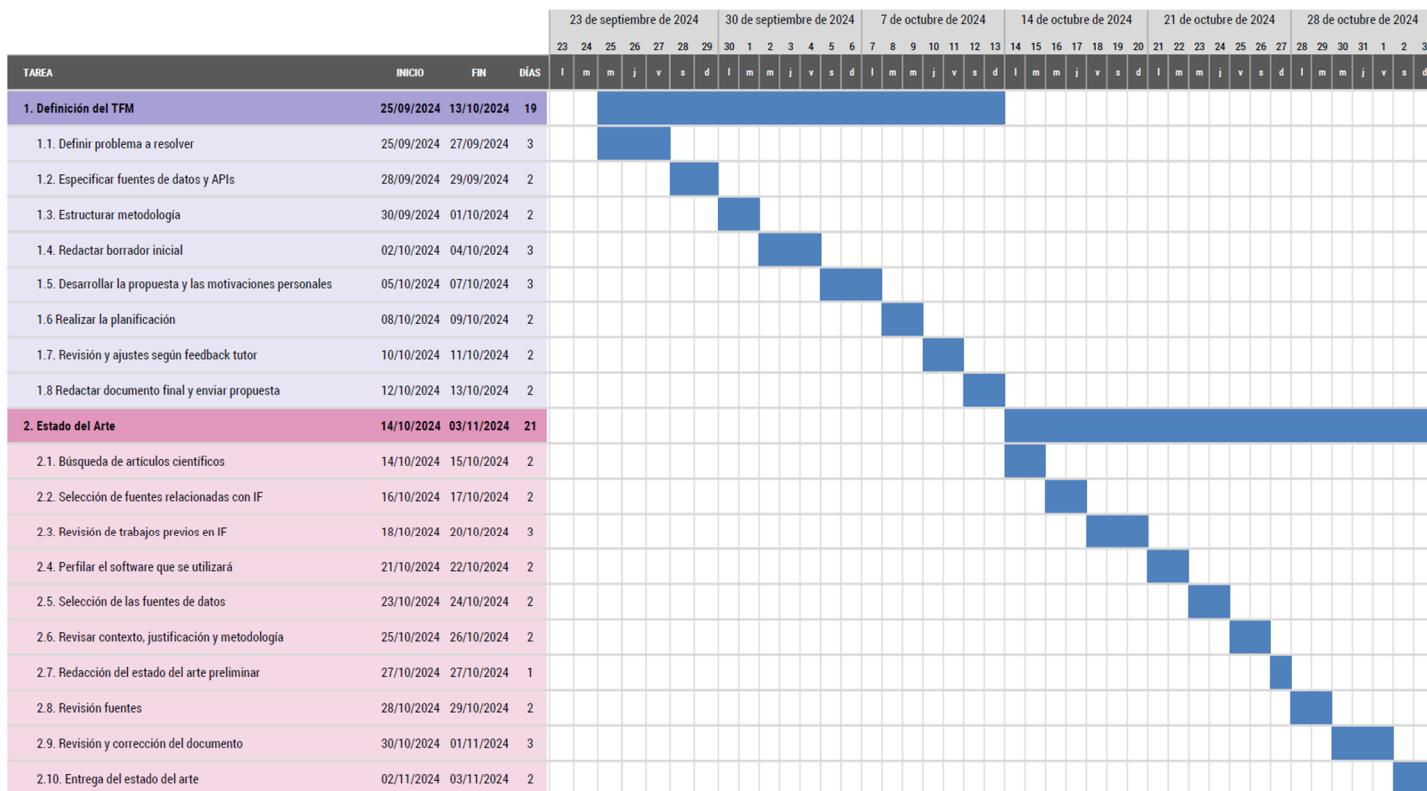


Ilustración 2. Planificación. Fases 1 y 2

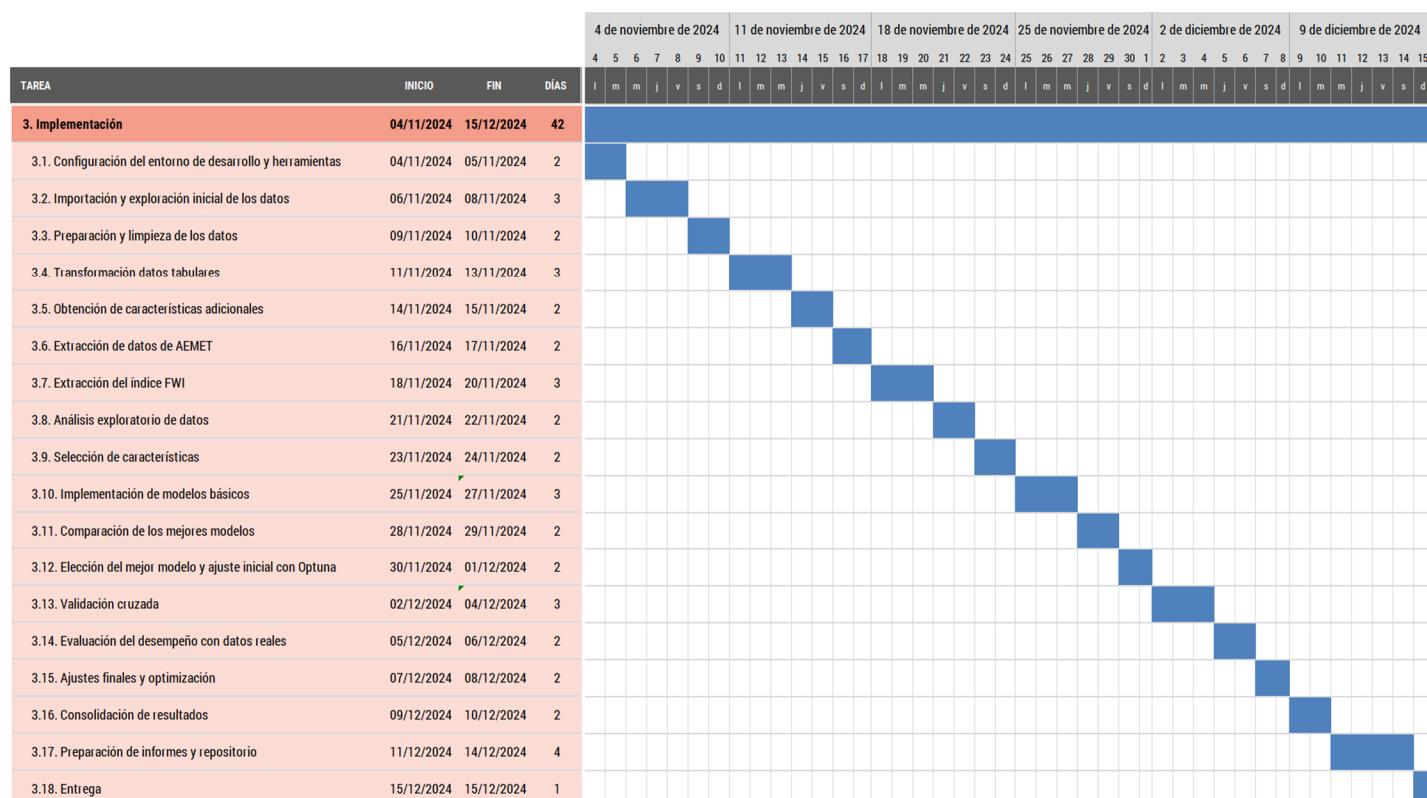


Ilustración 3. Planificación. Fase 3

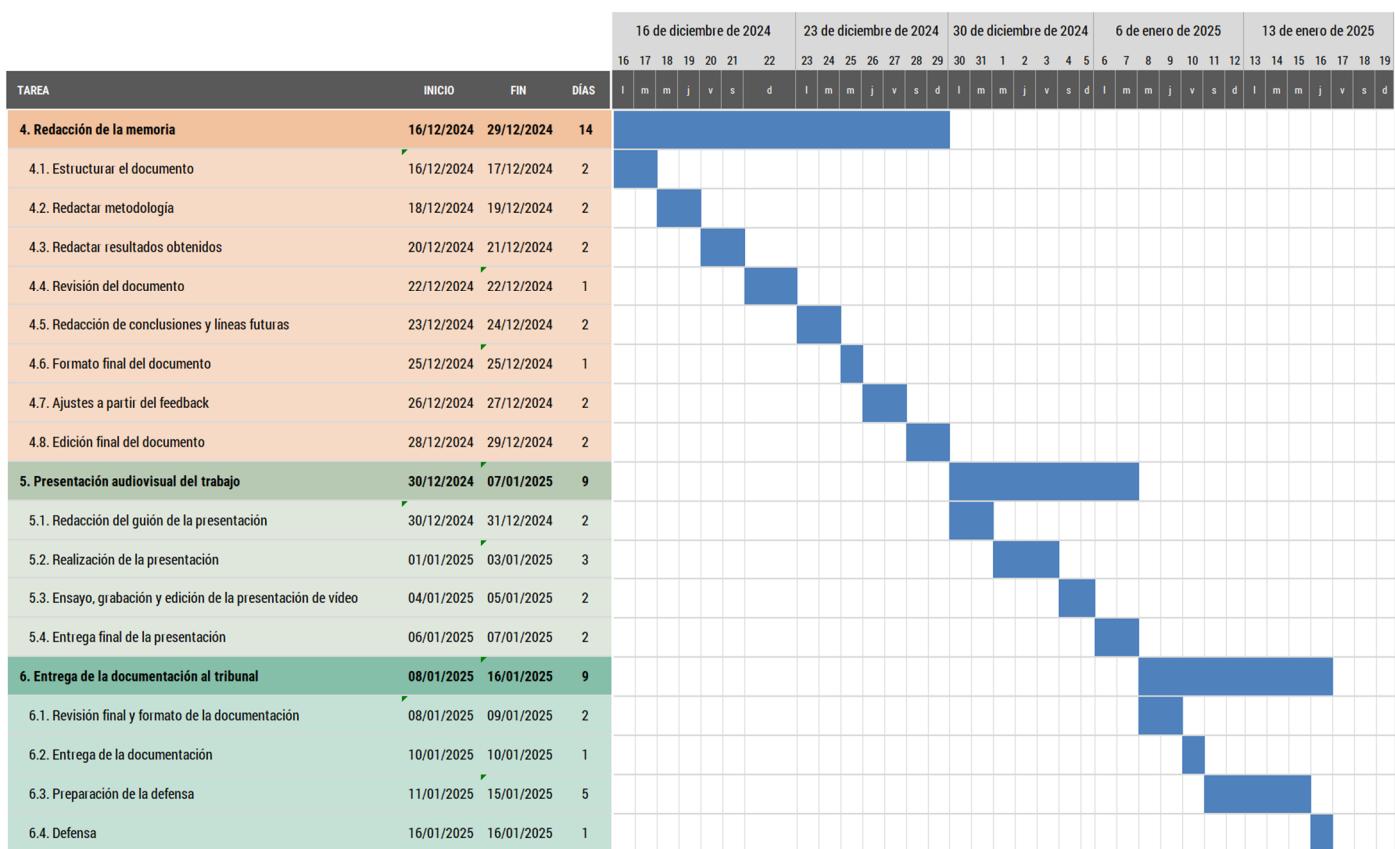


Ilustración 4. Planificación. Fases 4, 5 y 6

1. Estado del arte

2.1. Análisis de las principales variables que intervienen en los incendios forestales

A lo largo de toda la investigación llevada a cabo en torno al estado de arte se han identificado una serie de variables que se consideran determinantes en la modelización de los incendios forestales y que son ampliamente utilizadas en todos los estudios revisados. Por otro lado se incluye la utilización de algunas variables que no son tan comunes, así como la fuente y la razón por la que se incluye en esta relación. La clasificación se estructura de la siguiente manera:

Variables climatológicas, son básicas en cuanto a la modelización de incendios forestales ya que determinan las condiciones atmosféricas que influyen en la propagación e inflamabilidad de los incendios:

- **Temperatura.** Es uno de los factores relevantes en la propagación de incendios forestales; su incremento implica una disminución de la humedad, un aumento de la evaporación del agua lo que facilita la ignición y propagación del fuego.
- **Humedad relativa.** Cuanto más baja sea la humedad más seco es el ambiente lo que provoca que la vegetación se seque y se convierta en combustible con más facilidad para arder. Algunos estudios situaron esta variable como relevante en cuanto a la predicción de áreas quemadas en incendios forestales [4].
- **Velocidad del viento.** Actúa como impulsor haciendo que se extienda rápidamente acelerando su propagación.
- **Dirección del viento.** Influye en la dirección global de propagación del incendio.
- **Precipitación.** La lluvia o la ausencia de esta en la zona puede promover o no el incendio forestal, ya que la precipitación acumulada influye en la humedad del combustible.
- **Radiación solar.** Si el área está expuesta a una radiación solar intensa, seca la vegetación y se incrementa la inflamabilidad.
- **FWI (Fire Weather Index o Índice Meteorológico de Incendios).** Se introduce en algunos estudios [5] [6] entre las variables climatológicas y mide el riesgo de incendios forestales basado en variables como la temperatura, humedad, velocidad del viento y precipitaciones recientes, lo que en algunos casos puede ayudar a la reducción de variables de entrada en los modelos.
- **Frecuencia de rayos.** Suponen una importante fuente natural de generación de incendios y en algunos estudios [7] se incluye en la modelización, aunque hay que tener en cuenta que puede existir relación con otros factores como la precipitación.

Variables topográficas que nos ayudan a entender como la topografía del terreno influye en la propagación de un incendio siendo un factor importante en zonas montañosas o de difícil acceso:

- **Pendiente.** Influye en la velocidad de propagación del incendio, ya que aquellas pendientes más pronunciadas son las que favorecen una mayor velocidad de propagación.
- **Altitud.** Influye en la temperatura y humedad lo que afecta a la probabilidad de incendios.
- **Orientación.** Esta variable hace referencia a la dirección hacia la que se inclina la ladera y permite determinar la dirección de propagación del fuego. Las laderas orientadas al sol son más secas y por tanto más propensas a la combustión [\[8\]](#).

Indicadores de estado de la vegetación, hacen referencia al combustible vegetal; vegetaciones secas o densas facilitan la expansión de los incendios:

- **Tipo de vegetación.** Se considera un factor determinante ya que dependiendo su tipología vamos a tener distinta inflamabilidad [\[8\]](#).
- **Índice de sequía.** Un índice que combina información sobre precipitación y temperatura para evaluar la sequedad del combustible y que en algunas fuentes ha resultado ser un factor relevante [\[4\]](#) [\[9\]](#).

Variables temporales, nos permiten entender cómo el comportamiento de los incendios varía en función de la época del año y pueden revelar patrones estacionales o patrones relacionados con las actividades humanas:

- **Mes.** La época del año influye en el comportamiento del fuego, de hecho, algunas metodologías realizan una división en estaciones, estación seca, estación lluviosa, etc. [\[10\]](#).
- **Día.** Día de la semana en el que tuvo lugar el incendio donde se estudia si tiene alguna implicación el día de la semana en el que se produce lo que está más relacionado con actividades humanas [\[11\]](#).
- **Hora del día.** La humedad varía según la hora del día.
- **Frecuencia de incendios anteriores.** Hay estudios que los incluyen para comprender los patrones de ocurrencia y comportamiento de los incendios en el pasado [\[10\]](#).

Variables geográficas, nos permiten ubicar con exactitud el incendio y contextualizarlo:

- **Coordenadas geográficas del punto de inicio del incendio**, son variables difíciles de obtener con precisión [\[12\]](#).

Variables socioeconómicas, van a influir tanto en la ocurrencia de los incendios como en su rápida intervención:

- **Densidad de población.** Esta variable tiene una doble lectura, por un lado, una mayor densidad de población implica que exista una mayor probabilidad de incendios provocados por la acción humana, accidental o intencionadamente; por otro lado, implica una mayor capacidad de respuesta a la extinción de incendios debido a la existencia de mayores recursos [\[6\]](#).

- **Uso del suelo.** Algunos estudios clasifican el tipo de suelo en bosque, urbano, agrícola, etc. a partir de imágenes satelitales [8].
- **Actividades humanas.** Se puede incrementar la frecuencia de incendios debido a prácticas agrícolas como la quema de rastrojos [7].
- **Densidad de la red de carreteras.** A mayor cantidad de carreteras mejor acceso hay a zonas remotas para la extinción de incendios; por contra, se puede incrementar la probabilidad de incendios accidentales provocados por el propio tráfico [7].

Variables derivadas de imágenes satelitales, nos permiten monitorizar de manera más detallada la vegetación, la severidad de los incendios o la energía liberada a gran escala:

- **NDVI (Normalized Difference Vegetation Index o Índice de Vegetación por Diferencia Normalizada).** Si tenemos una vegetación densa el fuego se puede propagar rápidamente. Es un indicador muy utilizado en los análisis que se realizan después del incendio [13].
- **NBR (Normalized Burn Ratio o Índice Normalizado de Área Quemada).** Es un índice espectral utilizado para evaluar la severidad de los incendios forestales [14].
- **FRP (Fire Radiative Power o Índice de Energía Liberada en el Fuego):** Mide la energía liberada durante la combustión. A medida que aumenta el FRP, también lo hace la densidad y tamaño del fuego [6].
- **LST (Land Surface Temperature o Temperatura de la Superficie Terrestre):** En algunos estudios [6] que incluyen imágenes satelitales se introduce este dato que se basa en que a medida que la temperatura superficial terrestre es más elevada se agrava la severidad del fuego y su capacidad de propagación.

2.2. Análisis de los principales modelos de predicción utilizados

Si nos centramos en aquellos estudios que tienen como propósito la predicción de la extensión del área quemada o la severidad alcanzada en un incendio forestal, la predicción de incendios, la predicción de la propagación de los incendios o la simple detección de áreas quemadas encontramos que se han utilizado una variedad de modelos distintos. Desde modelos centrados en *machine learning* empleando datos tabulares a modelos de *deep learning* utilizando imágenes satelitales pasando por la implementación de modelos multimodales. Realizamos una revisión de los principales estudios llevados a cabo en los últimos años con datos pertenecientes a distintas partes del mundo.

Turquía

En el estudio realizado en Turquía [6] se utilizaron datos tabulares de incendios forestales ocurridos entre 2015 y 2019 donde su variable objetivo fue la predicción del área quemada considerándola en este caso como una variable continua con lo que el estudio se centra en resolver un problema de regresión.

Para ello implementa diferentes modelos como *Decision Tree* (DT), *Support Vector Regression* (SVR), *Gaussian Process Regression* (GPR) y *Feedforward Neural Networks* (FNN) introduciendo la optimización bayesiana para optimizar los hiperparámetros de los modelos. El estudio concluye que el GPR y la FNN son las que logran mejores rendimientos superando a otros estudios.

Hay que destacar que el modelo se utilizó con otro conjunto de datos correspondiente a una región geográfica totalmente distinta como la de Portugal obteniendo buenos resultados que demuestran la robustez del modelo.

Como limitaciones a tener en cuenta, utiliza un número limitado de variables predictoras, por lo que se podrían incluir aquellas relacionadas con la topografía o densidad poblacional e incluso como futura línea de investigación propone incluir imágenes de teledetección para mejorar la precisión.

Portugal

Son varios los estudios realizados sobre el *dataset* de incendios forestales en el Parque Natural de Monteshino en Portugal que está formado por datos tabulares de incendios que tuvieron lugar entre 2000 y 2003, entre ellos destacamos los siguientes:

En un primer estudio [15] cuyo objetivo era predecir el tamaño del área quemada en los incendios se evaluó el rendimiento los siguientes algoritmos de *machine learning*, *Multilayer Perceptron* (MLP), *Linear Regression* (LR), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT) y técnicas de *Stacking* combinando distintos modelos.

El modelo que mejores resultados obtuvo fue el SVM ya que según el *Mean Absolute Error* (MAE) obtenido, su rendimiento predictivo fue el mejor de entre todas las implementaciones.

Una de las limitaciones de este estudio se debe a que el conjunto de datos utilizados tiene pocos registros. Además, el uso de un software especializado del que no se han compartido los detalles completos de implementación dificulta el análisis de los parámetros utilizados en los algoritmos.

Un segundo estudio [16] realizado con el *dataset* de incendios de Portugal donde el objetivo es predecir el área quemada en un incendio forestal. En este caso el estudio se centró en tres algoritmos específicos regresión lineal, regresión *ridge* y regresión *lasso* donde este último es el que obtuvo mejores resultados.

Un tercer estudio [11] centrado en la predicción de incendios forestales pero en este caso investigando en más profundidad la influencia de la heterogeneidad espaciotemporal en la ocurrencia de los incendios forestales. Este estudio refuerza la idea de que están influenciados por patrones y factores específicos pertenecientes a cada zona geográfica y período de tiempo.

En este caso se emplearon cuatro modelos de clasificación, *XGBoost*, *Support Vector Machine* (SVM), *Random Forest* (RF) y *Decision Tree* (DT). Destacó el desempeño de *XGBoost* y se llegó a la conclusión de que factores como el mes, la ubicación geográfica y el FWI fueron las variables con mayor impacto en el modelo.

Entre las limitaciones se hace referencia al área de estudio que es muy limitada y la no inclusión de factores que pudieran ser relevantes como las actividades humanas.

Estados Unidos

En el estudio realizado en la región de Virginia (EE. UU.) [17] se utilizaron datos de incendios forestales que tuvieron lugar entre 1992 y 2015. En este caso el objetivo se basó en predecir la extensión de área quemada en un incendio forestal. Se utilizó por un lado el dataset original de *Kaggle* y otro mejorado incluyendo más características obtenidas del conjunto de datos de *UCI Machine Learning Repository*.

Se implementaron los siguientes algoritmos de *machine learning*, *Multilayer Perceptron* (MLP), *Linear Regression* (LR), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT) y técnicas de *Stacking* combinando distintos modelos.

Se concluyó que el DT fue el mejor algoritmo para el conjunto de datos mejorado, mientras que el MLP fue el mejor para el conjunto de datos original. En este estudio se emplearon mapas de calor de correlación para seleccionar las mejores características.

Otro estudio [18] realizado en EE. UU. realizado con este mismo conjunto de datos también tuvo como objetivo predecir la extensión de área quemada en los incendios forestales. Entre las variables predictoras se introduce la causa del incendio.

Se implementaron los siguientes algoritmos de *machine learning*, *Multilayer Perceptron* (MLP), *Linear Regression* (LR), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT), *Random Forest* (RF) y *Stacked Regressor*.

El estudio concluyó con que el modelo más preciso fue el SVM con un *Mean Absolute Percentage Error* (MAPE) del 41% en el conjunto de prueba.

Una de las limitaciones principales que se describe es que el conjunto de datos presenta grandes desequilibrios en cuanto al tamaño de los incendios, ya que la mayor parte de los incendios registrados tiene un tamaño pequeño de 10 hectáreas, mientras que hay incendios que se pueden extender hasta las 600 mil hectáreas. Esto puede dificultar la capacidad de los modelos para predecir con precisión los GIF (Grandes Incendios Forestales).

