A data pipeline for forest fire prediction in Pinamar

Ana Martínez Saucedo[1] and Pablo Ezequiel Inchausti[1]

Universidad Argentina de la Empresa (UADE). Instituto de Tecnología (INTEC).
Buenos Aires, Argentina
{anmartinez,pinchausti}@uade.edu.ar

Abstract. In recent years, the severity of forest fires has reached worrying levels both internationally and nationally. However, thanks to the advance of technology, it is possible to predict forest fires occurrence and magnitude through Machine Learning models specially developed for this purpose. To achieve this goal, this paper describes the development of an automated data pipeline in the Python programming language that generates a forest fires dataset specific to Pinamar area, thus allowing the subsequent training of predictive fire models. It is also configurable to gather meteorological, topographical and fuel data from other geographical areas.

Keywords: incendios forestales · medio ambiente· datos abiertos· machine learning· remote sensing

1 Introducción

Por definición un incendio forestal es un incendio no controlado en zonas cubiertas por vegetación y originado por causas naturales o humanas. Las consecuencias ambientales, económicas y sociales que estos provocan en el mundo [8] llevaron a gobiernos e investigadores a estudiar diversas maneras de prevenirlos, sobre todo en los últimos tiempos donde cada vez se torna más difícil controlarlos. En efecto, las pérdidas que ocasionan los incendios forestales en el país a nivel tanto ecológico como económico y social han obligado a distintas entidades gubernamentales a destinar en los últimos años grandes sumas de dinero tanto para prevenirlos a través de campañas de concientización [3], como para paliar las consecuencias económicas en las diversas regiones donde se han producido [2]. Tan solo en el 2020 en Argentina se quemaron más de 1,1 millones de hectáreas a causa de incendios forestales [7]. Además, se registraron en ese año más de 74.113 focos activos, una cifra récord que representa un incremento del 251,9% con respecto al año anterior [12].

Esta tendencia nacional también se vio reflejada en distintos puntos del país. La ciudad balnearia de Pinamar, conocida por sus extensos bosques de pino y dunas de arena, ha perdido según datos de incendios provistos por los bomberos locales más de 3,5 kilómetros cuadrados de bosques en los últimos seis años a causa de incendios forestales. Esta cifra representa aproximadamente el 10%

Received January 2023 Accepted April 2023; Published May 2023

de la superficie total cubierta por vegetación del partido, confirmando de esta manera la tendencia creciente de incendios forestales en la zona detectada por bomberos y autoridades locales.

Si bien en Argentina el 95% de los incendios forestales son originados por el hombre [1], la magnitud y desarrollo de estos dependen en gran medida de las condiciones climáticas específicas al momento y lugar donde se desarrollan. En efecto, el Servicio Nacional de Manejo del Fuego elabora diariamente mapas de peligro de incendio en base a las condiciones meteorológicas previstas con el objetivo de alertar a la población y contribuir a la prevención de incendios originados por el hombre.

En línea con diversos trabajos realizados con el fin de determinar la relación que existe entre la superficie final quemada a causa de incendios forestales y las condiciones ambientales circundantes, se desarrollaron diversos modelos de *Machine Learning* (ML) que contribuyan a la prevención de incendios forestales estableciendo como área de interés a la ciudad de Pinamar, provincia de Buenos Aires. Para ello fue necesario construir el *dataset* de incendios forestales de la zona que describa las condiciones topográficas, climáticas y ambientales que se dieron durante el desarrollo de cada incendio para así lograr inferir patrones que permitan prevenirlos.

En las próximas secciones se detalla la arquitectura del sistema que posibilitó reunir datos meteorológicos, topográficos y de combustible geográfica y temporalmente acotados para posteriormente generar el *dataset* requerido en la fase de entrenamiento de modelos predictivos.

2 Datos

Al iniciarse un incendio forestal existen factores como el área, la velocidad de propagación, y la longitud y altura de las llamas que modifican el comportamiento del mismo, permitiendo a los bomberos evaluar qué tácticas utilizar para combatirlos. Sin embargo, los factores ambientales también influyen en el comportamiento del fuego y se resumen en el denominado "Triángulo de comportamiento del fuego", compuesto por la meteorología, la topografía y el combustible. Considerando estos factores, resulta necesaria la previsión de las condiciones favorables de ocurrencia para poder prevenir y combatir oportunamente incendios forestales. Por esta razón se han desarrollado diversos índices de peligro de incendio que toman como variables algunos de los factores mencionados anteriormente.

Particularmente Argentina utiliza el Índice Meteorológico de Peligro de Incendio o FWI (por sus siglas en inglés *Forest fire Weather Index)* [9]. Este índice se calcula a partir de distintas variables meteorológicas (la temperatura, humedad relativa, precipitaciones de las últimas 24 horas y velocidad del viento) con el objetivo de describir el contenido de humedad de los distintos tipos de combustibles y el efecto que el viento produce sobre el comportamiento del fuego [17].

