Universidad de Alcalá

Artículo

Predicción de incendios forestales en Cataluña mediante el uso de Redes Neuronales Artificiales LSTM y el Modelo de Machine Learning XGBoost.

Nathalia Fernández Rodrigues*

*Correspondencia: nathaliafernandezr95@gmail.com

Resumen: El presente estudio está basado en el análisis de diversos factores meteorológicos, tipográficos, clasificaciones de niveles de riesgo y la movilidad urbana, todos estos relacionados con la ocurrencia de incendios forestales en Cataluña, España, los cuales en su mayoría son de causa humana. El principal objetivo es predecir los incendios forestales y para ello se desarrollan modelos de aprendizaje supervisado, el primero de ellos es de redes neuronales recurrentes LSTM debido a que las variables son dependientes del tiempo, por otro lado, se crea un modelo de machine learning Xgboost para comparar los resultados obtenidos en ambos. Finalmente, los modelos alcanzaron una precisión en sus predicciones de (97% el Xgboost y 66% el LSTM).

Palabras Clave: red neuronal, predicción; aprendizaje automático; incendios forestales, movilidad urbana, LSTM. Xgboost.

1. Introducción

Entre los efectos ecológicos que más repercuten sobre la diversidad biológica se encuentran los incendios forestales, estos son una fuente de emisión de carbono que contribuye al calentamiento global, alteran el ciclo hidrológico, modifican la biodiversidad, y el humo producido afecta la salud de humanos y animales.

Estos incendios forestales incrementan la probabilidad de que se produzcan nuevamente en los años subsiguientes, ya que al caer arboles la luz del sol reseca el bosque y se produce una acumulación de lo que se denomina combustible, uno de los tres elementos necesarios para que se produzca la ignición (oxigeno, combustible, y calor). Estos tres componentes presentes en la superficie terrestre incrementan posibilidad de que exista fuego y también determinan la continuidad del incendio.

De acuerdo con el reporte "Spreading like Wildfire: The Rising Threat of Extraordinary Landscape Fires" [1, p. 38], habrá un aumento de área quemada del 22 % para 2050 y un aumento del 39% para 2100, lo que indica la importancia de limitar el calentamiento global a 1,5°C para finales del siglo y también para minimizar los costos ambientales y sociales asociado con incendios forestales

Los incendios forestales y el incremento de estos es uno de los problemas más recurrentes en el territorio español, la mayoría de ellos son causados por personas y en ocasiones avanzan sin control, quemando miles de hectáreas de bosques y otros tipos de vegetación. "En Cataluña, el 80% de los incendios están causados por la mano del hombre. La causa principal de incendio forestal es la negligencia de actividades humanas" [2].

"Se reconoce que uno de los factores que dificultan la disponibilidad de los valores correctos es la naturaleza de los parámetros: hay parámetros totalmente dinámicos, que cambian constantemente (velocidad y dirección del viento), otros que cambian con frecuencia (humedades del combustible, las cuales varían con los ciclos día-noche y con el clima del lugar) y otros que cambian poco a poco como el tipo de combustible" [3].

Para obtener la información de los parámetros dinámicos mencionados anteriormente, Cataluña cuenta con gran parte de estaciones que miden los valores climatológicos debido a que su superficie forestales de un 63,8% del territorio según Instituto de estadística de Cataluña [4], estas estaciones nos aportan la información necesaria para incluir variables que son totalmente dinámicas y cambian de manera constante como el clima, por otro lado se incluyen las que varían poco en el tiempo como la pendiente o inclinación de la superficie y la altitud del terreno, el peligro de incendio y la vulnerabilidad según categorías basadas en un estudio del tipo de combustible y contenido del agua del suelo. También, se considera que los datos necesarios para la evaluación del riesgo en un área, como son el número de personas presentes un día determinado, o las actividades a que se dedican, no están generalmente disponibles [5]. Es por ello que

se añade la variable movilidad urbana, para determinar si un número de personas presentes en determinado lugar puede elevar el riesgo de incendio en esa área específica, con los datos disponibles de movilidad urbana desde el año 2020 que han surgido como necesidad de evaluar la movilidad urbana en la pandemia del covid-19.

Si analizamos las estadísticas proporcionadas por el Instituto de estadística de Cataluña, se muestra que en el año 2019 hubo un total de 545 incendios que afectó a 5.077 hectáreas, mientras que en el año 2020, las cifras bajaron a 289 incendios con solo 132,4 hectáreas afectadas, podría deberse a las restricciones de movilidad por la pandemia del coronavirus, posteriormente en el año 2022 se detectó una tendencia al alza tras flexibilizar las medidas de movilidad reportándose un total 607 incendios que afectaron 2.422,3 hectáreas [6].

Es por este motivo que surge la necesidad de evitar que se produzcan estos incendios forestales y para ello prever una distribución de recursos de extinción más eficiente con el fin de reducir costes y evitar daños y pérdidas, para ello se plantea el desarrollo de un modelo de predicción de incendios forestales y además incluir datos de la movilidad urbana como variable para determinar si la influencia humana tendría incrementa este riesgo de producción de incendios forestales, específicamente en la comunidad de Cataluña.

2. Metodología

En los últimos años, los métodos de inteligencia artificial (IA) han demostrado ser muy efectivo para predecir peligros naturales. Además, los métodos de IA con frecuencia se han utilizado en el contexto del modelado de incendios forestales y han superado a los métodos estadísticos convencionales en muchas casillas.

En nuestro estudio, que utilizó datos de series de tiempo para predecir incendios forestales, implementamos un modelo de memoria a largo plazo (LSTM) para predecir los incendios forestales. El modelo LSTM es un tipo especial de recurrente red neuronal (RNN) que conserva información histórica en datos utilizando una unidad de memoria interna selectiva. LSTM ha demostrado ser más eficaz para el análisis de series temporales problemas de predicción que otros métodos de IA. Ya que las ocurrencias de incendios tienen reglas y tendencias obvias y LSTM tiene ventajas en la predicción de tendencias de ocurrencia [7, p. 2], como se indica en el trabajo de investigación donde se comparan distintos modelos, donde se ha concluido que "el modelo LSTM obtenido el mejor rendimiento predictivo de los tres modelos, teniendo la precisión promedio más alta de 90.9%" [7, p. 7].

En el presente trabajo se plantea una predicción de ocurrencia de incendios forestales en Cataluña, utilizando en particular el modelo de red neuronal LSTM (del inglés "Long-Short Term Memory"), como se ha mencionado ésta es una red neuronal útil en la predicción de series de tiempo dado que son capaces de capturar las relaciones lineales y no lineales entre los datos debido a su estructura no lineal que permite un modelo con más grados de libertad.