Indonesia

En un estudio realizado en Indonesia [13] se utilizó un conjunto de datos formado por imágenes que presentaban información de las condiciones de vegetación previas al incendio y la severidad del incendio posterior. El objetivo es por tanto el de predecir esta severidad utilizando el BAIS2 (Índice de Área Quemada para *Sentinel-2*).

Se desarrolló un modelo basado en la construcción de una *Artificial Neural Network* (ANN) obteniendo precisiones superiores al 90%. Se llegó a la conclusión de que se podía utilizar como

modelo de predicción de la severidad de los incendios forestales a corto plazo así como las ventajas en la utilización de sistemas de alerta temprana.

Algunas de las limitaciones tienen que ver con el tiempo de ejecución ya que si se tienen recursos de hardware limitados puede ser un problema.

Otro estudio [19] realizado en la provincia de *Kalimantan del Sur* (Indonesia) utiliza un conjunto de datos obtenidos de dos fuentes distintas, por un lado obtiene la variable del área quemada del incendio de *Global Wildfire Information System* (GWIS) y para el resto de las variables se usa el obtenido de la fuente *SeasFire Cube* que abarcan un periodo de 2010 a 2020. En este caso la variable objetivo, área quemada, se realizó una transformación a variable binaria, cambiando el objetivo final a un problema de clasificación en incendio o no incendio.

Se implementó un *Random Forest* (RF) para predecir el área potencial quemada. Aunque el modelo obtuvo un buen resultado en el área de estudio, cuando se probó a nivel nacional, la precisión del modelo se redujo.

Perú

En el estudio [20] realizado en Cusco (Perú) se utilizó un conjunto de datos formado por imágenes satelitales con zonas afectadas por incendios forestales entre 2017 y 2021 obtenidas del sensor del *Sentinel-2*. En este caso el objetivo del estudio se basa en delimitar con mayor precisión la zona quemada en un incendio forestal mediante la utilización de imágenes satelitales.

Para resolver el problema implementa un modelo de segmentación semántica basado en una *Convolutional Neural Network* (CNN). Para la segmentación de las imágenes satelitales utiliza el modelo UNet mientras que para la CNN usa una arquitectura *Encoder-Decoder*.

El modelo obtuvo buenos resultados demostrando la efectividad del modelo *UNet* para la segmentación semántica en la delimitación de áreas quemadas y la importancia de las bandas B12, B11, B8 y B4 del sensor *Sentinel-2* como las bandas más relevantes.

En este caso el modelo tiene una dependencia bastante alta de la obtención de imágenes de calidad, por esta razón utiliza *Sentinel* frente a *Landsat*; por otro lado, depende también de la cantidad de datos que se puedan obtener ya que si no son representativos de la zona de estudio puede que el modelo presente dificultades para generalizar. Otra limitación a tener en cuenta es la complejidad computacional que supone entrenar este tipo de modelos.

India

En el estudio [21] realizado en *Uttarakhand* (India) se utilizó un conjunto de imágenes creado por los autores del estudio que contenían fotografías de incendios forestales obtenidas de los propios motores de búsqueda que formaban un *dataset* equilibrado entre las dos clases “incendio” y “no incendio”, lo que hace que el objetivo de este estudio sea el de resolver un problema de clasificación cuyo objetivo es la detección de incendios.

Para dar solución al problema planteado el estudio se centró en el uso de *MobileNetV2* que es un modelo preentrenado con arquitectura CNN la cual fue modificada para adaptarla a la detección de incendios forestales.

Las limitaciones se basaron en la obtención de imágenes para crear el conjunto de datos y en el uso de *transfer learning* donde hay que tener en cuenta la similitud de tareas a realizar, entre la tarea de origen que era la de la clasificación de imágenes con la de destino que consistía en detectar imágenes clasificadas como incendios. Esta metodología resulta interesante en el caso de tener conjuntos de datos reducidos.

En el estudio se llegó a la conclusión de que el aprendizaje por transferencia en combinación con el uso de modelos preentrenados puede suponer un enfoque prometedor y escalable en la detección de incendios forestales.

Rusia

Por último, analizamos un estudio [22] realizado en las regiones del norte de Rusia, *Krasnoyarsk*, la República de *Sakha* y la región de *Irkutsk* que tiene como objetivo predecir la propagación de incendios forestales que sea capaz de cubrir territorios extensos teniendo en cuenta parámetros ambientales relevantes en la dirección y propagación del incendio. El conjunto de datos recoge los incendios acaecidos entre los años 2019 y 2021. Está formado por imágenes satelitales obtenidas de *Sentinel-2* y *Landsat-8* y datos meteorológicos y geográficos de alta precisión.

Se utilizó el modelo de red neuronal profunda *MA-Net* que consiste en una CNN con estructura codificador-decodificador que procesa información espaciotemporal de manera eficiente y que los autores del estudio adaptaron para que pudiera recibir como entradas datos multimodales, es decir, imágenes y datos tabulares.

Para el primer día de pronóstico se obtuvo un resultado de *F1-score* de 0,68 mientras que para el quinto día se obtuvo un 0,65. Se realizaron diferentes pruebas para determinar las variables clave mediante la exclusión y se observó que características como el viento o la cobertura del suelo resultaron ser determinantes para el modelo.

Las limitaciones que se encontraron es que el modelo no funciona tan bien en las regiones del sur de Rusia al variar las condiciones ambientales lo que dificulta la generalización del modelo.

Tabla 1. Resumen de estudios sobre Modelos de Predicción de Incendios Forestales

Estudio	Año	Ubicación	Objetivo	Modelos	Limitaciones	Conclusión
<i>A comparative Bayesian optimization-based machine learning and artificial neural networks approach for burned area prediction in forest fires: an application in Turkey</i>	2023	Turquía	Predicción del área quemada como variable continua, resolviendo un problema de regresión.	DT, SVR, GPR, FNN	Número limitado de variables predictoras, podrían incluirse variables topográficas, de densidad poblacional y futuras líneas de teledetección.	GPR y FNN lograron mejores rendimientos, mostrando robustez incluso con datos de Portugal.
<i>Comparison of the Machine Learning Methods to Predict Wildfire Areas</i>	2022	Portugal	Predicción del tamaño del área quemada en incendios forestales.	MLP, LR, SVM, KNN, DT, Stacking	Datos limitados y falta de detalles completos de implementación del software.	SVM obtuvo mejor rendimiento predictivo en MAE.
<i>Machine Learning Regression Techniques to Predict Burned Area of Forest Fires</i>	2020	Portugal	Predicción del área quemada en incendios forestales utilizando técnicas de regresión.	Linear Regression, Ridge Regression, Lasso Regression	No especificadas.	Lasso Regression mostró mejores resultados en la predicción del área quemada.
<i>Wildfire Prediction Model Based on Spatial and Temporal Characteristics: A Case Study of a Wildfire in Portugal's Montesinho Natural Park</i>	2022	Portugal	Investigación sobre la influencia de la heterogeneidad espaciotemporal en la ocurrencia de incendios forestales.	XGBoost, SVM, RF, DT	Área de estudio limitada; falta de factores adicionales como actividades humanas.	XGBoost destacó, indicando que factores como mes, ubicación y FWI tienen gran impacto.
<i>Prediction of Forest Fire Area Using Machine Learning Algorithms</i>	2021	Estados Unidos	Predicción de la extensión del área quemada usando datos originales y mejorados.	MLP, LR, SVM, KNN, DT, Stacking	Desequilibrios en tamaño de incendios, lo que afecta la precisión en predicciones de grandes incendios forestales.	Decision Tree fue el mejor para datos mejorados; MLP destacó con datos originales.
<i>Wildfire Burn Area Prediction</i>	2019	Estados Unidos	Predicción de la extensión del área quemada, incluyendo la causa del incendio como variable predictor.	MLP, LR, SVM, KNN, DT, RF, Stacked Regressor	Desequilibrio en tamaño de incendios y limitaciones en la representación de incendios de gran tamaño.	SVM fue el más preciso con un MAPE del 41%.

<i>GeoAI for Disaster Mitigation: Fire Severity Prediction Models using Sentinel-2 and ANN Regression</i>	2022	Indonesia	Predicción de la severidad del incendio mediante el BAIS2 utilizando imágenes satelitales.	Artificial Neural Network (ANN)	Limitaciones en tiempos de ejecución debido a recursos de hardware limitados.	ANN mostró precisión superior al 90%, útil para modelos de predicción y alertas tempranas.
<i>Inteligencia Artificial aplicada a la predicción del Dengue e incendios forestales en Indonesia</i>	2024	Indonesia	Clasificación de área quemada en incendios mediante transformación binaria del objetivo.	RF	Precisión reducida a nivel nacional, aunque buena a nivel local.	El modelo mostró buena precisión en área de estudio, pero menos precisión nacional.
<i>Delimitación de áreas afectadas por incendios forestales mediante aprendizaje profundo en imágenes satelitales</i>	2024	Perú	Delimitar con precisión la zona quemada mediante imágenes satelitales.	Convolutional Neural Network (CNN) con arquitectura Encoder-Decoder y modelo UNet	Dependencia de la calidad de imágenes y representatividad de datos; complejidad computacional.	UNet fue efectivo para segmentación semántica, identificando bandas relevantes para delimitar zonas quemadas.
<i>Utilizing Transfer Learning and pre-trained Models for Effective Forest Fire Detection: A Case Study of Uttarakhand</i>	2024	India	Detección de incendios forestales mediante clasificación de imágenes.	MobileNetV2 (preentrenado y adaptado para detección de incendios)	Limitaciones en obtención de imágenes y transfer learning; requerimientos de similitud de tareas entre origen y destino.	Transfer learning y modelos preentrenados son prometedores para detección de incendios.
<i>Wildfire spreading prediction using multimodal data and deep neural network approach</i>	2024	Rusia	Predicción de la propagación de incendios forestales en áreas extensas.	MA-Net (CNN con estructura codificador-decodificador, multimodal)	Precisión reducida en regiones del sur de Rusia debido a variación en condiciones ambientales, lo que afecta la generalización.	MA-Net logró F1-score de 0.68 para el primer día de pronóstico, destacando la importancia del viento y la cobertura del suelo en la predicción.

2.3. Conclusiones del estado del arte

Tras realizar el análisis de los diferentes estudios sobre la predicción de incendios forestales se identifican una serie de puntos clave a tener en cuenta en el desarrollo de nuestro modelo que se desarrollan a continuación:

Modelos

Se utilizan diferentes enfoques cuando el objetivo es predecir la extensión de área quemada, como un problema de regresión donde la variable objetivo es una variable continua o como un problema de clasificación donde se suele transformar la variable objetivo en una variable binaria. Los modelos de regresión como *Support Vector Regression* (SVR) y *Gaussian Process Regression* (GPR) han obtenido buenos desempeños.

En general no se aprecia una uniformidad en los modelos utilizados, se usan diferentes enfoques, mostrando por un lado el buen rendimiento de modelos de *machine learning* como *Support Vector Machine* (SVM) o el uso de redes neuronales como *Artificial Neural Network* (ANN) o modelos más avanzados que admiten entradas de datos multimodales como el *MA-Net*.

El uso de *transfer learning* y modelos preentrenados como *MobileNetV2* resultó útil para el caso en que el conjunto de datos es precario en cuanto al número de muestras.

Factores clave

Se destaca el uso de condiciones meteorológicas en todos los casos de estudio resaltando la relevancia del uso de la temperatura, humedad, viento y precipitación. En algunos casos se introdujo el *Fire Weather Index* (FWI) que mostraron una buena contribución a los modelos finales.

Ha quedado ampliamente demostrado que los incendios forestales tienen una influencia directa con las zonas geográficas y también con los factores estacionales, es decir, que debemos tener en cuenta los patrones espaciotemporales en el modelo.

Otras características relacionadas con la cobertura del suelo, la tipología de vegetación y su densidad son factores relevantes en la caracterización de un incendio forestal.

En diversos estudios se insta a la inclusión de otros factores en los modelos como factores antropogénicos o de ordenación del territorio. Y en general, se hace hincapié en la importancia de la calidad de los datos para obtener buenas precisiones.

En cuanto al uso de imágenes satelitales, se usan predominantemente las proporcionadas por *Sentinel* y *Landsat* y su inclusión puede ayudar a mejorar la precisión del modelo sobre todo a gran escala.

Limitaciones

En la mayor parte de los estudios se da mucha importancia al conjunto de datos. Tienen que ser datos lo suficientemente representativos, algo que en general es difícil ya que suele haber muchas más muestras de incendios de pequeño tamaño frente a pocas muestras de incendios de gran tamaño, lo que puede dificultar la generalización del modelo.

Adaptar el modelo a las distintas regiones es una limitación recurrente en todos los estudios, ya que las condiciones ambientales, geográficas y topográficas pueden variar mucho de unas zonas a otras provocando que el modelo no sea igual de preciso en todas las zonas.

Otra limitación tiene que ver con los recursos computacionales, ya que el hecho de utilizar modelos avanzados que incluyan imágenes satelitales puede ralentizar los tiempos de entrenamiento.

Metodologías

El uso de técnicas de apilamiento y la combinación de distintos modelos ha sido eficaz para capturar la complejidad de los modelos, mejor que con el uso de un solo modelo.

La utilización de técnicas para optimizar los hiperparámetros como la optimización bayesiana permite incrementar el rendimiento del modelo.

Hay un interés subyacente en la utilización de modelos cada vez más complejos que sean capaces de abordar la naturaleza compleja de los incendios forestales.

11 estudios

Ubicación



Datos



Software



Modelos



Fechas
[2019, 2024]



VARIABLES

Ilustración 5. Esquema visual estado del arte

3. Materiales y métodos

3.1. Descripción de la base de datos del EGIF

La Estadística General de Incendios Forestales (EGIF) es la base de datos nacional que recopila la información sobre los incendios forestales ocurridos en España desde 1968 y que contiene más de 600 mil registros.

La base de datos está integrada por la información de cada uno de los incendios forestales que se producen en España y que se recogen en el PIF (Parte de Incendio Forestal) [23] que en la actualidad se encuentra en su novena actualización [24], en el cual se incluyen más de 150 campos de datos para cada incendio. Los datos son mecanizados por parte de cada una de las comunidades autónomas quienes los facilitan al ADCIF (Área de Defensa Contra Incendios Forestales) para incorporarlos a la base de datos.

La información consolidada de la EGIF está disponible hasta el año 2016 a nivel nacional. A partir de 2017 se cuenta con datos de algunas provincias que mantienen sus estadísticas actualizadas. La base de datos incluye variables como la superficie afectada, las condiciones meteorológicas, los medios empleados en la extinción, entre otras. Está compuesta por 108 tablas que se distribuyen de la siguiente manera:

AgrupacionMunicipio	CodIdioma	CodTipoEfecto	pif_incidencias	RelModeloCombustionPif
AgrupacionMunicipioDetalle	CodImpactoAmbiental	CodTipoEspacioProtegido	pif_localizacion	RelNoArboladoHerbaceoParteMonte
Ambito	CodIncidenciaProtecCivil	CodTipoFuego	pif_medios	RelNoArboladoLeniosoParteMonte
CatalogoMonte	CodIniciadoJuntoA	CodTipoProducto	pif_perdidas	RelNoForestalAfectadoParteMonte
CodAprovechamiento	CodInvestigacionCausa	CodTipoRenta	pif_propagacion	RelOtraPerdidaParteMonte
CodAtaque	CodMedioAereo	CodTipoVictima	pif_tecnicas	RelPerdidaMontePif
CodAutorizacionActividad	CodMedioPersonalExt	CodTitularidadMedio	pif_tiempos	RelRetardantePif
CodCausa	CodMedioPesado	CodTitularidadMonte	Provincia	RelTeselaAfectadaPif
CodCausante	CodMfFormacionArborea	CodTransportePersonal	RelArboladoAfectadoParteMonte	RelTipoArealIniciadoPif
CodCertidumbreCausa	CodMfTipoEstructural	CodVigilanteFijo	RelAsociadoPif	RelTipoAtaqueIndirectoPif
CodClaseDia	CodModeloCombustion	Comarcalsla	RelAtaquePif	RelTipoFuegoPif
CodConsecuenciaPersona	CodMotivacion	Comunidad	RelEspacioProtegidoPif	RelTransportePersonalPif
CodDatum	CodNivelGravedad	EntidadMenor	RelFactorCalculoPerdidaParteMonte	RelVictimaPif
CodDemaniajMonte	CodNoArboladoHerbaceo	EspacioProtegido	RelFactorProductoOtrosParteMonte	TipoClasificacion_JFN_MFE
CodDetectadoPor	CodNoArboladoLenioso	EspeciesValoracion	RelFactorRentaParteMonte	TipoProductoValoracion
CodEspecieArbol	CodNoForestal	Municipio	RelGrupoCausa	TipoRentaValoracion
CodEstacionMeteorologica	CodPeligro	ParteMonte	RelGrupoMedioRetardantePif	
CodEstadoMasa	CodPorcentajeAutoRegenerable	Pif	RelGrupoMotivacion	
CodEstadoPif	CodPropiedadMonte	pif_anexo	RelIncidenciasProtecCivilPif	
CodGradoResponsabilidad	CodRetardante	pif_causa	RelIniciadoJuntoAPif	
CodGrupoCausa	CodRiesgo	pif_comun	RelMedioAereoPif	
CodGrupoMedioRetardante	CodTipoArea	pif_condiciones	RelMedioPersonalPif	
CodGrupoMotivacion	CodTipoAtaqueIndirecto	pif_deteccion	RelMedioPesadoPif	

Ilustración 6. Relación de tablas en la base de datos del EGIF

- Las **tablas PIX_XXXXX** que contienen los datos generales del incendio y se corresponden con las dos primeras hojas del PIF. Estas tablas contienen un solo registro para cada incendio y pueden contener valores faltantes.
- Las **tablas ParteMonte** que contienen campos incluidos en el apartado de datos particulares del monte del PIF.
- Las **tablas CodXXXXX** que contienen las etiquetas de los códigos numéricos utilizados en los campos.
- Las **tablas RelXXXXX** que contienen datos del PIF con registro múltiple y que se relacionan con las tablas PIF_XXXX.
- **Resto de tablas**, permiten realizar la localización o extraer otro tipo de información acerca del monte.

3.2. Tratamiento de la base de datos del EGIF

La base de datos se proporciona en formato Access, por lo que las transformaciones iniciales se realizaron en este entorno utilizando SQL como herramienta principal para manipular los datos.

Tablas tipo Rel

Todas las tablas con prefijo Rel se corresponden con apartados del PIF donde es posible seleccionar más de una opción a la vez. Esto supone la aparición de números de parte repetidos en función del número de opciones que el director de extinción haya marcado. Para solucionar este problema, se han hecho algunas transformaciones en los datos. El objetivo es eliminar repeticiones innecesarias y guardar solo los valores únicos, junto con las combinaciones posibles de las opciones seleccionadas en cada apartado. Así, se evita duplicar registros y se facilita el análisis posterior de los datos.

Tipo de combustible en la zona del incendio

Explicamos a continuación, el proceso seguido para la tabla RelModeloCombustionPif que recoge el apartado 5.2. del PIF y que hace referencia al tipo de combustible en la zona del incendio.

La tabla está formada por el número de PIF, el número de parte y la columna IdModeloCombustion que recoge el tipo de combustible y tiene la siguiente categorización:

- 1: pastizales
- 2: matorrales
- 3: bosques
- 4: restos

Al realizar las consultas se verifica que hay más de 200 mil partes que se repiten. Se generan cuatro columnas que recogen la categorización y se codifica de forma binaria indicando la opción correspondiente que fue seleccionada en el PIF. Tras las transformaciones se obtienen los siguientes resultados:

pastizales	matorrales	bosques	restos	frecuencia	porcentaje	codificación
0	1	0	0	285584	47,31%	matorrales
0	1	1	0	95841	15,88%	mat_bos
1	0	0	0	60095	9,96%	pastizales
1	1	0	0	47574	7,88%	pas_mat
0	0	1	0	46466	7,70%	bosques
1	1	1	0	36158	5,99%	pas_mat_bos
1	0	1	0	11072	1,83%	pas_bos
0	1	0	1	5151	0,85%	matorrales
0	0	0	1	4853	0,80%	matorrales
1	0	0	1	2755	0,46%	pas
0	1	1	1	2442	0,40%	mat_bos
0	0	1	1	1801	0,30%	bosques
1	1	0	1	1764	0,29%	pas_mat
1	1	1	1	1514	0,25%	pas_mat_bos
1	0	1	1	511	0,08%	pas_bos
				603581	100,00%	

Tabla 2. Relación de combinaciones de los tipos de combustibles

Se obtienen hasta 15 combinaciones distintas de combustibles. Para reducirlas se descarta la opción “restos” dado que tiene muy poco peso en todo el conjunto de datos; la suma de todos los registros que contienen “restos” tiene un peso de un 3.4%, así que se considera “restos” como si fuera un 0 y se procede a la codificación del resto de combinaciones menos frecuentes. Por último, los registros que contienen solamente “restos” como opción de combustible pasan a codificarse como la opción más frecuente que contenía “restos”, es decir, su inmediato superior en la tabla.

Lugar del inicio del incendio

En el caso de la tabla RellIniciadoJuntoAPif, nos está indicando el lugar representativo del inicio del incendio en el que tenemos 9 variantes que se recogen en el apartado 3.3.2. en el que se pueden marcar diferentes casillas simultáneamente. Tras realizar la consulta de las 512 posibles combinaciones obtenemos 90. Mostramos las más frecuentes a continuación:

autovia_carretera	pista_camino	senda	lineas_electricas	excursionistas	vias_ferreas	edificaciones	vertederos	otros	frecuencia	porcentaje	codificación
0	0	0	0	0	0	0	0	1	265401	43,97%	otros
1	0	0	0	0	0	0	0	0	124024	20,55%	autovia_carretera
0	0	1	0	0	0	0	0	0	116838	19,36%	senda
0	1	0	0	0	0	0	0	0	57323	9,50%	pista_camino
0	0	0	1	0	0	0	0	0	21241	3,52%	lineas_electricas
0	0	0	0	1	0	0	0	0	6780	1,12%	excursionistas
0	0	0	0	0	0	1	0	0	3770	0,62%	edificaciones
0	0	0	0	0	0	0	1	0	3289	0,54%	vertederos
0	0	0	0	0	1	0	0	0	3040	0,50%	vias_ferreas
0	1	0	0	0	0	0	0	1	268	0,04%	otros
0	1	1	0	0	0	0	0	0	228	0,04%	senda
1	1	0	0	0	0	0	0	0	192	0,03%	autovia_carretera
0	1	0	0	0	0	1	0	0	136	0,02%	pista_camino
1	0	0	0	0	0	1	0	0	104	0,02%	autovia_carretera
0	1	0	0	1	0	0	0	0	93	0,02%	pista_camino
0	1	0	1	0	0	0	0	0	84	0,01%	pista_camino
1	0	0	0	0	0	0	0	1	83	0,01%	otros
0	0	0	1	0	0	0	0	1	76	0,01%	otros
											602970 99,90%

Tabla 3. Relación de combinaciones del lugar de inicio del incendio

Se observa que el grueso de registros que se obtienen han sido seleccionados de forma única en el parte. Por orden de importancia, otros, autovia_carretera, senda, pista_camino, lineas_electricas, excursionistas, edificaciones, vertederos, vias_ferreas. A partir de la siguiente fila, la primera combinación que aparece es pista_camino y otros con 268 registros frente a los más de 600 mil que contiene la base de datos, así que para simplificar se reemplazan todas las combinaciones con la opción única más frecuente.

