**Imagen que contiene Sitio web

El contenido generado por IA puede ser incorrecto.**

**Resumen**

Este trabajo presenta un modelo predictivo para la optimización de la gestión de recursos en la extinción de incendios forestales en la Comunidad Valenciana, utilizando técnicas de **machine learning**. El objetivo principal es desarrollar un sistema que prediga los recursos necesarios (terrestres y aéreos) basándose en datos históricos de incendios y variables meteorológicas y geoespaciales obtenidas de **Google Earth Engine**. El modelo busca mejorar la eficiencia operativa, reducir costos y minimizar el impacto ambiental y social de los incendios.

Para la modelización, se emplearon varios algoritmos de clasificación y regresión, como **Random Forest**, **XGBoost**, **LightGBM** y **CatBoost**, los cuales fueron evaluados y ajustados mediante técnicas de **validación cruzada** y **ajuste de hiperparámetros**. Las métricas utilizadas para evaluar los modelos incluyen **AUC**, **F1-score** y **recall**, determinando que los modelos más efectivos para la predicción de la necesidad de medios de extinción fueron **Random Forest** y **LightGBM**. Además, se realizó un análisis de **umbral económico** para equilibrar la eficiencia de los recursos con los costos operativos, lo que permitió una asignación más precisa de los recursos disponibles.

Utilizando la metodología **CRISP-DM**, se realizaron etapas de recopilación, tratamiento de datos, modelización y evaluación. La solución final incluye una **plataforma web en tiempo real** que permite a los operadores de emergencias tomar decisiones informadas sobre el despliegue de medios, basándose en las predicciones del modelo. Se demostró que el modelo puede optimizar la asignación de recursos y, al mismo tiempo, se establecieron estrategias para su integración en sistemas de gestión de incendios en fases posteriores del proyecto.

**Palabras clave:** Modelo predictivo, Optimización de recursos, Incendios forestales, Google Earth Engine, CRISP-DM, Random Forest, XGBoost, Asignación de recursos, Costos operativos, Impacto ambiental

Tabla de contenido

[INTRODUCCIÓN 4](#_Toc191471180)

[INCENDIOS EN EUROPA 4](#_Toc191471181)

[INCENDIOS EN LA COMUNIDAD VALENCIANA 5](#_Toc191471182)

[OBJETIVOS DEL PROYECTO 9](#_Toc191471183)

[METODOLOGÍA 10](#_Toc191471184)

[FASE: COMPRENSIÓN DEL NEGOCIO. 10](#_Toc191471185)

[1. Determinar los objetivos del negocio. 10](#_Toc191471186)

[2. Diseño de la solución operativa. 12](#_Toc191471187)

[3. Evaluación de la situación. 12](#_Toc191471188)

[4. Determinación de los objetivos analíticos. 16](#_Toc191471189)

[5. Plan de proyecto. 17](#_Toc191471190)

[FASE: COMPRENSIÓN DE DATOS. 18](#_Toc191471191)

[1. Recolección de datos iniciales. 18](#_Toc191471192)

[2. Descripción de los datos. 19](#_Toc191471193)

[3. Exploración de los datos 23](#_Toc191471194)

[FASE: PLATAFORMA TECNOLÓGICA. 34](#_Toc191471195)

[1. Diseño de la Arquitectura de Referencia del Modelo. 34](#_Toc191471196)

[2. Arquitectura de Referencia para el despliegue 36](#_Toc191471197)

[3. Estrategia tecnológica. 37](#_Toc191471198)

[4. Despliegue de la Arquitectura. 38](#_Toc191471199)

[FASE: TRATAMIENTO DE DATOS. 40](#_Toc191471200)

[1. Selección de datos. 40](#_Toc191471201)

[2. Limpieza de datos. 40](#_Toc191471202)

[3. Construcción de nuevos datos 41](#_Toc191471203)

[4. Integración de datos. 43](#_Toc191471204)

[5. Formateo de los datos. 43](#_Toc191471205)

[FASE: MODELIZACIÓN. 46](#_Toc191471206)

[1. Selección de Técnica de modelado 46](#_Toc191471207)

[2. Diseño de evaluación 46](#_Toc191471208)

[3. Construcción del modelo 47](#_Toc191471209)

[4. Evaluación 48](#_Toc191471210)

[5. Implementación y Uso Práctico 50](#_Toc191471211)

[FASE: EVALUACIÓN Y PRESENTACIÓN DE RESULTADOS. 53](#_Toc191471212)

[# AVIÓN ANFIBIO DE DESCARGA 53](#_Toc191471213)

[# AVIÓN CARGA TERRESTRE 56](#_Toc191471214)

[# AUTOBOMBA 58](#_Toc191471215)

[# HELICÓPTERO DE TRANSPORTE 60](#_Toc191471216)

[# BULLDOZER 62](#_Toc191471217)

[FASE: DESPLIEGUE TECNOLÓGICO. 64](#_Toc191471218)

[1. Planificación del Despliegue 64](#_Toc191471219)

[2. Planificación del Control, Seguimiento y Mantenimiento 66](#_Toc191471220)

[3. Implementación Tecnológica 66](#_Toc191471221)

[FASE: DESPLIEGUE DE NEGOCIO. 67](#_Toc191471222)

[1. Planificación del Despliegue de Negocio 67](#_Toc191471223)

[2. Planificación del Control y Seguimiento 67](#_Toc191471224)

[3. Implementación de la Solución 68](#_Toc191471225)

[CONCLUSIONES Y LÍNEAS FUTURAS. 69](#_Toc191471226)

[Conclusiones 69](#_Toc191471227)

[Líneas Futuras 69](#_Toc191471228)

[REFERENCIAS BIBLIOGRÁFICAS. 71](#_Toc191471229)

[Glosario de términos 72](#_Toc191471230)

[Términos del ámbito forestal 72](#_Toc191471231)

[Términos de Ciencia de Datos 73](#_Toc191471232)

# INTRODUCCIÓN

Las altas temperaturas que se están registrando debido al cambio climático cada vez son más fuertes y duraderas en el tiempo, llegando a producir un grave peligro ya que esta subida de temperaturas desencadena en un estrés por calor y además de los temidos incendios forestales. 1,2 Estos incendios en un *90% de las veces tienen una causa por la actividad humana*, mientras que el 10% restante se atribuye a causas naturales, aunque el hecho de que se produzca un incendio es algo que no depende únicamente de un factor, si no que se atribuye a varios de ellos como la fuente de ignición, la composición del combustible, el clima y la topografía, etc. 3

Los incendios forestales suponen un alto impacto tanto a nivel medioambiental ya que suponen un alto coste en la flora, fauna y los recursos naturales de los que nos lucramos todos en nuestro día a día4,5, ya que cada año se estima que aproximadamente 400 Mha de bosque se han consumido bajo la llama de incendios forestales3, A todos estos costes también se le suman los costes humanos con el cobro de vidas ya que en el año 2017 se contabilizaron un total de 140.000 muertes en Europa y 65.000 muertes que se ha producido solo en la región de california, estados unidos en el año 20196. A parte de los daños que suponen a nivel físico y psicológico que se arrastrarán a lo largo de los años por estas personas que han sufrido estas catástrofes de la naturaleza, incluyendo el personal cualificado que ha tenido que intervenir y ha sufrido picos de estrés que puede desencadenar en episodios de depresión y ansiedad1,7,8

El coste económico supone un coste alto que se suma a todo los anteriores, el coste de los incendios de 2017 que hemos referenciado arriba supuso a parte de la pérdida humana y los destrozos de la naturaleza y la fauna una inversión de 2732 millones de dólares para reparar los daños que se produjeron en estos grandes incendios registrados a lo largo de Europa1.

## INCENDIOS EN EUROPA

Aunque los incendios, en el marco de Europa y más concretamente en el marco del *sur de Europa* constituyen un fenómeno más en la configuración de los ecosistemas, por su posición geográfica que supone unas condiciones climatológicas sesgadas hacia una *mayor probabilidad de incidencia de incendios*, esta ha ido siendo más severa encareciendo de una manera más peligrosa a lo largo de los últimos años2,9

Mapa

El contenido generado por IA puede ser incorrecto.

**Figura 1 Representa los puntos donde se encontró mayor poder radioactivo en los incendios registrados (FRP)**

## INCENDIOS EN LA COMUNIDAD VALENCIANA

La Comunitat Valenciana es una de las regiones más afectadas por los incendios forestales debido a su clima mediterráneo, su geografía montañosa y su vegetación altamente inflamable. Entre 2016 y 2017, se registraron **690 incendios forestales**, que arrasaron un total de **8.811,9 hectáreas**, con una afectación especialmente significativa en las zonas del interior, donde predominan los espacios forestales y montañosos, lo que favorece la propagación del fuego (Radín et al., 2018). A pesar de que la mayoría de los incendios se produjeron en días con **alerta 1 (peligro medio-bajo)**, el **70 % de la superficie total quemada se concentró en sólo el 4 % de los días del año**, cuando la alerta de peligro extremo (nivel 3) estaba activa (Radín et al., 2018). En este contexto, una gestión eficiente de los recursos de extinción resulta esencial no solo para mitigar el impacto ecológico y económico de los incendios, sino también para optimizar la asignación de medios y reducir los costes operativos asociados a la lucha contra el fuego.

Los medios utilizados en la extinción de incendios forestales son diversos y cada uno tiene un papel crucial en la lucha contra el fuego, dependiendo de las características del incendio y la geografía del terreno. Los **helicópteros** y **aviones** son fundamentales, especialmente en terrenos de difícil acceso o cuando se requiere una respuesta rápida para evitar la propagación del fuego. Los **helicópteros biturbina tipo medio** y los **helicópteros de gran capacidad** se emplean para lanzar agua o retardantes de fuego desde el aire. Los **aviones anfibios**, por su parte, son utilizados en zonas más remotas, ya que tienen la capacidad de aterrizar en cuerpos de agua para recoger y descargar grandes cantidades de agua en el incendio. En el terreno, las **autobombas** son vehículos equipados con grandes mangueras y depósitos de agua que pueden entrar en zonas de difícil acceso para combatir el fuego de cerca, mientras que los **bulldozers** y **tractores** se utilizan para cortar la vegetación y crear líneas de defensa, deteniendo el avance del fuego.

|  |  |  |  |
| --- | --- | --- | --- |
| Tipo de Medio | Detalles | Cantidad | Horas Anuales de Disponibilidad |
| Medios Aéreos | | | |
| Helicópteros biturbina tipo medio (Bell 212/412) | Helicóptero biturbina de tipo medio, con "bambi bucket" | 3 unidades | 558 horas del 16/06 a 15/10 |
| Helicópteros gran capacidad (Kamov- Ka-32) | Helicóptero de gran capacidad para extinción | 1 unidad | 150 horas del 01/05 a 30/09   |  | | --- | |  | |
| Helicóptero de Coordinación | Helicóptero para coordinación de operaciones | 1 unidad | 220 horas de junio a agosto |
| Brigadas helitransportadas | Brigadas transportadas por helicópteros | 3 unidades | 300 h del 16/03 al 30 /11 |
| Aviones terrestres | Aviones para apoyo terrestre y extinción | 3 unidades | 304 h de mayo a octubre |
| Aviones anfibios | Aviones para la extinción en zonas de difícil acceso | 2 unidades | 150 h de mayo a octubre |
| Medios Terrestres | | | |
| Autobombas forestales |  |  |  |
| Provincial de Alicante | Vehículos destinados a la extinción de incendios | 23 unidades |  |
| Provincial Castellón | Vehículos destinados a la extinción de incendios | 14 unidades |  |
| Provincial de Valencia | Vehículos destinados a la extinción de incendios | 8 unidades |  |

Tabla . Medios disponibles de la comunidad Valenciana Fuente : Auditoría Operativa de los Servicios de Extinción de Incendios Forestales en la Comunitat Valenciana (2011-2015)​

Los **medios aéreos** como helicópteros y aviones son fundamentales en las operaciones de extinción, pero su **coste de operación es elevado**. Por lo que constituyen una parte considerable del presupuesto destinado a la extinción de incendios, lo que subraya la necesidad de **optimizar su uso** para mejorar la eficiencia económica del sistema

|  |  |
| --- | --- |
| Medio de Extinción | Coste (€/hora) |
| Aeronave de coordinación | 1.585,26 |
| Avión anfibio | 4.208,61 |
| Avión de carga en tierra | 855,55 |
| Helicóptero de extinción | 1.096,88 |
| Helicóptero de transporte | 1.096,88 |
| Helicóptero de coordinación | 1.585,26 |
| Autobomba | 29,98 |
| Bulldozer | 67,80 |
| Motobomba | 29,98 |
| Tractor cisterna agrícola | 42,00 |
| Brigada de extinción  (capataz+peon conductor+3 peones) | 91,40 |
| Técnico forestal | 28,84 |
| Agente forestal | 24,72 |

Tabla . Coste por horas de los efectivos para la mitigación de incendios de la comunidad valenciana. Fuente: Vázquez Vázquez, M. C., Chas Amil, M. L., & Touza, J. M. (2014)

En la **Comunitat Valenciana l**os **costes directos** de extinción de incendios durante el periodo 2011-2015 ascendieron a **320,3 millones de euros**. Esta cifra incluye los gastos derivados de la operación de los medios de extinción, tanto **terrestres** como **aéreos**, utilizados para controlar los incendios forestales. Además, en 2015, la Generalitat destinó **16,1 millones de euros** a la prevención de incendios forestales, lo que refleja el gasto significativo en la **prevención** y **extinción** de incendios (Castellano, E., Alum, C., & Rodríguez, C. 2003).

La alta variabilidad en la frecuencia e intensidad de los incendios forestales hace que la planificación presupuestaria para su extinción sea un desafío. No obstante, una **gestión eficiente de los medios disponibles** puede contribuir significativamente a la **reducción de los costes operativos**. En este sentido, la propuesta de este trabajo es el desarrollo de un **modelo predictivo** que permita la **asignación eficiente de recursos** en función de diversas variables, tales como la **superficie quemada**, la **intensidad del fuego** y las **condiciones meteorológicas**. Este modelo permitiría **ajustar el despliegue de medios**, evitando el uso innecesario de recursos y reduciendo así los costes.

El análisis de los **costes operativos** y la identificación de los **factores clave que influyen en los incendios** son esenciales para determinar la **estrategia óptima** de respuesta ante emergencias. El **uso de medios aéreos** y **terrestres** debe ajustarse de acuerdo con la **prioridad operativa** de cada incendio, para maximizar la **eficacia de las intervenciones** y reducir el impacto económico.

Con el desarrollo de un **modelo predictivo**, es posible asignar los medios de extinción de manera **más eficiente**, teniendo en cuenta la magnitud del incendio, las condiciones meteorológicas y la capacidad de los medios. Esto no solo permitirá reducir los costes operativos, sino también mejorar la eficacia de la **gestión de emergencias**, minimizando los daños a la **biodiversidad** y **bienes materiales**.

Actualmente la utilización a nivel global de inteligencia artificial, el machine learning y análisis de los datos en un tiempo más real se está introduciendo en situaciones como los grandes incendios donde se intenta detectar a través de variables que podemos obtener de las agencias aeroespaciales, o terminales medioambientales en un tiempo real5 y que nos puede ayudar a predecir posibles catástrofes naturales que ocurrirán y que podemos conseguir controlar de una manera más eficiente.