También, se plantea la realización de modelos de machine learning (ML) como el Xgboost (Extreme Gradient Boosting), considerado entre los mejores de ML para la clasificación binaria y que obtiene resultados bastante eficientes. Ya que es un algoritmo predictivo supervisado que utiliza el principio de boosting, generando múltiples modelos de predicción secuencialmente, y que cada uno de estos tome los resultados del modelo anterior, para generar un modelo más completo, con mejor dominio predictivo y mayor seguridad en sus resultados. En el estudio realizado sobre "Grid-based Urban Fire Prediction Using Extreme Gradient Boosting (XGBoost)", se muestra como ejemplo una tabla comparativa "Performance characteristics of different models" donde el modelo Xgboost es el que mejor resultado arroja [8, p. 11]

Para ambos modelos, se utilizan datos entre los años 2015 y 2022 de variables climatologías dinámicas medidas por estaciones de AEMET y METEOCA. Así mismo, se realiza un estudio doble, donde se incluye la variable movilidad urbana para hipotétizar como afectaría ésta a la ocurrencia de los incendios en los municipios. Los grupos de variables explicativas mencionadas en el apartado anterior corresponden a los inputs layers, mientras que el output será el valor predicho indicando la ocurrencia o no de incendios forestales.

Para la construcción del modelo, los datos meteorológicos a considerar son: humedad relativa, temperatura, precipitaciones, velocidad del viento y racha máxima de viento, duración del sol en horas, presión mínima y máxima; se tiene en cuenta el factor climático como dinámico y con un componente estacional, por lo que es fundamental para la construcción del modelo, obtener el registro histórico de las variaciones meteorológicas a través del tiempo entre 2015-2022. Se añaden variable proveniente de datos topográficos que varían muy poco en el tiempo, como la pendiente la cual influye en la transferencia de calor y radiación, y la altitud, la cual a medida que aumenta la temperatura disminuye gradualmente. Se incluye también variables categóricas de peligro de incendio y vulnerabilidad, las cuales se han obtenido mediante un estudio de combustibilidad del suelo catalán teniendo en cuenta la cobertura de vegetación y contenido de agua, categorizando en niveles de riesgo de incendio bajo, moderado, alto y muy alto y por ultimo incluimos la variable alto riesgo, la cual nos categoriza en "Si" o "No" si el municipio correspondiente es de alto riesgo a partir de la

Base Municipal y con la información del Decreto 64/95. Disponemos de la variable fecha, la cual se utiliza para el cálculo y obtención de las variables día del año, fin de semana, laborable y mes. También, disponemos de la variable "Ocurrencia" que será nuestro objeto de estudio y por lo tanto nuestra variable a predecir, en la que se encuentran el valor "Si" o "No" en caso de que haya habido un incendio en esa fecha y en ese municipio.

La metodología mencionada en el presente estudio se encuentra en siguiente repositorio disponible para su consulta : https://gitlab.com/Nathifer/incendios-forestales-en-cataluna/

2.1. División del conjunto de datos

En un primer modelo, se propone dividir el conjunto de datos en dos partes, los datos de entrenamiento estarán comprendidos entre el periodo entre 01-01-2015 y 31-12-2020 y los datos de prueba estarán comprendidos entre el 01-01-2021 y el 31-12-2022 en los notebooks, sin embargo, se debe tener en cuenta que nuestro conjunto de datos completo después de realizar el análisis, transformaciones y limpiezas quedan para el periodo comprendido entre 01-01-2015 y 01-02-2022, por lo que realmente el conjunto de prueba es entre el 01-01-2021 y 01-02-2022. Por otro lado, en los notebooks que incluyen la movilidad urbana, se disponen de estos datos hasta el 09-05-2021.

Se decide probar los modelos con distintas divisiones del dataset entre las que se contemplan:

1. División aleatoria utilizando train_test_split con un 20% de datos para prueba y un 80% de datos para el entrenamiento incluyendo el parámetro shufle =True para aleatorizar los datos
2. División aleatoria utilizando train_test_split con un 20% de datos para prueba y un 80% de datos para el entrenamiento incluyendo el parámetro shufle =False para mantener el orden delas fechas en los datos
3. División manual utilizando 9 meses para los datos de entrenamiento y los otros 6 meses para datos de prueba, teniendo en cuenta la estacionalidad.
4. División mediante undersampler manual, donde se iguala la clase mayoritaria '0' a la clase minoritaria '1'

2.2 Limitaciones

La principal limitación de esta investigación está referida a los datos de Movilidad Urbana, ya que están disponibles en un espacio de tiempo comprendido entre 2020 y 2021, por lo que solo se dispone de un año que coincida con la estacionalidad en la que se producen los incendios forestales (meses de verano). Sin embargo, la viabilidad de la construcción de la propuesta presentaba se debe a la existencia de los datos abiertos de las diversas variables que afectan el comportamiento de los incendios forestales. Esta situación, permitió visualizar la potencialidad del modelo y la posibilidad de realizar un estudio comparativo de varios modelos donde se incluyan distintos periodos de tiempo y la variable movilidad.

2.3. Extraccion y Transformación

En este trabajo de investigación se llevó a cabo la extracción, transformación y carga de datos; la extracción se llevó a acabo de los distintos sitios web mediante API, también se descargan datos en formato CSV, y otros en formato SHP del cual se extraen las tablas de atributos mediante la herramienta QGIS. Las fuentes de las cuales se realizan las extracciones se encuentran detalladas en el apartado "Apéndice"

Se realizaron distintas transformaciones en los sets de datos, entre las cuales podemos mencionar: las transformaciones de la variable fecha para obtener variables como el mes, día del año, laborable y fin de semana, además se realizan uniones entre todas las tablas obtenidas de las fuentes de datos, filtrado de los datos relevantes, en este caso solo los correspondientes a la comunidad autónoma de Cataluña, imputación de datos faltantes con distintas técnicas según el tipo de variable, creación de índice con la variable fecha, se realizan reemplazos de puntos, guiones , entre otros caracteres, de lo que resulta el siguiente dataframe:

Variables	Ejemplo de Valores
Fecha	2020-02-28

Prec (precipitación)	0.2
Tmax (Temperatura máxima)	17.0
Tmed (Temperatura media)	10.3
Tmin (Temperatura mínima)	3.6
Velmedia (Velocidad media del viento)	5.0
Racha (Racha del viento)	12.8
Sol (Índice de sol)	8.6
presMax (Presión máxima)	1006.2
presMin (Presión mínima)	1000.3
rhum (Humedad relativa)	54.0
altitud (Altitud del terreno)	70
porcentaje_pendiente_total (Pendiente del terreno)	2.9
alto_riesgo (Si existe o No alto riesgo en el municipio)	Si
peligro (Nivel de peligro de incendio del municipio)	Molt Alt
vulner (Nivel de vulnerabilidad de incendio del municipio)	Mitja
total_viajes_estación (Total de viajes como destino a ese municipio)	232066.57
laborable (Día laborable (L-V))	4
dayofweek (Día de la semana (0-6))	52
weekend (Si es o No fin de semana (0-1))	0
mes (Mes del año (1-12))	2

Tabla 1. Ejemplo del esquema final del conjunto de datos.

2.2. Analisis Exploratorio de Datos (EDA)

3.1.1. Balanceo de clases

Se comprueba el desbalanceo de la clase ocurrencia de incendio, por lo que se decide realizar el balanceo de estas mediante dos técnicas distintas, la primera de ellas Smote realizando un oversampling y la segunda con RandomUnderSampler para undersampling, después de aplicar ambas técnicas, vemos con cual técnica el modelo nos arroja mejor resultado, si es creando datos sintéticos o igualando la clase mayoritaria a la minoritaria.