Tipo de fuego que se da en la zona del incendio

Explicamos a continuación, el proceso seguido para la tabla RelTipoFuegoPif que recoge el apartado 5.2. del PIF y que hace referencia al tipo de fuego que se genera en función del entorno.

La tabla está formada por el número de pif, el número de parte y la columna IdTipoFuego que recoge el tipo de fuego y tiene la siguiente codificación:

- 1: de superficie
- 2: de copas
- 3: de subsuelo
- 8: focos secundarios

Como tenemos partes repetidos se generan cuatro columnas que recogen la categorización del tipo de fuego y se realiza la codificación binaria. Tras las transformaciones se obtienen las siguientes combinaciones:

superficie	copas	subsuelo	focos_secundarios	frecuencia	porcentaje	codificación
1	0	0	0	541184	89,66%	superficie
1	1	0	0	42235	7,00%	sup_cop
0	1	0	0	10123	1,68%	copas
1	0	1	0	6005	0,99%	superficie
0	0	1	0	2175	0,36%	superficie
1	0	0	1	670	0,11%	superficie
1	1	1	0	668	0,11%	sup_cop
0	1	1	0	428	0,07%	copas
1	1	0	1	46	0,01%	sup_cop
1	0	1	1	27	0,00%	superficie
1	1	1	1	19	0,00%	superficie
				603580	100,00%	

Tabla 4. Relación de combinaciones de los tipos de fuego

Se observa que las opciones de “subsuelo” y “focos_secundarios” no tiene apenas relevancia, así que se simplifican las combinaciones posibles a tres: “superficie”, “sup_cop” y “copas”, reemplazando el resto de las combinaciones con las más frecuentes.

Obtención del conjunto de datos preliminar

Una vez que hemos escogido y tratado las tablas de la base de datos que nos interesaban, realizamos un último conjunto de operaciones con código SQL que consistirán en la unión de todas las tablas mediante el uso de LEFT JOIN de manera sucesiva hasta obtener finalmente un conjunto de datos preliminar a partir de la base de datos del EGIF que está formada por 27 variables y 603.582 de los cuales 2154 pertenecen a GIF.

3.3. Preprocesamiento de los datos

Fuentes de datos:

- **EGIF (Estadística General de Incendios Forestales):** Contiene 603580 registros con información histórica sobre incendios forestales ocurridos en España.
- **NGMEP (Nomenclátor Geográfico de Municipios y Entidades de Población) [25]:** Proporcionada por el CNIG (Centro Nacional de Información Geográfica), incluye datos como coordenadas, superficie, población y altitud de municipios y entidades de población.
- **Códigos INE proporcionados por el INE (Instituto Nacional de Estadística) [26].** Se empleó para obtener el **código INE**, que es un identificador único para cada municipio.
- **CLIMATE DATA STORE, Fire danger indicators for Europe from 1970 to 2098 derived from climate projections [27].** Se empleó para extraer el **FWI** que es un indicador de peligro de incendio basado en el sistema canadiense de índices meteorológicos de incendios.
- **Series de población por provincia** de 1900 a 2001 [28] y principales series de población desde 1998 [29] del INE: Se empleó para calcular la densidad de población.

Enriquecimiento con la base de datos de Municipios y Entidades de Población

El objetivo fue complementar la base de datos EGIF con la información geográfica y demográfica de NGMEP, por un lado para enriquecer la base de datos pero por otro lado, para tratar de ubicar los incendios con coordenadas geográficas fiables, ya que se encontraron muchas discrepancias en cuanto a los datos tal y como se puede ver en la imagen.

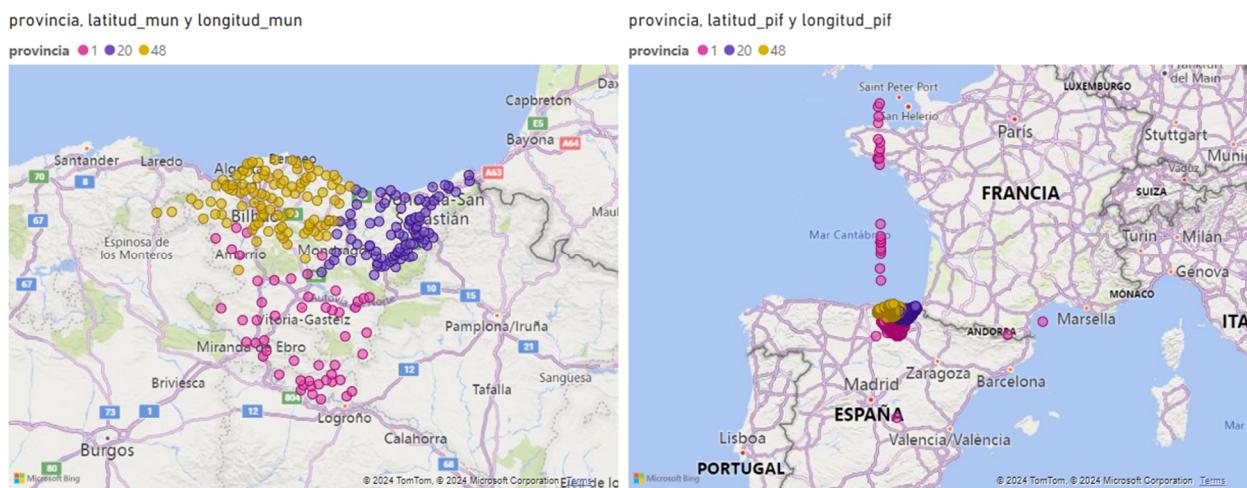


Ilustración 7. Discrepancias en las coordenadas geográficas.

Para poner ambas en común se ha utilizado una tercera base de datos del INE utilizando el código INE como clave común lo que permitió añadir información relacionada con la superficie, población, altitud y coordenadas de los municipios.

Después de realizar esta modificación encontramos que 67544 registros no pudieron localizarse geográficamente.



Ilustración 8. Registros de incendios forestales que carecen de coordenadas geográficas.

Son incendios que se registraron entre 1968 y 1982, un período en el que la localización se realizaba mediante el sistema de hoja y cuadrícula del mapa militar 1:250.000, un sistema actualmente en desuso.

Esto se debe a que en aquellos primeros años la forma de localización de los incendios se realizaba mediante hoja y cuadrícula de la división del mapa militar 1:250.000, actualmente en desuso, que acotaba los incendios a una cuadrícula de 10 x 10 km

Recuperación de registros no localizados geográficamente

Para poder ubicar geográficamente esos 67544 registros se extrajeron las coordenadas obtenidas del sistema de cuadrícula militar para asignar los datos del municipio más cercano.

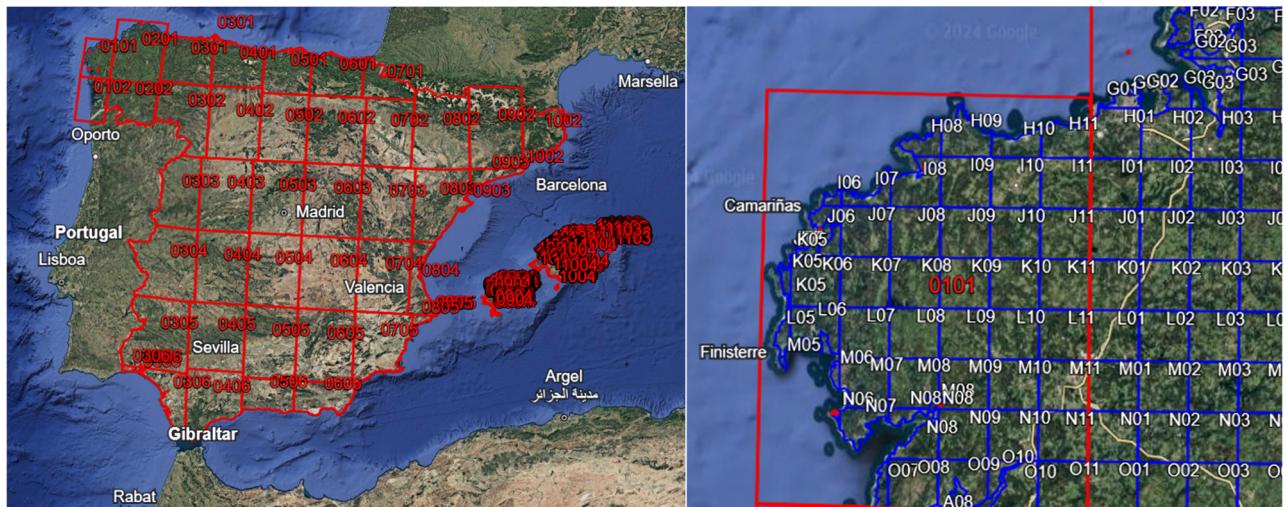


Ilustración 9. Sistema de Hoja (izquierda) y Cuadrícula Militar (derecha)

Para ello se realizaron los siguientes pasos:

Paso 1. Conversión del sistema de cuadrículas:

- Se partió de archivos KMZ con la información de las cuadrículas y etiquetas del mapa militar.
- Utilizando la app online MyGeodata Converter [30], se transformaron estos archivos en Excel.
- Se realizaron las transformaciones necesarias hasta obtener una tabla que relacionara cada hoja y cuadrícula militar con sus coordenadas geográficas en formato decimal.

Paso 2. Resolución de duplicados del sistema de cuadrículas:

- En las zonas de costa del mapa militar, algunas etiquetas representaban áreas irregulares o superpuestas que habían sido nombradas con la misma etiqueta en función de su cercanía.
- Para evitar duplicados, se calculó la media de las coordenadas de todas las etiquetas idénticas [31].

Paso 3. Imputación de datos geográficos y demográficos [32]. Una vez generadas las coordenadas del sistema de cuadrículas militares se incluyen en la base de datos. Para estos incendios que solo han podido ser localizados mediante este sistema se le imputan los datos del NGMEP utilizando la proximidad geográfica al municipio más cercano:

- Se utilizó la librería *SciPy* que implementa un algoritmo de vecinos más cercanos (KNN) con el que se calculan las distancias entre las coordenadas procedentes del mapa militar y las de los municipios en NGMEP.
- Se asignan los datos del municipio más cercano a cada registro: las coordenadas del municipio, población, superficie y altitud.

Quedaron sin identificar 15027 registros, correspondientes a incendios entre 1968 y 1973. Estos fueron descartados por falta de información suficiente para su localización.

En este punto del preprocesamiento, la base de datos contiene 26 variables y 588.554 registros de los cuales 2039 son de GIF.

Revisión de registros con datos climatológicos faltantes

Tal y como se identificó a lo largo de todo el estado del arte, las variables climatológicas tienen gran importancia a la hora de modelar los incendios forestales, así que se realiza un primer acercamiento a cómo se encuentran estas variables en la base de datos.

En el diagrama de barras se observa que hay una gran cantidad de valores faltantes y los que hay contienen bastantes datos atípicos que además carecen de legitimidad.

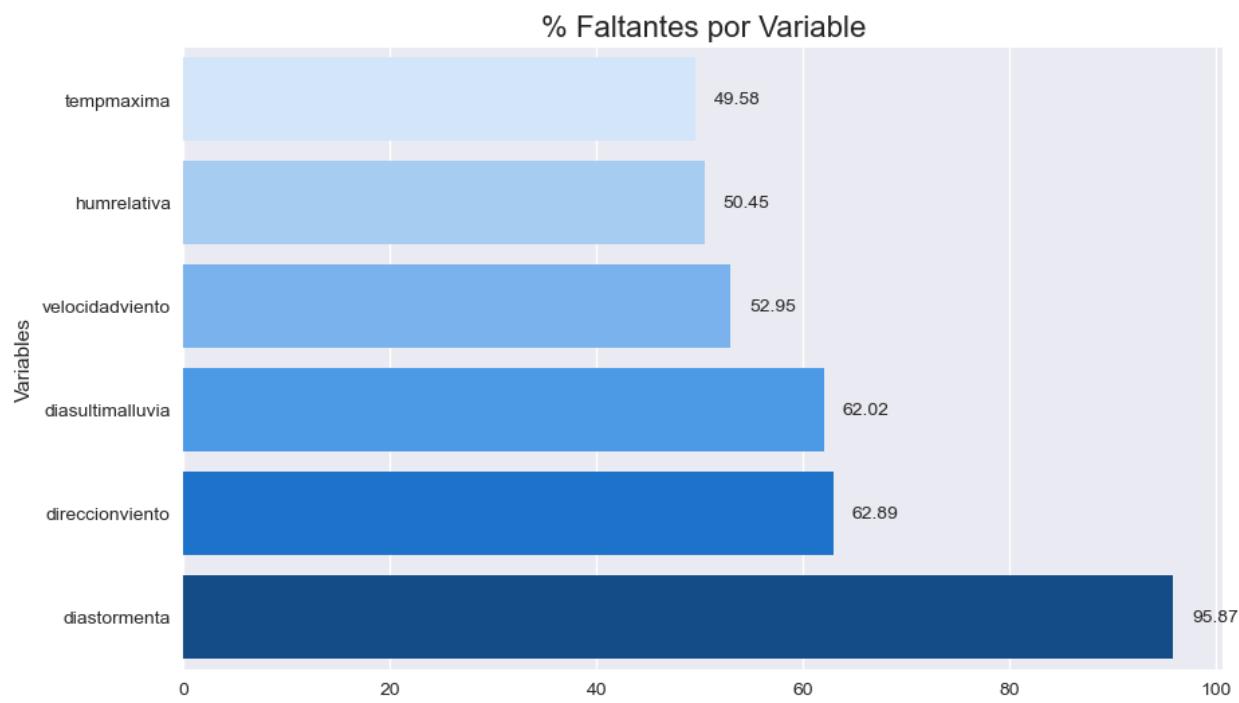


Ilustración 10. Porcentaje de datos faltantes en la base de datos de la EGIF

Sabiendo que los datos meteorológicos con los que se rellena el PIF se obtienen de la estación meteorológica más cercana se decide extraer estos datos de la propia AEMET.

Proceso de extracción de datos climatológicos faltantes [33]

AEMET OpenData [34] conserva datos históricos climatológicos de un total de 947 estaciones meteorológicas distribuidas por toda España. El proceso que se ha seguido para realizar la recopilación de datos ha sido el siguiente:

- La información meteorológica histórica se puede extraer por fecha [35]. Para cada una de las fechas se obtiene un archivo con los datos recogidos por todas las estaciones meteorológicas. Así que recopilamos todas las fechas únicas en las que se ha producido un incendio y extraemos los datos de AEMET.
- Obtenemos un listado de estaciones meteorológicas que pone a disposición la propia AEMET del que obtenemos información del indicativo, que es un código que identifica la estación y las coordenadas geográficas en grados, minutos y segundos.
- Necesitamos realizar una transformación de coordenadas así que generamos una función, “conversion_a_grados”, que nos permita realizar la conversión de coordenadas a grados decimales.
- Tenemos que asignar a cada incendio la estación meteorológica más cercana, para ello implementamos la función “estacion_cercana”, que dadas las coordenadas del incendio nos devolverá un *dataframe* con las estaciones ordenadas por distancia más cercana.
- El proceso de extracción de datos se realiza mediante la función “extraer_datos” que dada la fecha del incendio y el código de la estación extrae los datos meteorológicos que haya en el archivo. La información extraída se va guardando junto al número de parte del incendio y el indicativo para generar un *dataframe* de todos los incendios y sus variables meteorológicas.

El último paso será añadir esta información a nuestra base de datos que estará formada por 34 variables y 588.554 registros.

Proceso de extracción del FWI

Se realiza el proceso de extracción del índice de peligro de incendios ampliamente utilizado en el ámbito de los incendios forestales para incorporarlo a la base de datos, para ello se han extraído los datos de *Climate Data Store* localizando el FWI correspondiente a la fecha y las coordenadas del incendio e incorporándolo a nuestra base de datos. La interpretación del índice es la siguiente:

- Muy bajo: <5,2
- Bajo: 5,2 - 11,2
- Moderado: 11,2 - 21,3
- Alto: 21,3 - 38,0
- Muy alto: 38,0 - 50
- Extremo: >=50,0

Una vez añadido el FWI se tienen 35 variables y 588.554 registros, de los cuales 2039 son de GIF [36].

3.4. Análisis Exploratorio de Datos

A continuación se detalla el proceso seguido en el análisis exploratorio de datos [\[37\]](#).

Creación de nuevas variables

Introducimos la variable **densidad de población** como un indicador poblacional que modele la influencia de núcleos más o menos grandes de población en las zonas cercanas a los incendios forestales. Para ello, se introduce el número de habitantes del municipio entre su superficie.

Para que el cálculo de la densidad de población sea más realista, se ha tenido en cuenta que la población de los municipios ha sufrido cambios desde 1974, así que se ha realizado una estimación lineal de la evolución de la población en las provincias que después se ha extrapolado a los municipios de forma proporcional. El proceso de cálculo de la densidad de población se puede encontrar el notebook 04 del preprocesamiento: estimación de la población [\[38\]](#).

Por otro lado, introducimos la variable **densidad de incendios por municipio** realizando el cálculo del número acumulado de incendios ocurridos en el mismo municipio en los últimos cinco años dividido por la superficie del municipio en ha, para normalizar las diferencias en el tamaño superficial de los distintos municipios. La ventana escogida de cinco años para calcular la densidad de incendios en el municipio pretende reflejar los cambios que puede haber habido en cuanto a políticas de prevención de incendios forestales, variaciones climáticas. Incendios anteriores o alteraciones en el uso del suelo.

Creamos una nueva variable categórica del **rango horario de detección del incendio** a partir de la variable de fecha y hora de detección del incendio.

- Madrugada: 00 - 06
- Mañana: 06 -12
- Tarde: 12 - 18
- Noche: 18 – 00

Creamos la **variable objetivo a modelar, “claseincendio”** que recoge la severidad de los distintos tipos de incendios forestales en función de la extensión del área quemada [\[39\]](#):

- Conatos <= 1 ha
- Incendios > 1 ha y < 500 ha
- Gran Incendio Forestal (gif) >= 500 ha

Identificación y Manejo de Datos Faltantes y Anomalías

El conjunto de datos tiene algunos valores ausentes tal y como se muestra en el gráfico de barras:

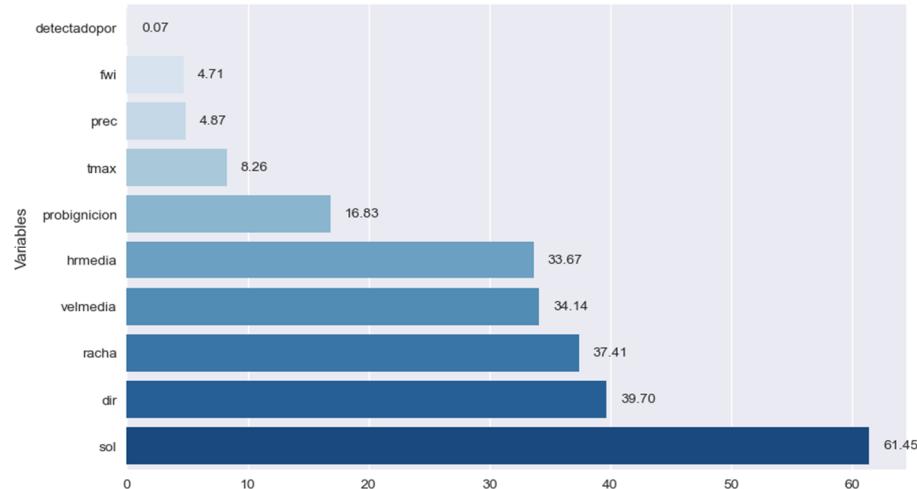


Ilustración 11. Porcentaje de datos faltantes después de la incorporación de los datos de AEMET

Se valoran las siguientes estrategias para resolver el problema de los datos faltantes. Eliminar aquellos registros que tengan alguna variable nula, imputarlos aplicando KNN o imputarlos mediante la moda o la media.

Mostramos la distribución de datos faltantes por clase de incendio:

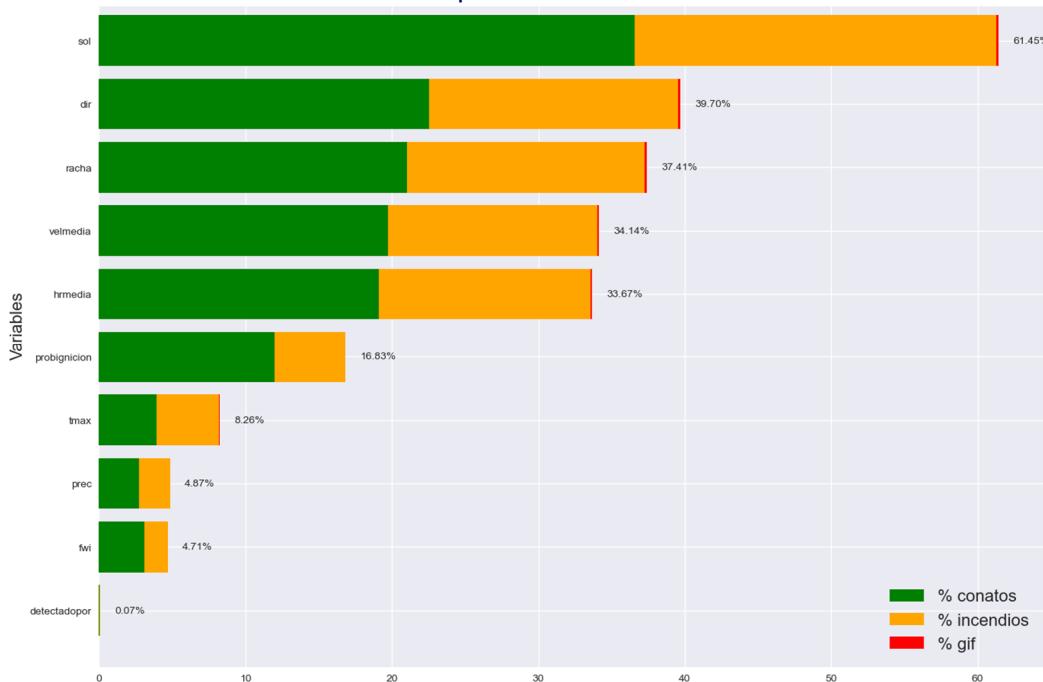


Ilustración 12. Porcentaje de datos faltantes en función de la clase de incendio

Se observa que la mayoría de los datos faltantes corresponden a conatos e incendios, mientras que los GIF representan una proporción mucho menor. Sin embargo, debido a la escasez de registros de GIF, cualquier pérdida en esta categoría puede ser crítica, ya que el modelado de los GIF es un objetivo clave.