Para poder producir ese control eficiente de los fuegos y poder realizar decisiones rápidas y sin apenas error en ellas hemos decidido la realización de un trabajo de fin de máster donde nos centramos no tanto en la predicción de los incendios si no más en la posibilidad de la toma de decisiones de envío o no de unidades a la zona donde se ha declarado el incendio

# OBJETIVOS DEL PROYECTO

* Desarrollar un modelo predictivo basado en machine learning para estimar:
  + Recursos terrestres necesarios: autobombas y bulldozer
  + Recursos aéreos: helicópteros y aviones.
* Desarrollar un prototipo de aplicación para la toma de decisiones en tiempo real.

# 

# METODOLOGÍA

El proyecto sigue la metodología CRISP-DM, (CRosss Industry Standard Process for Data Mining) que garantiza un enfoque estructurado e iterativo, la cual incluye las siguientes fases:

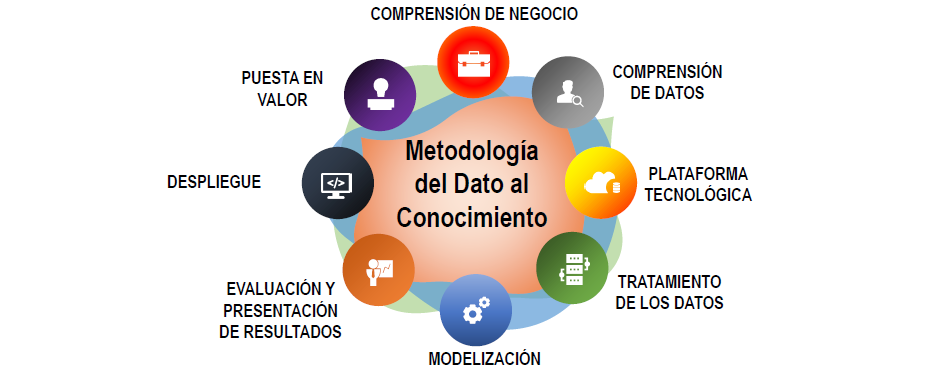


Figura 2 Esquema de la metodología seguida en el desarrollo del proyecto

## FASE: COMPRENSIÓN DEL NEGOCIO.

### Determinar los objetivos del negocio.

#### Antecedentes.

La gestión eficiente de incendios forestales es un desafío crítico en la Comunidad Valenciana, una región con alta incidencia de estos siniestros debido a factores climáticos, geográficos y vegetativos. La toma de decisiones en la fase inicial de un incendio es clave para minimizar daños y optimizar los recursos disponibles. Para ello, en este proyecto se lleva a cabo un análisis detallado del negocio, comprendiendo los procesos actuales de respuesta ante incendios, las variables clave en la toma de decisiones y las necesidades operativas de los equipos de extinción.

#### Objetivos del negocio.

En este contexto, el *negocio*, busca optimizar la gestión de incendios forestales mediante tecnología avanzada para mejorar la eficacia, reducir costos y minimizar el impacto ambiental y social. A largo plazo, pretende crear un modelo replicable para la gestión de emergencias en distintas áreas, contribuyendo a la mejora continua de estrategias de prevención y extinción. En la Comunidad Valenciana, los objetivos deben centrarse en:

**Reducción de Costos Operativos**: Se busca disminuir el uso innecesario de recursos durante las operaciones de extinción, lo que implica una mejor asignación de los medios disponibles (como medios de transporte, autobombas y recursos aéreos), asegurando que los costos sean eficientes en la fase de **comprensión y planificación del negocio**.

**Mejora en la Precisión de la Asignación de Recursos**: Utilizando modelos predictivos, se puede estimar con mayor precisión qué recursos son necesarios en cada incendio, ajustándose a las condiciones específicas del terreno y meteorológicas. Este objetivo se integra en la **fase de modelado** de CRISP-DM, mejorando la toma de decisiones.

**Integración de Ciencia de Datos**: La incorporación de tecnologías de ciencia de datos para fortalecer la capacidad predictiva y preventiva es fundamental para el éxito del proyecto. Se espera que, mediante la integración de datos históricos de incendios y condiciones meteorológicas, sea posible predecir mejor la asignación de recursos en los incendios futuros. Esto es clave en la **fase de preparación de los datos** y **modelado**.

**Impacto Ambiental y Social Reducido**: Al ser más eficientes en la asignación de recursos, se espera no solo reducir los costos operativos, sino también minimizar el impacto ambiental y social de los incendios, controlando los focos de manera más efectiva y rápida, y evitando la expansión descontrolada de los mismos. Este aspecto será evaluado durante la **fase de evaluación**, midiendo los resultados de la estrategia implementada.

#### Criterios de éxito del negocio.

Los criterios de éxito son parámetros clave que permiten evaluar el rendimiento y la eficacia del proyecto. Proporcionan medidas cuantificables del progreso del análisis hacia la consecución de los objetivos y sirven como referencia final para determinar si se han alcanzado las metas establecidas.

En este caso, el éxito del proyecto se medirá en función de su capacidad para optimizar el uso de los medios enviados en la respuesta inicial a incendios forestales. Un criterio fundamental será la reducción del envío innecesario de medios, con un objetivo inicial de disminuir su despliegue en al menos un 15% sin comprometer la eficacia en la extinción.

Al final de la campaña, se calculará una métrica comparativa entre las predicciones del modelo y los recursos realmente utilizados, permitiendo evaluar el impacto de la herramienta en la toma de decisiones operativas.

### Diseño de la solución operativa.

La solución operativa propuesta tiene como objetivo modernizar y optimizar la gestión de incendios forestales a través de la integración de tecnología avanzada, como modelos predictivos y sistemas de optimización de recursos. Este sistema no solo mejorará la eficacia y eficiencia en la respuesta a los incendios, sino que también contribuirá a la mejora continua de las estrategias de prevención y extinción en la Comunidad Valenciana. A continuación, se describen las funcionalidades clave de la Solución Operativa propuesta:

* Optimización de Recursos Operativos:

Algoritmos que permitan predecir la asignación de recursos para la lucha contra un incendio forestal, basados en datos históricos de incendios en la Comunidad Valenciana, y siguiendo criterios de eficiencia, minimizando el uso innecesario de medios.

* Plataforma Predictiva en Tiempo Real:

Una aplicación web para los operadores del Centro de Coordinación de Emergencias (CCE) y los responsables del Puesto de Mando Avanzado (PMA), donde podrán visualizar las predicciones sobre la necesidad de recursos (medios terrestres, medios aéreos, autobombas) en el momento de detección de un incendio forestal.

### Evaluación de la situación.

#### Inventario de recursos.

* Recursos de Hardware
  + PC’s con procesador Intel(R) Core(TM) i3-5010U CPU @ 2.10GHz 2.10 GHz.
  + Sistema operativo Windows 10.
* Fuentes de Datos:
  + Bases de datos históricas sobre incendios en la Comunidad de Valencia..
  + Datos meteorológicos proporcionados por AEMET.
  + Información geográfica de la Comunidad Valenciana en formato shapefile.
* Lenguajes y plataformas:
  + Python.
* Software y entornos de desarrollo:
  + QGIS.
  + Google Earth Engine.
  + Jupyter Notebook.
  + Google Colab.
  + GitHub.
* Otra información relevante:
  + Informes históricos sobre la gestión de incendios.
  + Manuales operativos y protocolos establecidos bajo el Plan Especial frente al Riesgo de Incendios Forestales (PEIF).

#### Requerimientos, suposiciones y restricciones.

* Para garantizar el correcto desarrollo del modelo, es fundamental establecer de manera clara los requisitos, supuestos y restricciones del proyecto. Analizar las posibles limitaciones permitirá comprender qué aspectos son esenciales para alcanzar los objetivos y qué premisas deben cumplirse para asegurar su éxito.
* El éxito del modelo predictivo depende de ciertos **requisitos** fundamentales:
* Calidad de los datos: Los datos utilizados deben ser completos, consistentes y libres de errores. La veracidad de la información es crucial, ya que una mala calidad en los datos puede afectar significativamente la precisión de las predicciones.
* Acceso a los datos: Los datos deben ser abiertos y accesibles, permitiendo su uso, reutilización y redistribución sin restricciones excesivas.
* Para el correcto desarrollo del modelo, se establecen ciertos **supuestos** que se consideran válidos:
* Los datos históricos son suficientes y representan fielmente el comportamiento de los incendios forestales en la región.
* Las variables meteorológicas y de terreno influyen directamente en la asignación de recursos y permiten hacer predicciones precisas.
* El sistema será utilizado exclusivamente por personal autorizado y capacitado en la gestión de emergencias.
* Durante la operación del sistema, los datos estarán disponibles en tiempo real desde fuentes oficiales.
* En el corto plazo, no se esperan cambios drásticos en el comportamiento de los incendios forestales que afecten la validez del modelo.
* Sin embargo, dado el alcance y características del proyecto, existen ciertas limitaciones a considerar:
* Solo se emplearán datos oficiales aprobados por las autoridades competentes, lo que puede limitar la disponibilidad de información en algunos casos.
* En una primera fase, el modelo no tendrá que integrarse completamente con las plataformas tecnológicas existentes de gestión de emergencias. Sin embargo, su diseño debe ser intuitivo y accesible para el personal operativo.
* Se establecerá un control sobre los costos de implementación y mantenimiento, evitando soluciones tecnológicas excesivamente costosas.
* El modelo estará limitado geográficamente a la **Comunidad Valenciana**, sin capacidad de extrapolación a otras regiones sin modificaciones adicionales.
* Se excluirán variables que no estén disponibles en tiempo real, como la superficie quemada o la cantidad de agua descargada durante las operaciones de extinción.
* Estas consideraciones permitirán estructurar un modelo realista y funcional, garantizando su aplicabilidad en el contexto operativo de la gestión de incendios forestales en la región.

#### Riesgos y contingencias

Tras un análisis detallado, se ha determinado que no existe un riesgo significativo de competencia con herramientas similares en el mercado, ya que actualmente no hay soluciones con las mismas características y enfoque específico que este modelo. Por lo tanto, no se considera necesario evaluar planes de contingencia en este sentido.

#### Terminología

En un anejo correspondiente se compilará un glosario de términos relevantes que permita estandarizar y facilitar la comunicación entre los participantes del proyecto de desarrollo del modelo predictivo de recursos para incendios forestales en la Comunidad Valenciana.

El glosario incluirá:

1. Terminología de negocio: Se definirán términos clave relacionados con el ámbito forestal y operativo, tales como "incendio forestal", "recursos de extinción", "combustible forestal", "medios aéreos" y "brigada helitransportada", entre otros. Esto garantizará que todos los involucrados compartan un mismo marco de referencia al discutir sobre la problemática y las estrategias de respuesta.
2. Terminología de ciencia de datos: Se explicarán los conceptos técnicos utilizados en el desarrollo del modelo predictivo, tales como "data leakage", "modelo de clasificación", "algoritmo de regresión", "validación cruzada" y "importancia de variables".

Este glosario servirá como una herramienta de referencia para asegurar una comunicación clara y coherente entre los diferentes equipos involucrados en el desarrollo e implementación del modelo.

#### Costes y Beneficios

En esta fase del proyecto, no se contempla un análisis detallado de costes y beneficios, ya que el enfoque está en evaluar la precisión y utilidad operativa del modelo predictivo.

Actualmente, el modelo no está integrado en los procesos oficiales de gestión de incendios, sino que se utilizará como herramienta de apoyo en la toma de decisiones. Sin una implementación real, no es posible medir su impacto económico con precisión.

Además, la evaluación del modelo se hará manualmente al final de la temporada de incendios, sin un sistema formal de medición de ahorro en costes. También, el modelo sigue en fase de ajuste, por lo que cualquier análisis económico prematuro podría quedar obsoleto ante futuras optimizaciones.

En fases posteriores, si el modelo demuestra ser eficaz, se analizará su impacto económico en términos de ahorro en recursos, reducción de tiempos de respuesta y costes de implementación.

Por ahora, la prioridad es validar su precisión y utilidad operativa, dejando el estudio de costes y beneficios para cuando se disponga de datos más sólidos sobre su desempeño real.

### Determinación de los objetivos analíticos.

#### Objetivos analíticos.

El objetivo analítico de este proyecto es desarrollar un modelo predictivo que permita estimar con precisión los recursos necesarios para la extinción de incendios forestales en la Comunidad Valenciana. Se busca determinar si un incendio requerirá medios aéreos, terrestres o una combinación de ambos, así como predecir la cantidad específica de autobombas, helicópteros, aviones y brigadas necesarias en cada caso.

#### Criterios de éxito de Ciencia de Datos

Desde la perspectiva de ciencia de datos, el éxito del modelo se medirá en función de su precisión y capacidad de generalización, asegurando que pueda predecir correctamente los recursos necesarios en incendios no vistos previamente. Se establecerán métricas como precisión y F1-score para la clasificación, y MAE y RMSE para la predicción de cantidades específicas. Además, se garantizará la interpretabilidad del modelo, permitiendo a los responsables de emergencias comprender los factores que influyen en las predicciones. El modelo deberá integrarse de manera eficiente en la toma de decisiones operativas, proporcionando resultados en tiempo real. Finalmente, su desempeño será evaluado al final de la temporada de incendios, ajustándolo para mejorar su eficacia en futuras campañas.

### Plan de proyecto.

#### Plan de Proyecto

Este proyecto se desarrolla siguiendo las etapas del proceso CRISP-DM, pero es relevante resaltar que, esta metodología no es lineal, sino que es iterativa, de modo que, según los requerimientos del modelo y los avances obtenidos, puede ser necesario volver a etapas previas para ajustar y mejorar el enfoque a lo largo del proceso. De manera muy genérica el proyecto contempla las siguientes etapas:

**Recolección de datos oficiales de** incendios forestales desde la web de la Conselleria de Medio Ambiente de la Comunidad Valenciana, que incluirá una depuración inicial de los datos y el almacenamiento en una base de datos en formato CSV en la nube.

**Exploración y verificación de los datos e incorporación de datos relevantes** mediante consultas en Google Earth Engine, lo que permitirá enriquecer las variables disponibles para el análisis.

**Preparación de los datos para el análisis** lo que incluirá la limpieza de registros, la transformación de variables y la eliminación de datos que se encuentren fuera del perímetro de la Comunidad Valenciana.

**Modelado** donde se emplearán herramientas de Python para desarrollar modelos predictivos.

**Evaluación de los resultados** utilizando métricas adecuadas según el tipo de problema.

**Implementación y futuras líneas de investigación**

#### Evaluación inicial de técnicas analíticas

Dado que el objetivo del trabajo es predecir el uso de distintos medios aéreos y terrestres en la extinción de incendios forestales se seleccionarán las herramientas y enfoques más adecuados para el problema a resolver. En este caso:

En el análisis exploratorio de datos (EDA) se han aplicado las siguientes técnicas:

* **Distribución de variables:** Se analizaron histogramas y diagramas de caja para identificar distribuciones sesgadas, valores atípicos y posibles problemas de calidad de datos.
* **Correlación entre variables:** Se generaron matrices de correlación y mapas de calor para identificar relaciones fuertes entre variables predictoras y la variable objetivo.
* **Análisis de valores faltantes:** Se identificaron patrones de valores nulos y se evaluaron estrategias para su imputación o eliminación.
* **Segmentación de datos:** Se analizaron los incendios en función de factores clave como tipo de terreno, condiciones meteorológicas y ubicación geográfica.

Dado que el problema puede abordarse desde una perspectiva de clasificación y regresión, se han considerado diferentes modelos:

* **Random Forest:** Proporciona interpretabilidad y buen rendimiento en datos estructurados.
* **Gradient Boosting (XGBoost):** Eficiente en la detección de patrones complejos y captura relaciones no lineales.
* **Regresión Lineal Múltiple:** Fácil de interpretar y establecer relaciones lineales.
* **XGBoost Regressor:** Para capturar relaciones más complejas con alta capacidad predictiva.