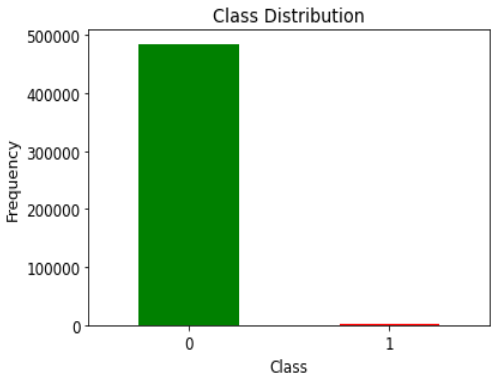

Figura 1. Frecuencia por número de observaciones: Desbalanceo de las clases de ocurrencia de incendio, donde la clase "1" representa solo un 0,62%.

Los siguientes gráficos muestran cómo se han balanceado las clases con una cantidad de incendios que antes no teníamos. Esos son todos los puntos de datos sintéticos que se ha creado con Smote, mientras que la otra imagen muestra los datos que se eliminan de la clase mayoritaria para igualarla a la clase minoritaria mediante RandonUnderSampler.

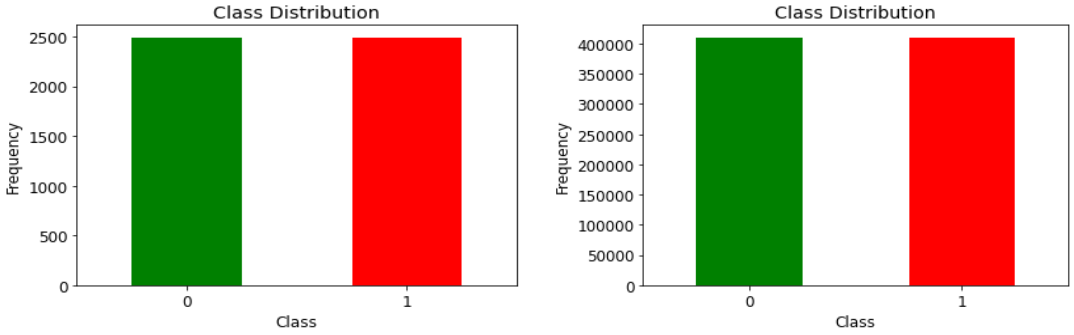

Figura 2. Frecuencia por número de observaciones: Balanceo de Clases con RandomUnderSampler y Balanceo de Clases con Smote.

3.1.2. Componente Estacional

Claramente al tratarse de valores climatológicos se disponen de un componente estacional en esta serie temporal, para comprobar esta estacionalidad se realiza el trazado de los datos y comprobar si existe una estructura estacional en los datos.

A modo de ejemplo, podemos ver en la trama de la variable temperatura, que en su mayoría tiene un nivel estable, es decir, es una serie estacionaria donde vemos que cada año se repite un patrón muy similar en cada una de ellas.

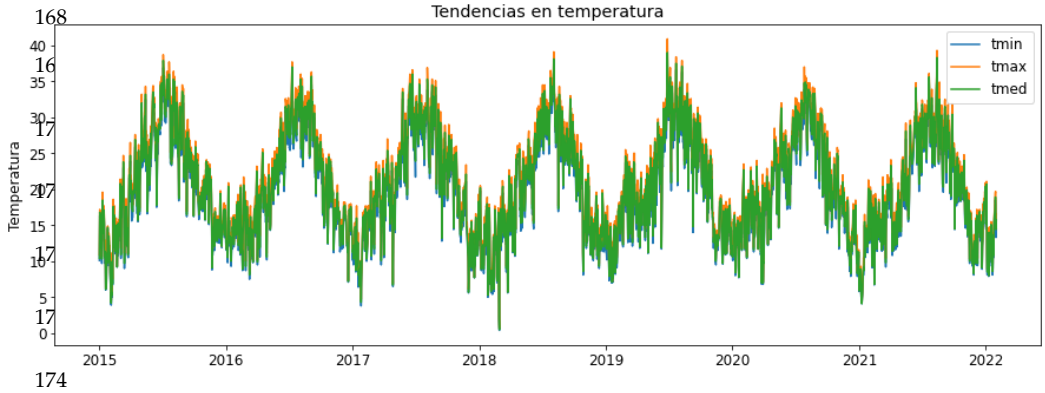

Figura 3. Valores Climatológicos: Temperatura.

3.1.3. Autocorrelación

Analizamos si la serie es estacionaria, mediante la visualización de correlograma y autocorrelograma de cada una de las variables, a continuación, a modo de ejemplo se muestra la variable temperatura media. En este caso como cada valor de autocorrelación está por encima del área sombreada de azul, nos indica que los coeficientes de autocorrelación son significativos. Lo que significa que hay autocorrelación para los retrasos (lags). Esto ocurre con el resto de variables analizadas.

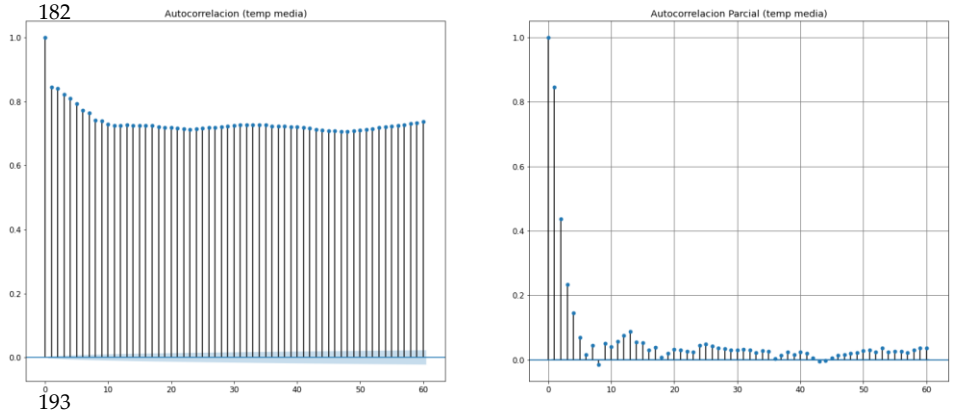

Figura 4. Autocorrelación y Autocorrelación Parcial de la variable temperatura media

3.1.4. Correlación

Debido a que los predictores no correlacionados o con mínima correlación funcionan mejor para una serie temporal, realizamos un diagrama de correlación entre las variables y posteriormente aplicamos un test de levene para conocer si las varianzas de todos los grupos son iguales y de esta manera poder determinar cuáles de las variables podemos prescindir para el modelo.

Test levene:
H0: Las varianzas de todos los grupos son iguales.
H1. Al menos una varianza es distinta entre todos los grupos.