Finalmente se decide imputar los valores faltantes, en los casos de variables categóricas por la moda o valor más frecuente. En el caso de la probabilidad de ignición se imputa por la media de la provincia. Y para el resto de las variables numéricas, se imputa en cascada empezando por la media de año, mes y provincia, en una segunda iteración por la media del año, mes y comunidad autónoma y finalmente, por la media de mes y la provincia.

En la revisión de duplicados comprobamos que no hay duplicidades en los registros.

3.4.1. Análisis Univariante

Análisis de las variables numéricas

Mostramos los descriptivos básicos

	count	mean	std	min	25%	50%	75%	max
probignicion	570372.0	14.771302	21.657798	0.000000	0.000000	0.000000	30.000000	1.000000e+02
poblacion	570372.0	13823.670034	63961.083329	1.000000	1061.037720	2677.232632	7999.623597	3.034878e+06
superficie	570372.0	13844.766023	16738.152259	1.260000	5159.767500	9364.851100	16795.857000	1.750229e+05
altitud	570372.0	444.235716	322.496031	1.000000	158.000000	423.000000	676.000000	1.695000e+03
lon	570372.0	-5.721403	3.078424	-17.998381	-8.035272	-6.571112	-4.112524	4.289666e+00
lat	570372.0	41.581651	1.992805	27.756103	40.719346	42.234413	42.853151	4.374035e+01
prec	570372.0	0.632092	3.208286	0.000000	0.000000	0.000000	0.000000	1.600000e+02
tmax	570372.0	24.560029	7.215267	-8.500000	19.700000	25.000000	30.000000	4.660000e+01
dir	570372.0	19.092042	8.848017	0.000000	13.000000	19.796117	26.000000	3.600000e+01
velmedia	570372.0	2.759745	1.708335	0.000000	1.700000	2.500000	3.340000	3.140000e+01
racha	570372.0	9.573598	3.388180	0.000000	7.500000	9.025000	10.952941	5.440000e+01
sol	570372.0	8.952031	2.741758	0.000000	7.479592	9.273134	10.900000	1.490000e+01
hrmedia	570372.0	58.818690	13.379487	4.000000	50.000000	59.000000	68.000000	1.000000e+02
denpoblacion	570372.0	1.412760	5.579249	0.000017	0.110435	0.260299	0.846355	2.484758e+02
denincendios	570372.0	0.013596	0.019278	0.000008	0.001980	0.006039	0.018188	7.936508e-01
fwi	570372.0	18.885638	20.087228	1.000000	1.000000	10.204297	34.207344	1.625351e+02
areaquemada	570372.0	12.963650	163.677376	0.000000	0.100000	0.780000	3.030000	3.069139e+04

Tabla 5. Descriptivos básicos de las variables numéricas

Hay variables que presentan una gran variabilidad como es el caso de la población, la superficie de los municipios, la altitud, las precipitaciones y el área que se ha quemado en el incendio que está muy sesgada ya que la mayoría de los incendios afectan a áreas pequeñas.

Variables como la precipitación y la temperatura máxima tienen valores elevados. También sorprende el hecho de que haya temperaturas máximas negativas que son legítimas. El área quemada además de tener una gran dispersión tiene valores máximos muy altos pero legítimos.

En general, las condiciones climáticas y geográficas son factores importantes que podrían influir en la propagación del incendio. Los factores demográficos como el tamaño de la población más cercana a los incendios podrían resultar relevante en cuanto a la extinción.

Visualización de datos atípicos

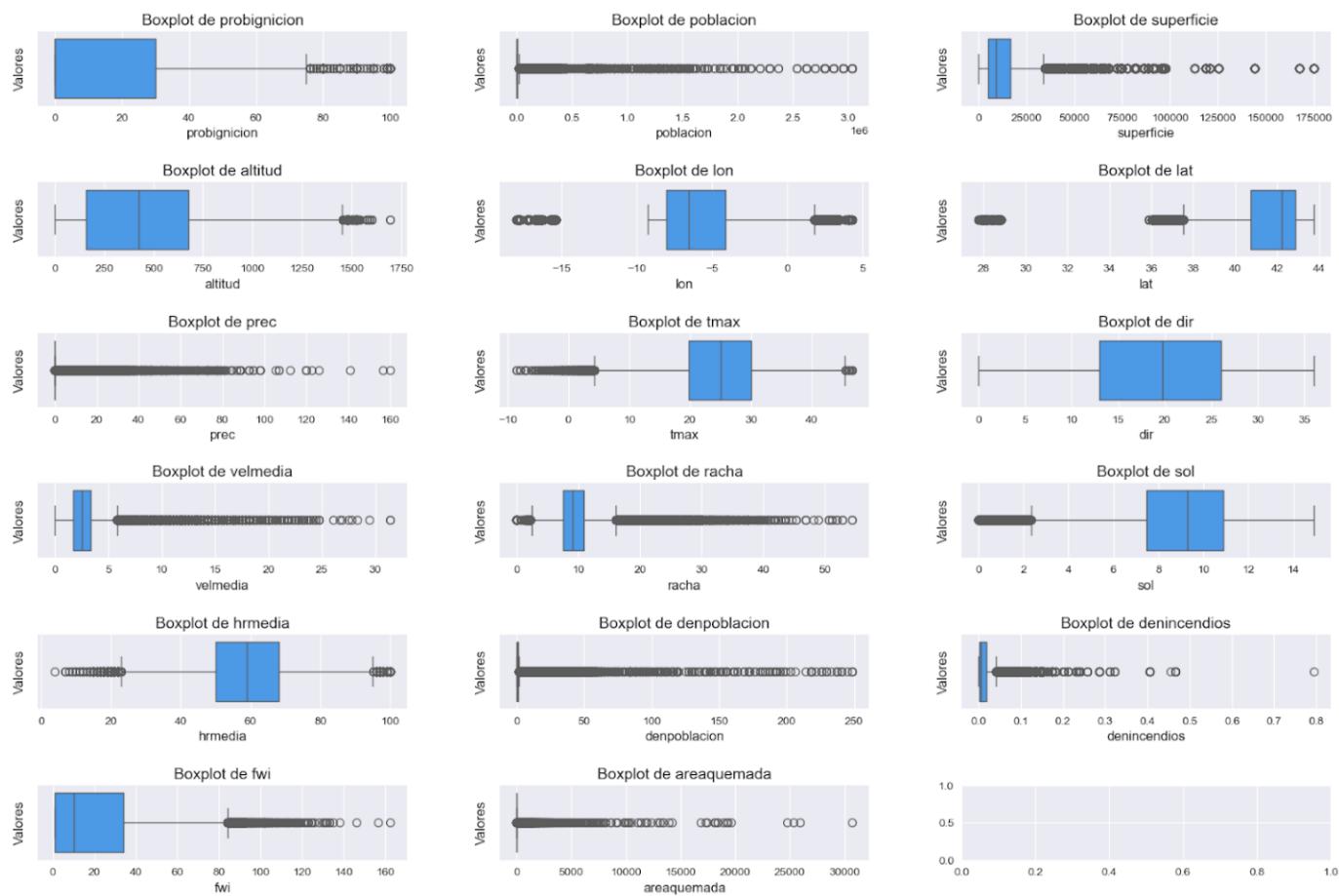


Ilustración 13. Visualización de los datos atípicos de las variables numéricas

Aunque a simple vista parece haber muchos datos atípicos son datos legítimos y que se mantienen en la base de datos, ya que van a contribuir al complejo modelado de los incendios forestales en general pero de los GIF en particular.

Visualización de la distribución

Mostramos a continuación la distribución de las variables numéricas en todo el conjunto.

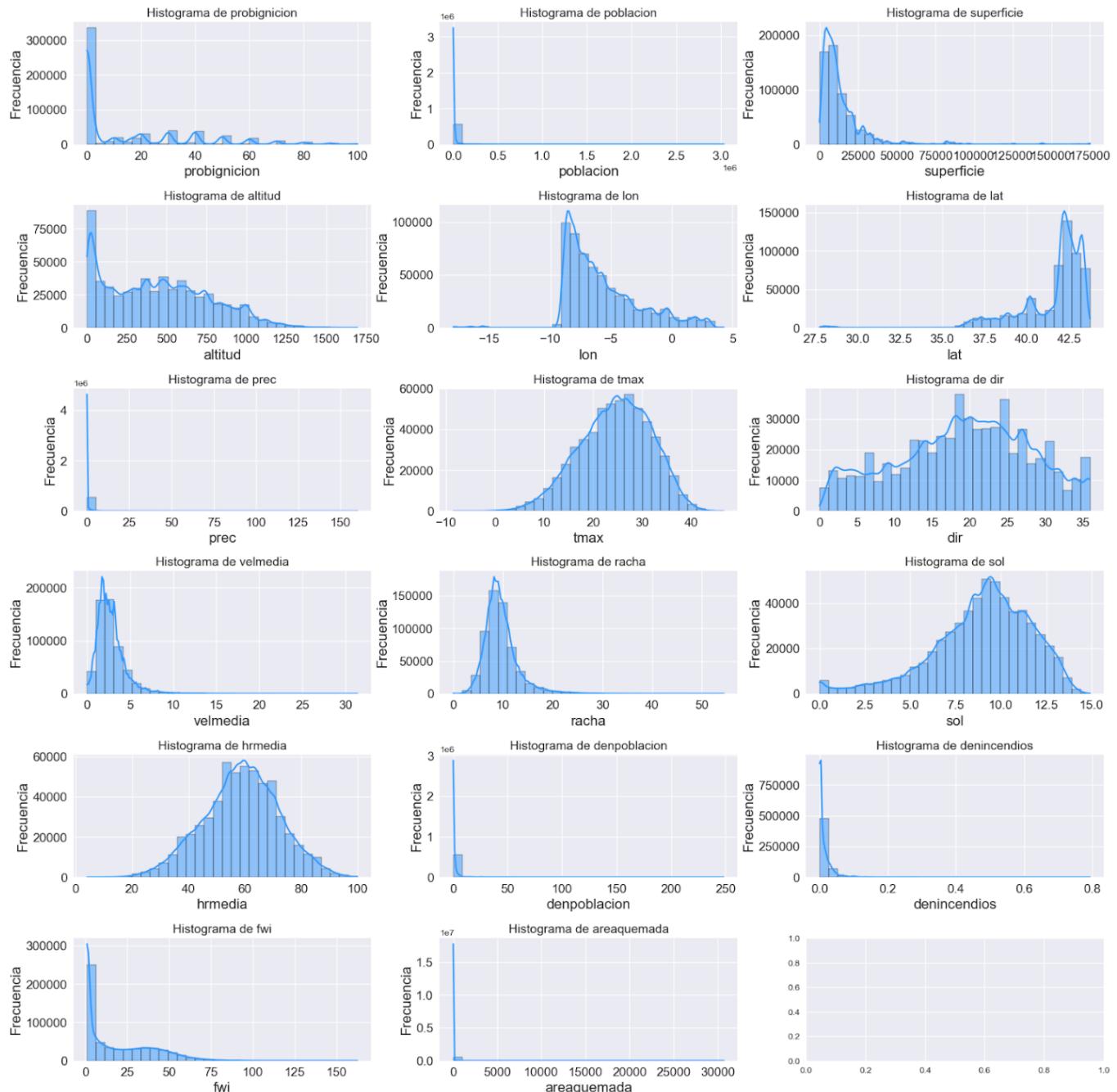


Ilustración 14. Visualización de las distribuciones de las variables numéricas

Observamos en las distribuciones que las variables no siguen distribuciones normales.

A continuación, mostramos la distribución de las variables numéricas por clase de incendio.

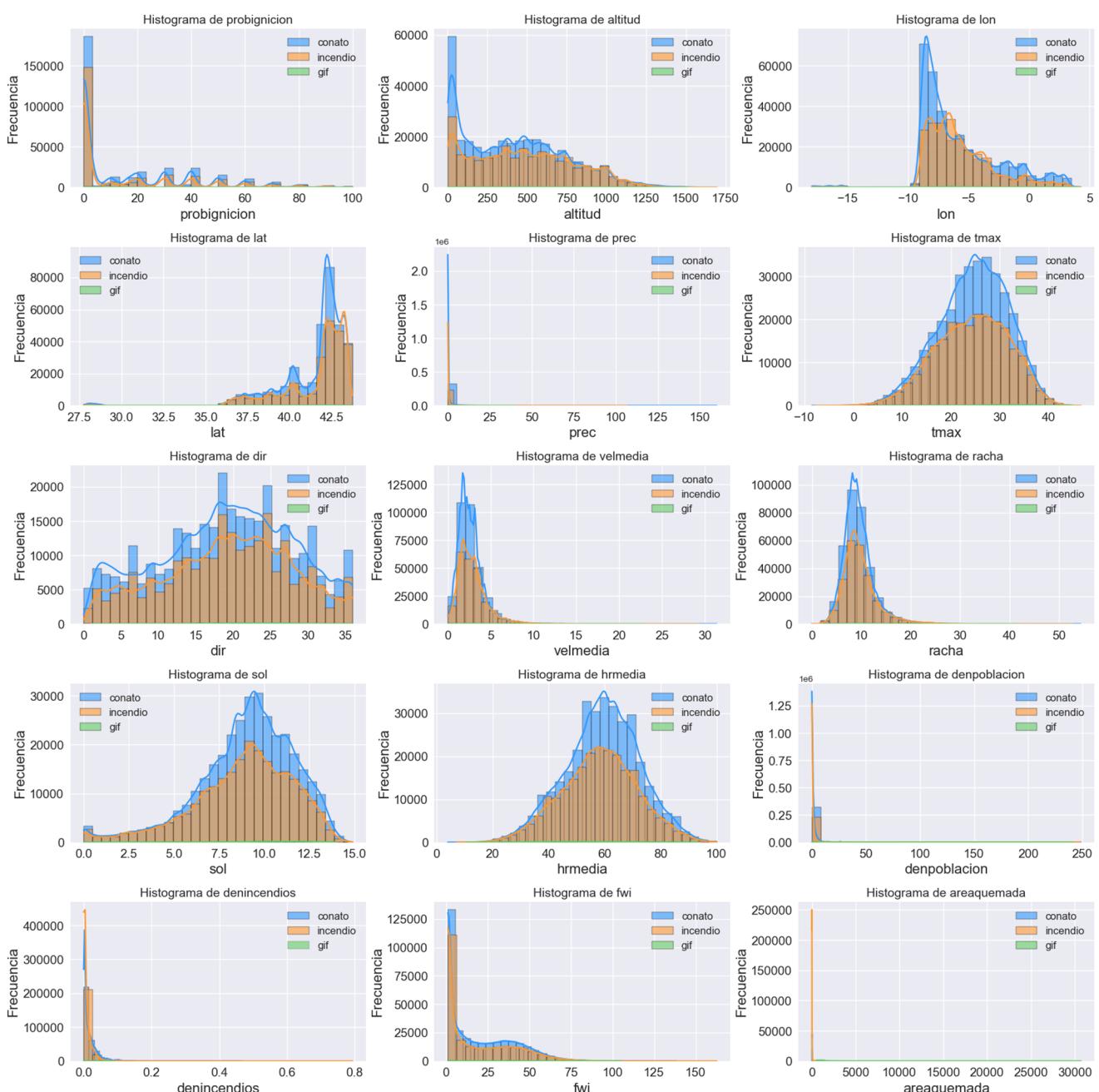


Ilustración 15. Visualización de las distribuciones de las variables numéricas por clase de incendio

Observamos que las variables numéricas siguen distribuciones similares en las clases conato e incendio y lo único que las diferencia son la cantidad de ocurrencias, algo normal, ya que tenemos más muestras de conatos que de incendios.

La clase GIF es tan pequeña con respecto a las otras dos que no se aprecia en el gráfico y por esta razón la mostramos por separado.

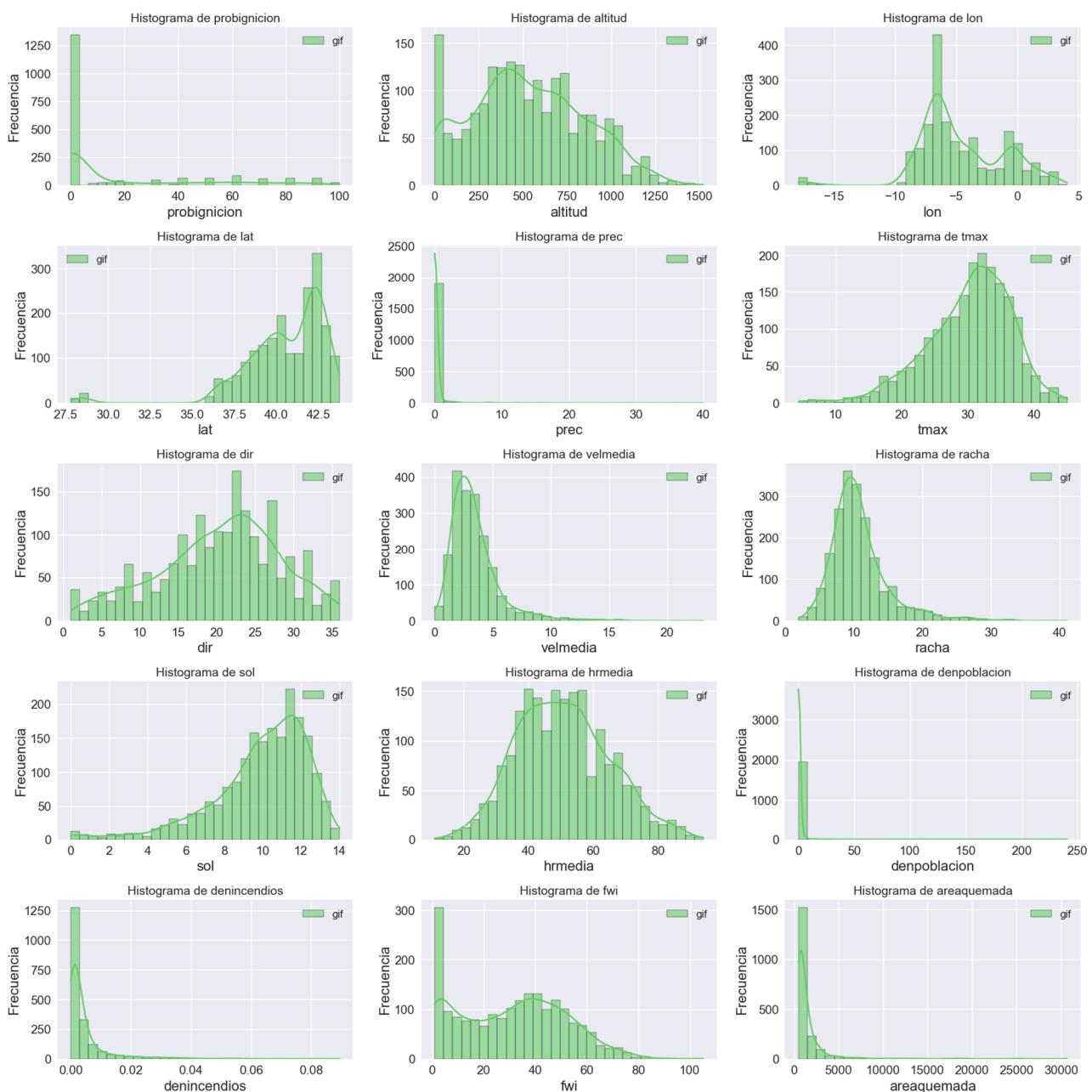


Ilustración 16. Visualización de las distribuciones de las variables numéricas en clase Gif

En el caso de los Gif se observa que en la distribución de la temperatura máxima su máximo se encuentra entre 30 y 35ºC mientras que en la distribución de conatos e incendios este máximo está desplazado a una temperatura entre 20 y 30ºC.

Con la longitud ocurre algo similar ya que el máximo de la distribución se desplaza hacia la derecha en comparación con las clases conato e incendio.

En el caso de la insolación, en los Gif predomina una mayor cantidad de horas de insolación.

Análisis de las variables categóricas

Mostramos a continuación la distribución de las variables categóricas en todo el conjunto.

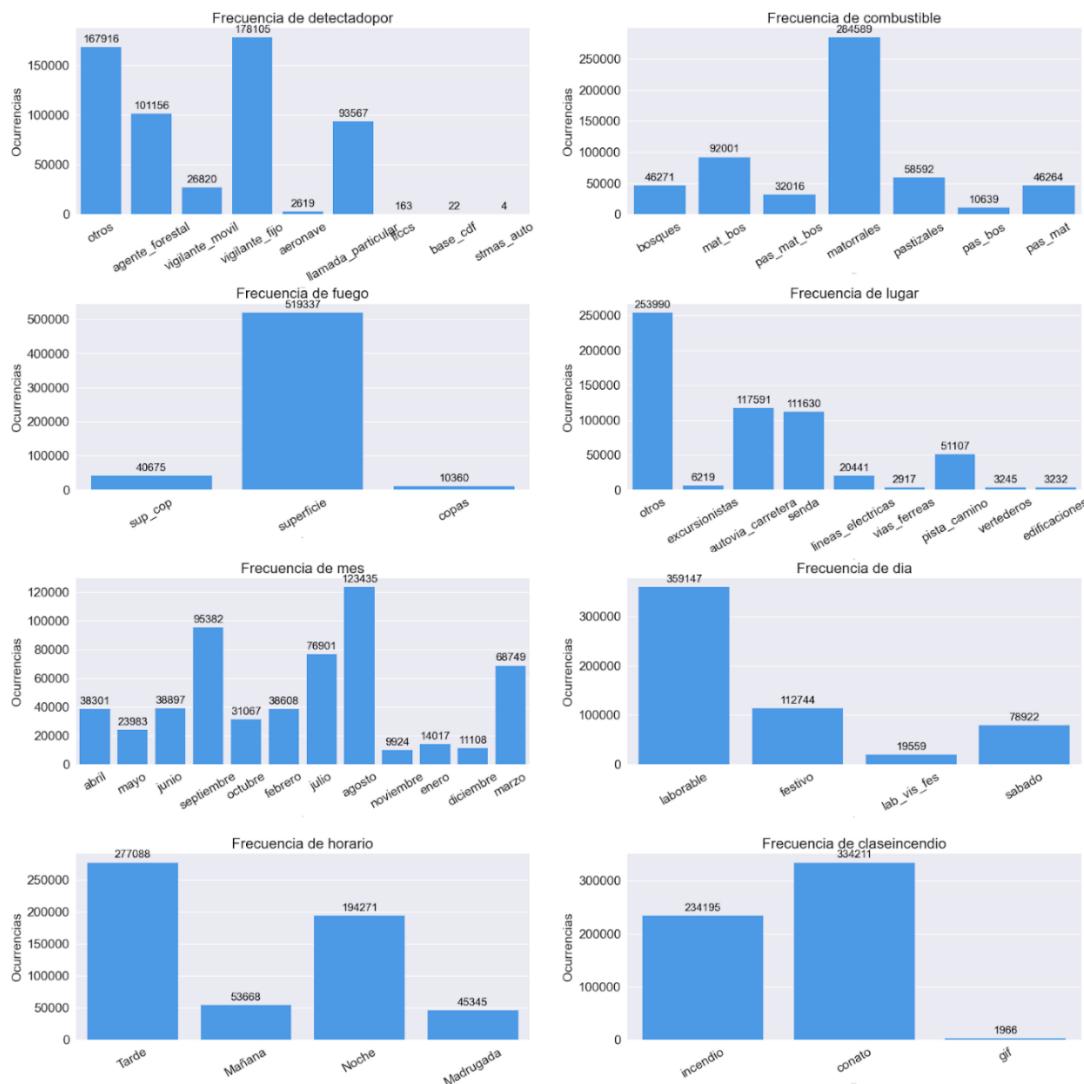


Ilustración 17. Visualización de las ocurrencias de las variables categóricas

Al examinar los diagramas de frecuencia de las variables categóricas, se identifican categorías con pocas ocurrencias y que son susceptibles de ser eliminadas o agrupadas para simplificar el análisis y el modelo. Se observan los siguientes casos:

Forma de detección del incendio. Categorías como “aeronave”, “ffccs”, “base_cdf” y “sistemas automáticos” tienen pocas ocurrencias lo que implica que podrían ser poco significativas para el modelo.