Los modelos se evaluarán utilizando:

* **Para clasificación:** Precisión, recall, F1-score y matriz de confusión.
* **Para regresión:** Error absoluto medio (MAE) y error cuadrático medio (MSE).

## FASE: COMPRENSIÓN DE DATOS.

### Recolección de datos iniciales.

En esta fase se recopilan, analizan, describen y explican las fuentes de datos que más tarde nos serán de utilidad para la creación de los modelos predictores. El objetivo de esta etapa **es evaluar los datos existentes para determinar su calidad y poder establecer una estrategia de limpieza y preparación** para la siguiente fase.

|  |  |  |
| --- | --- | --- |
| **Fuente de datos** | **Naturaleza** | **Origen** |
| Estadistica\_completa.csv | Base de datos abierta | https://mediambient.gva.es/ |
| df\_coordenadas.csv | Transformación a latitud y longitud | https://mediambient.gva.es/ |
| df\_GEE.csv | webscrapping | Google Earth Engine |
| df\_grid\_20x20.csv | Qgis | https://mediambient.gva.es/ |

Figura 3. Fuentes de datos con su naturaleza de origen e ubicación de extracción

### Descripción de los datos.

***Estadistica\_completa.csv***

Tras realizar un estudio profundo de las fuentes de datos, se ha identificado como más relevante la que podemos descargar desde la web de la Conselleria de Medio Ambiente de la Comunidad Valenciana. Los datos proporcionados en este archivo abarcan los incendios en la Comunidad de Valencia en el periodo de 1997 a 2015. El conjunto de datos está estructurado en un archivo CSV con 10,722 registros y 159 columnas. A continuación, se presentan las categorías principales que agrupan los datos disponibles:

**Causas de Incendios**: Este grupo agrupa las variables relacionadas con la causa del inicio del incendio. Sin embargo, estas variables no se incluyen en el modelo, ya que las causas son desconocidas en el momento en que se inicia el incendio.

**Condiciones del Terreno**: Agrupa las variables relacionadas con el terreno afectado por el incendio. En este caso, solo se incluye la variable "EROSIÓN", que indica la vulnerabilidad del terreno a este factor.

**Condiciones del Combustible**: Este grupo abarca todas las variables que afectan directamente a los combustibles presentes en la zona del incendio, como el tipo de combustible y los índices de propagación del fuego.

**Condiciones Meteorológicas**: Incluye todas las variables meteorológicas relacionadas con los datos de viento, temperatura y humedad relativa en el momento del inicio del incendio. Este grupo es fundamental para el modelo, ya que las condiciones climáticas son determinantes en la propagación del fuego.

**Daños**: En este grupo se encuentran las variables sobre los daños a la infraestructura y la superficie quemada. No se incluyen en el modelo, ya que estos datos solo se pueden conocer después de que el incendio haya sido extinguido.

**Localización y Fecha**: Agrupa las variables relacionadas con la localización del incendio, como las coordenadas de inicio, la población, el término municipal, y las fechas y horas de detección y llegada de los medios de extinción. Algunas variables, como la hora de extinción, no se consideran en el modelo, ya que son datos posteriores.

**Métodos de Lucha**: Este grupo incluye información sobre los métodos de control utilizados, como los ataques directos o indirectos y el uso de retardantes. Estas variables no se incluyen en el modelo, ya que corresponden a decisiones posteriores a la detección inicial del incendio.

**Recursos**: Engloba las variables objetivo del modelo, que buscan determinar los recursos necesarios, tales como medios aéreos, terrestres y humanos para la extinción del incendio.

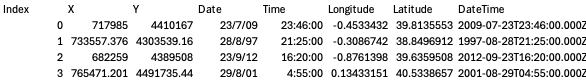
**Otros**: Contiene variables que no son útiles para el estudio y, por lo tanto, no se consideran en el análisis.

***df\_coordenadas.csv***

Esta base de datos es una transformación de los datos originales de *Estadistica\_completa.csv*, donde hemos cogido las variables de ‘X’ e ‘Y’ en formato UTM para transformarlas a latitud y longitud, ya que es la manera de poder extraer los datos en google earth engine.

Para eso se desarrolló un pipeline en Python que cogía cada línea de coordenadas X e Y y las transformaba a latitud y longitud (nom en formato dms).

Después se hacía otra transformación a formato ‘DataTime’ legible por GEE, para que pudiera extraer el dato concreto de un día y hora específico, este formato se rige de la siguiente manera *ISO ‘2003-01-09T15:20:00.000Z’*

**

***df\_GEE***

A través de la base de datos df\_coordenadas, se introduce en el sistema asset manager de GEE y se escribe el código para poder hacer las llamadas a la base de datos de GEE y recopilar datos específicos de nuevas variables que nos van a ser de utilidad a la hora de desarrollar el modelo. Las nuevas variables que se obtienen son de tipo geoespaciales, meteorológicos y operacionales que cubren una amplia gama de aspectos relacionados con los incendios forestales con la API de Google Earth Engine.

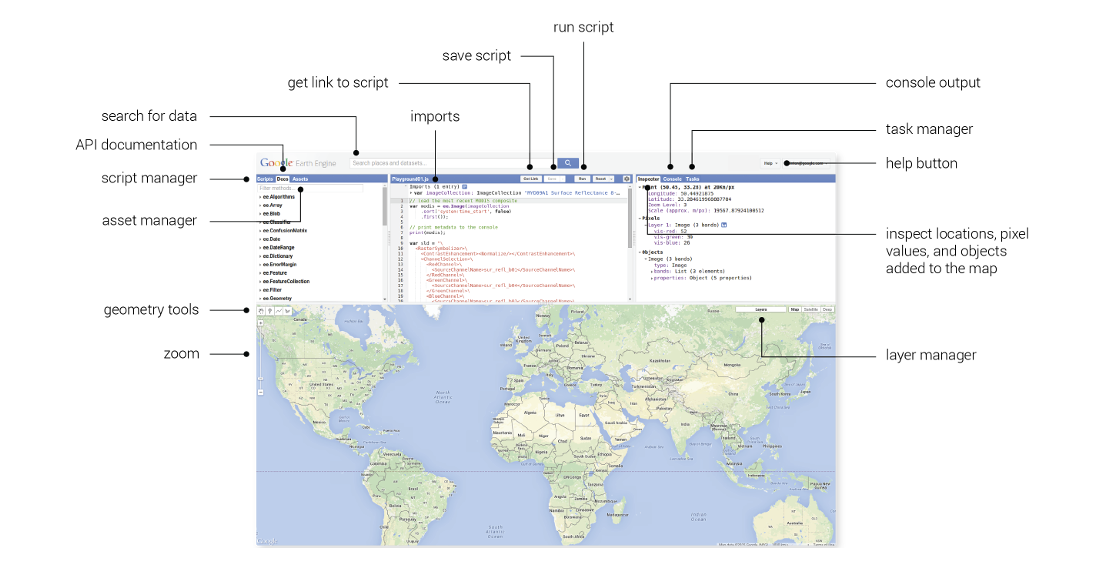


Figura 4 Imagen esquemática de la plataforma de Google earth engiene

La incorporación de nuevas variables en los modelos de predicción de incendios es fundamental para mejorar la precisión en la estimación del comportamiento del fuego y optimizar la asignación de recursos. Estas variables proporcionan información adicional o información más precisa, sobre los factores que influyen en la propagación del fuego, como la **Elevación, FuelLoadIndex, LandCover, DifficultyIndex, temperatura**, **velocidad del viento**, **humedad** (a través de temperatura del rocio), **pendiente**. Al considerar estos aspectos, se obtiene un panorama más completo y se mejora la capacidad de predicción.

El uso de variables como el **Índice de Dificultad de Acceso** y el **FuelLoadIndex** permite identificar áreas de difícil acceso y zonas con alta carga de combustible, lo que facilita la planificación de recursos. Esto asegura que se prioricen las zonas más críticas y se asignen los medios adecuados para su control.

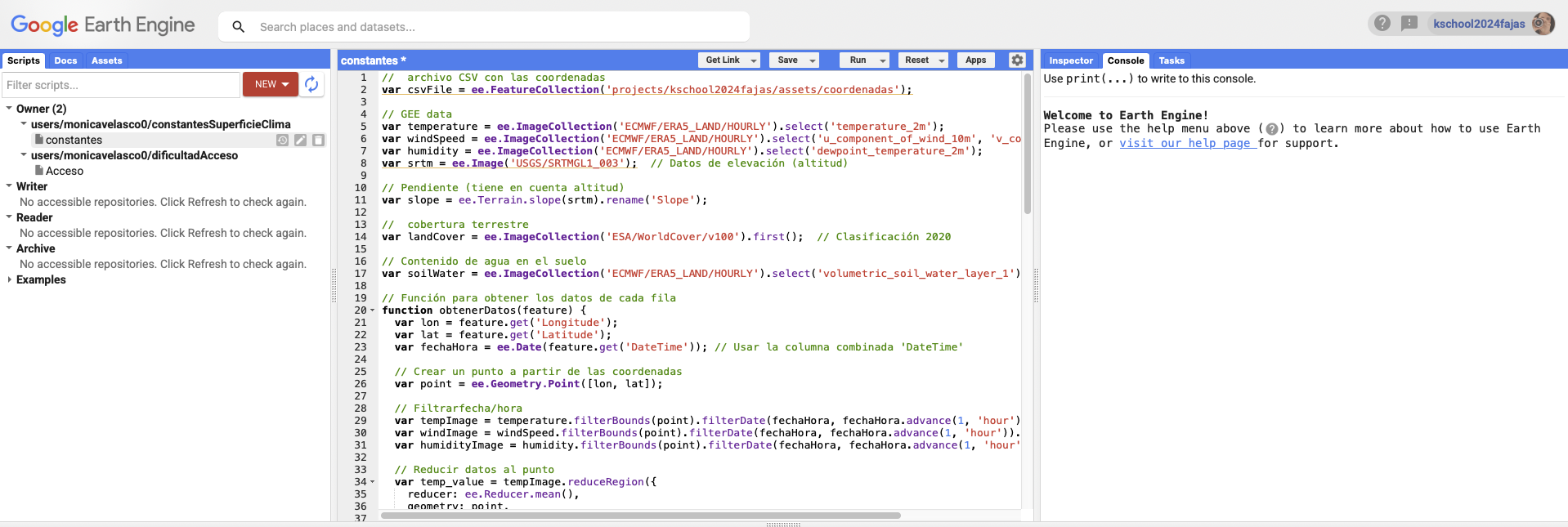


Figura 5. Imagen de la pantalla de script usada para la extracción de las variables

***df\_grid\_20x20.csv***

El archivo df\_grid\_20x20.csv contiene los datos de una cuadrícula de 20x20 km que cubre el total de la Comunidad Valenciana. El archivo consta de 170 registros con las siguientes características:

id: Identificador de cada celda en la cuadrícula.

left, top, right, bottom: Coordenadas geoespaciales que delimitan cada celda.

row\_index, col\_index: Índices de fila y columna en la cuadrícula.

name2: Un identificador adicional para cada celda formado por el índice de la fila y la columna.

### Exploración de los datos

#### Exploración de dataset inicial (df\_original.csv)

En una primera fase haremos una exploración del conjunto inicial de datos provenientes de la descarga desde la web de la Conselleria de Medio Ambiente de la Comunidad Valenciana. Más adelante se hará una exploración del dataset final con las variables preprocesadas además de las provenientes de Google Earth Engine.

**El dataset inicial (df\_original.csv) consta de 10722 filas y 142 columnas**. La descripción de todas las variables se puede encontrar en *resumen\_variables\_tfm.xlsx* del repositorio de Github.

El dataset inicial incluye dos columnas, ‘X’ e ‘Y’, que representan la ubicación del foco del incendio en coordenadas UTM. Dado que la Comunidad Valenciana se encuentra en la zona UTM 30N, utilizaremos este parámetro para convertirlas a coordenadas geográficas (latitud/longitud). Sin embargo, en buena parte de los registros, ambos valores aparecen en cero, lo que equivale a carecer de localización precisa. Este problema se abordará más adelante. A continuación, se muestra una visualización de los fuegos cuyos datos de localización sí están disponibles desde el inicio:

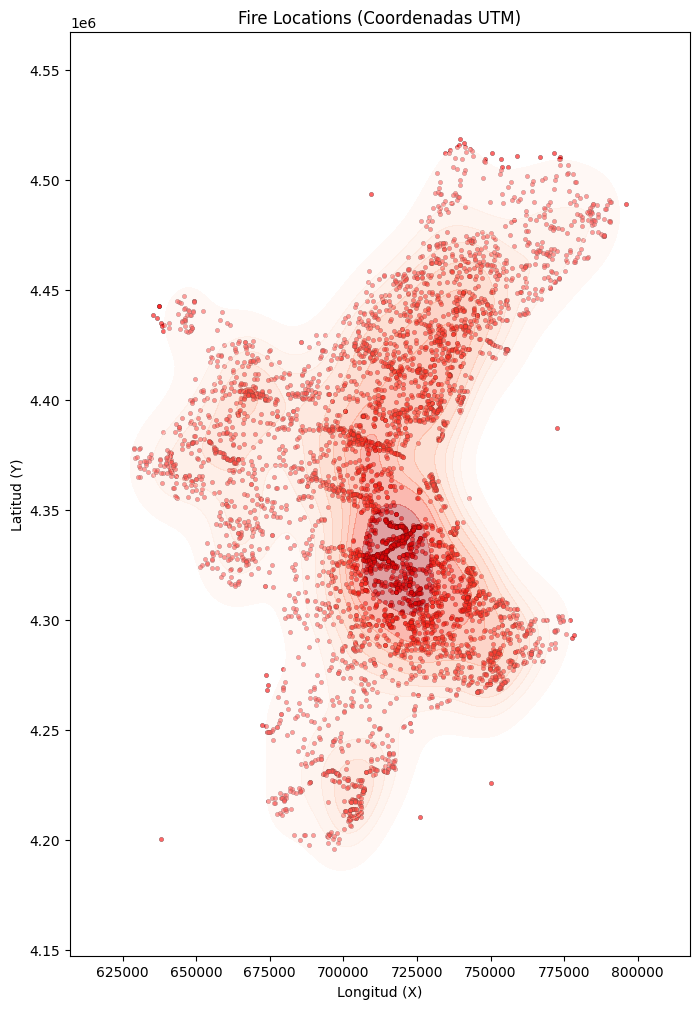


Figura 6. Representación de los fuegos localizados en la comunidad valenciana

Vamos a ver los incendios por provincias y comarcas para ver las zonas más afectadas:

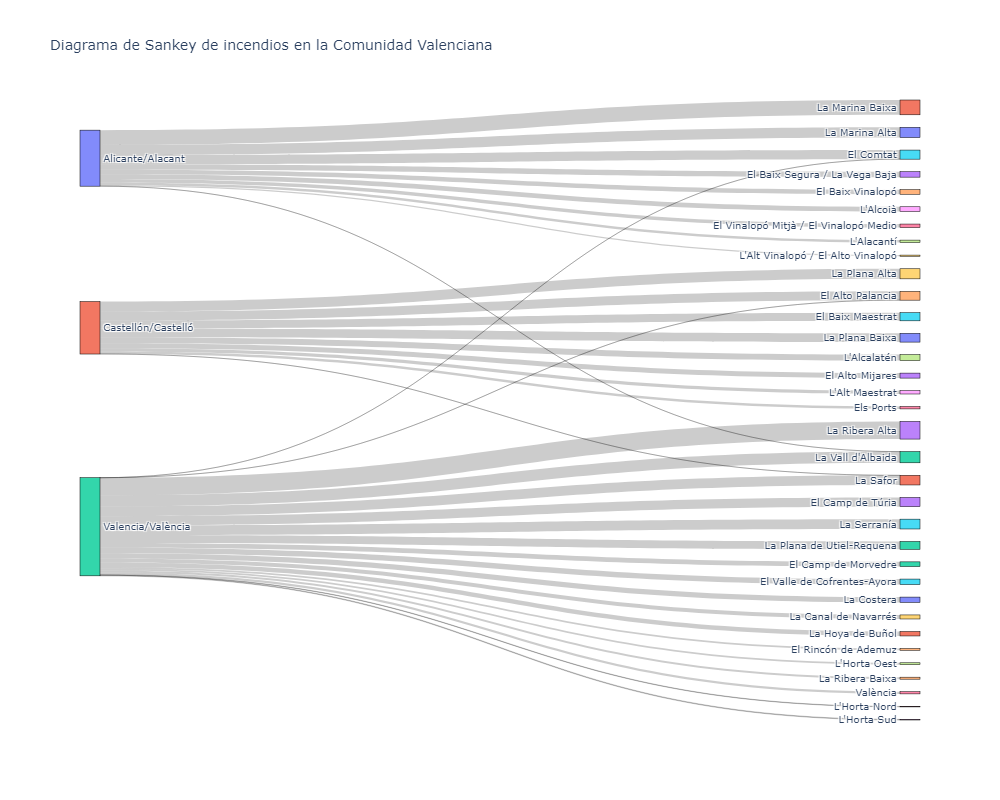


Figura 7 Diagrama de incendios de la comunidad valencia dependiendo de las provincias y las comarcas valencianas

Se puede observar que, si bien, tenemos incendios distribuidos por toda la región se puede observar una gran concentración de éstos en la zona centro de la provincia de Valencia y norte-centro de la provincia de Alicante.