Se muestran evidencias para rechazar la hipótesis de que los dos grupos tienen la misma varianza, por lo que son heterocedásticos, ya que obtenemos para la mayoría un p-value inferior a 0.05. Por lo tanto, la hipótesis nula de igualdad de varianzas no se acepta y se concluye que hay una diferencia entre las variaciones. Solo la variable "laborable" nos arroja un pvalue=0.85, lo que nos asegura su correlación y podríamos eliminarla del modelo.

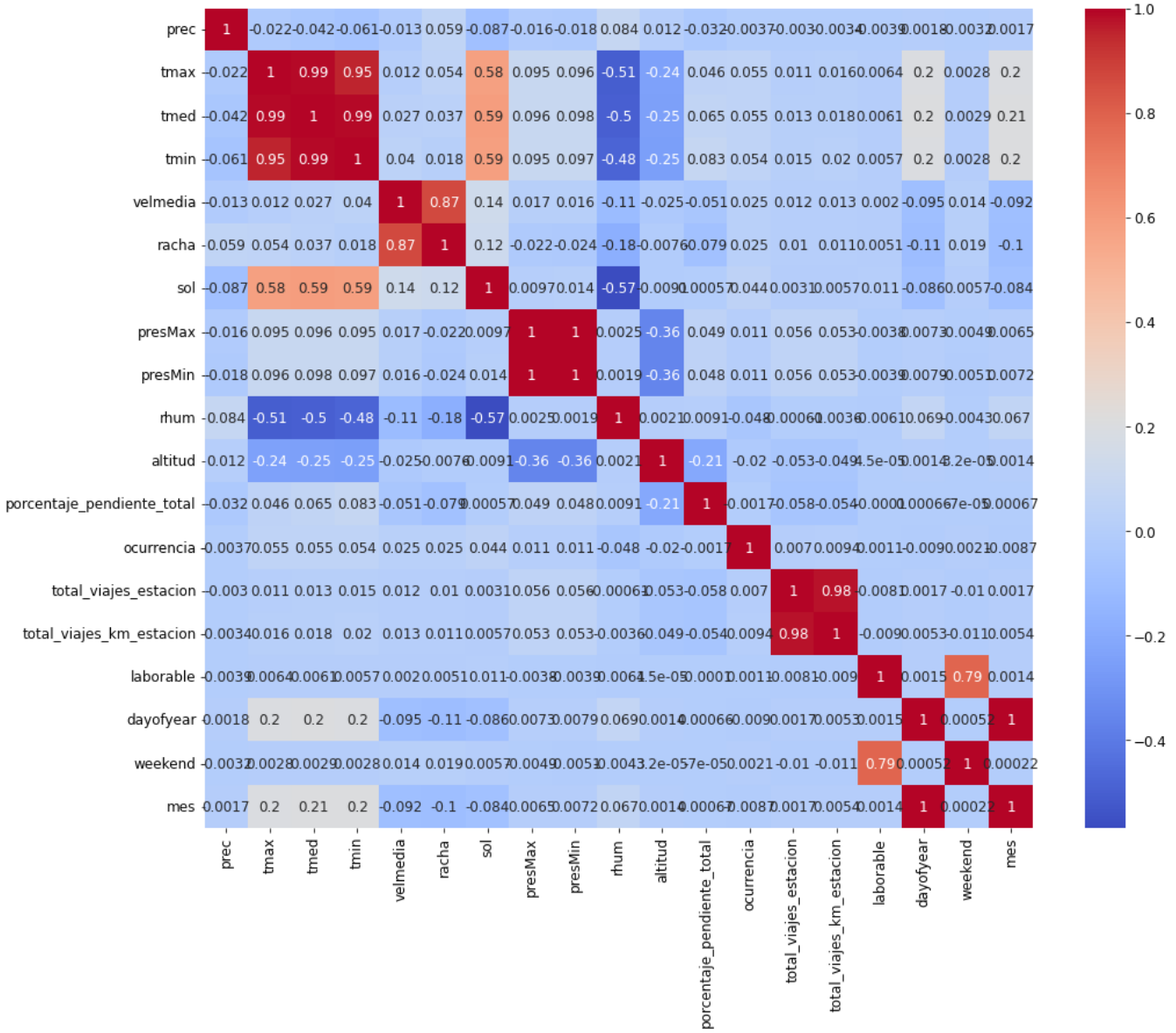

Figura 5. Correlación entre las variables

2.3. Descripción de los modelos

2.3.1 Modelo LSTM

Se presentan modelos con una capa de entrada de distintas dimensiones dependiendo de la cantidad de variables que se utilice y si este incluye o no la movilidad, se debe también tener en cuenta que se muestra un número mayor al realizar la normalización de las variables categóricas, seguidamente se añade la capa del modelo LSTM con 200 neuronas, luego dos capas de apagado de neuronas (drop_out) que se para la desactivación de un porcentaje de neuronas aleatorias de una capa de una red neuronal y así reducir la cantidad de sobreajuste de la red neuronal al conjunto de datos de entrenamiento de forma que ninguna neurona memorice parte de la entrada, también se añaden dos capas con activación sigmoid las cuales una vez se asocia el coste a cada neurona dentro de una iteración se realiza la optimización de los parámetros asociados a esa neurona con el propósito de reducir la función de coste final. El algoritmo de optimización utilizado en este caso es el Adam (adaptive moment estimation). Finalmente compilamos el modelo con loss='binary_crossentropy' ya que es la función de pérdida predeterminada que se utiliza para problemas de clasificación binaria donde los valores objetivo son 0 y 1. También se incluye un batch_size de 32 ejemplos que se introducen en la red para que entrene de cada vez.

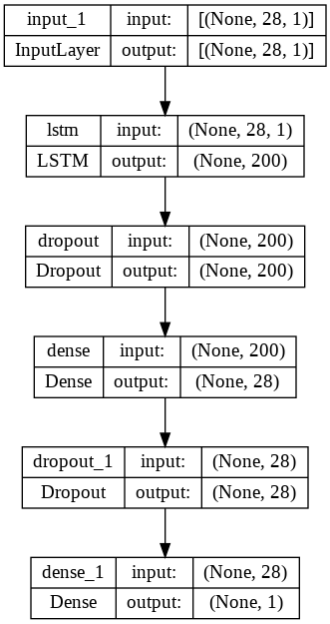

Figura 6. Arquitectura de la red neuronal LSTM

2.3.2 Modelo XGBoost

Extreme Gradient Boosting (XGBoost), consiste en la construcción de un modelo a partir de muchos modelos de árboles de decisión, los cuales se agregan para ajustar los errores de predicción cometidos por modelos anteriores

Para este modelo hemos utilizado número de estimadores 500, lo que quiere decir la cantidad de árboles que utiliza el modelo, la profundidad máxima que pueden alcanzar los árboles se ha fijado entre 5 y 20, la tasa de aprendizaje la hemos fijado en 0.0001 para reducir el riesgo de overfitting y considerando que la cantidad de árboles fijados es suficiente. También, se especifica el objetivo de aprendizaje correspondiente a binary:logistic para clasificación binaria.