Lugar representativo de inicio del incendio. Algunas categorías como las inmediaciones de zonas por donde pasan excursionistas, las vías férreas, los vertederos y las edificaciones, tienen una baja representación así que se podrían reagrupar.

El tipo de día. Los días laborables que son víspera de festivos tienen un número de ocurrencias bastante menor al resto de días con lo que podría indicar una baja influencia en cuanto al comportamiento de los incendios.

El tipo de fuego. Presenta un dominio por parte de la categoría fuegos de superficie que hace pensar que esta variable tendrá una influencia baja en el modelado de los incendios.

El mes de ocurrencia de los incendios. Como era de esperar la temporada estival destaca por su mayor concentración de incendios, no obstante, destaca el mes de marzo por presentar una frecuencia significativa de incendios.

El horario de detección de los incendios se produce en la franja de tarde y noche mayoritariamente.

Mostramos las variables categóricas en función de la clase de incendio.

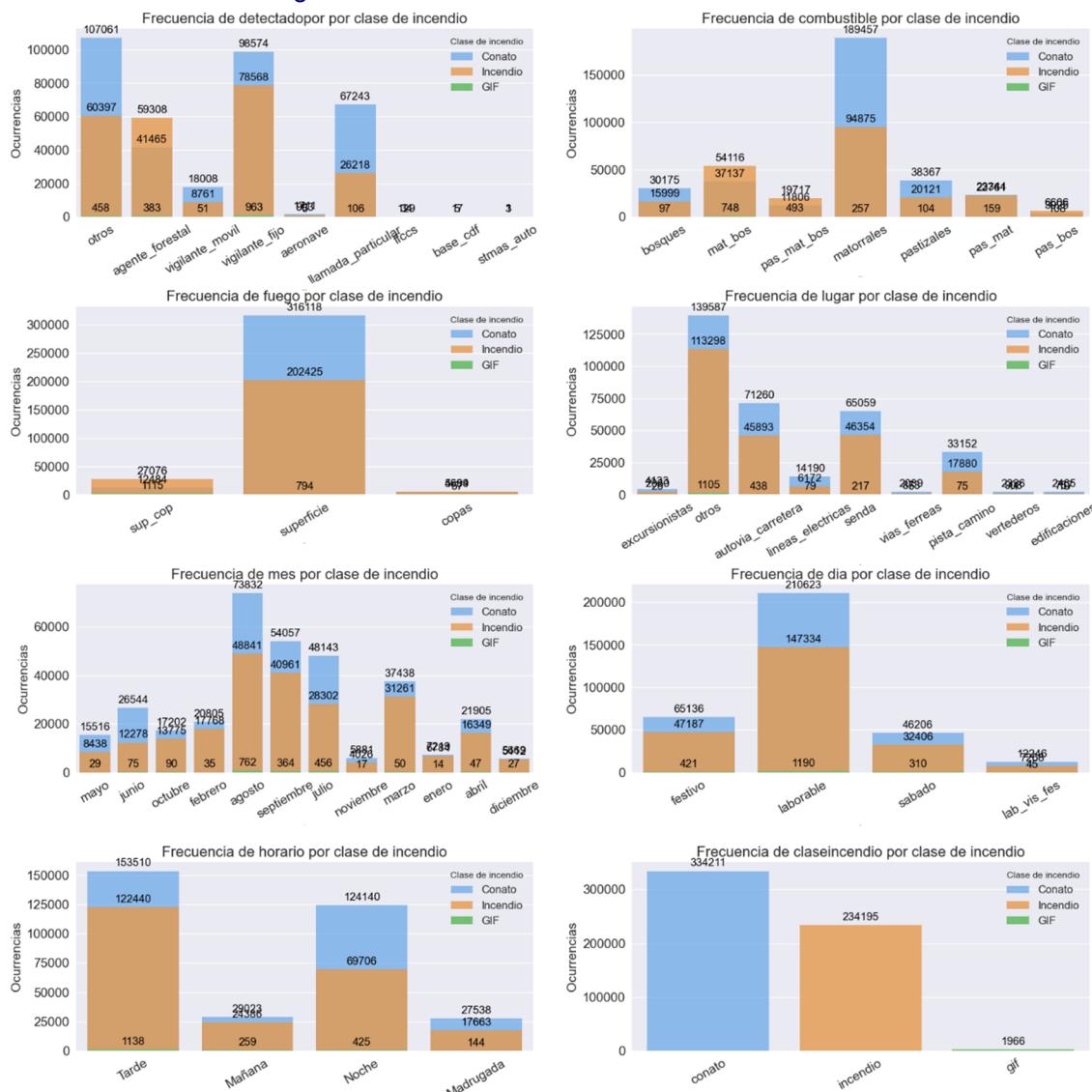


Ilustración 18. Variables categóricas en función de la clase de incendio

En cuanto a las clases incendio y GIF, vemos que hay algunas diferencias en el tipo de detección, donde es más frecuente en los incendios que sean detectados por un agente forestal o que el tipo de combustible de combinación entre matorrales y bosques sea más significativo de incendio; también el tipo de fuego que pertenece a copas o la combinación de estas con superficie sea indicativo de la clase de incendio más que de la de conatos. Finalmente, destacan los meses de marzo y abril como meses fuera de temporada activa de incendios donde también se producen una gran cantidad de conatos e incendios y no tanto de GIF como puede verse en los siguientes gráficos:

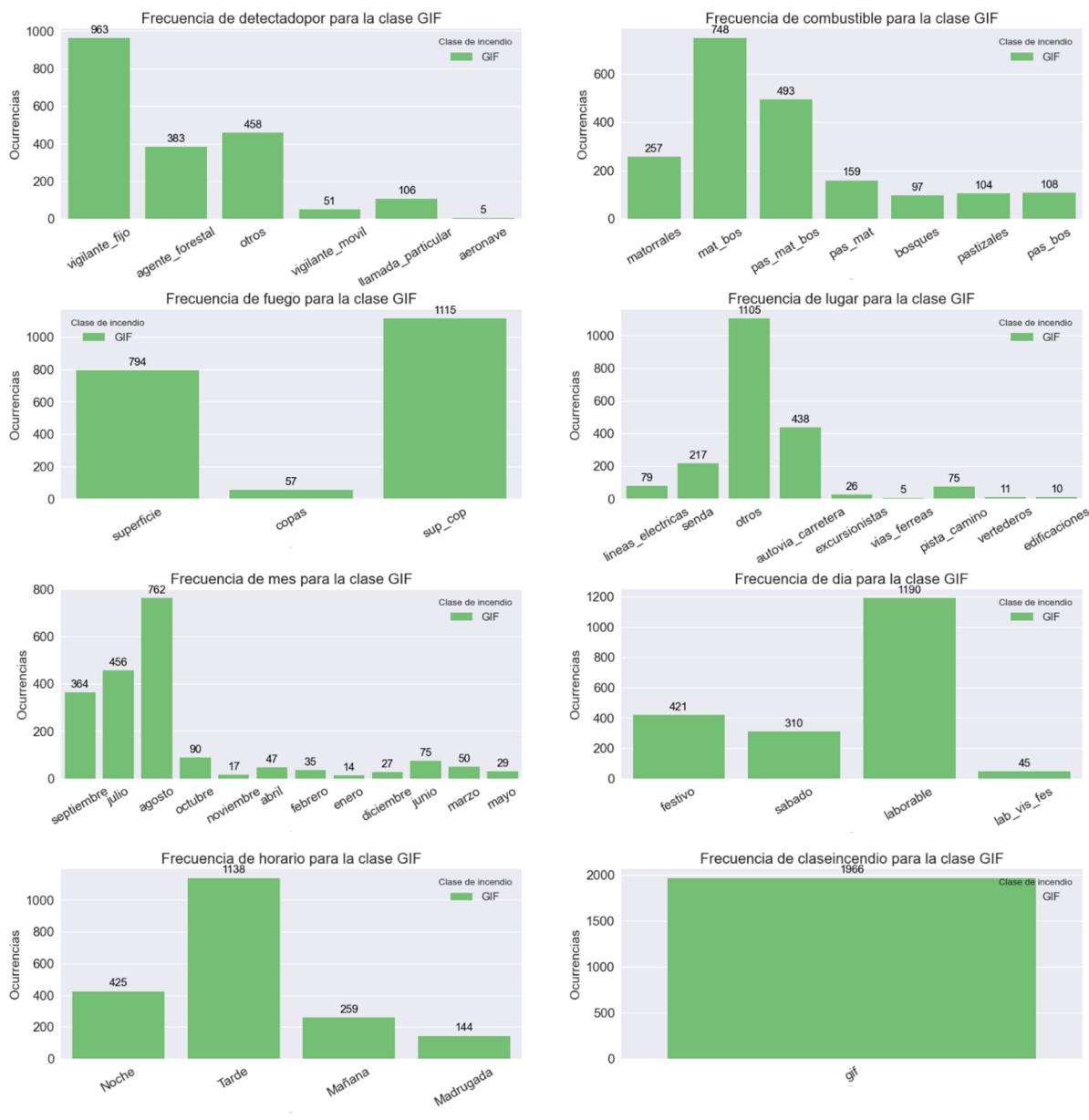


Ilustración 19. Variables categóricas en la clase GIF

Destaca el tipo de fuego donde el de superficie o la combinación de superficie y copas sea significativo de ser un GIF, aunque también coincide con la clase incendio. Así como el tipo de combustible con la combinación de matorrales y bosques, matorrales y bosques.

Análisis de la variable objetivo categorizada

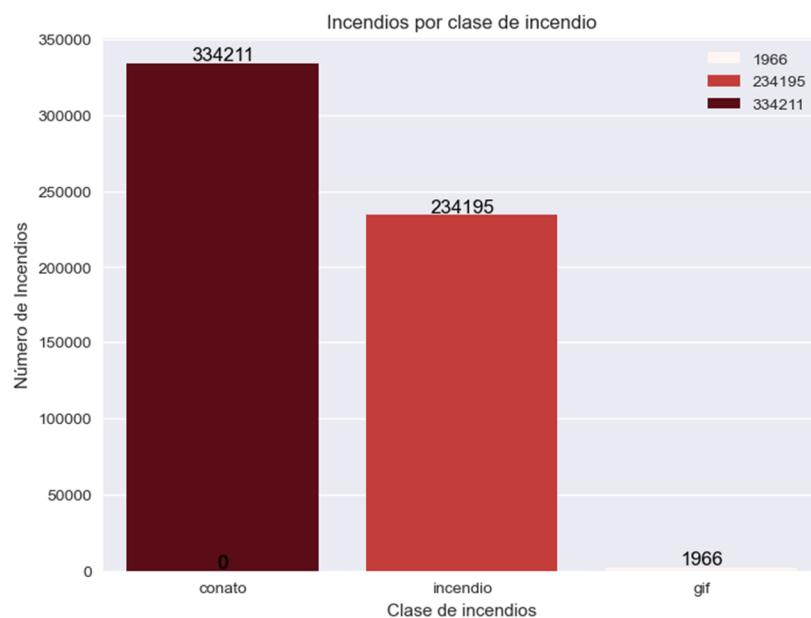


Ilustración 20. Distribución del número de ocurrencias por clase de incendio

En este diagrama de frecuencias se representan los incendios forestales ocurridos en España entre 1974 y 2016. La mayoría de los casos corresponden a conatos, es decir, incendios cuya superficie quemada es menor a 1 hectárea. Sin embargo, existe una gran variabilidad en los incendios restantes: una parte significativa abarca superficies quemadas desde 1 hasta 500 hectáreas, mientras que los grandes incendios forestales, que constituyen la clase minoritaria con solo 1.966 registros presentan áreas quemadas que oscilan entre 500 y 31.000 hectáreas.

Esta distribución plantea un problema para modelar el comportamiento de los incendios. Los grandes incendios forestales, por su baja frecuencia y gran variabilidad, podrían ser interpretados por el modelo como datos atípicos. No obstante, es importante desarrollar un modelo que sea capaz de caracterizar esta clase de incendios, ya que son los más destructivos y virulentos.

3.4.2. Análisis Multivariante

En este análisis se explora cómo las variables se relacionan entre sí, tanto para variables numéricas como categóricas.

Análisis de Variables Numéricas

Construimos la matriz de *Spearman* ya que no requiere que haya una distribución normal en las variables y es más robusta frente a valores atípicos.

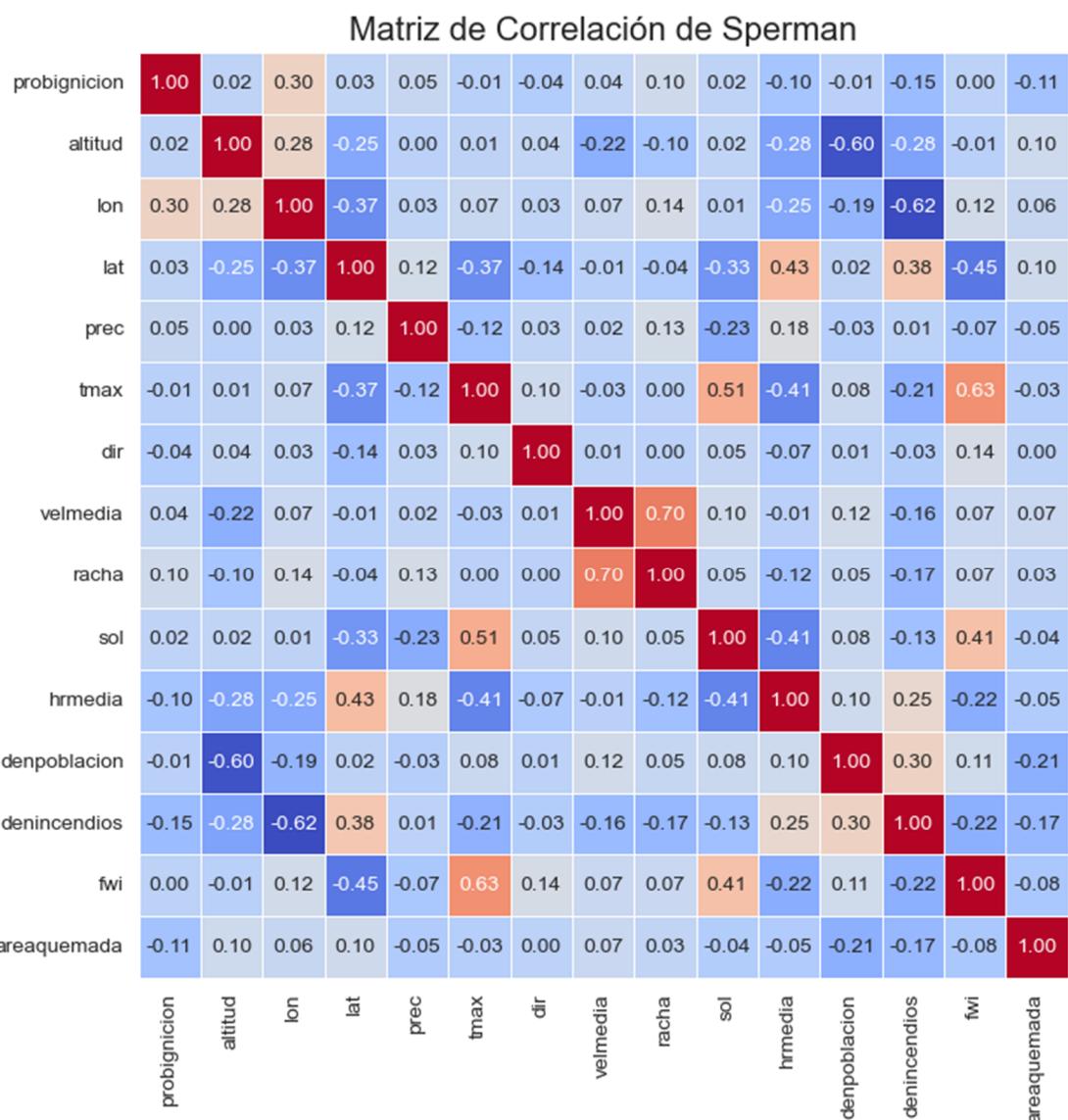


Ilustración 21. Matriz de correlación de Spearman

En el análisis de correlación de *Spearman* se observa el comportamiento monótono de algunas variables, es decir de forma consistente pero no necesariamente proporcional. Analizamos a continuación algunas de ellas:

La variable objetivo, **área quemada**, tiene una correlación negativa de -0,21 con la densidad de población, según el coeficiente de Spearman. Esto sugiere que los núcleos con menor densidad de población tienden a estar asociados con incendios de mayor magnitud. Observamos una correlación negativa de -0,17 con la densidad de incendios lo que puede indicar que aquellas zonas con una mayor densidad de incendios registran áreas quemadas más pequeñas.

El **índice FWI** presenta correlaciones significativas con algunas variables meteorológicas. Por ejemplo, tiene una correlación positiva de 0,63 con la temperatura máxima y de 0,41 con la insolación, mientras que muestra una correlación negativa con la humedad relativa media. Estos

valores son coherentes, ya que el índice FWI se calcula a partir de datos meteorológicos históricos de días previos, como la temperatura, la humedad, el viento y las precipitaciones, para estimar las condiciones favorables para la propagación de incendios. Destaca también la correlación negativa de -0,45 con la latitud, lo que sugiere que las zonas más al sur de España tienden a experimentar valores más altos de FWI.

La **densidad de incendios** presenta una correlación negativa con la altitud de -0,28 lo que sugiere que las zonas más bajas pueden ser más propensas a tener una mayor densidad de incendios. La correlación negativa de -0,62 con la latitud y la positiva de 0,38 con la longitud refleja que la densidad de incendios es mayor en las zonas más al sur y este de España. Con la temperatura máxima tiene una correlación negativa de -0,21 lo que sugiere que lugares con temperaturas máximas más moderadas tienen una densidad más elevada de incendios. Una correlación positiva de 0,25 con la humedad relativa media indica que las zonas con mayor humedad relativa registran mayores densidades de incendio. Al igual que con la densidad de población que tiene una correlación de 0,30 lo que sugiere que las zonas más pobladas presentan mayor densidad de incendios.

La **humedad relativa** presenta correlaciones significativas con varias variables. Tiene una correlación negativa de -0,41 con la temperatura máxima, lo que indica que a medida que la temperatura aumenta, la humedad relativa tiende a disminuir, ya que el aumento de la temperatura genera un ambiente más seco. Por el contrario, presenta una correlación positiva de 0,18 con la precipitación, lo que sugiere que a mayor precipitación, mayor es la humedad relativa. Además, muestra una correlación positiva de 0,43 con la latitud, indicando que a medida que nos desplazamos hacia el sur de España, la humedad relativa tiende a aumentar. En cambio, tiene una correlación negativa de -0,25 con la longitud, lo que implica que la humedad relativa disminuye al desplazarnos hacia el este. Del mismo modo, la humedad relativa disminuye con el aumento de la altitud, como lo indica la correlación negativa de -0,28.

La **insolación** presenta diversas correlaciones con variables meteorológicas. Tiene una correlación negativa de -0,33 con la latitud, lo que sugiere que a medida que nos desplazamos hacia el norte de España (latitudes más altas), la insolación tiende a disminuir. Por otro lado, presenta una correlación positiva de 0,51 con la temperatura máxima, lo que refleja que un aumento en la insolación generalmente se asocia con un aumento en la temperatura. También tiene una correlación positiva de 0,41 con la humedad relativa media, lo que sugiere que mayores niveles de insolación pueden estar asociados con un aumento en la humedad relativa, posiblemente por el efecto de la evaporación. Finalmente, la insolación muestra una correlación positiva de 0,41 con el índice FWI, lo que implica que un mayor nivel de insolación puede estar relacionado con condiciones más favorables para la propagación de incendios.

La **racha** presenta una correlación positiva muy fuerte con la velocidad media, lo que sugiere que cuando las rachas de viento son intensas, la velocidad media del viento también tiende a aumentar.

La **velocidad media** muestra una correlación negativa de -0,22 con la altitud, lo que sugiere que en las zonas de mayor altitud, la velocidad media del viento tiende a ser menor.

La **temperatura máxima** presenta las siguientes correlaciones: una correlación negativa de -0,37 con la latitud, lo que sugiere que a medida que nos desplazamos hacia el norte, la temperatura máxima tiende a disminuir; una correlación positiva de 0,51 con la insolación, indicando que mayores niveles de insolación se asocian con temperaturas máximas más altas; una correlación negativa de -0,41 con la humedad relativa media, lo que implica que a medida que la humedad relativa aumenta, la temperatura máxima tiende a ser más baja; y finalmente, una correlación positiva de 0,63 con el índice FWI, lo que refleja que temperaturas máximas más altas están asociadas con condiciones más favorables para la propagación de incendios.

La **precipitación** presenta correlaciones moderadas con otras variables. La más significativa es la de la insolación con la que tiene un valor de -0.23 y que indica de forma general que las zonas con menos insolación podrían experimentar mayores precipitaciones.