Respecto a los incendios por años no hay un patrón que nos hable de una tendencia de crecimiento/decrecimiento de número de incendios por años. **Podemos destacar 1993, 1994 y 2005** como años donde ocurrieron un número de incendios altamente inusual.

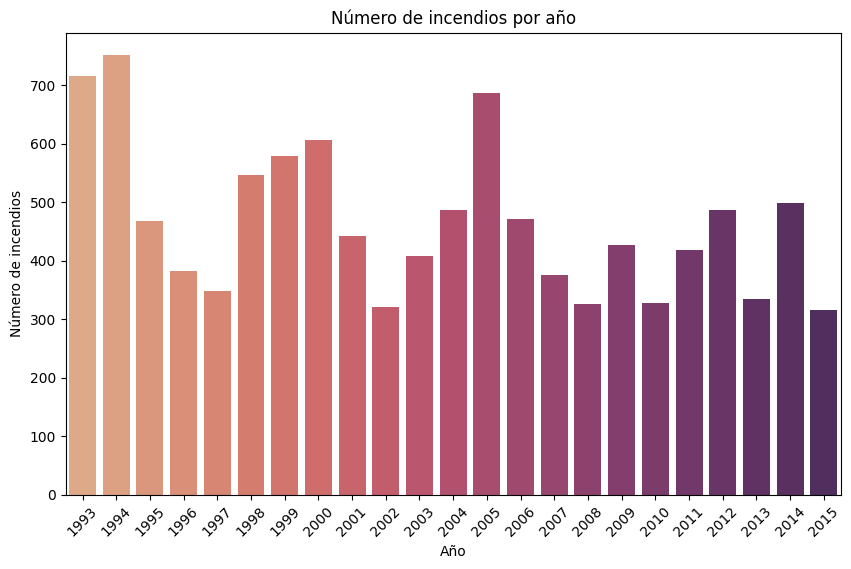


Figura 8. Número de incendios en la comunidad Valenciana dependiendo de los años en los que lo estudiamos

Veamos la distribución de incendios por mes del año:

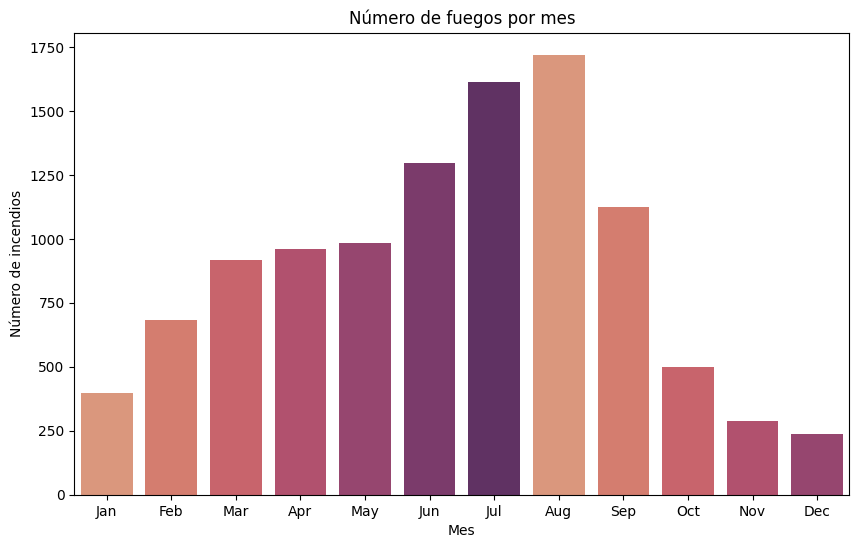


Figura 9. Distribución de incendios dependiendo de los meses del año

Se observa una tendencia previsible de un incremento notable de incendios a partir de la primavera, **alcanzando su máximo en julio y agost**o, los meses más calurosos del año. Durante este periodo, la vegetación se encuentra más seca, lo que proporciona mayor cantidad de combustible para el inicio y desarrollo de los incendios.

Representamos el número de incendios por día de la semana:



Figura 10. Distribución por días de la semana

Vamos a estudiar ahora la superficie quemada en estos incendios por año:

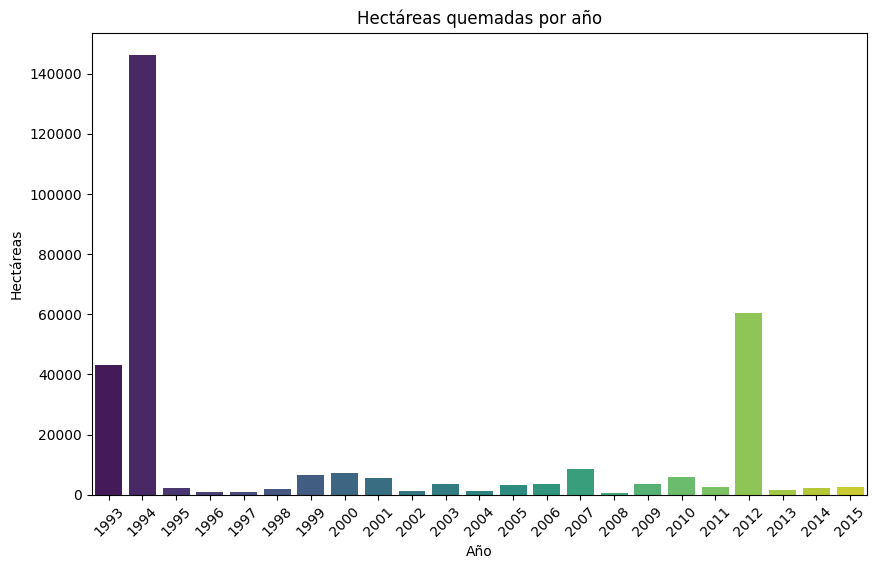


Figura 11. Área registrada quemada tras el incendio a lo largo de los años de los que tenemos información.

1993 y 1994 vuelven a ser años especialmente nefastos en cuanto a incendios se refiere. 2012 a pesar de no destacar especialmente en el anterior gráfico de número de incendios se erige como el peor año de los últimos años. Podemos corroborarlo con [este link](https://es.wikipedia.org/wiki/Incendio_forestal_de_la_Comunidad_Valenciana_de_2012) a la página de la Wikipedia donde nos habla de un incendio especialmente cruento durante ese año.

Los incendios pueden estar provocados por causas muy diferentes. Veamos en una visualización las causas principales:

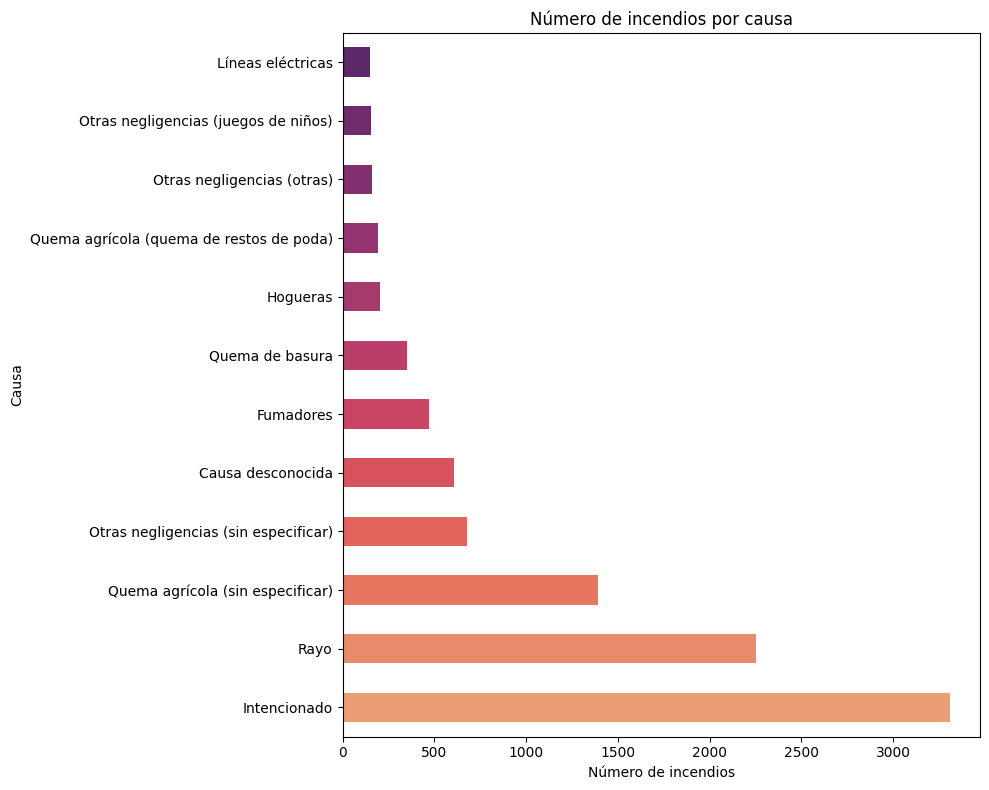
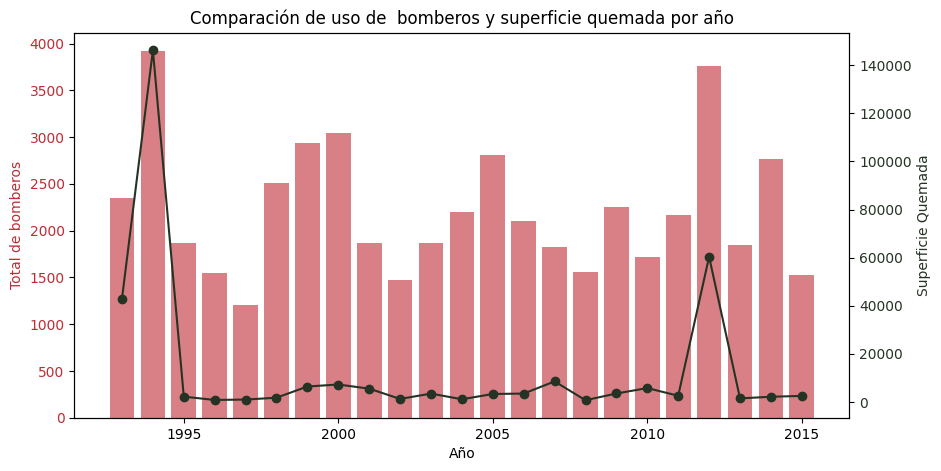


Figura 12. Tipo de causa que provocó el inicio del incendio

**Como muestra la visualización, una gran cantidad de incendios son intencionados o debido a negligencias del ser humano**. También es de reseñar los incendios producidos por un rayo.

Vamos ahora a cruzar el número de medios técnicos con el número de hectáreas quemadas por año para ver la consistencia de nuestros datos:

Gráfico, Gráfico de barras, Histograma

El contenido generado por IA puede ser incorrecto.

Figura 13. Uso de bomberos (izquierda) uso de autobombas (derecha) en relación con el are final quemada.

No podemos pasar por alto el coste personal en vidas humanas perdidas en este tipo de catástrofes, donde destacan los años previamente mencionados.



Figura 14. Personas que han fallecido a causa del incendio de manera directa o indirectamente

#### Exploración de dataset final (df\_con\_grid\_fire.csv)

En este dataset tenemos nuevas variables meteorológicas procedentes de Google Earth Engine como “WindSpeed\_U", "WindSpeed\_V", "Temperature (C)" y "Humidity (%)”.

Represento sus distribuciones:

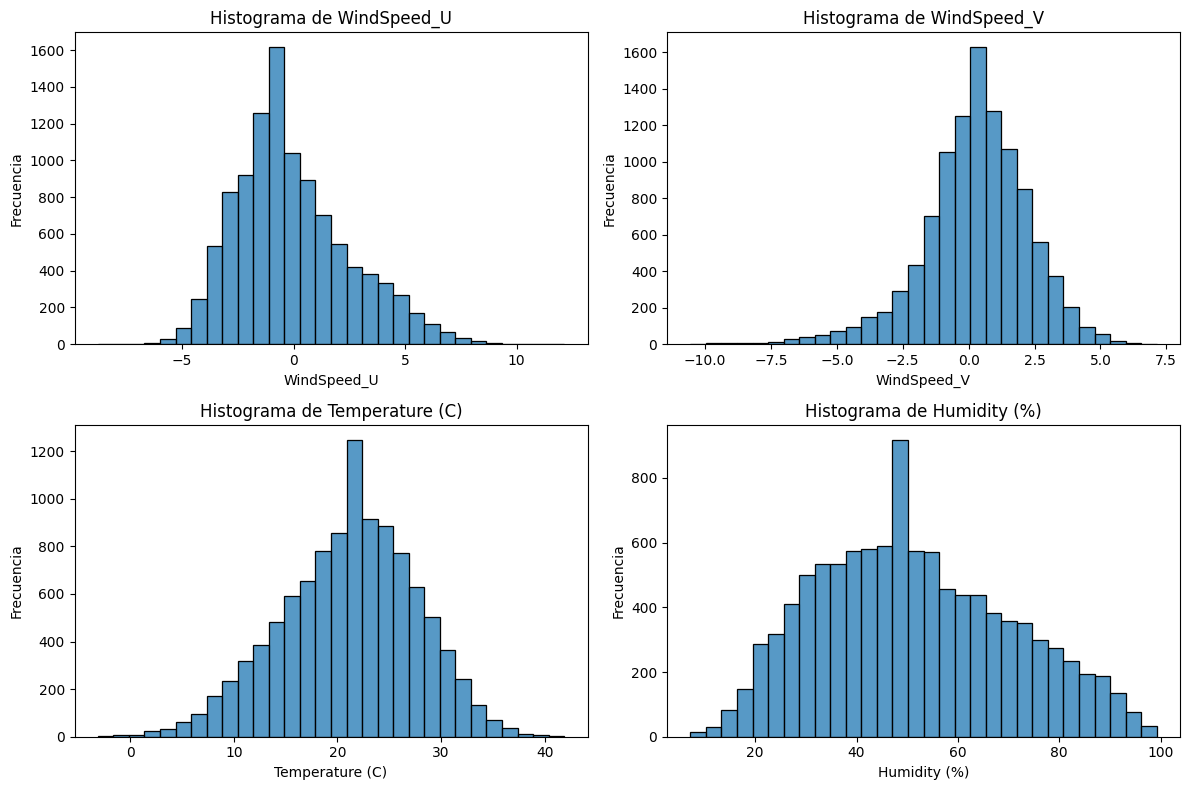


Figura 15 Representación en forma de histograma de las variables extraídas de GEE que se usan en el modelo

A partir de los datos de “WindSpeed\_U" y "WindSpeed\_V" creamos un gráfico tipo rosa de los vientos para ver los vientos predominantes durante los fuegos. Incluso se podría deducir que estos incrementan la probabilidad de que un fuego se propague.



