Índice general

Capítulo 1

Conclusiones, aportaciones y trabajo futuro

En este último capítulo se va a realizar una recapitulación de las conclusiones extraídas en cada una de las secciones del presente trabajo de fin de estudios. En primer lugar, se presentarán de forma resumida las conclusiones obtenidas a lo largo del trabajo. A continuación, se detallarán las aportaciones realizadas en el campo de la predicción de incendios forestales. Por último, se indicarán algunas lineas de investigación, dentro del campo de la predicción de incendios forestales y la inteligencia artificial, que permitirían profundizar en el desarrollo de la metodología presentada de cara a obtener mejores modelos con utilidad práctica.

1.1. Conclusiones

El control y extinción de los incendios forestales requiere del desplazamiento de un gran número de efectivos, y el éxito en la operación depende en muchos casos de la velocidad en la respuesta. En el presente trabajo se ha desarrollado una metodología completa para dotar de herramientas que permitan predecir eficazmente las zonas en riesgo de verse afectadas por un incendio forestal en la Comunidad Autónoma de Andalucía, facilitando así la toma de decisiones y permitiendo una asignación de los recursos más eficiente.

A lo largo de esta memoria se ha desarrollado una metodología completa para la construcción de modelos de predicción de incendios forestales en la Comunidad Autónoma de Andalucía mediante el uso de técnicas de *Machine Learning* y procesamiento de datos geoespaciales, adoptando un enfoque dinámico y global. Dinámico, pues se predice el el riesgo de incendio forestal para una localización específica en un día concreto. Global, pues se han considerado 27 variables que abarcan las 5 dimensiones principales : antropológica, demográfica, meteorológica, topográfica y de vegetación.

Se ha comenzado introduciendo el problema, justificando su relevancia y estableciendo 3 tareas claras para alcanzar el objetivo del trabajo. La primera de las tareas implicaba la construcción de un conjunto de datos adecuado para el análisis estadístico y la construcción de modelos de ML. Esta tarea se abordó en el capítulo 2, donde no solo se ha generado un conjunto de 20.000 muestras sobre el que se ha desarrollado el trabajo, si no que se ha implementado un algoritmo para tomar muestras aleatorias de casos positivos y negativos

dentro del marco del estudio y asociar a cada observación los valores correspondientes de todas las variables consideradas. Se ha explorado el uso de estratificación por mes en la selección de la muestra, con el objetivo de entrenar modelos que sean capaces de detectar relaciones más profundas en los datos relacionadas con la aparición de los incendios forestales. Se ha visto que, por ejemplo, en el caso del *Random Forest* esta estratificación ha conducido a que la variable *uso_suelo* haya tomado demasiada importancia en la clasificación, conduciendo a modelos que no parecen comportarse adecuadamente. ||| Queda, por tanto, abierta la pregunta de cuál sería la mejor manera de tomar la muestra de casos negativos de forma que se encuentre un equilibrio entre la sensibilidad del modelo a las variables meteorológicas sin perder la influencia de las demás variables. |||

A continuación, se ha analizado en profundidad el conjunto de datos generado, recurriendo principalmente a representaciones gráficas, aunque también se han usado métodos numéricos. La complejidad de esta fase radica en que para llegar a obtener información relevante es necesario tener en cuenta las dimensiones espacial y temporal de los datos. Este proceso ha permitido llegar a un mayor conocimiento acerca del conjunto de datos, caracterizado por una gran presencia de valores *outlier*, por distribuciones asimétricas hacia la derecha en casi todas las variables numéricas y correlaciones generalmente pequeñas en valor absoluto. A nivel gráfico, se ha observado que las variables que muestran mayores diferencias entre ambas clases son *PRECTOTCORR* (el total diario de precipitaciones), *WS10M* (la velocidad del viento a 10 m) y *uso_suelo* (la clasificación de uso de suelo). Un buen reflejo de la complejidad del conjunto de datos es que para explicar el 80 de la varianza de las 18 variables numéricas en la muestra se necesitan 11 componentes principales.

La siguiente tarea planteada se ha desarrollado en el capítulo 5, donde se han construido distintos modelos de ML de clasificación binaria. Se ha usado el flujo de trabajo propuesto por el paquete de R *tidymodels* para preprocesar los datos, entrenar los modelos, ajustar los valores de los hiperparámetros y evaluar sus rendimientos en los datos test. Los modelos considerados han sido: regresión logística con penalización, regresión logística con penalización usando PCA, k-*Nearest Neighbours*, SVM lineal, SVM radial, árbol de decisión y *Random Forest*. Se ha usado una partición temporal en entrenamiento-validación-test para entrenar los modelos, ajustar los parámetros y evaluar su capacidad de generalización sobre nuevos datos. Los mejores resultados sobre el conjunto de datos generado con estratificación por mes, los han dado los modelos de regresión logística con penalización, SVM lineal y SVM radial, los cuales han mostrado un comportamiento muy similar tanto sobre el conjunto de validación como sobre el conjunto test. Los valores más elevados en las métricas de de rendimiento sobre los datos test los ha alcanzado el modelo de regresión logística lasso con un coeficiente de penalización de 0.000464, que ha alcanzado un ROC-AUC de 0.795, una tasa de acierto de 0.710, 0.726, una especificidad de 0.693 y una presición de 0.718.