Análisis de Variables Categóricas

Para analizar las variables categóricas se ha utilizado el test de Chi-cuadrado para la independencia que determina si hay una relación significativa entre las variables categóricas a estudiar. Para ello se ha creado la siguiente matriz:

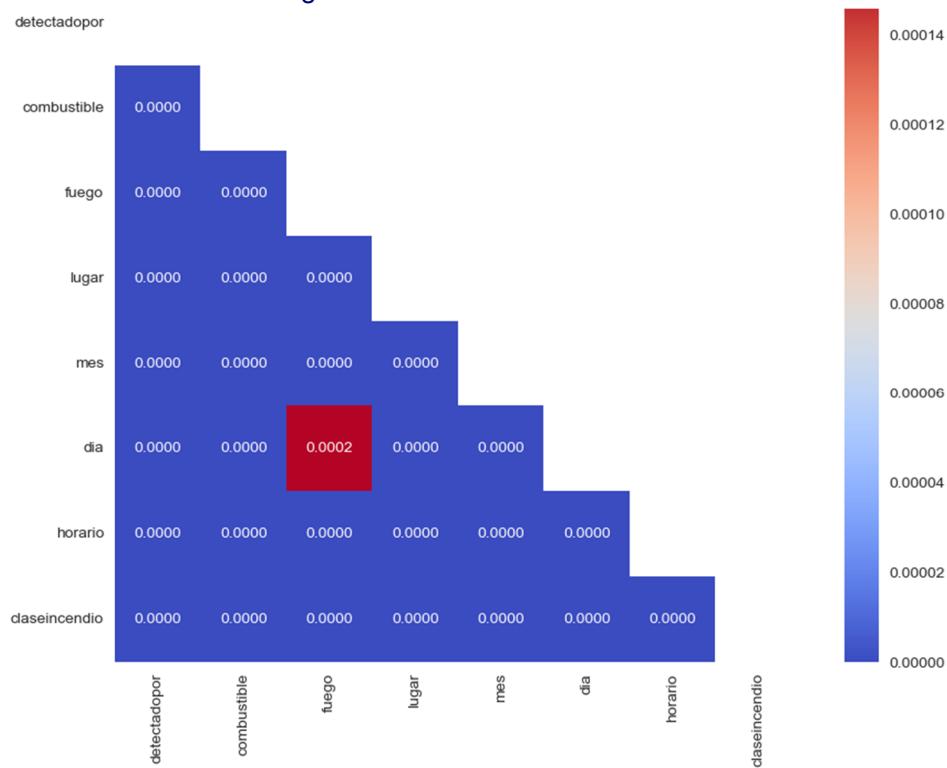


Ilustración 22. Test Chi-cuadrado para las variables categóricas.

Los valores de p obtenidos en el análisis de las relaciones entre las variables categóricas (incluyendo la variable objetivo, que representa las diferentes clases de incendios) son todos menores a 0.05. Esto indica que las relaciones observadas son estadísticamente significativas, lo que sugiere que existe una relación no aleatoria entre ellas, es decir, las asociaciones entre las variables no son fruto de simples coincidencias, sino que reflejan relaciones reales entre las variables.

Gráfico de barras agrupadas

En los diagramas de barras agrupados por clase de incendio se muestra el porcentaje que cada uno de los atributos pertenecientes a cada variable tiene sobre cada clase de incendio.

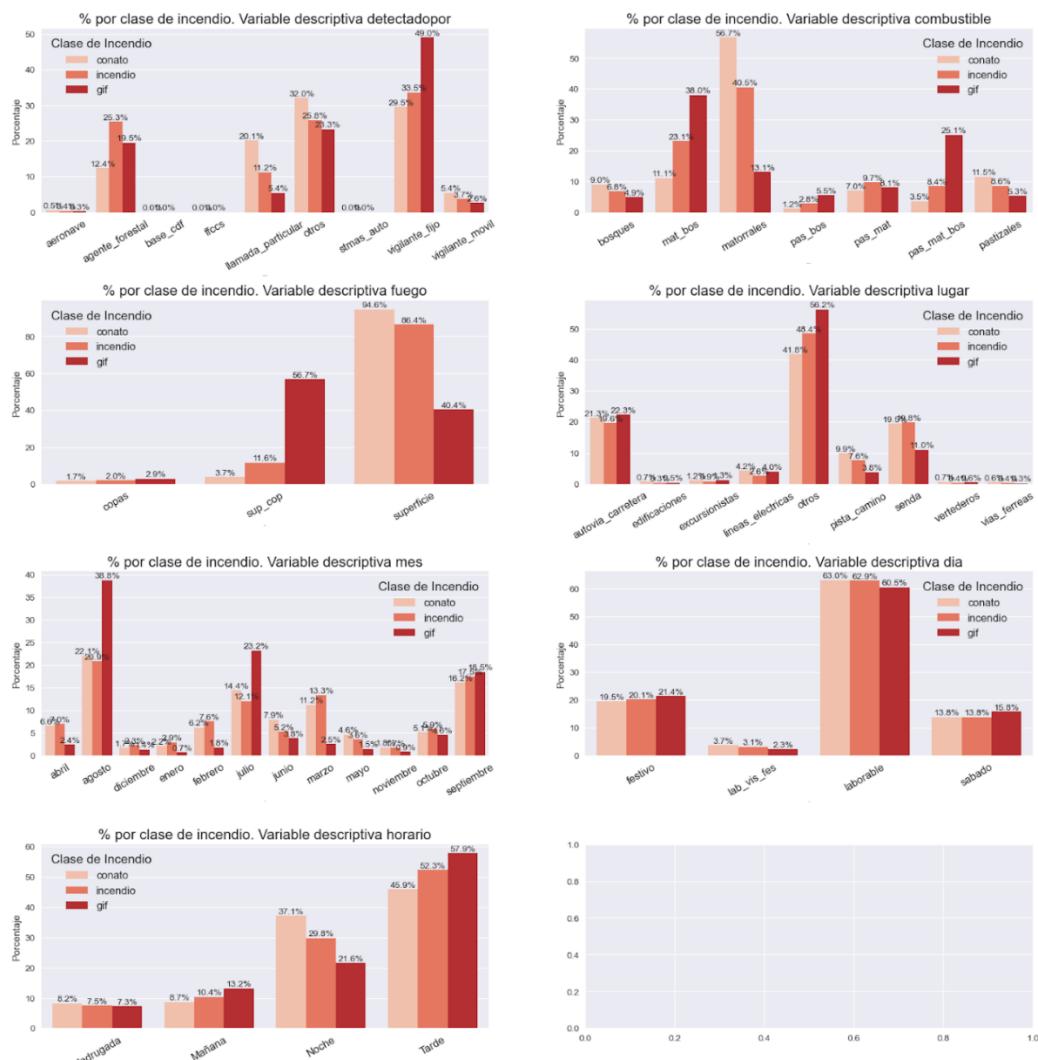


Ilustración 23. Visualización de las variables categóricas en función de la clase de incendio.

En los gráficos de barras destaca que el atributo de la variable descriptiva detectado por el vigilante fijo aparece en un 49% de las ocurrencias de los GIF. En cuanto al combustible destaca que en los GIF tiene especial relevancia la mezcla de matorral y bosques con un 38% y la mezcla de matorral, pastizal y bosques, con un 25.1% de sus ocurrencias mientras que en los conatos es muy significativo el combustible de tipo matorral en su gran mayoría de ocurrencias. En el tipo de fuego también es muy significativo en los GIF el tipo combinado de superficie y copas, mientras que el que solo se produce en superficie se asocia más a los incendios y conatos, aunque también presenta un alto porcentaje en el caso de los GIF, pero tal y como se puede ver en el gráfico el atributo superficie y copas combinado para el tipo de fuego es bastante descriptivo de un GIF. Por último los meses de julio y agosto también suponen ser meses frecuentes en los que se han producido grandes incendios forestales.

3.5. Conclusiones del EDA

Creación de nuevas variables

- La densidad de población por municipio.
- La densidad de incendios por municipio.
- El rango horario de detección del incendio.

Manejo de datos faltantes

En este aspecto se optó por imputar los valores faltantes en lugar de eliminar registros debido a la baja cantidad de registros de grandes incendios forestales que hay en comparación con las otras clases y que se consideran clave para el modelado de los incendios forestales.

Análisis univariante

Se observan valores atípicos entre las variables numéricas que se mantienen al valorar que son datos legítimos. Las distribuciones de muchas de las variables no son normales, lo que implica que habrá que realizar transformaciones en los datos o utilizar modelos que no asuman normalidad como es el caso de los árboles de decisión y las redes neuronales.

Algunas variables categóricas tienen ciertos atributos con pocas ocurrencias como la forma de detección del incendio o el tipo de fuego que podría sugerir una eliminación o reagrupación por su baja relevancia, pero que finalmente fue descartada en el análisis multivariante.

Con respecto a la variable objetivo, la mayor parte de los incendios son conatos de menos de 1 hectárea de área quemada y el resto corresponde en su gran mayoría a incendios de entre 1 y 500 hectáreas quemadas, algo menos de 2 mil registros que suponen solo un 0,3% de la base de datos se corresponden con grandes incendios forestales que son los más destructivos y su baja ocurrencia dificulta el modelado.

El sesgo hacia los conatos implica que tendremos que usar técnicas de balanceo de clases como submuestreo o sobremuestreo al centrarnos en resolver el problema como un problema de clasificación pues buscamos predecir la severidad a alcanzar por el incendio forestal.

Análisis multivariante

En el análisis multivariante de las variables numéricas no se han encontrado relaciones relevantes con respecto a la variable objetivo lo que indica que modela los incendios forestales para predecir la severidad del incendio implica una alta complejidad, ya que, aunque se identifican algunas tendencias, no son lo suficientemente fuertes como para resolver el problema con modelos sencillos.

Las relaciones entre las variables categóricas resultaron ser estadísticamente significativas, por tanto, las relaciones observadas no son aleatorias.

En cuanto a la relación entre las variables categóricas se analizaron especialmente las ocurrencias con respecto a las clases de incendios y se puso especial atención a los GIF. Son representativas el tipo de fuego que combina la superficie y las copas al mismo tiempo; en global, tiene pocas ocurrencias y casi todos los incendios presentan fuego de superficie, pero la mitad de los GIF han presentado este tipo de fuego.

Descripción de la base de datos para el modelo

Las variables descritas a continuación serán las que se introduzcan en los diferentes modelos con el objetivo de predecir la severidad del incendio forestal:

- **areaquemada** (float64): Área quemada en hectáreas. Esta es la variable objetivo en el modelo que describe el tamaño del incendio.
- **probignicion** (float64): Probabilidad de ignición en la zona del incendio.
- **altitud** (float64): Altitud de la zona afectada. Se mide en metros.
- **lon** (float64) y **lat** (float64): Coordenadas geográficas (longitud y latitud)
- **prec** (float64): Precipitación diaria en milímetros.
- **tmax** (float64): Temperatura máxima registrada en el día en °C.
- **dir** (float64): Dirección de la racha máxima en decenas de grado.
- **velmedia** (float64): Velocidad media del viento en m/s.
- **racha** (float64): Racha máxima de viento en m/s.
- **sol** (float64): Insolación en horas.
- **hrmedia** (float64): Humedad relativa media en porcentaje.
- **fwi** (float64): Índice de riesgo de incendio es adimensional.
- **denpoblacion** (float64): Densidad de población en la zona en habitantes/ha.
- **denincendios** (float64): Densidad de incendios en la zona en los últimos cinco años en incendios/ha.
- **detectadopor** (object): Variable categórica que indica quién detectó el incendio. Con los siguientes atributos: 'otros', 'agente_forestal' , 'vigilante_movil' , 'vigilante_fijo', 'aeronave', 'llamada_particular' , 'ffccs', 'base_cdf', y 'stmas_auto'.
- **combustible** (object): Tipo de combustible presente en la zona afectada por el incendio. Con los siguientes atributos: 'bosques', 'mat_bos', 'pas_mat_bos', 'matorrales' , 'pastizales', 'pas_bos' y 'pas_mat'.
- **fuego** (object): Tipo de fuego registrado. Con los siguientes atributos: 'sup_cop', 'superficie' y 'copas'.
- **lugar** (object): Ubicación del incendio, que puede referirse a áreas específicas o regiones. Con los siguientes atributos: 'otros', 'excursionistas', 'autovia_carretera', 'senda', 'lineas_electricas', 'vias_ferreas', 'pista_camino', 'vertederos' y 'edificaciones'.
- **mes** (object): Mes en el que ocurrió el incendio.
- **dia** (object): Tipo de día en el que se inició el incendio. Con los siguientes atributos: 'laborable', 'festivo', 'lab_vis_fes' y 'sabado'.
- **horario** (object): Rango horario en que se registró el incendio. Con los siguientes atributos: 'Tarde', 'Mañana', 'Noche' y 'Madrugada'
- **claseincendio** (object): Categoría de incendio según la clasificación oficial en conatos, incendios y GIF.

4. Resultados

4.1. Selección de características

Para realizar la selección de características [40] y reducir la dimensionalidad del modelo se utilizó una Regresión Logística Multiclasa y un *Random Forest*.

Con la Regresión Logística se buscan relaciones lineales entre características y clases. A partir de la regresión se obtuvo una evaluación preliminar de la importancia de características y de las dificultades a abordar en el modelo, ya que nos encontramos ante un conjunto de datos muy desbalanceado con una clase minoritaria Gif especialmente difícil de clasificar.

					MATRIZ DE CONFUSIÓN Regresión Logística Multiclasa			
					conato	15901	6102	
					Incendio	21350	9230	
					gif	30	53	298
precision	recall	f1-score	support		conato	44800	15901	6102
0	0.73	0.67	0.70	66803	Incendio	16311	21350	9230
1	0.57	0.46	0.51	46891				
2	0.02	0.78	0.04	381				
accuracy			0.58	114075	gif	30	53	298
macro avg	0.44	0.64	0.41	114075				
weighted avg	0.66	0.58	0.62	114075				

Tabla 6. Selección de características. Resultados de una Regresión Logística Multiclasa

Posteriormente, se implementó un *Random Forest* con el que capturar las relaciones más complejas y no lineales. El modelo proporcionó una selección de características con un gran número de coincidencias con la Regresión Logística, en concreto 26 características en común. Finalmente, se aplicó un criterio de selección basado en el 95% de importancia acumulada de las variables con el que se consiguió disminuir la dimensionalidad. Además, el *Random Forest* obtuvo en general mejores resultados por lo que finalmente se realizó la selección de variables con este modelo.

					MATRIZ DE CONFUSIÓN Random Forest			
					conato	10830	1	
					Incendio	27592	12	
					gif	29	360	4
precision	recall	f1-score	support		conato	56012	10830	1
0	0.74	0.84	0.79	66843	Incendio	19235	27592	12
1	0.71	0.59	0.64	46839				
2	0.24	0.01	0.02	393				
accuracy			0.73	114075	gif	29	360	4
macro avg	0.56	0.48	0.48	114075				
weighted avg	0.73	0.73	0.73	114075				

Tabla 7. Selección de características. Resultados de Random Forest.

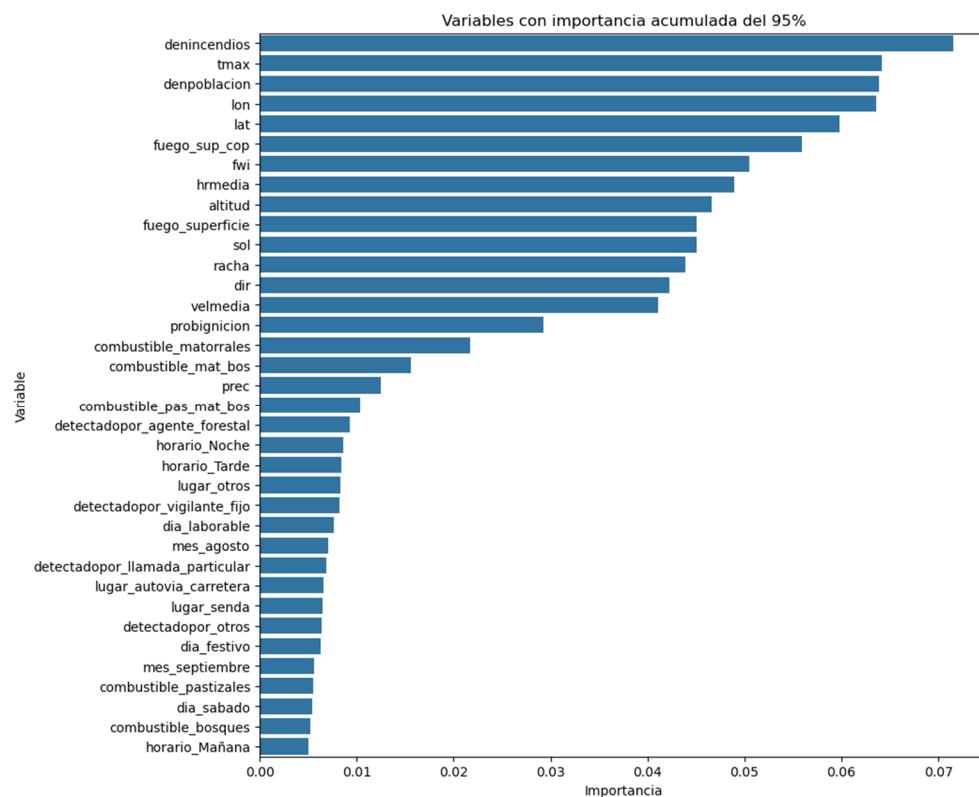


Ilustración 24. Selección de características con Random Forest.

4.2. Pruebas preliminares con *Random Forest*

Se continúa con *Random Forest* como modelo inicial y se realizan una serie de pruebas preliminares del desempeño de los datos con la selección de características escogidas. Se entrena el modelo bajo las siguientes premisas:

- Con el conjunto de datos desbalanceado
- Realizando un submuestreo aleatorio de las clases mayoritarias conato e incendio al nivel de la clase minoritaria GIF
- Un sobremuestreo de la clase minoritaria GIF a las mayoritarias conato e incendio.

En la siguiente tabla se pueden observar los resultados:

Balanceo de clases	Clase	Precision	Recall	F1-Score		Precision	Recall	F1-Score
Sin balanceo	conato	0,74	0,84	0,79	accuracy			0,73
	incendio	0,71	0,59	0,65	macro avg	0,56	0,48	0,48
	gif	0,28	0,01	0,02	weighted avg	0,73	0,73	0,72
Submuestreo	conato	0,63	0,68	0,66	accuracy			0,65
	incendio	0,57	0,52	0,54	macro avg	0,65	0,65	0,65
	gif	0,77	0,76	0,77	weighted avg	0,65	0,65	0,65
Sobremuestreo	conato	0,77	0,77	0,77	accuracy			0,73
	incendio	0,67	0,66	0,67	macro avg	0,51	0,52	0,52
	gif	0,09	0,13	0,11	weighted avg	0,73	0,73	0,73

Tabla 8. Resultados de Random Forest: desbalanceado, submuestreado y sobremuestreado

Cuando no se realiza ningún tipo de balanceo, la clase conatos está bien representada, frente a la clase incendios que tiene más dificultades para ser identificada correctamente. Por último, la clase GIF tiene un rendimiento muy bajo en la que apenas puede predecirse.

Al aplicar un submuestreo, se produce una disminución de la capacidad de predicción para las clases conato e incendio en favor de la clase GIF que muestra una mejora significativa ya que el modelo pasa a enfocarse también en esta clase.

Por otro lado, el sobremuestreo de la clase minoritaria genera mejoras en las clases mayoritarias pero la clase GIF no es capaz de mejorar a pesar del balanceo y sigue siendo difícil de predecir.

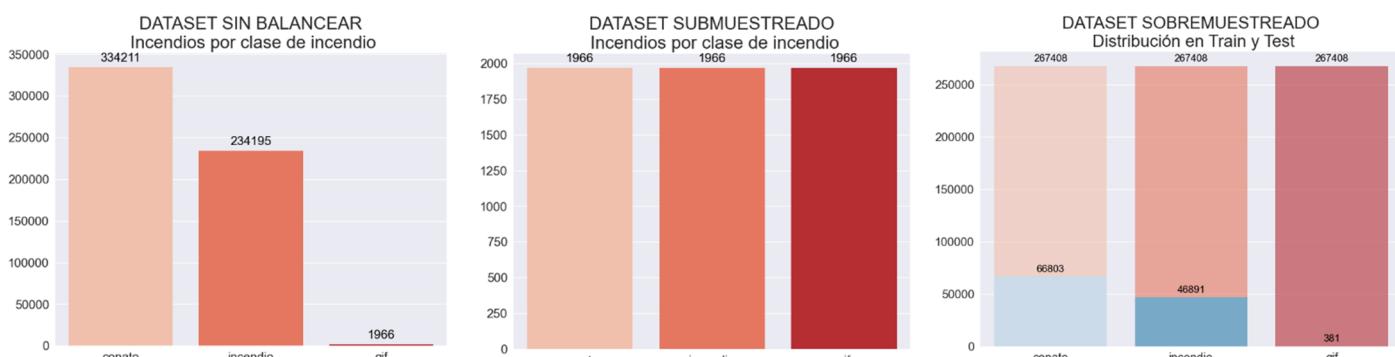


Ilustración 25. Gráficos comparativos del número de registros en el dataset desbalanceado, submuestreado y sobremuestreado

Estos resultados preliminares indican que el conjunto de datos sin balancear no permite identificar de manera efectiva una clase tan importante como la de los grandes incendios forestales. Además, el sobremuestreo sin control no mejora los resultados debido a la gran cantidad de muestras sintéticas generadas para las clases minoritarias. Por lo tanto, se opta por realizar un submuestreo del conjunto de datos que permite obtener un mejor equilibrio de las clases.

4.3. Submuestreo de las clases conato e incendio

Para realizar el submuestreo de las clases mayoritarias, conato e incendio, se utilizaron tres implementaciones:

- *Cluster-Based*
- *Tomek Links*
- *Edited Nearest Neighbors (ENN)*

Submuestreo Clases Mayoritarias	Clase	Muestras	Precision	Recall	F1-Score		Precision	Recall	F1-Score
Cluster-Based	conato	3000	0,65	0,75	0,69	accuracy			0,62
	incendio	3000	0,54	0,49	0,51	macro avg	0,63	0,62	0,62
	gif	1966	0,70	0,62	0,66	weighted avg	0,62	0,62	0,62
Tomek Links	conato	288570	0,69	0,75	0,72	accuracy			0,67
	incendio	188155	0,61	0,56	0,58	macro avg	0,68	0,67	0,38
	gif	1966	0,74	0,71	0,72	weighted avg	0,67	0,67	0,67
Edited Nearest Neighbors (ENN)	conato	70626	0,89	0,95	0,92	accuracy			0,83
	incendio	26405	0,81	0,80	0,81	macro avg	0,81	0,81	0,81
	gif	1966	0,75	0,67	0,71	weighted avg	0,82	0,83	0,83
ENN + Tomek Links	conato	70482	0,90	0,94	0,92	accuracy			0,84
	incendio	26177	0,82	0,82	0,82	macro avg	0,82	0,82	0,82
	gif	1966	0,74	0,70	0,72	weighted avg	0,83	0,84	0,84

Tabla 9. Resultados de Random Forest con diferentes técnicas de submuestreo de las clases mayoritarias conato e incendio

El submuestreo con *Cluster-Based* consigue una mejora de la clase GIF, pero tiene algunos problemas para identificar la clase conato y la clase incendio que también ha disminuido sus métricas con respecto al submuestreo realizado aleatoriamente en el paso anterior.

En el caso de *Tomek Links* obtenemos una mejora en los resultados con respecto a *Cluster-Based* pero siguen siendo moderados. Donde se experimenta una mejoría es en la clase GIF que tiene buenos valores de precisión y *recall*.

El submuestreo con ENN mejora significativamente las métricas de las clases conato e incendio lo que indica que el submuestreo está permitiendo predecir mejor las clases mayoritarias. Por otro lado, la clase minoritaria GIF experimenta un mejor rendimiento pero el *recall*, que sigue siendo bajo, indica que sigue teniendo dificultades para ser predicha.

Con la combinación de ENN y *Tomek Links* obtenemos una precisión muy buena para la clase conato junto con la clase incendio que también presenta un buen rendimiento, mientras que la clase minoritaria sigue experimentando algunas dificultades para ser identificada.