Figura 16. Representación de viento, dependiendo de la velocidad y la dirección de este

Podemos ver que los vientos imperantes son los de tipo sudeste que traen humedad y altas temperaturas del Mediterráneo.

Creamos una nueva rosa de los vientos solo con los datos de los top 200 incendios que han quemado más hectáreas.

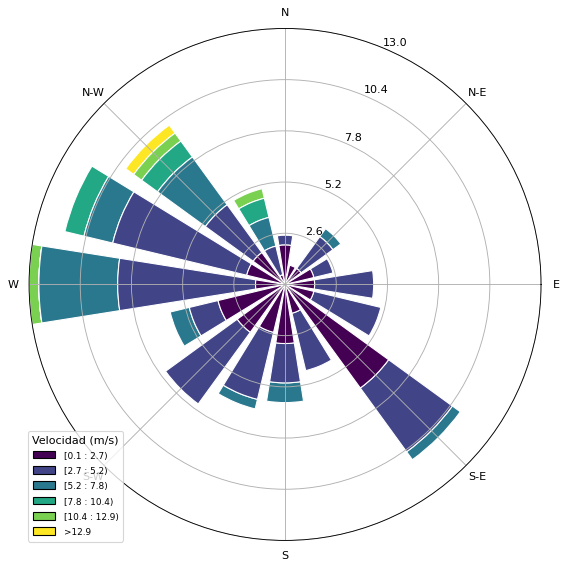
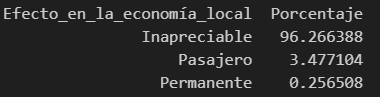


Figura 17. Representación del viendo en dirección y velocidad solo de los 200 incendios que más superficie quedó quemada tras el incendio

En esta ocasión los vientos predominantes son los vientos del este que disminuyen la humedad además de intuir una correlación entre alto valor de viento y grandes incendios.

Los fuegos, además de tener un coste de extinción y en vidas humanas tienen unas consecuencias económicas en la zona en la que suceden. En nuestro dataset tenemos información sobre esto.



1. Verificación de la calidad de los datos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset : Estadistica\_completa.csv | | | | |
| ***Dataset statistics*** |  |  | ***Variable types*** |  |
| Number of variables | 142 |  | Numeric | 79 |
| Number of observations | 10722 |  | Categorical | 44 |
| Missing cells | 112262 |  | Text | 4 |
| Missing cells (%) | 7.4% |  | DateTime | 14 |
| Duplicate rows | 0 |  | Unsupported | 1 |
| Duplicate rows (%) | 0.0% |  |  |  |
| Total size in memory | 11.6 MiB |  |  |  |
| Average record size in memory | 1.1 KiB |  |  |  |

Figura 18. Descripción de los datos que tenemos del primer *DataSet*

|  |  |  |
| --- | --- | --- |
| ***Categoría*** | ***Variable*** | ***Descripción*** |
| *Variables Constantes* | C\_A | Valor constante "9" |
| CCAA\_NOM | Valor constante "Comunitat Valenciana" |
| *Valores Nulos Altos (>50%)* | HUSO | 50.8% valores nulos |
| FPM\_AEEXT | 56.1% valores nulos |
| HPM\_AEEXT | 57.4% valores nulos |
| FPBRIGHEL | 84.4% valores nulos |
| HPBRIGHEL | 84.7% valores nulos |
| MOTIV\_TXT | 93.1% valores nulos |
| NCAUSANT | 100.0% valores nulos |
| RETARD\_TXT | 90.2% valores nulos |
| TATACI\_TXT | 98.8% valores nulos |
| ENP\_1\_TXT | 93.2% valores nulos |
| ID\_ENP\_TXT | 93.2% valores nulos |
| OBSERV | 81.4% valores nulos |
| *Variables con Alto Sesgo* | HUSO | γ1 = 45.18 (Sesgado) |
| AGENTES\_F | γ1 = 95.04 (Sesgado) |
| BULDOZZER | γ1 = 58.07 (Sesgado) |
| HELEXTNUM | γ1 = 29.94 (Sesgado) |
| HELEXTDES | γ1 = 33.52 (Sesgado) |
| MUERTOS | γ1 = 63.77 (Sesgado) |
| HERIDOS | γ1 = 28.19 (Sesgado) |
| A\_EST\_CA | γ1 = 69.57 (Sesgado) |
| S\_ARB | γ1 = 61.89 (Sesgado) |
| SUP\_F | γ1 = 37.40 (Sesgado) |
| *Valores Únicos* | ET\_ID | Todos los valores son únicos |
| NUMPARTE | Todos los valores son únicos |
| *Problema de Compatibilidad* | NCAUSANT | Tipo no compatible, necesita limpieza |
| *Variables con Alto Número de Ceros (>50%)* | DTORMENTA | 99.0% valores en 0 |
| VIGFIJ | 85.2% valores en 0 |
| MOTIVAC | 93.1% valores en 0 |
| TECNICOS | 92.0% valores en 0 |
| VOLUN\_ORG | 85.3% valores en 0 |
| BULDOZZER | 99.5% valores en 0 |
| RETARD | 90.2% valores en 0 |
| MUERTOS | 99.9% valores en 0 |
| HERIDOS | 99.5% valores en 0 |
| C\_TLF | 99.8% valores en 0 |
| ENP\_1 | 93.2% valores en 0 |

Figura 19. Descripción de las variables dentro del primer DataSet. Centrado en el valor de nulos.

## FASE: PLATAFORMA TECNOLÓGICA.

El éxito del sistema predictivo propuesto depende de una **plataforma tecnológica** robusta que permita la recopilación, procesamiento y análisis de datos en tiempo real, además de garantizar la disponibilidad del modelo de aprendizaje automático para su consulta a través de la **aplicación web.**

La arquitectura de la plataforma se divide en dos partes fundamentales:

**Arquitectura del modelo**: responsable del tratamiento de datos, entrenamiento y ejecución del modelo de predicción.

**Arquitectura de despliegue**: Encargada de exponer los resultados del modelo a través de una aplicación web, garantizando que los organismos encargados de la gestión de incendios puedan realizar consultas en tiempo real.

### Diseño de la Arquitectura de Referencia del Modelo.

Esta etapa se centra en la adquisición, preparación y procesamiento de los datos, así como en el desarrollo y entrenamiento del modelo de Machine Learning.



Figura 20. Esquematización de la arquitectura que tendremos en cuenta para el desarrollo del modelo

* Origen:
  + Archivo CSV con datos históricos de incendios en la Comunidad de Valencia.
  + Información geográfica de la Comunidad Valenciana (coordenadas) procesados con QGIS.
  + Datos geoespaciales (temperatura, humedad, accesibilidad…) obtenidos mediante Google Earth Engine.
* Ingesta
  + ETL (Extracción, Transformación y Carga): Procesos implementados en Python con librerías como pandas, geopandas y rasterio para limpieza y preparación de datos.
* Almacenamiento:
  + El almacenamiento de los datos será en un archivo .csv.
  + **Google Drive** se utiliza como repositorio central, permitiendo compartir y acceder a los datos desde distintas ubicaciones.
* Data Science:
  + El modelo de predicción se desarrollará y entrenará en Python con librerías como Scikit-learn, TensorFlow o XGBoost, dependiendo de la metodología más adecuada tras la experimentación.
  + Para evitar data leakage, solo se usarán variables disponibles en el momento en que se detecta un incendio.
  + Se evaluarán distintos enfoques, combinando modelos de clasificación (para determinar si se requieren medios de extinción) y modelos de regresión (para estimar la cantidad de medios necesarios).
* Visualización:
  + Aplicación web de acceso privado
* Procesamiento:
  + Con Python se aplicarán técnicas de limpieza, filtrado y transformación para estandarizar las diferentes fuentes de información. Se eliminarán valores atípicos y se imputarán valores faltantes cuando sea necesario.
* Gobierno del dato:
  + Se aplicarán normas básicas de organización, trazabilidad y seguridad para asegurar la fiabilidad de la información.
* Infraestructura: El uso de una infraestructura SaaS (Software as a Service) en la nube es la mejor estrategia para este proyecto, ya que proporciona escalabilidad y flexibilidad, permitiendo gestionar grandes volúmenes de datos sin necesidad de reconfiguración manual.

### Arquitectura de Referencia para el despliegue

La arquitectura de referencia para ell despliegue define cómo se integra y pone en producción el modelo de machine learning para que pueda ser utilizado en un entorno real, ya sea por otros sistemas o por usuarios finales.

En el contexto de este proyecto, dado que en esta fase el modelo es solo una herramienta de apoyo y no se integrará en los procesos operativos, se establece el siguiente despliegue.



Figura 21. Esquematización de la arquitectura que tendremos en cuenta para el desarrollo del despliegue

* **Origen:** Los datos utilizados en la predicción provienen de múltiples fuentes en tiempo real y bases de datos históricas:
  + Datos meteorológicos de estaciones cercanas al incendio.
  + Registros históricos de incendios utilizados como referencia para el modelo predictivo.
  + Coordenadas de inicio del incendio.
* **Ingesta:** La aplicación recibe solicitudes con información actualizada sobre un incendio (coordenadas, velocidad viento, época del año…) que utilizará para ejecutar el modelo.
* **Almacenamiento:** Para almacenar los datos utilizados por el sistema se implementa.
  + CSVpara almacenar información de incendios históricos.
* **Data Science:** El modelo de Machine Learning, previamente entrenado en la fase de desarrollo, se integra en la infraestructura de despliegue a través de un archivo. pkl (pickle)
* **Visualización:** La aplicación devuelve el resultado de ejecutar el modelo indicando que medios terrestres y/o aéreos son necesarios para luchar contra el incendio forestal.
* **Procesamiento:** El procesamiento en esta arquitectura de referencia se encarga de la transformación, limpieza y análisis de los datos antes de generar predicciones con el modelo entrenado. En primer lugar, los datos provenientes de fuentes externas, como archivos CSV y GEE, se preprocesan utilizando Pandas y NumPy, manejando valores nulos, convirtiendo variables categóricas en numéricas y escalando las características si es necesario.

### Estrategia tecnológica.

Para garantizar un funcionamiento eficiente del sistema predictivo, se ha adoptado una estrategia tecnológica diferenciada para las dos fases clave del proyecto:

* **Fase de desarrollo y entrenamiento (Arquitectura del Modelo).**
* **Fase de despliegue y operación en producción (Arquitectura de Explotación).**

**Arquitectura del Modelo (desarrollo y entrenamiento)**En esta etapa, se utilizará una infraestructura **SaaS (Software as a Service)** en la nube, que ofrece escalabilidad y flexibilidad para gestionar grandes volúmenes de datos y facilitar la colaboración entre los miembros del equipo. Concretamente:

* **GitHub** se empleará como repositorio central para el control de versiones y el desarrollo colaborativo del código.
* **Google Drive** se utilizará para el almacenamiento y la gestión de conjuntos de datos y documentación, aprovechando la facilidad de acceso y la sincronización automática de archivos.

**Arquitectura de Explotación (despliegue y operación)**Una vez finalizada la fase de desarrollo, la implementación en producción se realizará en una infraestructura **On-Premise**, con el fin de:

* **Garantizar la independencia de proveedores** y minimizar la dependencia de servicios externos.
* Reforzar la **seguridad y el control absoluto** de los datos y de la infraestructura, lo cual resulta crítico en entornos con altos requisitos de protección de la información.
* **Reducir riesgos** ante posibles emergencias o incidencias de disponibilidad de servicios en la nube.

Para la interacción con el sistema y la visualización de los resultados del modelo, se ha elegido **Streamlit** como interfaz principal, debido a su facilidad de uso y a la posibilidad de desarrollar paneles de control y visualizaciones de forma rápida e intuitiva. De esta manera, los usuarios finales pueden analizar la información y tomar decisiones de manera ágil, garantizando la **usabilidad** y el **mantenimiento** a largo plazo en el entorno de producción.

### Despliegue de la Arquitectura.

Durante esta etapa de desarrollo, no será necesario realizar un despliegue complejo. La ejecución del modelo se llevará a cabo localmente mediante **Jupyter Notebook**, utilizando **Python** con las librerías específicas del proyecto que se puede encontrar dentro de los .ipynb

Esto permitirá la validación y prueba del modelo en un entorno controlado, asegurando que las predicciones sean fiables antes de su explotación.

En la fase de explotación, la responsabilidad del despliegue del modelo recaerá sobre el **equipo de sistemas de la Conselleria de Medio Ambiente de la Comunidad Valenciana**. Este equipo se encargará de integrar la solución en la infraestructura existente, asegurando su accesibilidad para los organismos responsables de la gestión de incendios forestales.

## FASE: TRATAMIENTO DE DATOS.

En esta fase se ha procedido a la selección, limpieza, transformación y preparación de datos para el análisis.

### Selección de datos.

Se identificaron las variables relevantes, considerando la naturaleza de los conjuntos de datos, para garantizar que el análisis y el modelo predictivo estén basados en las características adecuadas. Como se ha comentado en apartados anteriores todos los datos que no tienen que ver con datos futuros, datos de personal humano que intervino o datos que son redundantes son eliminados de primeras.

### Limpieza de datos.

Se realizaron tareas de limpieza para eliminar los valores faltantes, corregir errores y tratar los valores atípicos que podrían afectar la calidad de los datos. Así se eliminan los registros que presenten valores atípicos en las variables X, Y, que corresponde con las coordenadas UTM de inicio del incendio, asegurándonos que las coordenadas están dentro de las coordenadas que ostenta la Comunidad de Valencia.