Finalmente, se ha evaluado el desempeño de los modelos en dos casos prácticos, cumpliendo así con la tercera tarea planteada. Primero, se han usado los modelos de regresión logistica con penalización, SVM lineal y RF para predecir la probabilidad estimada de incendio el día 15 de cada mes de 2022 en toda Andalucía. Y por último, se han usado el modelo de regresión logística lasso y el SVM lineal para para predecir el incendio provocado de Sierra Bermeja, en Málaga. Aunque la baja resolución de las variables meteorológicas impide llegar a un mayor nivel de detalle, ambos modelos están indicando una situación de riesgo elevado de incendio forestal en la zona el día 8 de septiembre de

2023, al observarse un incremento significativo de las probabilidades de incendio estimadas en todo el área en estudio, como consecuencia de unas condiciones meteorológicas especialmente favorables para un incendio forestal.

Es necesario mencionar también las limitaciones de los modelos construidos, aunque algunas de ellas ya han sido comentadas a lo largo del trabajo. Por un lado, hay que tener ciertas consideraciones sobre los datos considerados para el trabajo. Aunque las variables consideradas son fruto de un amplio estudio previo y cobren las 5 dimensiones consideradas, son solo una primera aproximación aproximación al problema y probablemente en futuros estudios se necesite considerar un número mucho más elevado de variables, dada la complicada naturaleza del problema. Las información de los incendios de la que se ha podido disponer solo cubre los incendios que calcinaron una extensión superior a 100 ha, no siendo así exhaustiva, por lo que las conclusiones que puedan extraerse de los modelos construidos son limitadas. Además, como ya se ha indicado, la calidad de la información meteorológica disponible es limitada al proceder de modelos construidos a partir de observaciones satelitales.

Por otro lado, desde un punto de vista estadístico, es importante mencionar que la hipótesis de independencia entre las observaciones, sobre la que se sustentan los modelos construidos es violada fuertemente por los datos, al estar correlacionados espacial y temporalmente. Sin embargo, esto no impide su utilización, ya que de hecho han alcanzado resultados bastante satisfactorios. Otro comentario en esta línea es que todos los modelos construidos realizan la predicción de forma local, sin tener en cuenta la estructura de correlaciones presente en los datos, lo que inevitablemente conlleva un pérdida de información importante. Por ello, explorar el uso de modelos más complejos que sí consideren las correlaciones temporales y espaciales existentes entre las observaciones podría llevar a alcanzar mejores resultados.

Cabe también mencionar que las estimaciones de error con las que se han evaluado los modelos son medidas globales obtenidas en una muestra concreta, por lo que deben ser tomadas con precaución, tanto en este estudio, como en todos los que se aborden problemas similares. Por ello se han realizado análisis prácticos del desempeño de los modelos, para evaluar su rendimiento en la realidad. Por último, se debe considerar también que al ser un problema tan complejo en el que influyen tantas dimensiones diferentes, la significación que pueda tener un solo valor es limitada, por muy bueno que sea el modelo. Así, las conclusiones de los modelos deberán ser siempre tomadas con precaución y bajo el asesoramiento de un experto en el incendios forestales.

Pese a todo ello, los resultados obtenidos en los modelos construidos a lo largo de este trabajo son prometedores, ilustrando el potencial que podría llegar a tener la aplicación de la Inteligencia Artificial en la predicción de incendios forestales. Sin embargo, se trata de una investigación introductoria limitada por los recursos disponibles. Aumentar el número de variables consideradas, obtener fuentes de información de mayor calidad, considerar modelos más complejos, estudiar la forma óptima de generar la muestra de casos negativos para entrenar los modelos son algunas de las tareas que deberán llevarse a cabo en futuras investigaciones para permitir que estas tecnologías tengan un impacto en la lucha contra el fuego, permitiendo una mejor gestión de los recursos y llegando a salvar vidas.