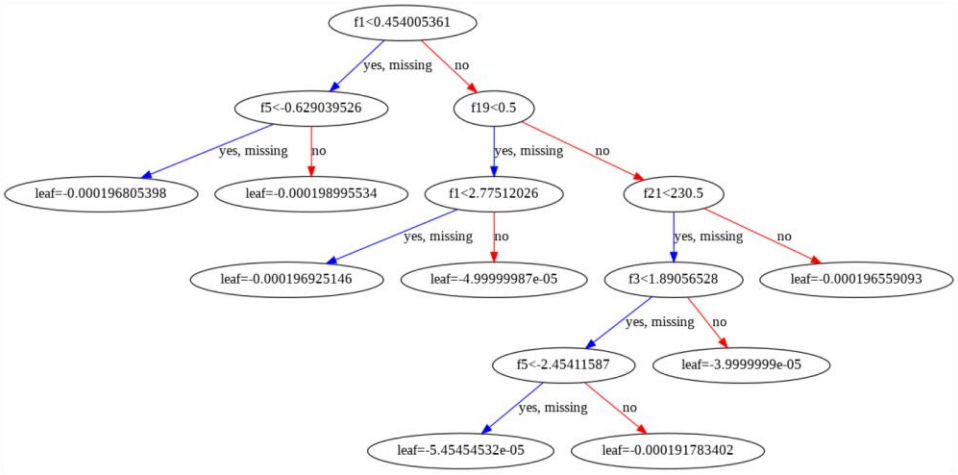

Figura 7. Arquitectura del modelo XGboost

2.4. Métricas

Para este estudio, la métrica de exactitud (accuracy) no es la más idónea debido al desbalanceo de clases, ya que podría arrojarnos un 99% de exactitud si predice siempre como la clase mayoritaria, es decir, que no hay incendios, sin embargo, hemos utilizado el reporte de clasificación (clasification_report), disponible en la librería de sklearn, el cual nos muestra el accuray para cada clase.

Para tener una mejor representación de la calidad del modelo, estudiaremos la métrica de exhaustividad (recall), ya que esta proporciona información sobre la cantidad de ocurrencia de incendios o no que el modelo es capaz de identificar, es decir, nos indica los verdaderos positivos clasificados correctamente.

También evaluaremos el valor de f1, como una de las métricas muy utilizadas en problemas con clases desbalanceada, ya que combina la precisión y exhaustividad en un sólo valor mucho más objetivo.

3. Resultados y Discusiones

3.1.1 Modelo LSTM (2015-2022) con Movilidad Urbana

En la construcción de este modelo se incluyen las 15 variables dinámicas de los valores climatológicos, variables que varían poco en el tiempo como la altitud y pendiente y variables categóricas como riesgo, peligro, vulnerabilidad, laborable, día del año, mes y fin de semana para el periodo comprendido entre los años 2015 y 2022. Además, se añade la Movilidad Urbana (total_viajes_estación) y en base a los resultados se analiza si es la movilidad urbana es una variable representativa en la ocurrencia de los incendios y como estas podrían cambiar nuestras métricas de evaluación del modelo. Este nos servirá para realizar un estudio comparativo con los otros modelos, y determinar si un mayor periodo de datos de las variables dinámicas y la inclusión de la variable movilidad urbana, puede mejorar la métrica.

El mejor resultado de este modelo se obtuvo al realizar una división aleatoria de los datos con 20% para prueba 80% para entrenamiento utilizando el parámetro shuffle=True para aleatorizar los datos, posteriormente se realizó un balanceo con la técnica de RandomUnderSampler igualando la clase mayoritaria a la minoritaria a 2419 valores. Finalmente se realizó la búsqueda de hiperparametros mediante gridsearch, con lo que se alcanzaron las siguientes métricas:

- Exactitud de 0.578
- Precisión de 0.648
- Exhaustividad de 0.576
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,92 y para clase 0 se obtuvo un 0,23, lo que nos da una exactitud de un 0,58 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,69 mientras que para la clase 0 un 0,35, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

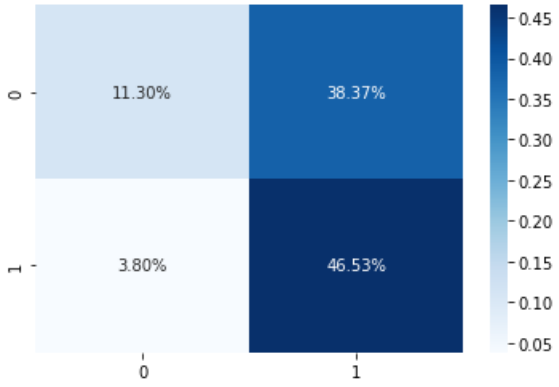

Figura 8. Matriz de confusión: Modelo LSTM (2015-2022) con Movilidad Urbana

3.1.2 Modelo LSTM (2015-2022) sin Movilidad Urbana

En la construcción de este modelo se incluyen las 15 variables dinámicas de los valores climatológicos, variables que varían poco en el tiempo como la altitud y pendiente y variables categóricas como riesgo, peligro, vulnerabilidad, laborable, día del año, mes y fin de semana; para el periodo comprendido entre los años 2015 y 2022.

El mejor resultado de este modelo se obtuvo al realizar una división aleatoria de los datos con 20% para prueba 80% para entrenamiento utilizando el parámetro shuffle=True para aleatorizar los datos, posteriormente se realizó un balanceo con la técnica de RandomUnderSampler igualando la clase mayoritaria a la minoritaria a 2409 valores. Finalmente se realizó la búsqueda de hiperparametros mediante gridsearch, con lo que se alcanzaron las siguientes métricas:

- Exactitud de 0,645
- Precisión de 0,659
- Exhaustividad de 0,641
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,81 y para clase 0 se obtuvo un 0,47, lo que nos da una exactitud de un 0,65 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,70 mientras que para la clase 0 un 0,57, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

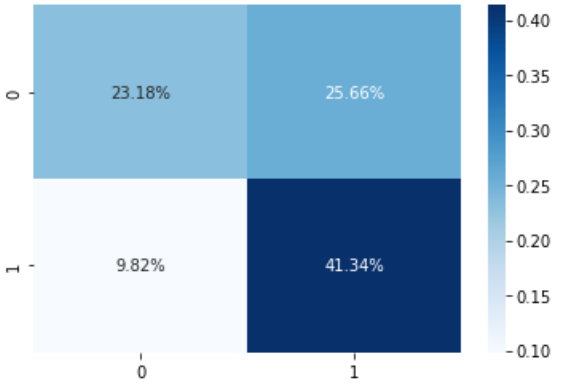

Figura 9. Matriz de confusión: Modelo LSTM (2015-2022) sin Movilidad Urbana

Comparándolo con el modelo anterior (Modelo LSTM (2015-2022) con Movilidad Urbana), se obtiene una mayor exhaustividad para la clase 0, sin embargo, se disminuye la exhaustividad para la clase 1 en un 0,11, sin embargo, en este se obtiene un modelo con mayor exactitud para ambas clases de un 0,65 es decir un 0,7 más. Aun así, nuestro modelo anterior que incluye la variable movilidad urbana predice con mayor exhaustividad la clase 1 (ocurrencia de incendio) el cual es nuestro objetivo, que el modelo detecte lo mejor posible esta clase.