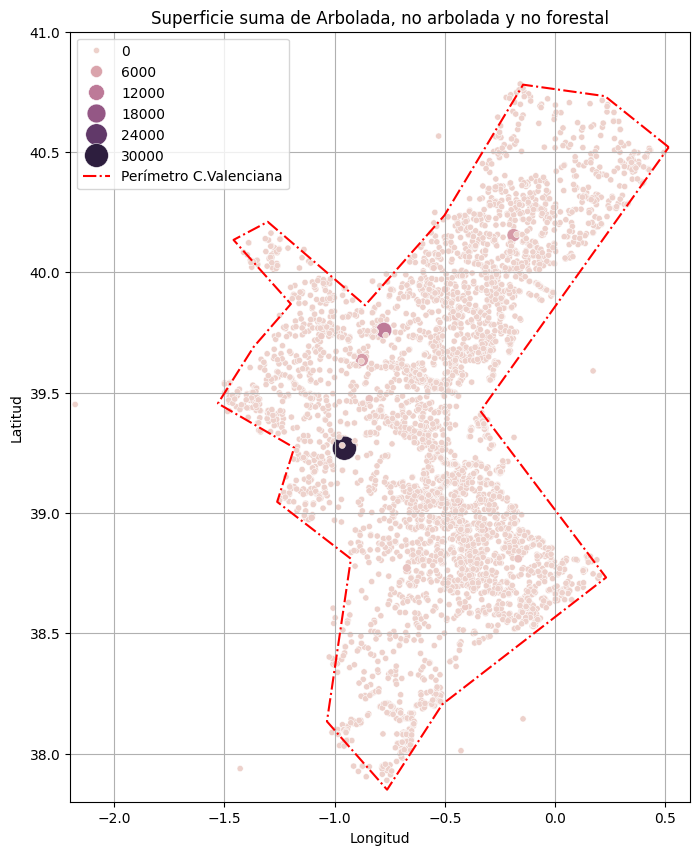


Figura 22. Representación del perímetro de la comunidad Valenciana con la variable de superficie quemada tras los incendios registrados

### Construcción de nuevos datos

#### Creación de coordenadas UTM faltantes

La mitad de nuestro dataset tenía valores faltantes de las coordenadas UTM por lo que nos era complicada la extracción de los datos a través de la plataforma de GEE, por lo que se hizo un tratamiento de estas líneas para poder obtener una coordenada UTM de la que se pudieran obtener las demás variables que extraemos de GEE.

Para ello, al tener datos de la comarca a la que pertenecían se han obtenido las coordenadas UTM de esa comarca y se ha podido de esta manera hacer todo el proceso de transformación en latitud y longitud y después se a podido dar el formato ‘DateTime’ propio de GEE para que puedan ser incluidas y procesadas en las llamadas a la base de datos de GEE.

#### Extracción de datos Google Earth Engine

Una vez que se han obtenido todas las coordenadas y se ha formateado el ‘DateTime’ se han realizado dos códigos en GEE para poder obtener los datos de las siguientes variables que se integrarán dentro de los códigos de modelos que realizaremos.

* Elevación
* Temperatura del rocío: después se transformará para poder obtener la humedad relativa
* LandCover
* Pendiente
* Temperatura: en Kelvin que se transformara para poder tenerla en grados centigrados.
* Velocidad del viento
* FuelLoadIndex
* DifficultyIndex

#### Creación de variables

Se procede a crear un índice de peligrosidad en la propagación del incendio que va a ser la variable de *fire\_risk\_index*, se tiene en cuenta una regla no escrita en el ámbito forestal como es la regla del 30/30/30, que dice que, si tenemos más de 30ºC de temperatura, menos de 30% de humedad relativa en el ambiente y más de una velocidad de 30Km/h entonces la peligrosidad del incendio es grande, ya que la probabilidad de crecer este es mucho mayor.

Para ello se usan las variables extraídas de Temperatura (ºC), Humedad relativa (%) y Velocidad del viento (Km/h) para el cálculo de este índice. Además, se les da diferente importancia a cada una de las variables que se normalizará dependiendo de los valores que hemos comentado.

def Temperatura\_normalizada(temperature):

return max(0, min(1, (temperature - 30) / (40 - 30)))

def Humedad\_normalizada(humidity):

return max(0, min(1, (30 - humidity) / 30))

def Viento\_normalizado(wind\_speed):

return max(0, min(1, (wind\_speed - 30) / (60 - 30)))

def calculate\_fire\_risk(row):

Temperature = Temperatura\_normalizada(row['Temperature (C)'])

Humedad = Humedad\_normalizada(row['Humidity (%)'])

Viento = Viento\_normalizado(row['V\_VIENTO'])

fire\_risk\_index = 0.4 \* Temperature + 0.4 \* Humedad + 0.2 \* Viento

A través de de este índice sintético que se ha creado, se crean unos índices de riesgo derivados de este pero dependientes de las otras variables que hemos obtenido como son el FuelLoadIndex, DifficultyIndex y LandCover

# Feature 1:

data['difficulty\_viento\_humedad'] = data['DifficultyIndex'] \* (data['V\_VIENTO'] + data['Humidity (%)'])

# Feature 2:

data['fire\_risk\_fuel'] = data['fire\_risk\_index'] \* (data['FuelLoadIndex'] / 100)

# Feature 3:

data['fire\_risk\_landcover'] = data['fire\_risk\_index'] \* data['LandCover']

### Integración de datos.

Se fusionaron los conjuntos de datos de cada fuente para crear un único conjunto de datos consolidado, eliminando aquellos registros innecesarios. Para esto se ha unido el *Dataset* inicial, con el *Dataset* después de la creación de nuevas coordenadas, transformación a latitud longitud y la extracción de GEE de las nuevas variables. de tal manera que tenemos las variables iniciales, las variables nuevas (GEE) y las variables obtenidas a partir de features engineering en un mismo .csv.

### Formateo de los datos.

Se realizaron diversas transformaciones en los datos, como la codificación de variables con el fin de facilitar el proceso de modelado. Para esto en los modelos de clasificación se realizan una serie de funciones en donde transformamos las variables de de un rango [0 - x\_n] a un rango de [0, 1], donde 0 es ausencia de efectivo y 1 efectivo enviado, después se verá en un análisis de regresión la cantidad total de efectivos, pero el objetivo principal es la predicción de si es necesario el efectivo no.

Quedándonos finalmente con las siguientes variables:

|  |  |
| --- | --- |
| **Nombre de la Variable** | **Descripción** |
| C\_A | Código de la Comunidad Autónoma del incendio |
| C\_MUNI | Código de término municipal de inicio registrado en la estadística. |
| CODMUN | Código de término municipal de inicio registrado en la estadística, junto con provincia |
| COM\_CODICV | Código de la comarca del incendio capas ICV |
| COMARCA | Código de la Comarca |
| CS\_ELEC | Cortes de suministro eléctrico |
| DifficultyIndex | Valor que mide la dificultad de acceder a una ubicación específica, basado en pendiente, elevación y cobertura del suelo. |
| CUAD | Cuadrícula del Mapa Militar 1:250.000 |
| D\_APIE | Distancia aproximada a pie (metros) |
| D\_VIENTO | Dirección del Viento. N 360, NE 45, E 90, SE 135, S 180, SO 225, O 270, NO 315. El valor 0 se interpreta como sin datos. |
| DTORMENTA | Días desde la última tormenta |
| DULLUVIA | Días desde la última lluvia |
| DifficultyIndex | Indice de Dificultad extraido desde GEE |
| difficulty\_viento\_humedad | Indice de dificultad dependiente combinado con viento y humedad |
| EROSION | Riesgo de erosión. 0 – Bajo;1- Moderado;2- Alto |
| ELEVATION | Elevacion del punto de coordenadas |
| ESTACION | Código de la estación meteorológica |
| Fire\_risk\_index | Riesgo de propagación de incendio dependiendo de la temperatura (30ºC) Humedad (30%) Viento (30km/h) |
| Fire\_risk\_Fuel | índice de dificultad combinado con la variable de FuelLoadIndex |
| Fire\_risk\_landcover | Índice de dificultad combinado con la variable de LandCover |
| FuelLoadIndex | Mide la cantidad de biomasa disponible para quemarse en una determinada área. (toneladas por hectaria) |
| Humidity (%) | Humedad Relativa |
| HOJA | Hoja del Mapa Militar 1:250.000 |
| IND\_PEL | Índice del Peligro Meteorológico. |
| LandCover | Clasificación del uso de la tierra en categorías como bosques, pastizales, cuerpos de agua, etc. |
| MOD\_COM | Modelos de Combustible en la zona de incendio, o combinación |
| PROV | Código de la Provincia |
| V\_VIENTO | Velocidad del Viento Km/h |
| X | Coordenada formato UTM |
| Y | Coordenada formato UTM |

Figura 23. Variables utilizadas para los primeros procesados del modelo y comprender como podemos realizar el desarrollo de este dependiendo del efectivo que queremos predecir.

**Exploración de Datos:** Una vez transformados los datos, se exploraron para comprender su distribución, identificar posibles relaciones entre variables y resaltar las características más importantes antes de proceder con el modelado.

## 

## 

## 

## FASE: MODELIZACIÓN.

En esta etapa, se han estudiado las técnicas de modelado para aplicar los algoritmos más adecuados al problema, considerando la naturaleza de los datos y los objetivos del proyecto.

Una vez revisado el objetivo del conjunto de datos para alinearlo con el objetivo de negocio, se construyeron los modelos y se ajustaron sus parámetros. Posteriormente, se evaluaron para determinar su calidad y seleccionar el más adecuado para su uso, en base al estudio económico y los objetivos principales del proyecto, para posteriormente introducirlo en la plataforma tecnológica.

### Selección de Técnica de modelado

Dado que el propósito del proyecto es predecir si será necesario utilizar un determinado medio de extinción, se optó por modelos de clasificación, incluyendo:

* **Random Forest**
* **XGBoost**
* **LightGBM**
* **CatBoost**

### Diseño de evaluación

#### Selección y Reducción de Variables

* Se evaluaron correlaciones entre variables para identificar posibles redundancias.
* Se eliminaron variables altamente correlacionadas para evitar problemas de multicolinealidad.
* Se identificaron las características más importantes utilizando un modelo de **Random Forest**.
* Se redujo el conjunto de variables para optimizar el tiempo de cómputo sin afectar el rendimiento del modelo.

En cada uno de los modelos lo que se realiza es una visión general del modelo con todas las variables posibles que se pueden usar para ver cómo se pueden comportar dependiendo del modelo que usemos. Se evalúa la métrica de AUC

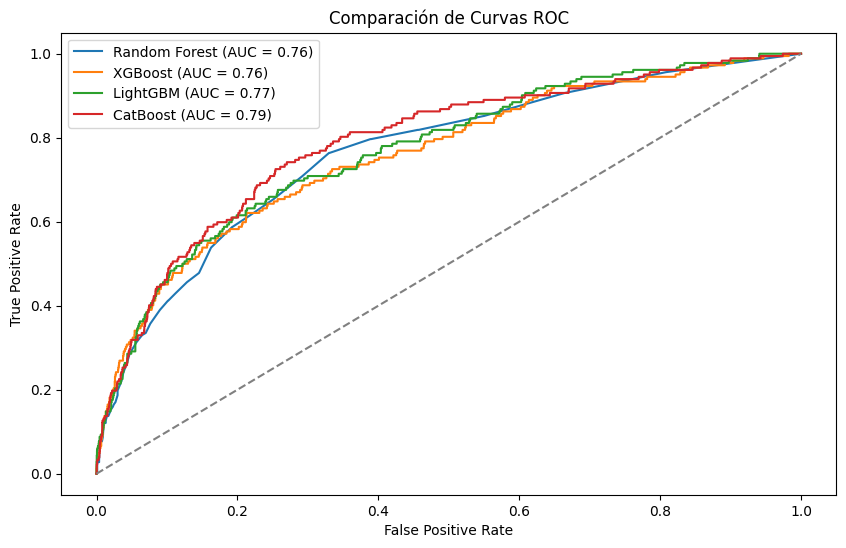


Figura 24. Gráfico de ejemplo de los posibles modelos que se pueden usar con todas las variables que luego vamos a ir ajustando, junto con los hiperparametros y los estudios económicos y de umbral

Después se comprueba con cada uno de los modelos la importancia que se le da a cada una de las variables y se retiran las últimas para que después el procesado del modelo sea mucho más rápido y se vuelve a comprobar el AUC para ver si ha habido alguna variación. Si se observa un empeoramiento se estudiaría qué variable ha podido provocar y se volvería a incluir en las variables.

### Construcción del modelo

#### División de Datos

Los datos fueron divididos en conjuntos de entrenamiento y prueba mediante la función train\_test\_split, aplicada al *Dataset* resultante del proceso de limpieza y codificación. Se hace un test\_size del 30% de la muestra.

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

#### Modelado

##### 3.1 Clasificación (Predecir si es necesario un medio de extinción)

Se probaron varios modelos de clasificación:

* *Random Forest Classifier*
* *LightGBM*
* *CatBoost*

Las métricas de evaluación utilizadas fueron:

* *AUC (Área bajo la curva ROC)*
* *Accuracy (Precisión global del modelo)*
* *Precission*
* *Recall*
* *F1-score*

Lo primero que se hace es un modelo lo más sencillo posible donde únicamente se ajusta (.fit) cada uno de los tres modelos que vamos a procesar para ver más o menos como funcionan con las métricas que hemos elegido.

Adicionalmente, se realizaron unos ajustes de hiperparámetros con **GridSearchCV** y **RandomizedSearchCV**, para encontrar los mejores parámetros de ajuste y así poder mejorar las métricas.

Se compararon las curvas **ROC** de los modelos para seleccionar el mejor, y además nos fijamos principalmente en el Recall y el F1-score, se comprueba tanto el macro como el weight y también se presta atención en las propias variables de ‘verdadero’ (1) o ‘Falso’ (0).

### Evaluación

Una vez se obtiene el mejor modelo luego se hace el *Estudio de umbral*, y se combina con el e*studio de coste económico*

##### 4.1 Estudio de umbral

El estudio del umbral se usa para poder comprobar si podemos mejorar las métricas del modelo creando un gráfico que use el mejor modelo creado, variando los diferentes umbrales en pasos de 0.1 para después poder seleccionar el mejor en combinación con el coste económico total que luego calcularemos.

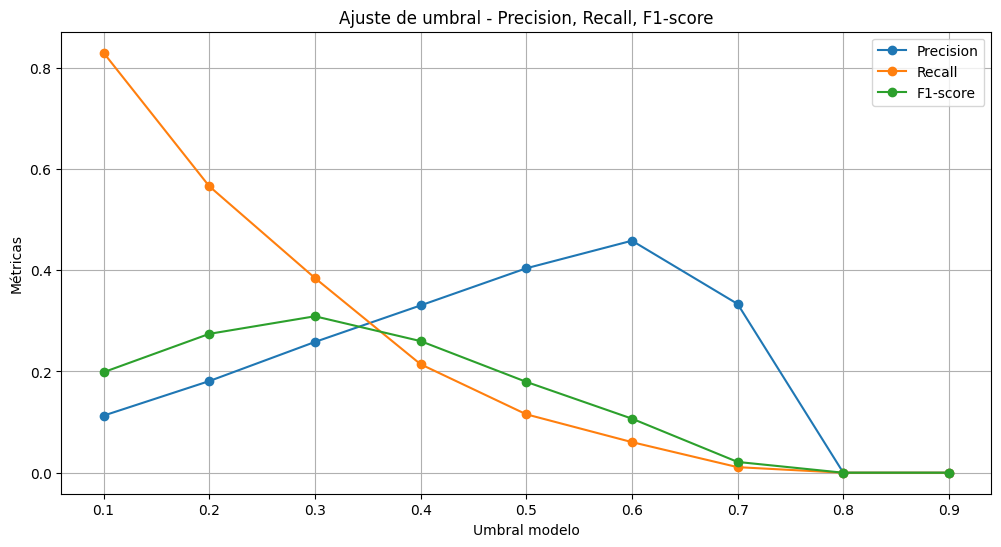


Figura 25. Ejemplo de grafico de estudio del umbral de un efectivo en específico un modelo en especifico

##### 4.2 Estudio de Coste económico

Para el cálculo del coste económico lo que vamos a tener en cuenta es el cálculo de los umbrales y de este cálculo sacaremos los *falsos positivos*, que son envíos de los efectivos sin ser necesarios y por tanto conlleva un coste por horas que está indicado en el apartado de introducción. En estos falsos positivos se usa el peor de los casos que es enviar la totalidad de las naves.