1.2. Aportaciones

A lo largo de la presente memoria se ha presentado una metodología completa para construir modelos de predicción de incendios forestales desde un enfoque dinámico y global. En este trabajo se ha aplicado al caso de la Comunidad Autónoma de Andalucía considerando un conjunto de 27 variables explicativas, aunque su extensión a otras regiones y a conjuntos de variables más amplios es relativamente sencillo. Las principales aportaciones del proyecto han sido:

- La recopilación de una gran cantidad de información relevante para el estudio de los incendios forestales a partir de fuentes oficiales (Tabla **??**).
- La implementación de métodos y funciones para procesar los conjuntos de datos espaciales recopilados y generar muestras útiles para el análisis estadístico y la construcción de modelos de clasificación binaria, asociando a cada observación los valores correspondientes de todas las variables explicativas consideras. Estas funciones pueden ser útiles también fuera del ámbito de estudio de los incendios forestales, ya que permiten conocer los valores de las 27 variables consideradas correspondientes a un día y una localización dentro de los límites del estudio.
- El uso de estratificación en el proceso de generación de la muestra, con el objetivo de evitar modelos superficiales y estimaciones del error sesgadas positivamente. Aunque los resultados obtenidos, no son del todo satisfactorios y requieren de un mayor estudio, de entre todos los trabajos similares consultados, este es el primero en el que se cuestiona la forma de tomar las muestras negativas.
- La generación de una gran cantidad de mapas y gráficos que permiten estudiar en profundidad la distribución espacio-temporal de las variables incluidas en el estudio.
- El entrenamiento, ajuste y comparación del rendimiento de distintos algoritmos de clasificación binaria dentro del ML.

1.3. Trabajo futuro

En el presente trabajo se han llegado a resultados prometedores en cuanto al potencial de las herramientas en él presentadas. Sin embargo, es incuestionable que las limitaciones en tiempo y recursos han obligado a adoptar un enfoque más global, tomando ciertas simplificaciones y dejando algunos caminos sin explorar o sin estudiar en toda la profundidad que requieren. Es por ello que este estudio no está completo. Para poder llegar a construir modelos que verdaderamente sean útiles en el campo de la predicción de incendios forestales es necesario ahondar en esta investigación y dedicarle una mayor cantidad de recursos.

A lo largo del trabajo se han justificado todas las decisiones metodológicas tomadas, indicando en muchos casos alternativas mejores a las empleadas pero inviables dadas las limitaciones de un trabajo de fin de estudios. A continuación se recopilan estas propuestas, añadiendo algunas otras, con el objetivo de mostrar lineas posibles de investigación para extender la metodología presentada, llegando así a construir modelos que puedan tener un impacto significativo en la lucha contra el fuego:

- Aumentar el número de variables consideradas, tomando como referencia trabajos como Martínez-Fernández et al. [2] y Vilar del Hoyo et al. [3], en los que se estudian las variables más influyentes en la predicción de incendios causados por el hombre.
- Explorar el uso de la EGIF para la obtención de la información relativa a los incendios forestales. Esto es de vital importancia de cara a extender el estudio, ya que en el estudio presente solo se han considerado los incendios mayores de 100 ha, puesto que son los únicos disponibles en la REDIAM. Además, esto permitiría añadir otras dimensiones al problema, como la predicción de la superficie afectada por los incendios forestales o de la propagación de los incendios a partir de los puntos de origen del fuego.
- Buscar fuentes de información meteorológica viables y de mayor calidad, a ser posible proveniente de estaciones meteorológicas y no de modelos basado en observaciones . En esta dirección podría ser interesante explorar en mayor profundidad el uso de la API de la AEMET.
- Sería necesario revisar el procedimiento de generación de la muestra de de casos negativos, ajustando convenientemente los parámetros considerados. Como ha quedado reflejado en el trabajo, la composición de la muestra de casos negativos usada para entrenar los modelos tiene un impacto directo en el funcionamiento de estos. De la misma forma aquí se ha considerado un muestreo con estratificación por mes y un muestreo completamente aleatorio en las fechas, sería interesante estudiar los efectos que pueden tener otras formas de seleccionar la muestra de casos negativos. Esto será especialmente relevante si los tamaños muestrales son reducidos.
- El uso de modelos de *Deep Learning* como redes convolucionales podría traer mejoras significativas en los modelos, al considerar la estructura de correlaciones espaciales presentes en los datos [1].

Bibliografía

[1] Liz-López, Helena; Huertas-Tato, Javier; Pérez-Aracil, Jorge; Casanova-Mateo, Carlos; Sanz-Justo, Julia y Camacho, David (2024). «Spain on fire: A novel wildfire risk assessment model based on image satellite processing and atmospheric information». *Knowledge-Based Systems*, **283**, p. 111198. ISSN 0950-7051. doi: 10.1016/j.knosys.2023.111198. https://www.sciencedirect.com/science/article/pii/S0950705123009486.

[2] Martínez-Fernández, Jesús; Vega-García, Cristina y Chuvieco, Emilio (2009). «Human-caused wildfire risk rating for prevention planning in Spain». doi: 10.1016/j.jenvman.2008.07.005.

[3] Vilar del Hoyo, L.; Isabel, Martín; M.P. y Martínez Vega, F.J. (2011). «Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data». *European Journal of Forest Research*, **130**, pp. 983–996. doi: 10.1007/s10342-011-0488-2.