Después de aplicar las diferentes técnicas de submuestreo, el *dataset* se reduce al número de muestras que aparece en la tabla para cada método.

Dados los resultados, se realiza un submuestreo del conjunto de datos aplicando una combinación de ENN y *Tomek Links* [41].

4.4. Sobremuestreo de la clase minoritaria GIF

Tras el submuestreo de las clases mayoritarias con ENN y Tomek Links el *dataset* continúa desbalanceado con alrededor de 70 mil muestras para la clase conato, más de 25 mil para la clase incendio y no llega a 2 mil muestras para la clase GIF [42].

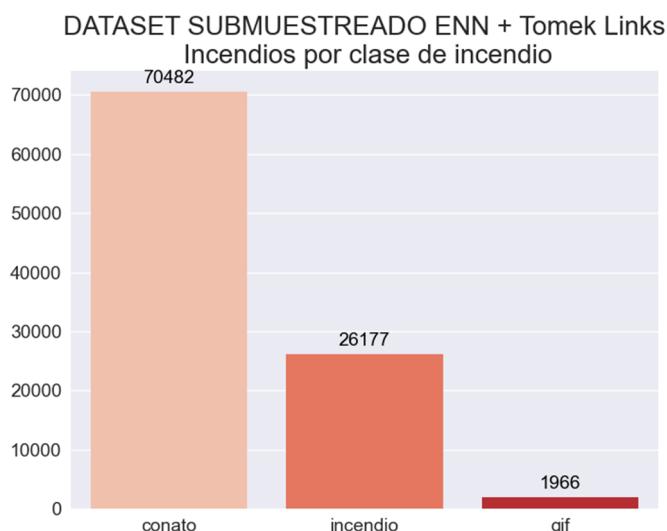


Ilustración 26. Gráfico con el número de registros por clase después de realizar el submuestreo con ENN y Tomek Links

Se aborda el problema realizando un sobremuestreo de la clase minoritaria GIF con diferentes generadores de muestras sintéticas para equilibrar el *dataset*. Los resultados se recogen en la siguiente tabla:

Sobremuestreo Clase Minoritaria	Clase	Precision	Recall	F1-Score		Precision	Recall	F1-Score
SMOTE	conato	0,97	0,99	0,98	accuracy			0,95
	incendio	0,91	0,92	0,92	macro avg	0,85	0,67	0,69
	gif	0,67	0,09	0,16	weighted avg	0,95	0,95	0,95
ADASYN	conato	0,98	0,97	0,98	accuracy			0,94
	incendio	0,90	0,91	0,91	macro avg	0,74	0,76	0,75
	gif	0,34	0,40	0,37	weighted avg	0,95	0,94	0,95
BorderlineSMOTE	conato	0,98	0,98	0,98	accuracy			0,95
	incendio	0,91	0,90	0,90	macro avg	0,75	0,77	0,76
	gif	0,37	0,42	0,40	weighted avg	0,95	0,95	0,95
RandomOverSampler	conato	0,97	0,98	0,98	accuracy			0,95
	incendio	0,90	0,93	0,92	macro avg	0,81	0,68	0,70
	gif	0,56	0,12	0,20	weighted avg	0,95	0,95	0,95

Tabla 10. Resultados de Random Forest con diferentes técnicas de sobremuestreo de la clase GIF

Dados los resultados se llega a la conclusión de que la generación de muestras sintéticas mejora en general el desempeño de las clases conato e incendio y la precisión global del modelo *Random Forest* que estamos utilizando para el análisis, pero no consigue mejorar la clase minoritaria GIF que es tan importante predecir.

4.5. Estrategia híbrida: submuestreo clases conato e incendio y sobremuestreo clase GIF

De las pruebas realizadas hasta el momento, el mejor resultado se obtiene con el submuestreo de las clases mayoritarias empleando ENN y *Tomek Links*, pero se continúa con un *dataset* desbalanceado.

Es por esa razón se decide realizar una combinación de submuestreo aleatorio de las clases mayoritarias al *dataset* ya reducido con ENN y *Tomek Links* para equilibrar las clases mayoritarias a la minoritaria junto con la generación de muestras sintéticas para la clase minoritaria tratando de buscar un equilibrio entre clases que nos permitan mejorar la minoritaria sin empeorar el desempeño de la mayoritarias.

Se lleva a cabo un submuestreo gradual de las clases mayoritarias. Los resultados son los siguientes:

Submuestreo aleatorio clases mayoritarias	Clase	Muestras	Precision	Recall	F1-Score		Precision	Recall	F1-Score
Dataset base: ENN + Tomek Links para la comparación	conato	70482	0,90	0,94	0,92	accuracy			0,84
	incendio	26177	0,82	0,82	0,82	macro avg	0,82	0,82	0,82
	gif	1966	0,74	0,70	0,72	weighted avg	0,83	0,84	0,84
Submuestreo clases mayoritarias a 26177 muestras	conato	26177	0,95	0,97	0,96	accuracy			0,93
	incendio	26177	0,91	0,95	0,93	macro avg	0,83	0,68	0,69
	gif	1966	0,61	0,11	0,19	weighted avg	0,92	0,93	0,92
Submuestreo clases mayoritarias a 13088 muestras	conato	13088	0,94	0,96	0,95	accuracy			0,90
	incendio	13088	0,87	0,94	0,91	macro avg	0,85	0,72	0,74
	gif	1966	0,73	0,24	0,36	weighted avg	0,90	0,90	0,89
Submuestreo clases mayoritarias a 6544 muestras	conato	6544	0,93	0,95	0,94	accuracy			0,87
	incendio	6544	0,84	0,91	0,87	macro avg	0,83	0,77	0,78
	gif	1966	0,73	0,43	0,54	weighted avg	0,86	0,87	0,86
Submuestreo clases mayoritarias a 3141 muestras	conato	3141	0,91	0,94	0,93	accuracy			0,85
	incendio	3141	0,84	0,84	0,84	macro avg	0,84	0,84	0,84
	gif	1966	0,77	0,74	0,75	weighted avg	0,85	0,85	0,85
Submuestreo clases mayoritarias a 1966 muestras	conato	1966	0,85	0,93	0,89	accuracy			0,81
	incendio	1966	0,83	0,72	0,77	macro avg	0,82	0,81	0,81
	gif	1966	0,76	0,80	0,78	weighted avg	0,81	0,81	0,81

Tabla 11. Submuestreo aleatorio de las clases conato e incendio

Se observa que a medida que se reduce el número de muestras de las clases mayoritarias desciende ligeramente su rendimiento, sin embargo, el de la clase minoritaria se incrementa ya que el modelo pasa a tenerlo en cuenta al ir equilibrándose el número de muestras por cada clase.

Así, el mejor resultado de submuestreo aleatorio se produce al reducir las clases mayoritarias a 3141 muestras donde además se mejoran ligeramente los resultados que habíamos conseguido con el submuestreo con ENN y *Tomek Links* con el *dataset* sin balancear, es decir, estamos obteniendo un *dataset* más reducido y balanceado con prácticamente los mismos buenos resultados que ya se habían alcanzado.

Finalmente, empleamos distintas estrategias para terminar de equilibrar el *dataset* utilizando diferentes técnicas de generación de muestras sintéticas para la clase minoritaria GIF. Los resultados pueden apreciarse en la siguiente tabla:

Sobremuestreo Clase Minoritaria	Clase	Precision	Recall	F1-Score		Precision	Recall	F1-Score
SMOTE	conato	0,90	0,92	0,91	accuracy			0,82
	incendio	0,84	0,75	0,80	macro avg	0,81	0,82	0,81
	gif	0,69	0,78	0,73	weighted avg	0,83	0,82	0,83
ADASYN	conato	0,92	0,93	0,93	accuracy			0,84
	incendio	0,86	0,79	0,82	macro avg	0,83	0,84	0,83
	gif	0,70	0,79	0,74	weighted avg	0,85	0,84	0,84
BorderlineSMOTE	conato	0,91	0,93	0,92	accuracy			0,84
	incendio	0,87	0,78	0,82	macro avg	0,83	0,84	0,83
	gif	0,71	0,80	0,75	weighted avg	0,85	0,84	0,84
RandomOverSampler	conato	0,91	0,93	0,92	accuracy			0,85
	incendio	0,86	0,81	0,83	macro avg	0,83	0,84	0,84
	gif	0,74	0,76	0,75	weighted avg	0,85	0,85	0,85
Dataset base: ENN + Tomek Links para la comparación	conato	0,90	0,94	0,92	accuracy			0,84
	incendio	0,82	0,82	0,82	macro avg	0,82	0,82	0,82
	gif	0,74	0,70	0,72	weighted avg	0,83	0,84	0,84

Tabla 12. Resultados de Random Forest con estrategia combinada de submuestreo de las clases conato e incendio y sobremuestreo de la clase GIF

Se escoge *RandomOverSampler* para generar muestras sintéticas sólo en la clase GIF y en la fase de entrenamiento. Con la introducción de estas muestras se consigue un *dataset* balanceado con un buen rendimiento para todas las clases.

4.6. Modelos

Una vez optimizado el *dataset* se implementan de una forma básica una serie de modelos para testear cuáles pueden ser los más adecuados para abordar nuestro problema.

En la tabla a continuación se pueden ver los distintos modelos y métricas obtenidas:

	Clase	Precision	Recall	F1-Score		Precision	Recall	F1-Score
RF	conato	0,91	0,94	0,93	accuracy			0,85
	incendio	0,84	0,84	0,84	macro avg	0,84	0,84	0,84
	gif	0,77	0,74	0,75	weighted avg	0,85	0,85	0,85
CatBoost	conato	0,94	0,94	0,94	accuracy			0,85
	incendio	0,86	0,79	0,83	macro avg	0,84	0,85	0,84
	gif	0,71	0,81	0,76	weighted avg	0,86	0,85	0,85
XGBoost	conato	0,94	0,94	0,94	accuracy			0,85
	incendio	0,84	0,83	0,84	macro avg	0,84	0,84	0,84
	gif	0,74	0,76	0,75	weighted avg	0,86	0,85	0,85
LightGBM	conato	0,94	0,93	0,94	accuracy			0,85
	incendio	0,85	0,80	0,83	macro avg	0,84	0,84	0,84
	gif	0,72	0,79	0,75	weighted avg	0,85	0,85	0,75
MLP Red neuronal feedforward	conato	0,88	0,95	0,91	accuracy			0,81
	incendio	0,81	0,78	0,80	macro avg	0,79	0,79	0,79
	gif	0,69	0,65	0,67	weighted avg	0,81	0,81	0,81
CNN	conato	0,92	0,90	0,91	accuracy			0,82
	incendio	0,79	0,81	0,80	macro avg	0,80	0,80	0,80
	gif	0,70	0,70	0,70	weighted avg	0,82	0,82	0,82
ANN	conato	0,88	0,94	0,91	accuracy			0,83
	incendio	0,83	0,84	0,83	macro avg	0,81	0,81	0,81
	gif	0,73	0,65	0,69	weighted avg	0,83	0,83	0,83
KNN	conato	0,84	0,89	0,86	accuracy			0,76
	incendio	0,72	0,79	0,75	macro avg	0,74	0,72	0,73
	gif	0,67	0,49	0,56	weighted avg	0,75	0,76	0,75
Decision Tree	conato	0,86	0,85	0,86	accuracy			0,75
	incendio	0,74	0,71	0,72	macro avg	0,73	0,73	0,73
	gif	0,59	0,65	0,62	weighted avg	0,75	0,75	0,75
SVM	conato	0,85	0,90	0,88	accuracy			0,79
	incendio	0,78	0,79	0,78	macro avg	0,78	0,77	0,77
	gif	0,71	0,62	0,66	weighted avg	0,79	0,79	0,79

Tabla 13. Comparativa de distintos modelos utilizando el dataset balanceado

En general, la mayor parte de los modelos clasifican bien la clase conato, sin embargo para la clase incendio y sobre todo para la clase GIF encontramos más dificultades.

Para la clase conato destacan modelos como *CNN*, *MLP*, *Random Forest*, *CatBoost*, *XGBoost* y *LightGBM*. En especial los tres últimos modelos tienen una *precision*, *recall* y *F1-Score* altos en torno a 0.94.

En el caso de la clase incendio destacan *ANN*, *Random Forest*, *CatBoost*, *XGBoost* y *LightGBM* con valores de *F1-Score* en torno a 0.83-0.84. Destaca *Random Forest* por tener la mejor relación entre *precision* y *recall* siendo la más equilibrada.

En general, *CatBoost*, *LightGBM* y *XGBoost* y *Random Forest* tienen buenas métricas para la clase *GIF*. *Random Forest* tiene la mejor *precision* de 0,77, mientras que *CatBoost* tiene el mejor *recall* de 0,81.

A continuación se optimizan los mejores modelos mediante *Optuna* con el que podremos explorar los mejores parámetros para cada uno de los modelos seleccionados *Random Forest* [42], *CatBoost* [43], *LightGBM* [44], y *XGBoost* [45] con los siguientes resultados.

Random Forest

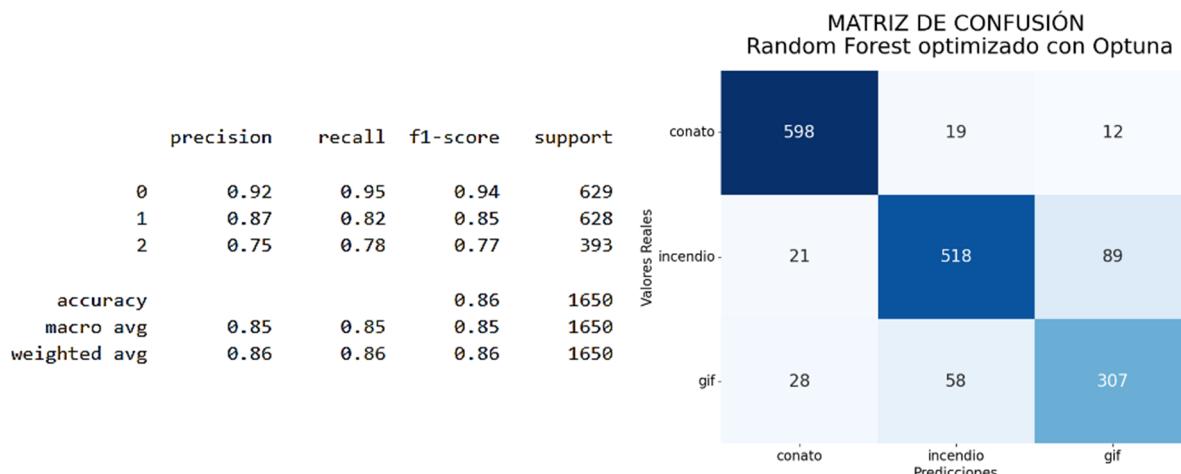


Tabla 14. Resultados de Random Forest optimizado con Optuna

XGBoost

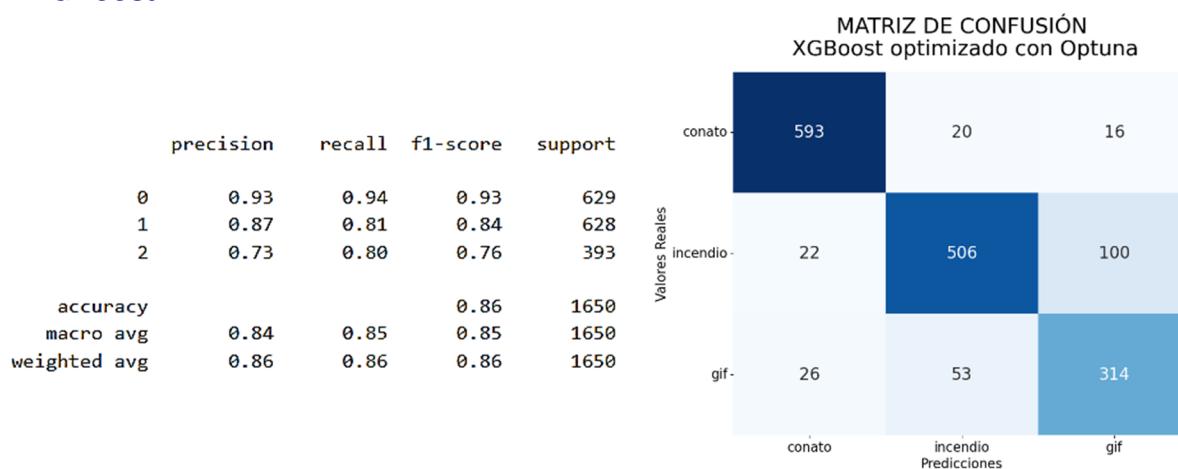


Tabla 15. Resultados de XGBoost optimizado con Optuna

LightGBM

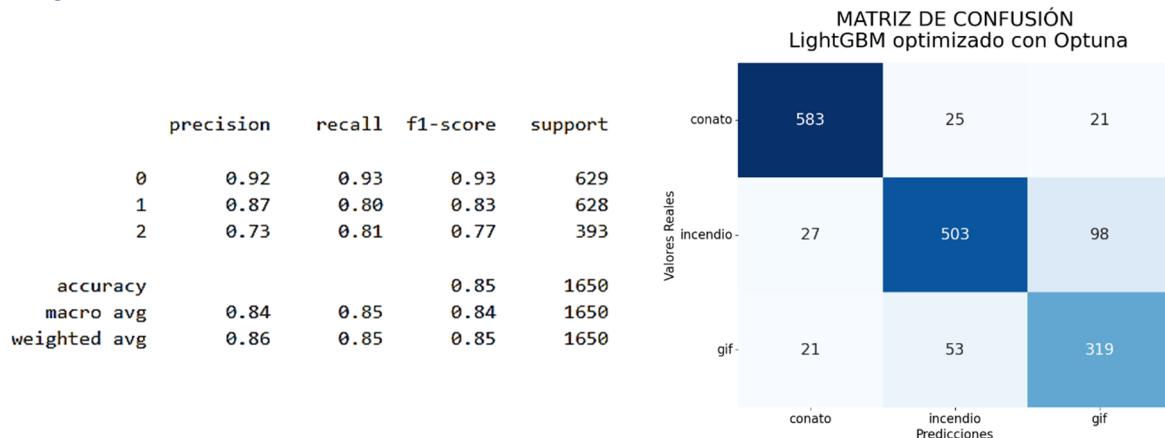


Tabla 16. Resultados de LightGBM optimizado con Optuna

CatBoost

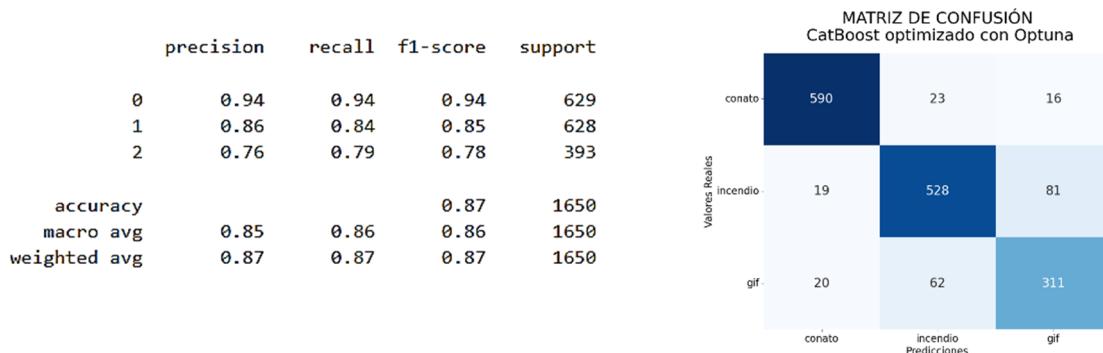


Tabla 17. Resultados de CatBoost optimizado con Optuna

En general todos los modelos han mejorado ligeramente sus métricas. La mejora más significativa se produce en la clase GIF con un incremento en el *F1-Score* tanto en *CatBoost* como en *LightGBM* y *XGBoost*. Además, en *CatBoost* se experimenta un incremento del *recall* de la clase incendio lo que indica que está identificando la clase de forma correcta.

Dado que las clases GIF e incendio son las más importantes elegimos el modelo que mejor equilibrio tenga en esas clases para el *F1-Score*. Para la clase GIF, *LightGBM* y *CatBoost* tienen un valor similar de *F1-Score*, 0.77-0.78, superando a los otros dos modelos por poca diferencia.

La clase incendio tiene un buen *F1-Score* de 0.85 para *CatBoost* y *Random Forest*, pero para este último tiene un desempeño más bajo para la clase GIF.

CatBoost tienen un *macro avg F1-Score* de 0.86 y un *weighted avg F1-Score* de 0.87 que indica que es un modelo robusto para todas las clases incluyendo incendio y GIF. Además tienen un *accuracy* alto de 0.87, aunque en este caso sea una métrica menos relevante.

Entre todos los modelos analizados existen pocas diferencias, destaca *CatBoost* como la mejor alternativa en global para todas las clases, en especial para incendio y GIF ya que tiene un buen

equilibrio entre *precision*, *recall* y *F1-Score*. Se escoge *CatBoost* como modelo final y realizamos una validación cruzada para comprobar lo robusto que es.