Y además se tiene en cuenta el falso negativo, que es el hecho de haber tenido que haber enviado un efectivo, pero no se realizó, por lo tanto se produjo una pérdida que puede englobar la pérdida de terreno, bienes individuales, muertes, y toda la repercusión social que puede conllevar unos costes económicos derivados. Para este cálculo se utiliza la mediana de Área afectada de todo el DataSet que tenemos y se le multiplica el coste calculado por Evenor-tech (<https://evenor-tech.com/el-coste-economico-de-los-incendios-forestales/> )

Se realizan los cálculos en relación con el umbral y los costes económicos calculados para un valor de falso positivo o falso negativo, y se saca una lista de costes asociados al umbral.

Umbral: 0.1, Coste total: 7684090.75 €

Umbral: 0.2, Coste total: 4300387.28 €

Umbral: 0.3, Coste total: 2882336.17 €

Umbral: 0.4, Coste total: 1929877.25 €

Umbral: 0.5, Coste total: 1322524.35 €

Umbral: 0.6, Coste total: 949077.57 €

Umbral: 0.7, Coste total: 569433.62 €

Umbral: 0.8, Coste total: 347496.80 €

Umbral: 0.9, Coste total: 141506.40 €

Después de la extracción de estos costes se vuelve a analizar en conjunto la gráfica de métricas dependiendo del umbral y los costes, **teniendo un mayor peso el conseguir una métrica buena**, independientemente de que el coste suba.

Se busca un equilibrio entre la obtención de una buena métrica sobre pensando que, **aunque los costes aumenten, no se elija el umbral menor que suponga un coste mayor** si no que busquemos también que los gastos totales a asumir, aunque sean altos, no sean excesivos

### Implementación y Uso Práctico

Se desarrolló una función (una para cada efectivo de mitigación de incendio) para predecir si es necesario o no enviar el efectivo al punto donde se ha declarado el incendio.

Para ello hemos previamente guardado el modelo y el umbral, que finalmente hemos considera óptimos tras el estudio económico y el estudio de umbral, en un formato .pkl para después a través de la función, que se llamara igual que la propia variable a predecir, podamos llamar al modelo y ajustarlo a unas nuevos valores de variables que introduciremos.

Hemos decidido que consideramos que hay que enviar un efectivo cuando la probabilidad de que el efectivo se envíe sea mayor del 0.80, (80%)

def BULDOZZER(data):

""" Función para predecir si se debe enviar un avión basado en las características dadas.

Args:

data (dict): dar un diccionario con las variables que hemos definido en el modelo """

# Asegurar orden igual que en el modelo porque si no en la lectura del .pkl dara error

features = ['Slope', 'TEMPMAX', 'V\_VIENTO', 'H\_RELAT', 'DifficultyIndex',

'Elevation', 'TOT\_EST', 'ENP\_1', 'EIMP\_GLOB', 'SHAPE\_AREA',

'BOMBEROS', 'GUARCIVIL']

try:

# ordeanar

input\_data = np.array([data[feature] for feature in features]).reshape(1, -1)

# Probabilidad segun el modelo que hemos creado y cargado luego como 'best\_model'

probability = best\_model.predict\_proba(input\_data)[0, 1]

# Metemos recomendación un umbral de probabilidad del 80%. Se pone esta probabilidad para se etractos ya que puede que en la creación del modelo hayamos metido algún falso positivo

recommendation = "Si, enviar" if probability >= 0.8 else "No, no enviar"

return {

"probability": probability,

"recommendation": recommendation

}

except KeyError as e:

return {"error": f"Falta la característica requerida: {e}"}

except Exception as e:

return {"error": f"Ocurrió un error: {e}"}

## 

## 

## FASE: EVALUACIÓN Y PRESENTACIÓN DE RESULTADOS.

La metodología que se ha descrito en el apartado anterior se ha repetido 5 veces, en cada uno de los efectivos que se han elegido para poder obtener un modelo de clasificación específico.

En cada uno de los efectivos se han desarrollado 3 tipos de modelos que van a ser RandomForestClassifier, lgb.LGBMClassifier, CatBoostClassifier y de cada uno de estos desarrollos se van a mostrar los resultados y además se va a elegir el modelo final que se utilizarían en la plataforma de despliegue tecnológico.

La estructura de los resultados que se muestran a continuación son primero los resultados de los tres modelos ajustados en hiper parámetros, donde se mide la matriz de confusión, la curva ROC y las métricas divididas.

Después, la siguiente tabla, muestra todos los análisis de umbral y económico, junto con la nueva matriz de confusión elegida para el umbral que hemos decidido por cada modelo.

### # AVIÓN ANFIBIO DE DESCARGA

En los primeros modelos que se intenta crear a partir del ajuste de los hiperparámetros muestran cómo la curva ROC es muy parecida en todos los modelos, dependiendo del ratio que queramos tener de verdaderos positivos, las dos mejores opciones estarían entre Random forest Classifier y el LGB

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
| Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. | Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. | Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. |
| Imagen de la pantalla de un celular con letras  El contenido generado por IA puede ser incorrecto. | Imagen de la pantalla de un celular con letras  El contenido generado por IA puede ser incorrecto. | Imagen de la pantalla de un celular con letras  El contenido generado por IA puede ser incorrecto. |

Figura 26. Resultados de Avión anfibio, para el modelo previo al estudio de umbral y económico con el ajuste de hiperparámetros obteniendo el mejor modelo. Se muestran los tres tipos de modelos RandomForest, Lgb y CatBoost

Con el ajuste del umbral lo que queremos es mejorar el recall sobre todo de los valores de ‘1’, para que esté un poco más equilibrado para que tengamos más verdaderos positivos. Para eso hemos modificado los umbrales elegidos han sido 0.38, 0.28 y 0.15 para cada uno de los modelos de izquierda a derecha.

Con estos resultados lo que podemos comprobar que el que da un equilibrio mayor entre coste económico y mejores resultados de Recall y F1-Score, es el **LGB, que se elige como el mejor resultado de modelo para los aviones de descarga anfibios.**

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
| Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. | Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. | Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. |
| Tabla  El contenido generado por IA puede ser incorrecto. | Imagen de la pantalla de un celular con letras  El contenido generado por IA puede ser incorrecto. | Imagen de la pantalla de un celular con letras  El contenido generado por IA puede ser incorrecto. |
| Texto  El contenido generado por IA puede ser incorrecto. | Texto  El contenido generado por IA puede ser incorrecto. | Texto  El contenido generado por IA puede ser incorrecto. |

Figura 27. Resultados del estudio económico y de umbral para tomar la decisión del mejor modelo para la plataforma tecnológica

Estos resultados son guardados en formato .pkl para poder usarlo en la plataforma tecnológica.

### # AVIÓN CARGA TERRESTRE

De igual manera que en el anterior se puede comprobar como las curvas ROC son muy parecidas en todos los modelos, con los valores de Acuracy mejor para el modelo de CatBoost, sin embargo los resultados de Recall y F1 Score son mejor en los otros dos modelos.

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |

Figura 28. Resultados de Avión de carga terrestre, para el modelo previo al estudio de umbral y económico con el ajuste de hiperparámetros obteniendo el mejor modelo. Se muestran los tres tipos de modelos RandomForest, Lgb y CatBoost

En esta ocasión los umbrales que hemos elegido han sido 0.38 , 0.43 y 0.33 para cada uno de los modelos de izquierda a derecha, donde volvemos a tener unas mejores resultados de Recall para el LGB, con mayor cantidad de ‘1’ acertados , aunque hay más falsos positivos que en resto de los modelos, además que el coste económico es el mayor de todos, por lo que se toma la decisión del **mejor resultado de modelo para los aviones de carga terrestre es el RandomForestClassifier.**

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |
|  |  |  |

Figura 29. Resultados del estudio económico y de umbral para tomar la decisión del mejor modelo para la plataforma tecnológica

Estos resultados son guardados en formato .pkl para poder usarlo en la plataforma tecnológica.

### # AUTOBOMBA

De igual manera que en el anterior se puede comprobar como las curvas ROC son muy parecidas en todos los modelos, con los valores de Acuracy mejor para el modelo de CatBoost, sin embargo los resultados de Recall y F1 Score son mejor en los otros dos modelos, ya que aunque sí que es capaz de obtener muy buenos resultados en en el hecho de ‘enviar un efectivo’ pero no en el hecho de no enviarlo, ya que obtiene muchos falsos positivos.

Los resultados entre Random y LGB son muy similares, por lo que a través de los estudios de umbral y económico se decidirá cual realmente de los tres es el más indicado para usar finalmente.

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
| Gráfico, Gráfico de líneas  El contenido generado por IA puede ser incorrecto. |  |  |
|  |  |  |

Figura 30. Resultados de Autobomba, para el modelo previo al estudio de umbral y económico con el ajuste de hiperparámetros obteniendo el mejor modelo. Se muestran los tres tipos de modelos RandomForest, Lgb y CatBoost

En esta ocasión los umbrales que hemos elegido han sido 0.45 , 0.45 y 0.80 para cada uno de los modelos de izquierda a derecha. CatBoost, que va a ser **el mejor resultado de modelo para los Autobombas,** aunque cualquiera de los otros dos también sería una opción ya que tanto a nivel económico como las métricas son muy parecidas en todos los casos estudiados**.**

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |
| Texto  El contenido generado por IA puede ser incorrecto. |  |  |

Figura 31. Resultados del estudio económico y de umbral para tomar la decisión del mejor modelo para la plataforma tecnológica

Estos resultados son guardados en formato .pkl para poder usarlo en la plataforma tecnológica.

### 

### # HELICÓPTERO DE TRANSPORTE

Lo mismo que hemos podido observar donde las curvas ROC son muy parecidas en todos los modelos que hemos comprobado, donde parece que de media todos tienen métricas muy parecidas y que parece que quizás los valores de Random Forest son un poco los más equilibrado, pero hacemos el estudio económico y de umbral para poder ajustar mucho mejor la decisión final

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |

Figura 32. Resultados de Helicoptero terrestre, para el modelo previo al estudio de umbral y económico con el ajuste de hiperparámetros obteniendo el mejor modelo. Se muestran los tres tipos de modelos RandomForest, Lgb y CatBoost

En esta ocasión los umbrales que hemos elegido han sido 0.38, 0.45 y 0.40 para cada uno de los modelos de izquierda a derecha. CatBoost se descarta por el sobrecoste que supone respecto al resto de modelos a nivel económico, siendo el menor coste el que tenemos con **LGB**, que va a ser **el mejor resultado de modelo para los Helicopteros,** con respecto al RandomForestClassifier, ya que las métricas que obtenemos tanto de Recall como f1-Score son mejores**.**

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |
| Texto  El contenido generado por IA puede ser incorrecto. |  |  |

Figura 33. Resultados del estudio económico y de umbral para tomar la decisión del mejor modelo para la plataforma tecnológica

Estos resultados son guardados en formato .pkl para poder usarlo en la plataforma tecnológica.

### # BULLDOZER

En el caso de los bulldozer no lo hemos tratado como balanced data a través del propio procesado interno del modelo, si no que hemos decidido hacer un tratamiento de datos externo a través de SMOTETomek donde intentamos igualar las dos categorías que están super desbalanceadas al haber muy poca cantidad de efectivos que se han registrado que se han enviado.

Tras intentar hacer un ajuste con los hiper parámetros los resultados que se obtienen son mucho peores que un simple ‘.fit’ con el modelo que hemos elegido en cada ocasión, por lo que los estudios económicos y de umbral de igual manera va a ser con los modelos más simples ya que son con los que obtenemos mejores resultados, donde de primeras LGB es el mejor modelo antes de poder hacer el estudio económico y de umbral.

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |

Figura 34. Resultados de Buldozzer, para el modelo previo al estudio de umbral y económico sin el ajuste de hiperparámetros. Se muestran los tres tipos de modelos RandomForest, Lgb y CatBoost

En esta ocasión los umbrales que hemos elegido han sido 0.4 , 0.1 y 0.20 para cada uno de los modelos de izquierda a derecha. CatBoost se descarta por el sobrecoste, siendo el menor coste el que tenemos con **LGB**, que va a ser **el mejor resultado de modelo para los Buldozzers,** con respecto al RandomForestClassifier, ya que las métricas que obtenemos f1-Score son mejoresconsiguiendo que los falsos negativos sean menores.

|  |  |  |
| --- | --- | --- |
| **RandomForestClassifier** | **lgb.LGBMClassifier** | **CatBoostClassifier** |
|  |  |  |
|  |  |  |
|  |  |  |

Figura 35. Resultados del estudio económico y de umbral para tomar la decisión del mejor modelo para la plataforma tecnológica

Estos resultados son guardados en formato .pkl para poder usarlo en la plataforma tecnológica.

## 

## FASE: DESPLIEGUE TECNOLÓGICO.

El despliegue tecnológico del modelo predictivo tiene como objetivo trasladar el modelo entrenado a un entorno de producción en el que se pueda utilizar eficazmente para optimizar la gestión de los incendios forestales. A continuación, se detallan los aspectos clave de la planificación del despliegue, el control, seguimiento y mantenimiento, y la implementación tecnológica de la solución.

### 1. Planificación del Despliegue

La estrategia de implementación en esta primera fase del proyecto consiste en desplegar la aplicación web como una herramienta adicional para la toma de decisiones en la gestión de incendios forestales. Al ser una primera fase del proyecto, el modelo se presenta como una herramienta adicional que proporciona recomendaciones sobre los medios más adecuados, basándose en las condiciones meteorológicas, las coordenadas geográficas y otros parámetros relevantes.

1. **Preparación del Modelo:**  El modelo predictivo ha sido desarrollado en Python, utilizando bibliotecas clave como scikit-learn, pandas y numpy para el modelado predictivo y la manipulación de datos. El modelo utiliza un algoritmo de clasificación que predice la necesidad de un medio aéreo y/o terrestre, dependiendo de factores como la temperatura, la humedad, la velocidad del viento y el área quemada. Este modelo se entrenó con datos históricos de incendios y se evaluó utilizando un conjunto de datos de prueba. Tras el entrenamiento, el modelo fue exportado como un archivo .pkl usando joblib para su integración posterior en la aplicación web.

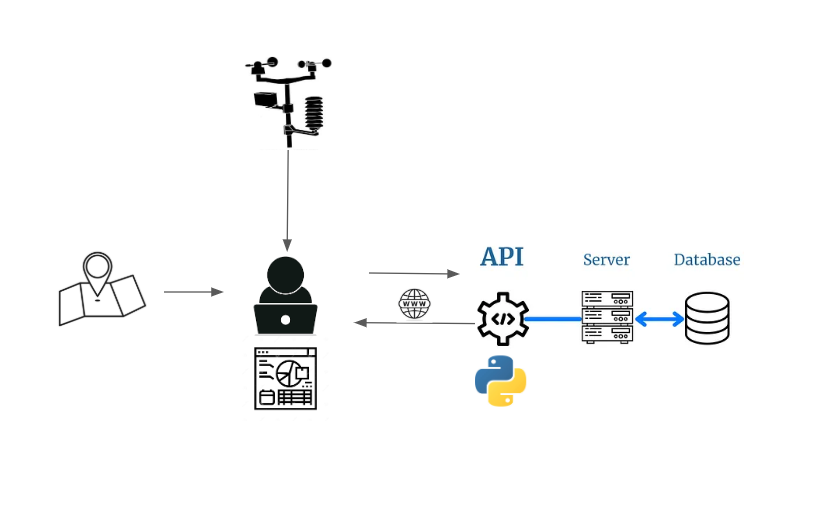


Figura 36. Esquema de la planificación del modelo

1. **Despliegue de la Aplicación Web**: La aplicación se desplegará en un entorno web accesible a los usuarios a través de **Streamlit**, proporcionando una interfaz intuitiva para la introducción manual de datos meteorológicos y coordenadas de incendios. El modelo de predicción, previamente entrenado, ha sido guardado en un archivo **.pkl** (Pickle), lo que permite su carga rápida y eficiente sin necesidad de volver a entrenarlo en cada ejecución.