3.1.3 *Modelo LSTM (2020-2022) con Movilidad Urbana*

En la construcción de este modelo se incluyen las 15 variables dinámicas de los valores climatológicos, variables que varían poco en el tiempo como la altitud y pendiente, variables categóricas como riesgo, peligro, vulnerabilidad, laborable, día del año, mes y fin de semana para el periodo comprendido entre los años 2020 y 2022. Además, se añade la Movilidad Urbana (total_viajes_estación) y en base a los resultados se analiza si es la movilidad urbana es una variable representativa en la ocurrencia de los incendios y como estas podrían cambiar nuestras métricas de evaluación del modelo.

El mejor resultado de este modelo se obtuvo al realizar una división aleatoria de los datos con 20% para prueba 80% para entrenamiento utilizando el parámetro shuffle=True para aleatorizar los, posteriormente se realizó un balanceo con la técnica de RandomUnderSampler igualando la clase mayoritaria a la minoritaria a 296 valores. Finalmente se realizó la búsqueda de hiperparametros mediante gridsearch, con lo que se alcanzaron las siguientes métricas:

- Exactitud de 0,640
- Precisión de 0,667
- Exhaustividad de 0,658
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,81 y para clase 0 se obtuvo un 0,51, lo que nos da una exactitud de un 0,64 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,66 mientras que para la clase 0 un 0,62, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

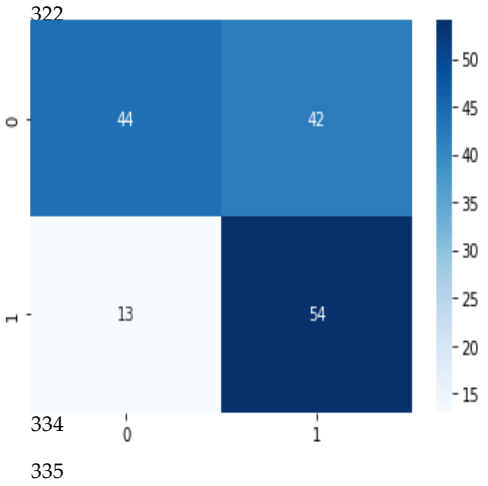

Figura 10. Matriz de confusión: Modelo LSTM (2020-2022) con Movilidad Urbana

Comparándolo con el modelo anterior (Modelo LSTM (2015-2022) con Movilidad Urbana) el cual que incluye la variable movilidad urbana y con un periodo de tiempo más amplio, por lo que dispone de más datos de valores climatológicos, se muestra unas métricas muy parecidas que parece no ser relevante disponer de más datos en las variables climatológicas para la predicción, con un año el modelo es capaz de obtener resultados muy similares.

3.1.4 *Modelo LSTM (2020-2022) sin Movilidad Urbana*

En la construcción de este modelo se incluyen las 15 variables dinámicas de los valores climatológicos, variables que varían poco en el tiempo como la altitud y pendiente, variables categóricas como riesgo, peligro, vulnerabilidad, laborable, día del año, mes y fin de semana; para el periodo comprendido entre los años 2020 y 2022.

El mejor resultado de este modelo se obtuvo al balancear primero los datos manualmente con undersampler igualando la clase mayoritaria a la minoritaria a 382 valores, posteriormente se utilizó la técnica de Smote para balancear la clase minoritaria a 310 valores. Finalmente se realizó la búsqueda de hiperparametros mediante gridsearch, con lo que se alcanzaron las siguientes métricas:

- Exactitud de 0,588
- Precisión de 0,637
- Exhaustividad de 0,603
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,86 y para clase 0 se obtuvo un 0,35, lo que nos da una exactitud de un 0,59 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,66 mientras que para la clase 0 un 0,47, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

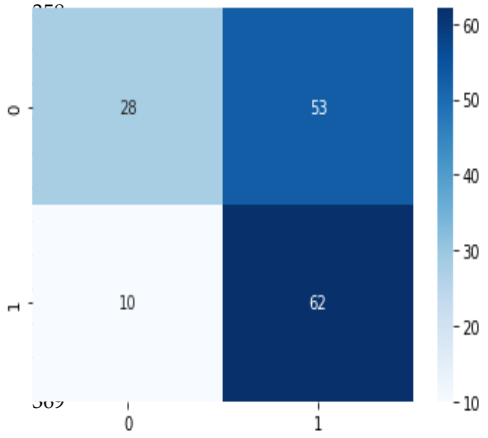

Figura 11. Matriz de confusión: Modelo LSTM (2020-2022) sin Movilidad Urbana

Comparando este modelo con el modelo "Modelo LSTM (2015-2022) sin Movilidad Urbana", en el que incluye un periodo más amplio de datos, que implica mayor cantidad de datos para el modelo, se evidencia una mejora en la exhaustividad de las clases de 0,58 a 0,64 y en la exactitud de 0,59 a 0,65.

Comparando este modelo con el modelo "Modelo LSTM (2020-2021) con Movilidad Urbana", en el que se incluye la movilidad urbana, se evidencia una mejora en la exhaustividad de las clases de 0,60 a 0,65 y en la exactitud de 0,58 a 0,64.

3.2.1 XGBoost (2015-2022) con Movilidad Urbana

El mejor resultado del modelo se ha obtenido luego de dividir los datos manualmente con undersampler igualando la clase mayoritaria a la minoritaria a 3029 datos correspondientes a los incendios ocurridos en ese periodo de 2015a 2022, del modelo obtuvimos las siguientes métricas:

- Exactitud de 0,682
- Precisión de 0,684
- Exhaustividad de 0,682
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,73 y para clase 0 se obtuvo un 0,63, lo que nos da una exactitud de un 0,68 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,70 mientras que para la clase 0 un 0,67, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

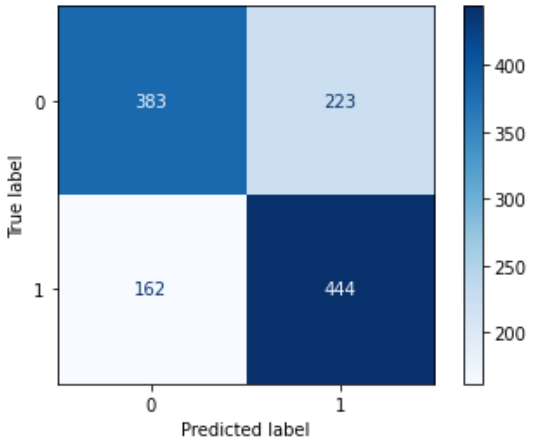

Figura 12. Matriz de confusión: Modelo XGBoost (2015-2022) con Movilidad Urbana

3.2.2 XGBoost (2015-2022) sin Movilidad Urbana

El mejor resultado del modelo se ha obtenido realizando una división aleatoria de los datos, con un 20% para datos de prueba y el 80% para datos de entrenamiento, se aplicó la técnica de undersampler manual, igualando la clase mayoritaria la minoritaria a 3029 datos, correspondientes a los incendios ocurridos en ese periodo de 2015 a 2022. Del modelo obtuvimos las siguientes métricas:

- Exactitud de 0,676
- Precisión de 0,681
- Exhaustividad de 0,676
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,76 y para clase 0 se obtuvo un 0,59, lo que nos da una exactitud de un 0,68 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,70 mientras que para la clase 0 un 0,65, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

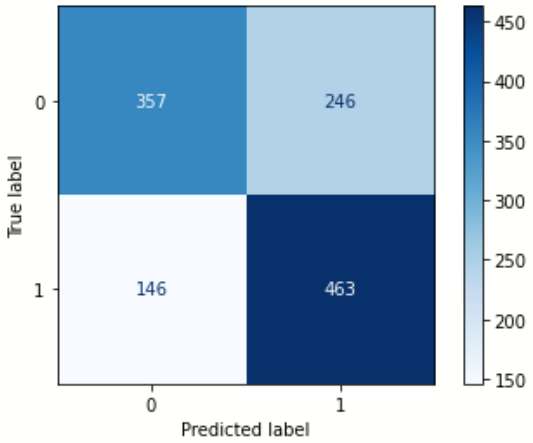

Figura 13. Matriz de confusión: Modelo XGBoost (2015-2022) sin Movilidad Urbana

3.2.3 XGBoost (2020-2022) con Movilidad Urbana

El mejor resultado de este modelo se obtuvo al realizar una división de los datos con 20% para prueba 80% para entrenamiento en el periodo de 2020 y 2022, posteriormente se realizó un balanceo con la técnica Smote igualando la clase minoritaria a la mayoritaria a 315 valores. Finalmente se realizó la búsqueda de hiperparametros mediante gridsearch, con lo que se alcanzaron las siguientes métricas:

- Exactitud de 0,640
- Precisión de 0,640
- Exhaustividad de 0,642
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0,63 y para clase 0 se obtuvo un 0,66, lo que nos da una exactitud de un 0,64 para ambas clases.
- Para la clase 1 se obtuvo de F1-score un 0,66 mientras que para la clase 0 un 0,62, lo que indica que para nuestra clase objetivo de predicción tiene mejor precisión y exhaustividad.

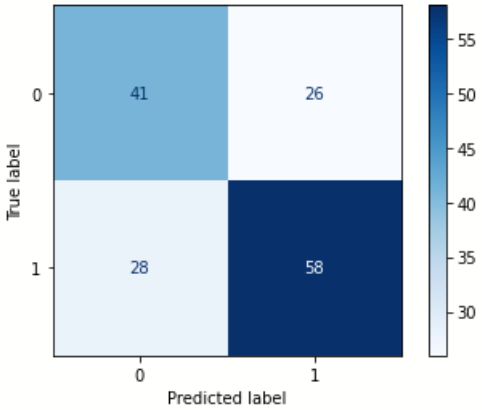

Figura 14. Matriz de confusión: Modelo XGBoost (2020-2022) con Movilidad Urbana

3.2.4 XGBoost (2020-2022) sin Movilidad Urbana

El mejor resultado de este modelo se obtuvo al realizar una división de los datos con 20% para prueba 80% para entrenamiento en el periodo de 2020 y 2022, posteriormente se realizó un balanceo con la técnica Smote igualando la clase minoritaria a la mayoritaria a 310 valores. Finalmente se realizó la búsqueda de hiperparametros mediante gridsearch, con lo que se alcanzaron las siguientes métricas:

- Exactitud de 0,96732
- Precisión de 0,96728
- Exhaustividad de 0,96744
- Para la clase 1 de ocurrencia de incendios se obtuvo una exhaustividad del 0, 96 y para clase 0 se obtuvo un 0,97, lo que nos da una exactitud de un 0,97 para ambas clases.
- Para ambas clases se obtuvieron F1-score de un 0,97 lo que indica que para nuestro modelo tiene muy buena precisión y exhaustividad.

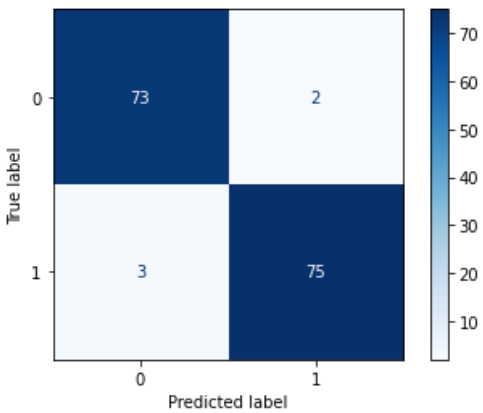

Figura 15. Matriz de confusión: Modelo XGBoost (2020-2022) sin Movilidad Urbana

Modelos	Exhaustividad (Recall)	Exactitud (Accuracy)	F1-Score (Clase 0 y 1)	Precisión
LSTM				
LSTM con movilidad (2015–2022)	0.576	0.578	[0: 0,65 1: 0,69]	0.648
LSTM sin movilidad (2015–2022)	0.641	0.645	[0: 0,57 1: 0,70]	0.659
LSTM con movilidad (2020–2022)	0,658	0,640	[0: 0,62 1: 0,66]	0,667
LSTM sin movilidad (2020–2022)	0.603	0.588	[0: 0,47 1: 0,66]	0.637
XGBoost				
XGBoost con movilidad (2015–2022)	0,682	0,682	[0: 0,67 1: 0,70]	0,684
XGBoost sin movilidad (2015–2022)	0,676	0,676	[0: 0,65 1: 0,70]	0,681
XGBoost con movilidad (2020–2022)	0,642	0,640	[0: 0,62 1: 0,66]	0,640
XGBoost sin movilidad (2020–2022)	**0,967**	0,967	[0: 0,97 1: 0,97]	0,967

Tabla 2. Resumen de métricas por modelo.

4. Conclusiones

En este trabajo presentamos un modelo predictivo basado en redes neuronales LSTM y un modelo de machine learning XGBoost para dar respuesta a nuestro principal objetivo, el cual es predecir los incendios forestales en Cataluña y determinar si la variable movilidad urbana podría ser relevante para la predicción de la ocurrencia de estos.

De los resultados obtenidos, se concluye que el modelo basado en redes neuronales LSTM muestra como mayor resultado un 66% de exhaustividad y específicamente un 70% para la clase 1, mientras que el modelo XGBoost es con el que mejores métricas hemos obtenido, después de realizar una búsqueda de sus hiperparametros, un 97% de exhaustividad.

De lo resultados expuestos podemos acotar que un modelo LSTM sin la variable movilidad urbana y con un periodo de datos de solo un periodo comprendido entre 2020 y 2022, no es el más idóneo para este estudio, mientras que el XGBoost nos proporciona las mejores métricas para estas mismas casuísticas.

El estudio concluye que el modelo de XGBoost sin la variable movilidad urbana y para el periodo de datos entre 2020 y 2022, desempeñó un mejor resultado de predicción en la ocurrencia de incendios forestales que los otros modelos estudiados.