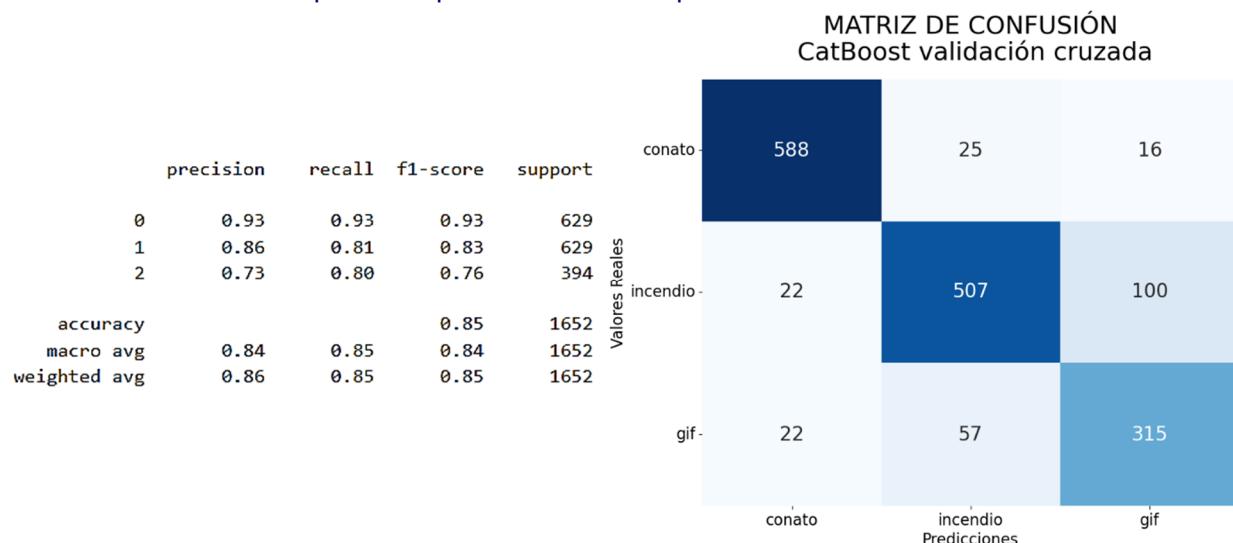


Tabla 18. Resultados de *CatBoost* tras validación cruzada

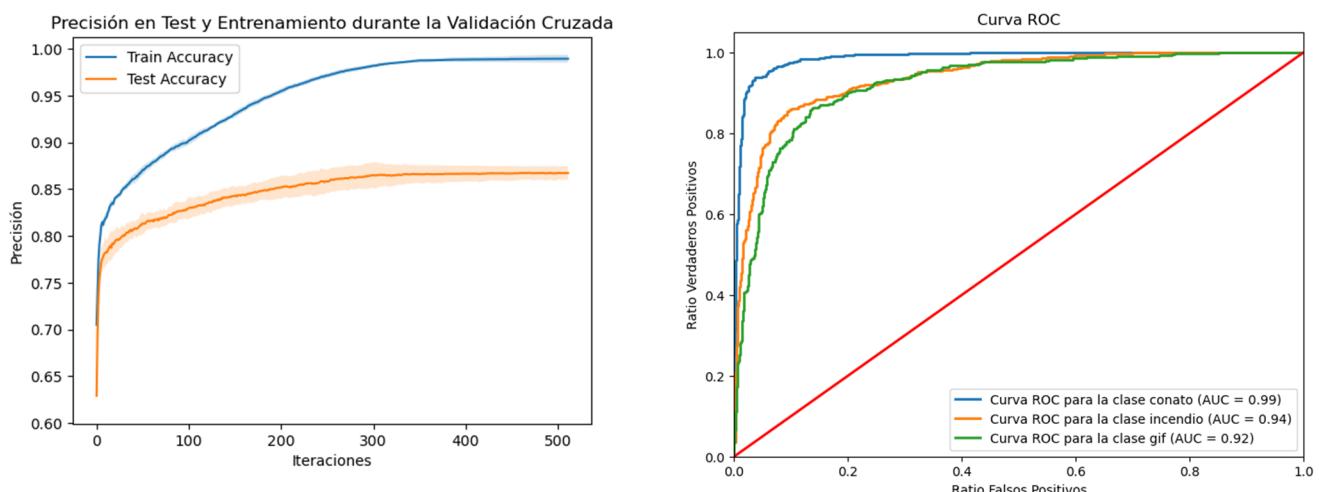


Ilustración 27. Curvas de precisión en test y train para *CatBoost* y Curvas ROC para cada clase de incendio

Se observa una disminución en el *accuracy* y los *F1-Score* lo que puede indicar un leve sobreajuste inicial además de la disminución en el *recall* de las clases incendio y GIF.

4.7. Desempeño de *CatBoost* con los incendios de 2017

Seleccionados los incendios forestales del año 2017, se les somete al mismo preprocesamiento y se introducen en el mejor modelo entrenado de *CatBoost*. Se obtienen los siguientes resultados:

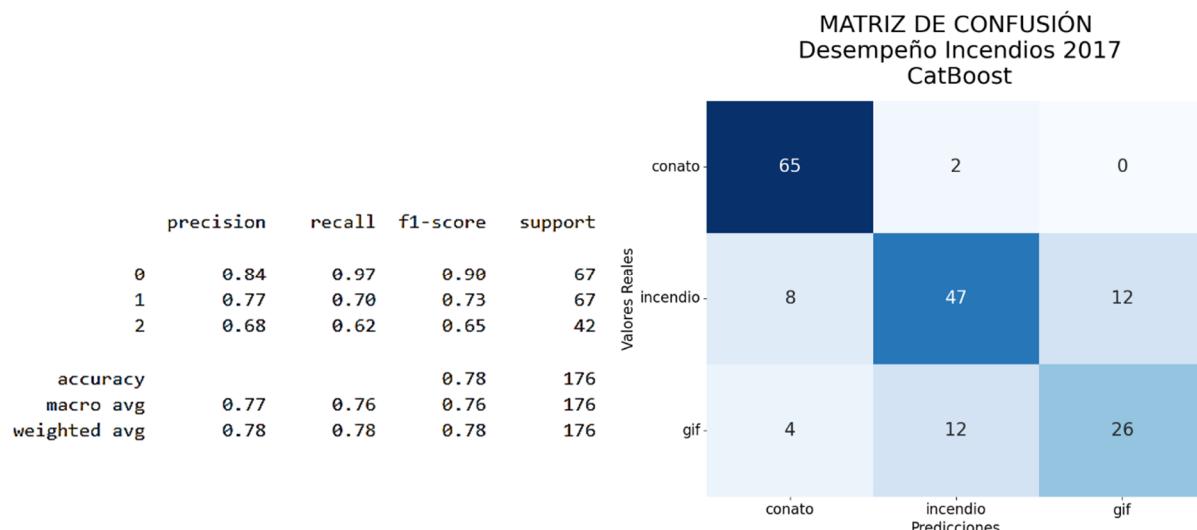


Tabla 19. Resultados de *CatBoost* con los incendios de 2017

En general, vemos que el modelo es capaz de generalizar bastante bien la clase conato con un *recall* muy alto, sin embargo, tanto en la clase incendio como en la clase GIF el *recall* cae dando lugar a predicciones incorrectas, las más llamativas serían los tres conatos que en realidad fueron GIF y los dos conatos que habían sido predichos como GIF.

5. Conclusiones y trabajos futuros

Dificultades en la modelización de los grandes incendios forestales

Una de las principales dificultades es la modelización de los Grandes Incendios Forestales debido a las pocas muestras de las que se dispone. A pesar de su menor número, son los incendios más relevantes ya que son los casos con mayor impacto económico, humano y ambiental, constituyendo los que más recursos necesitan para su extinción y siendo su capacidad destructiva la que los convierte en los de mayor interés.

La naturaleza poco frecuente de este tipo de incendios dificulta el entrenamiento de los modelos ya que tienden a tener sesgo hacia las clases más representadas como conatos o incendios. Sin embargo, dada su magnitud, los esfuerzos han de centrarse en mejorar su predicción desde una visión global de todas las clases de incendios; los GIFs deben ser el foco prioritario en cuanto a su predicción.

Respecto a las técnicas utilizadas para balancear las clases, se observa lo siguiente:

- Las técnicas de submuestreo han demostrado ser útiles para equilibrar las clases desbalanceadas, pero hay que aplicarlas con moderación para evitar la eliminación de muestras relevantes que pueden alejar nuestro modelo de la realidad y crear un modelo demasiado optimizado.
- Las técnicas de sobre muestreo funcionan mejor cuando se aplican en cantidades mínimas ya que replicar excesivamente las pocas muestras de la clase minoritaria introduce redundancia y reduce la capacidad predictiva.

Incorporación de nuevas variables relevantes

En el desarrollo de este trabajo se han introducido variables que han demostrado ser relevantes en la modelización de los incendios forestales:

- Densidad de incendios en el municipio en los últimos cinco años. Esta variable trata de recuperar patrones históricos que pueden influir en la recurrencia de incendios en una zona, desde el propio evento que cambia indiscutiblemente el entorno hasta la influencia de las políticas de extinción y prevención llevadas a cabo en esa ventana temporal.
- Densidad histórica de población en el municipio. Que recoge la interacción humana con el entorno o la capacidad de respuesta al incendio forestal en función de los núcleos de población.

Estas variables han resultado ser relevantes para la modelización de los incendios forestales, pero es necesario profundizar en su estudio con un mayor rigor metodológico, especialmente en investigaciones futuras que puedan dedicar más tiempo y recursos a su desarrollo.

Importancia de los conatos en la modelización

La idea de prescindir de los conatos bajo la premisa de que no llegan a ser considerados incendios perjudica la modelización, ya que contienen información relevante sobre las condiciones iniciales que pueden llevar a pasar de un conato de incendio a un incendio. Integrarlos en el análisis mejora la representatividad del modelo y permite una comprensión más profunda del problema.

Necesidad de ampliar y enriquecer la recopilación de datos

Hay una necesidad en cuanto a seguir trabajando en la recopilación de datos que caractericen los incendios forestales:

- Datos relacionados con el combustible, es decir, integrar datos históricos de series de vegetación procedentes del Banco de Datos de la Naturaleza.
- Datos topográficos como la pendiente y la orientación del terreno son variables relevantes por explorar. La dificultad radica en que en la mayor parte de los casos no se conocían las coordenadas de inicio del incendio, es por esta razón que habría que considerar la inclusión de imágenes que recojan variables topográficas de la zona.
- Indicadores satelitales como el NDVI (*Normalized Difference Vegetation Index*) antes y después del incendio pueden proporcionar información clave sobre el impacto en la vegetación.

Modelos multimodales y el uso de imágenes satelitales

La inclusión de imágenes satelitales antes y después del incendio permite explorar modelos multimodales que combinen datos tabulares con análisis de imágenes. La dificultad en este caso, además del coste computacional y el almacenamiento de las imágenes tiene que ver con las fuentes de información:

- Las imágenes de alta calidad de la misión *Sentinel* están disponibles desde 2017, mientras que la base de datos del EGIF está consolidada solo hasta 2016.
- La alternativa son las imágenes de la misión *Landsat*, que tienen menor resolución y frecuencia, pero podrían ser útiles para investigaciones preliminares.

Análisis espacial mediante clustering

Otro enfoque diferente para abordar el problema de la clasificación de la severidad de un incendio forestal es la aplicación de algoritmos de *clustering* de manera preliminar para agrupar aquellas zonas con características homogéneas para posteriormente modelar los incendios forestales para cada zona.

Aplicación práctica y futuras investigaciones

Los resultados obtenidos tienen aplicación en cuanto a la gestión y prevención de incendios forestales, concretamente:

- Prevención. Los patrones identificados podrían integrarse en sistemas de alerta temprana.
- Planificación. Las variables más relevantes podrían ser útiles para el diseño de estrategias forestales que mejoren la prevención e intervención en el transcurso de un incendio.

En futuras investigaciones habrá que combinar datos satelitales con tabulares siguiendo estrategias de creación de modelos multimodales que incluyan técnicas de *deep learning* para las imágenes junto con técnicas de *machine learning* para los datos tabulares. Esta combinación permitirá una modelización de los incendios forestales más completa y una mejora en la predicción de las clases incendio y GIF.

6. Referencias

- [1] Estrategia Forestal Española horizonte 2050 [Internet]. Ministerio Para la Transición Ecológica y el Reto Demográfico. [citado 10 de octubre de 2024]. Disponible en:
https://www.miteco.gob.es/es/biodiversidad/temas/politica-forestal/planificacion-forestal/politica-forestal-en-espana/pfe_estrategia_forestal.html
- [2] Jain P, Coogan SCP, Subramanian SG, Crowley M, Taylor S, Flannigan MD. A review of machine learning applications in wildfire science and management. *Environmental Reviews* [Internet]. 28 de julio de 2020 [citado 10 de octubre de 2024];28(4):478-505. Disponible en:
<https://doi.org/10.1139/er-2020-0019>
- [3] ESTADÍSTICA GENERAL DE INCENDIOS FORESTALES (EGIF) [Internet]. Ministerio Para la Transición Ecológica y el Reto Demográfico. [citado 10 de octubre de 2024]. Disponible en:
<https://www.miteco.gob.es/es/biodiversidad/temas/incendios-forestales/estadisticas-datos.html>
- [4] Wood DA. Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight. *Artificial Intelligence in Agriculture* [Internet]. 2021 Jan 1;5:24–42. Available from:
<https://www.sciencedirect.com/science/article/pii/S2589721721000118>
- [5] Cortez P, Morais AJR. A data mining approach to predict forest fires using meteorological data. In: Neves JM, Santos MF, Machado JM, editors. *New trends in artificial intelligence: proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007)*; 2007 Dec; Guimarães, Portugal. Lisboa: Associação Portuguesa para a Inteligência Artificial (APPIA); 2007. p. 512-523. Disponible en: <https://hdl.handle.net/1822/8039>
- [6] Yazici K, Taskin A. A comparative Bayesian optimization-based machine learning and artificial neural networks approach for burned area prediction in forest fires: an application in Turkey. *Natural Hazards* [Internet]. 21 de septiembre de 2023;119(3):1883-912. Disponible en:
<https://doi.org/10.1007/s11069-023-06187-4>
- [7] Joshi J, Sukumar R. Improving prediction and assessment of global fires using multilayer neural networks. *Scientific Reports* [Internet]. 8 de febrero de 2021;11(1). Disponible en:
<https://www.nature.com/articles/s41598-021-81233-4>
- [8] Karalidis K. Deep learning method for forest fire detection and simulation of wildfire expansion using Sentinel-2 images [Internet]. Utrecht: Utrecht University; 2023 [citado 2023 May 2]. Disponible en: <https://studenttheses.uu.nl/handle/20.500.12932/43836>
- [9] Castelli M, Vanneschi L, Popović A. Predicting Burned Areas of Forest Fires: an Artificial Intelligence Approach. *Fire Ecology* [Internet]. 1 de abril de 2015;11(1):106-18. Disponible en:
<https://doi.org/10.4996/fireecology.1101106>

- [10] Anzola JD, Fuentes LD, Rodríguez EM. Desarrollo de un modelo de estimación para la prevención de incendios forestales en Colombia [Internet]. 2024 Jun 11 [citado 2024 Oct 21]. Disponible en: <http://hdl.handle.net/10584/11968>
- [11] Dong H, Wu H, Sun P, Ding Y. Wildfire Prediction Model Based on Spatial and Temporal Characteristics: A Case Study of a Wildfire in Portugal's Montesinho Natural Park. *Sustainability* [Internet]. 15 de agosto de 2022;14(16):10107. Disponible en: <https://www.mdpi.com/2071-1050/14/16/10107>
- [12] Li X, Wang X, Sun S, Wang Y, Li S, Li D. Predicting the Wildland Fire Spread Using a Mixed-Input CNN Model with Both Channel and Spatial Attention Mechanisms. *Fire Technology* [Internet]. 20 de junio de 2023;59(5):2683-717. Disponible en: <https://doi.org/10.1007/s10694-023-01427-2>
- [13] Ali SD, Ridwan I, Pradana AF, Septiani W, Ulfah M, Sabian NT, et al. GeoAI for disaster mitigation: Fire severity prediction models using Sentinel-2 and ANN regression. In: 2022 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES); 2022 Nov 16-18; Yogyakarta, Indonesia. IEEE; 2022. p. 1-7.
doi:10.1109/ICARES56907.2022.9993515
- [14] Lee C, Park S, Kim T, Liu S, Md Reba MN, Oh J, Han Y. Machine learning-based forest burned area detection with various input variables: a case study of South Korea. *Appl Sci.* 2022;12:10077. <https://doi.org/10.3390/app121910077>
- [15] Bayat G, Yildiz K. Comparison of the Machine Learning Methods to Predict Wildfire Areas. *Turkish Journal Of Science And Technology* [Internet]. 2 de septiembre de 2022;17(2):241-50. Disponible en: <https://dergipark.org.tr/en/pub/tjst/issue/72762/1063284>
- [16] Elshewey A, Elsonbaty A. Forest fires detection using machine learning techniques. *Xi'an Jianzhu Keji Daxue Xuebao/Journal of Xi'an University of Architecture & Technology.* 2020; XII:1-8. Disponible en: <https://docsdrive.com/?pdf=medwelljournals/ijscomp/2021/1-8.pdf>
- [17] Mohajane M, Costache R, Karimi F, Pham QB, Essahlaoui A, Nguyen H, et al. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. *Ecological Indicators* [Internet]. 2021 Jun 7;129:107869. Available from: <https://www.sciencedirect.com/science/article/pii/S1470160X21005343>
- [18] Stanford-Moore A. Wildfire burn area prediction [Internet]. 2019. Disponible en: <https://www.semanticscholar.org/paper/Wildfire-Burn-Area-Prediction-Stanford-Moore/01a3c446ebcac31abb7859699a0d14315c4b01bf>
- [19] Magdalena VLFM, De Catalunya Departament de Teoria del Senyal I Comunicacions UP. Inteligencia artificial aplicada a la predicción del Dengue e incendios forestales en Indonesia [Internet]. 2024. Disponible en: <https://upcommons.upc.edu/handle/2117/406714>

- [20] Quispe Varillas JJ. Delimitación de áreas afectadas por incendios forestales mediante aprendizaje profundo en imágenes satelitales [Internet]. Lima: Universidad Nacional Federico Villarreal; 2024 [citado 2024 Oct 21]. Disponible en: <https://hdl.handle.net/20.500.13084/9326>
- [21] Gupta HP, Mishra R. Utilizing Transfer Learning and pre-trained Models for Effective Forest Fire Detection: A Case Study of Uttarakhand [Internet]. arXiv.org. 2024. Disponible en: <https://arxiv.org/abs/2410.06743>
- [22] Shadrin D, Illarionova S, Gubanov F, Evteeva K, Mironenko M, Levchunets I, et al. Wildfire spreading prediction using multimodal data and deep neural network approach. *Scientific Reports* [Internet]. 31 de enero de 2024;14(1). Disponible en: <https://www.nature.com/articles/s41598-024-52821-x>
- [23] Ministerio para la Transición Ecológica y el Reto Demográfico. (n.d.). *Parte de incendio forestal*. [PDF]. Ministerio para la Transición Ecológica y el Reto Demográfico. Recuperado el 15 de noviembre de 2024.
https://www.miteco.gob.es/content/dam/miteco/es/biodiversidad/temas/incendios-forestales/parteincendioforestal_web_tcm30-132604.pdf
- [24] Ministerio para la Transición Ecológica y el Reto Demográfico. (n.d.). *Instrucciones de relleno del parte de incendio forestal*. [PDF]. Ministerio para la Transición Ecológica y el Reto Demográfico. Recuperado el 15 de noviembre de 2024.
https://www.miteco.gob.es/content/dam/miteco/es/biodiversidad/temas/incendios-forestales/instrucciones_parte_incendio_tcm30-512355.pdf
- [25] Infraestructura de Datos Espaciales de España (IDEE). (n.d.). *Nomenclátor Geográfico de Municipios y Entidades de Población* [Base de datos]. Consejo Superior Geográfico. Recuperado el 15 de noviembre de 2024. <https://www.idee.es/csw-inspire-idee/srv/spa/catalog.search?#/metadata/spaignNomenclatorGeograficoMunicipiosEntPob>
- [26] Instituto Nacional de Estadística (INE). (n.d.). *Relación de municipios y sus códigos por provincias* [Archivo de datos]. INEbase. Recuperado el 15 de noviembre de 2024.
https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177031&idp=1254734710990
- [27] Climate Data Store. (n.d.) *Fire danger indicators for Europe from 1970 to 2098 derived from climate projections*. (2024, November 12). <https://cds.climate.copernicus.eu/datasets/sis-tourism-fire-danger-indicators?tab=overview>
- [28] Instituto Nacional de Estadística. (n.d.). *Población según sexo y edad desde 1900 hasta 2001*. Instituto Nacional de Estadística. Recuperado el 15 de noviembre de 2024.
<https://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=03003.px&L=0>

- [29] Instituto Nacional de Estadística. (n.d.). *Principales series de población desde 1998*. Instituto Nacional de Estadística. Recuperado el 15 de noviembre de 2024.
<https://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=03003.px&L=0>
- [30] MyGeodata Converter. (n.d.). *KMZ to XLSX converter* [Aplicación web]. MyGeodata Cloud. Recuperado el 15 de noviembre de 2024. <https://mygeodata.cloud/converter/kmz-to-xlsx>
- [31] Luengo, P. (2024). mapa_militar. [Repositorio GitHub].
https://github.com/patriciaLuca/DataSource_FireGroundAI/tree/main/mapa_militar
- [32] Luengo, P. (2024). *01_Preprocesamiento_imputacion_datos_geograficos_demograficos* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataSource_FireGroundAI/blob/main/01_Preprocesamiento_imputacion_datos_geograficos_demograficos.ipynb
- [33] Luengo, P. (2024). *02_Preprocesamiento_extraccion_datos_climatologicos_AEMET* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataSource_FireGroundAI/blob/main/02_Preprocesamiento_extraccion_datos_climatologicos_AEMET.ipynb
- [34] Agencia Estatal de Meteorología (AEMET). (n.d.). *Inventario de estaciones meteorológicas* [Dataset]. AEMET OpenData. Recuperado el 15 de noviembre de 2024.
[https://opendata.aemet.es/dist/index.html#/valores-climatologicos/Inventario%20de%20estaciones%20\(valores%20climatol%C3%B3gicos\)](https://opendata.aemet.es/dist/index.html#/valores-climatologicos/Inventario%20de%20estaciones%20(valores%20climatol%C3%B3gicos))
- [35] Agencia Estatal de Meteorología (AEMET). (n.d.). *Climatologías diarias (valores climáticos históricos)* [Dataset]. AEMET OpenData. Recuperado el 15 de noviembre de 2024.
<https://opendata.aemet.es/dist/index.html#/valores-climatologicos/Climatolog%C3%ADas%20diarias.1>
- [36] Luengo, P. (2024). *03_Preprocesamiento_extraccion_datos_FWI* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataSource_FireGroundAI/blob/main/03_Preprocesamiento_extraccion_datos_FWI.ipynb
- [37] Luengo, P. (2024). *05_Analisis_exploratorio_de_datos* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataSource_FireGroundAI/blob/main/05_Analisis_exploratorio_de_datos.ipynb
- [38] Luengo, P. (2024). *04_Preprocesamiento_estimacion_poblacion* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataSource_FireGroundAI/blob/main/04_Preprocesamiento_estimacion_poblacion.ipynb

- [39] Ministerio para la Transición Ecológica y el Reto Demográfico. (n.d.). *Indicador 49: Superficie arbolada afectada respecto al tamaño de los incendios*. [PDF]. Ministerio para la Transición Ecológica y el Reto Demográfico. Recuperado el 15 de noviembre de 2024.
https://www.miteco.gob.es/content/dam/miteco/es/biodiversidad/temas/inventarios-nacionales/indicador_49_superficie_arbolada_afectada_por_incendios_tcm30-207537.pdf
- [40] Luengo, P. (2024). *06_Seleccion_de_caracteristicas* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/06_Seleccion_de_caracteristicas.ipynb
- [41] Luengo, P. (2024). *08_RandomForest_Submuestreo* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/08_RandomForest_Submuestreo.ipynb
- [42] Luengo, P. (2024). *09_Random_Forest_Sobremuestreo* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/09_Random_Forest_Sobremuestreo.ipynb
- [43] Luengo, P. (2024). *10_CatBoost* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/10_CatBoost.ipynb
- [44] Luengo, P. (2024). *11_LightGBM* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/11_LightGBM.ipynb
- [45] Luengo, P. (2024). *12_XGBoost* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/12_XGBoost.ipynb
- [46] Luengo, P. (2024). *13_CatBoost_IF_2017* [Archivo Jupyter Notebook]. GitHub.
https://github.com/patriciaLuca/DataScience_FireGroundAI/blob/main/13_CatBoost_IF_2017.ipynb