Si llegara a ser una solución de gran fiabilidad y uso por diferentes actores de la gestión se generalizara habría que plantearse utilizar un backend para gestionar las peticiones de manera eficiente permitiendo la escalabilidad e integración con otros sistemas. Además se debería implementar una API REST garantizando tiempos de respuesta adecuados.

1. **Integración de Datos Meteorológicos**: En esta primera fase los datos meteorológicos ingresan manualmente en la aplicación que serán los facilitados por los servicios meteorológicos oficiales en el momento de la predicción.
2. **Integración de Datos de Nuevos incendios:** Estos datos provienen de la Consejería de Medio Ambiente de la Comunidad Valenciana y serán integrados al final de cada campaña de incendios, lo que alimentará el modelo y permitirá mejorar sus predicciones.
3. **Funcionalidad de Predicción**: La aplicación web permitirá, en función de los datos introducidos, calcular qué medios aéreos y terrestres serían necesarios para la extinción del incendio, basándose en las predicciones del modelo entrenado. Además, mostrará información sobre los incendios más cercanos y los medios que fueron utilizados en estos, proporcionando así un contexto adicional.
4. **Comunicación con la Base de Datos**: La base de datos remitirá la información disponible sobre el incendio y solicitará el cálculo del modelo a la aplicación web. La respuesta de la predicción será enviada de vuelta y presentada al usuario de manera clara y accesible.

La aplicación de Streamlit es accesible públicamente a través del siguiente enlace:   
  
<https://firecast.streamlit.app/>

### 

### 2. Planificación del Control, Seguimiento y Mantenimiento

La estrategia de seguimiento y mantenimiento del modelo consistirá principalmente en comparar las predicciones realizadas durante la campaña con los resultados reales al final de la temporada de incendios. Este proceso se llevará a cabo manualmente en esta fase inicial del proyecto.

1. **Registro de Predicciones**: El sistema registrará las predicciones realizadas por el modelo para cada incendio detectado durante la campaña.
2. **Comparación con Resultados Reales**: Al final de la campaña, se compararán las predicciones con los medios reales que fueron enviados a los incendios, con el objetivo de evaluar si las predicciones del modelo permitieron una gestión más eficiente, reduciendo el número de medios utilizados y optimizando los recursos.
3. **Evaluación de Ahorro Económico**: El análisis de los resultados tendrá como objetivo identificar posibles ahorros económicos. El objetivo planteado es alcanzar un ahorro del 15% en los costes operativos asociados a la extinción de incendios mediante la optimización de los recursos, basado en las predicciones del modelo.
4. **Mantenimiento del Modelo**: En esta fase inicial, el mantenimiento del modelo se centrará en recopilar los resultados de la comparación entre predicciones y resultados reales. Con base en estos análisis, se podrá ajustar y reentrenar el modelo en fases posteriores.

### 3. Implementación Tecnológica

La solución que se implementará en esta fase es una aplicación web sencilla, independiente del sistema operativo de los equipos de gestión de incendios, que proporcionará predicciones sobre los recursos necesarios para extinguir los incendios. A través de esta aplicación, los usuarios podrán ingresar los datos meteorológicos y las coordenadas geográficas, y la plataforma les mostrará los medios más adecuados a desplegar, junto con un histórico de los recursos utilizados en incendios cercanos.

* **Aplicación Web:** La aplicación es de fácil acceso a través de un navegador web y está diseñada para ser intuitiva. Los usuarios ingresarán los datos meteorológicos y de localización, y el sistema proporcionará una predicción sobre los medios que se deberían utilizar.
* **Visualización de Resultados Históricos:** Además de las predicciones, la aplicación mostrará los incendios recientes más cercanos y los recursos que se han utilizado en ellos, ayudando a contextualizar las recomendaciones del modelo.
* **Futuras Fases de Implementación:** Aunque en esta fase no se integrará con los sistemas operativos existentes de los equipos de gestión de incendios, se prevé que, en fases posteriores, la aplicación pueda integrarse con estos sistemas para automatizar el proceso y mejorar la respuesta ante los incendios.

## FASE: DESPLIEGUE DE NEGOCIO.

En este caso, la fase de puesta en valor se enfoca en garantizar que el modelo predictivo, desarrollado para la predicción de medios aéreos y terrestres necesarios para la extinción de incendios forestales, sea utilizado como una herramienta complementaria en el proceso de gestión de incendios.

### 1. Planificación del Despliegue de Negocio

La aplicación web será utilizada como una herramienta adicional para determinar qué medios aéreos y terrestres deben enviarse como primera respuesta cuando se detecta un incendio. Esta solución no se integrará directamente en los sistemas operativos de gestión de incendios en esta fase inicial.

El responsable del proyecto se encargará de interactuar con los stakeholders y definir cómo se utilizará la solución. La aplicación web proporcionará recomendaciones basadas en las predicciones del modelo y ayudará a tomar decisiones más informadas sobre los recursos a enviar.

### 2. Planificación del Control y Seguimiento

El seguimiento del modelo se llevará a cabo de manera manual en una primera fase. Al final de la temporada de incendios, se compararán las predicciones realizadas por el modelo con los medios aéreos y terrestres realmente utilizados en cada incendio. Esta comparación permitirá evaluar la efectividad del modelo y determinar si las predicciones contribuyen a una mejor toma de decisiones y optimización de recursos.

En cuanto a la medición del impacto, se establecerá una métrica al final de la campaña de incendios para analizar si el uso de la solución ha permitido reducir el número de medios enviados a los incendios, lo que podría generar un ahorro económico en las operaciones de extinción.

### 3. Implementación de la Solución

La solución implementada en esta fase es una aplicación web independiente, que proporciona a los usuarios las predicciones del modelo sobre los medios necesarios para la extinción de incendios. Esta herramienta se utilizará para asistir en la toma de decisiones, pero no reemplazará los sistemas existentes de gestión de incendios en esta fase inicial.

# 

# CONCLUSIONES Y LÍNEAS FUTURAS.

## Conclusiones

En el presente trabajo se ha desarrollado un modelo predictivo basado en Machine Learning para la optimización de la gestión de recursos de extinción de incendios forestales en la Comunidad Valenciana. A través del análisis de datos históricos de incendios, condiciones meteorológicas y variables geoespaciales, se ha logrado implementar una herramienta capaz de predecir la necesidad de medios terrestres y aéreos con un enfoque basado en clasificación y regresión.

Entre los principales hallazgos del estudio, destacan:

* Eficiencia del Modelo: Los modelos entrenados (Random Forest, y LightGBM) han demostrado ser efectivos en la predicción del uso de recursos, con resultados satisfactorios en métricas como AUC, F1-score y recall.
* Optimización de Costes: La aplicación del modelo permitiría reducir los falsos positivos en la asignación de recursos, evitando desplazamientos innecesarios de medios costosos, especialmente aeronaves.
* Interpretabilidad y Usabilidad: Se ha desarrollado una herramienta web para mejorar la interpretabilidad y usabilidad del modelo y garantizar que los resultados sean comprensibles para los responsables de la toma de decisiones.

Este trabajo constituye un avance significativo en la aplicación de la ciencia de datos para la gestión de incendios forestales, pero aún quedan aspectos por mejorar y explorar.

## Líneas Futuras

Para continuar con la evolución del proyecto y maximizar su impacto, se plantean las siguientes líneas de investigación y desarrollo:

1. **Integración con Datos en Tiempo Real:**
   * Conexión directa con fuentes meteorológicas (AEMET, GEE) para mejorar la predictibilidad del modelo.
   * Incorporación de datos satelitales, sensores, cámaras en tiempo real.
2. **Mejora del Modelo y Evaluación Avanzada:**  
   * Ampliación del modelo para estimar con mayor precisión la cantidad de efectivos necesarios en un incendio forestal, garantizando una asignación eficiente de los recursos.
   * Realización de mejora en los ajustes de regresión lineal que no se muestran en el documento porque los resultados obtenidos no son todo lo conformes que quisiéramos y se necesita de mayor tiempo para poder hacer una mejor implementación de la regresiones.
3. **Desarrollo de una Plataforma Interactiva:**
   * Implementación de una API optimizada para la captura en tiempo real de los parámetros clave en la gestión de incendios forestales, mejorando la interoperabilidad con otros sistemas.
   * Desarrollo de una interfaz web más intuitiva y funcional utilizando Streamlit u otras tecnologías avanzadas para facilitar su adopción operativa.
4. **Extensión a Otras Regiones y Validación Externa:**  
   * Aplicación del modelo en otras comunidades con alta incidencia de incendios forestales.
   * Validación con datos de nuevas campañas de incendios para ajustar la generalidad del modelo.
5. **Evaluación de Impacto Económico y Medioambiental**:  
   * Implementación de métricas de ahorro económico y optimización de recursos.
   * Análisis del impacto del modelo en la reducción de la superficie quemada y la respuesta operativa.

En conclusión, este trabajo ha sentado las bases para una gestión más eficiente de los recursos contra incendios forestales mediante la aplicación de inteligencia artificial. La evolución del modelo y su integración en procesos operativos podría representar una innovación clave en la lucha contra incendios, mejorando la capacidad de respuesta y reduciendo los impactos ambientales y económicos.

# REFERENCIAS BIBLIOGRÁFICAS.

1. Vitolo, C., Di Napoli, C., Di Giuseppe, F., Cloke, H. L., & Pappenberger, F. (2019). Mapping combined wildfire and heat stress hazards to improve evidence-based decision making. *Environment International*, *127*, 21–34. https://doi.org/10.1016/j.envint.2019.03.008

2. Ubeda Cartañà, X., Mataix-Solera, J., Francos, M., & Farguell, J. (2021). *Grandes incendios forestales en España y alteraciones de su régimen en las últimas décadas* (pp. 147–161). https://doi.org/10.34037/978-989-9053-06-9\_1.2\_11

3. Ahmad, S., Baklouti, M., Umirzakova, S., & Jamil, F. (2023). AI in Wild Fire Management. *Fire Ecology*.

4. Bloise, L. (2017). Surveillance, hygiène et entretien des lentilles de contact. *Journal Francais d’Ophtalmologie*, *40*(4), 329–337. https://doi.org/10.1016/j.jfo.2017.01.004

5. Pourghasemi, H. R., Gayen, A., Lasaponara, R., & Tiefenbacher, J. P. (2020). Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling. *Environmental Research*, *184*. https://doi.org/10.1016/j.envres.2020.109321

6. Sánchez, G. (2019). *Caracterización, evolucion e incidencia de los grandes incendios forestales en el noreste de España*. Universidad de Cantabria.

7. García-Heras Hernández, F., Arroyo, J. G., & Molinero González, O. (n.d.). *Ansiedad, estrés, y estados de ánimo del Personal Especialista en Extinción de Incendios Forestales Anxiety, stress, and mood states of wildland firefighters* (Vol. 41). https://recyt.fecyt.es/index.php/retos/index

8. Qin, H., Vickery, J., Brenkert-Smith, H., Bekee, B., & Prasetyo, Y. (n.d.). *Do Actions Reduce Perceived Risk? A Longitudinal Analysis of the Relationship between Risk Perception and Actions in Response to Forest Disturbance in Colorado*.

9. Vicente Oliver, J. (2018). *III congreso forestal de la Comunitat Valenciana* (III). Plataforma forestal valenciana.

# Glosario de términos

## Términos del ámbito forestal

* **Incendio forestal**: Fuego que se propaga sin control sobre áreas cubiertas de vegetación forestal.
* **Combustible forestal**: Material vegetal (hojas, ramas, troncos) susceptible de arder en un incendio.
* **Índice de peligro meteorológico**: Indicador basado en variables como temperatura, humedad y viento, que mide el riesgo de incendio en una zona.
* **Brigada helitransportada**: Equipo de extinción que se transporta en helicópteros para intervenir en incendios forestales de difícil acceso.
* **Autobomba forestal**: Vehículo equipado con depósito de agua y mangueras para la extinción de incendios en terrenos forestales.
* **Bulldozer**: Máquina pesada utilizada para remover vegetación y crear líneas de defensa contra incendios.
* **Avión anfibio**: Aeronave capaz de recoger agua de cuerpos de agua y descargarla sobre incendios.
* **Avión de carga en tierra:** Aeronave para transportar y descargar agua o retardantes sobre incendios forestales, obteniendo el recurso desde una base terrestre.
* **Retardante de fuego**: Sustancia química aplicada en la vegetación para reducir la velocidad de propagación de un incendio.
* **Fire Risk Index (Índice de Riesgo de Incendio)**: Indicador que mide la peligrosidad de propagación de un incendio en función de la temperatura, humedad y velocidad del viento.
* **Fuel Load Index**: Estimación de la cantidad de material combustible presente en una determinada área.
* **Difficulty Index**: Indicador que evalúa la dificultad de acceso al área de un incendio, considerando la topografía y la densidad de vegetación.
* **Land Cover**: Clasificación de la cobertura del suelo, como bosques, pastizales, cuerpos de agua, entre otros.

## Términos de Ciencia de Datos

* **Machine Learning**: Rama de la inteligencia artificial que desarrolla algoritmos capaces de aprender patrones a partir de datos.
* **Modelo de clasificación**: Algoritmo que asigna categorías a datos nuevos en función de patrones previamente aprendidos (ejemplo: determinar si se requieren medios aéreos o terrestres en un incendio).
* **Modelo de regresión**: Algoritmo que predice valores numéricos (ejemplo: estimar la cantidad exacta de autobombas necesarias en un incendio).
* **Data leakage**: Uso indebido de datos que contienen información futura en la fase de entrenamiento de un modelo, lo que distorsiona sus predicciones.
* **Validación cruzada**: Técnica utilizada para evaluar el rendimiento de un modelo dividiendo los datos en múltiples conjuntos de entrenamiento y prueba.
* **Métricas de evaluación**: Medidas utilizadas para evaluar el desempeño de un modelo predictivo, como precisión, recall, F1-score (para clasificación) y MAE o RMSE (para regresión).
* **Feature Engineering**: Proceso de creación de nuevas variables a partir de las existentes para mejorar el desempeño de un modelo.
* **Grid Search / Randomized Search**: Métodos para encontrar la mejor combinación de hiperparámetros en un modelo de aprendizaje automático.
* **ROC-AUC**: Métrica que evalúa la capacidad de un modelo de clasificación para distinguir entre clases positivas y negativas.
* **Overfitting (Sobreajuste)**: Situación en la que un modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización en datos nuevos.
* **Streamlit**: Framework en Python para la creación rápida de aplicaciones web interactivas utilizadas en ciencia de datos.
* **Google Earth Engine (GEE)**: Plataforma de análisis geoespacial que permite extraer y analizar datos satelitales para modelado ambiental.
* **QGIS**: Software de código abierto para el análisis y visualización de datos geoespaciales.
* **API**: Interfaz de programación que permite la comunicación entre diferentes sistemas y el acceso a datos desde fuentes externas.
* **Pickle (.pkl)**: Formato de archivo en Python utilizado para almacenar modelos de machine learning entrenados y reutilizarlos sin necesidad de reentrenamiento.