Finalmente, empleando este modelo enfocado en el machine learning, este estudio fue capaz de proporcionar clarificaciones para predecir los incendios forestales con precisión y tiempo de ejecución apropiados. Para futuras investigaciones el trabajo se puede ampliar utilizando más datos de la movilidad urbana y poder compararlo nuevamente y otros conjuntos de datos para mejorar y confirmar la fiabilidad del modelo propuesto.

5. Apéndice

Los datos utilizados en el en la investigación provienen de las siguientes fuentes de datos abiertas y anonimizadas las cuales son:

- **Servicio Meteorológico de Cataluña (Meteocat)**
  - Datos meteorológicos: variables medidas con una frecuencia diaria, registrados en todas las estaciones de la red de estaciones meteorológicas automáticas de Cataluña. [9]
  - Ocurrencia de incendios: registros de los incendios forestales reportados de Cataluña, el municipio y comarca donde se produce, la fecha y las hectáreas afectadas. Se crea la variable Incendio que nos dirá si para un día determinado hubo incendio "1" o no "0". [10]
- **Agencia Estatal de meteorología (Aemet)**
  - Valores meteorológicos diarios en distintas estaciones de toda España, en las cuales se filtran solo las correspondientes a Cataluña. [11]

- **Instituto de Estadística de Cataluña**
  - Superficie y Pendientes de las Comarcas de Cataluña. [12]
  - Altitud, superficie y población. Municipios de Cataluña. [13]
- **Instituto Cartográfico y Geológico de Cataluña:**
  Proporciona un mapa de Peligro y Vulnerabilidad de Incendio Forestal de Protección Civil de Cataluña [14]
- **Departamento de Agricultura de la Generalitat de Cataluña**
  Proporciona datos sobre los municipios de Cataluña con alto riesgo de incendio forestal basado en el Decreto 64/95. [15]
- **Instituto Nacional de Estadística:**
  Proporciona tablas con las áreas de movilidad (INE) en el apartado de "Descripción de las áreas de movilidad y su población a 1 de enero de 2019", donde nos permite realizar la conversión de municipio de la tabla MITMA (movilidad urbana) con los municipios INE. [16]
- **Ministerio de Transporte y Movilidad Urbana:**
  Proporciona datos sobre la movilidad urbana durante el periodo de 02/2020 – 05/2022 y está constituida datos anonimizados asociados a los registros de conexión de los dispositivos móviles con la red de telefonía móvil diarios. [17]

Conflicto de Interés:

El autor declara no tener conflicto de intereses.

Referencias

[1] N. Unidas., «Spreading like Wildfire: The riding Threat of Extraordinary Landscape Fires,» 2022.

[2] Generalitat de Catalunya, «Bosque y Riesgo de incendios,» 16 06 2014. [En línea]. Available: https://web.gencat.cat/es/actualitat/reportatges/incendis/prevencio-i-proteccio/bosc-i-risc-dincendi/. [Último acceso: 12 2021].

[3] M. M. Denham, «Prediccion de Incendios Forestales Basada en Algoritmos Evolutivos Guiados por los Datos,» 10 07 2007. [En línea]. Available: https://ddd.uab.cat/pub/trerecpro/2007/hdl_2072_5208/TreballDeRecerca.pdf. [Último acceso: 11 2022].

[4] Instituto de Estadística de Cataluña, «Utilización del suelo,» 09 06 2022. [En línea]. Available: https://www.idescat.cat/indicadors/?id=anuals&n=10547&tema=terri&lang=es. [Último acceso: 12 2022].

[5] C. García Vega, «Dos modelos para la predicción de incendios forestales en Whitecourt Forest, Canadá,» 11 12 1998. [En línea]. Available: https://www.researchgate.net/profile/Cristina-Vega-Garcia/publication/28052588_Dos_modelos_para_la_prediccion_de_incendios_forestales_en_Whitecourt_Forest_Canada. [Último acceso: 12 2022].

[6] Instituto de Estadística de Cataluña, «Incendios forestales,» 16 05 2022. [En línea]. Available: https://www.idescat.cat/indicadors/?id=anuals&n=10545&lang=es. [Último acceso: 12 2022].

[7] H. Wang, M. Zhang y H. Liang, «A Neural Network Model for Wildfire Scale Prediction Using Meteorological Factors,» 21 11 2019. [En línea]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8924693. [Último acceso: 12 2022].

[8] . J. Meong-Hun y Haeng Yeol Oh , «Grid-based Urban Fire Prediction,» 24 10 2022. [En línea]. Available: https://sensors.myu-group.co.jp/sm_pdf/SM3149.pdf. [Último acceso: 12 2022].

[9] Servicio Meteorológico de Cataluña, «Dades meteorològiques de la XEMA,» 11 03 2019. [En línea]. Available: https://analisi.transparenciacatalunya.cat/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee. [Último acceso: 12 2022].

[10] Departamento de Acción Climatica, Alimentacion y Agenda Rural, «Incendis forestals a Catalunya. Anys 2011-2021,» 21 05 2021. [En línea]. Available: https://analisi.transparenciacatalunya.cat/Medi-Rural-Pesca/Incendis-forestals-a-Catalunya-Anys-2011-2021/bks7-dkfd. [Último acceso: 12 2022].

[11] Agencia Estatal de Meteorología, «Agencia Estatal de Meteorología - Centro de Descargas - Acceso General,» [En línea]. Available: https://opendata.aemet.es/centrodedescargas/productosAEMET?. [Último acceso: 12 2022].

[12] Instituto de Estadística de Cataluña, «Superficie y pendientes. Comarcas y Aran, y ámbitos,» 17 01 2020. [En línea]. Available: https://www.idescat.cat/indicadors/?id=aec&n=15181&lang=es. [Último acceso: 12 2022].

[13] Instituto de Estadística de Cataluña, «Altitud, superficie y población. Municipios,» 27 12 2021. [En línea]. Available: https://www.idescat.cat/indicadors/?id=aec&n=15903&lang=es. [Último acceso: 12 2022].

[14] Departamento de Interior. Dirección General de Protección Civil, «Mapa de Protecció Civil de Catalunya: Risc d'incendis forestals,» 21 03 2022. [En línea]. Available: https://analisi.transparenciacatalunya.cat/Seguretat/Mapa-de-Protecci-Civil-de-Catalunya-Risc-d-incendi/m9sy-395b. [Último acceso: 12 2022].

[15] Departamento de Acción Climática, Alimentación y Agenda Rural, «Municipios con alto riesgo de incendio forestal,» 21 11 2016. [En línea]. Available: https://agricultura.gencat.cat/es/serveis/cartografia-sig/bases-cartografiques/boscos/municipis-alt-risc-incendi-forestal/. [Último acceso: 12 2022].

[16] Instituto Nacional de Estadística, «Datos de movilidad,» 02 12 2020. [En línea]. Available: https://www.ine.es/covid/covid_movilidad.htm#tablas_resultados. [Último acceso: 12 2022].

[17] Ministerio de Transportes, Movilidad y Agenda Urbana, «Open Data Movilidad 2020-2021,» 14 02 2020. [En línea]. Available: https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data/opendata-movilidad. [Último acceso: 12 2022].

497

498