Para determinar el riesgo de incendio el FWI se vale de cinco componentes relacionados jerárquicamente [9, 16]. Los primeros tres son códigos que representan categorías de humedad en combustibles, siendo los mismos:

- Código de humedad de combustibles finos (o FFMC por sus siglas en inglés *Fine Fuel Moisture Code*): estima la humedad de los combustibles finos.
- Código de humedad del mantillo (DMC por sus siglas *Duff Moisture Code*): estima cuán húmeda es la materia orgánica que se encuentra en los primeros 7 centímetros de suelo.
- Código de sequía (DC por sus siglas en inglés *Drought Code*): estima la humedad de la materia orgánica a mayor profundidad (más de 18 centímetros).

Una vez obtenidos los códigos descritos se calculan los siguientes índices intermedios que caracterizan el comportamiento del fuego:

- Velocidad de propagación del incendio (o ISI por sus siglas en inglés *Initial Spread Index*): estima el potencial de propagación en base a la velocidad del viento y el FFMC.
- Combustible disponible (o BUI por sus siglas en inglés *Buildup Index*): estima el calor que producirían combustibles pesados combinando los códigos DMC y DC.

Por último, estos dos últimos índices se combinan linealmente para obtener el FWI, indicador de la intensidad potencial del fuego. Este es el índice que se utiliza para determinar el riesgo de incendio forestal, en donde un valor de FWI alto indica condiciones meteorológicas favorables para desencadenarlo.

Así como el FWI y sus índices derivados han demostrado tener influencia en la magnitud de los incendios forestales ocurridos [15], el Índice de Vegetación de Diferencia Normalizada o NDVI (por sus siglas en inglés *Normalized Difference Vegetation Index*) ha evidenciado ser un índice sumamente útil para observar la salud y existencia de vegetación en una zona dada. En efecto, según bomberos locales las zonas con mayor superficie cubierta por vegetación son aquellas más susceptibles de ocurrencia de incendios de gran magnitud.

El objetivo del sistema desarrollado es reunir a partir de diversas fuentes de datos los atributos mencionados y otros que en la literatura se han probado ser influyentes en la ocurrencia y magnitud de incendios forestales. En particular, para el área de interés establecida el sistema reúne por cada registro de incendio forestal los atributos descritos en la Tabla 1.

3 Arquitectura

El sistema en cuestión es un *pipeline* de datos desarrollado en Python y compuesto por cuatro fases para cada fuente de datos que se ejecutan de forma secuencial y reproducible, guardando en cada etapa parcial los datos con los que se trabajan para que la posterior etapa utilice como entrada. Estas etapas (Fig.1) consisten en la extracción de datos crudos (*extractors*), su transformación (*transformers*), procesamiento (*processors*) y posterior ingeniería de

Atributo	Descripción	Unidad de medida
Día	Día del mes	N/A
Mes	Mes del año	N/A
Día no laboral	Feriados y fines de semana	N/A
Hora	Hora de ocurrencia de incendio	N/A
X	Eje X (Longitud)	N/A
Y	Eje Y (Latitud)	N/A
LST	Temperatura de Superficie del Suelo	K (Kelvin)
NDVI	Índice de Vegetación de Diferencia Normalizada	N/A
Elevación	Elevación	m (Metros)
DC	Código de sequía	N/A
DMC	Código de humedad del mantillo	N/A
FFMC	Código de humedad de combustibles finos	N/A
ISI	Velocidad de propagación del incendio	N/A
BUI	Combustible disponible	N/A
FWI	Índice meteorológico de peligro de incendio	N/A
Temperatura	Temperatura (dato horario)	°C (Grados centígrados)
Humedad	Humedad relativa (dato horario)	% (Porcentaje)
Viento	Velocidad del viento (dato horario)	km/h (Kilómetros por hora)
Precipitaciones	Precipitaciones (últimas 24 horas)	mm (Milímetros)

Table 1. Atributos recopilados y generados por el sistema.

características (*feature engineering*). Una vez ejecutadas las cuatro etapas se genera el archivo fires.csv conteniendo los registros de incendios forestales con los atributos definidos en la Sección 2.

Dado que cada etapa tiene objetivos específicos, en las siguientes secciones se detallan los mismos.

3.1 Extracción

Para obtener los datos requeridos para generar el *dataset* de incendios forestales correspondientes a los años 2015 – 2019 se debieron consultar distintas fuentes. En primer lugar fue necesario obtener los registros de incendio del Partido de Pinamar. Estos registros fueron provistos por la Asociación de Bomberos Voluntarios de Pinamar (ABVP) como planillas mensuales en formato Excel. Cada registro contiene la fecha, hora y ubicación del siniestro, los recursos materiales y humanos requeridos en el combate del incendio, y la superficie total quemada medida en hectáreas.

De igual manera, para obtener los datos climáticos para los años en estudio se debió realizar un pedido de información meteorológica de la estación meteorológica de Villa Gesell al departamento del Centro de Información Meteorológica (CIM), perteneciente al Servicio Meteorológico Nacional (SMN). Esto fue necesario ya que actualmente la API del SMN provee los registros climáticos horarios de diversas estaciones meteorológicas a partir del 26 de noviembre de 2017. Entre los datos climáticos horarios que recopilan estas estaciones se encuentran la

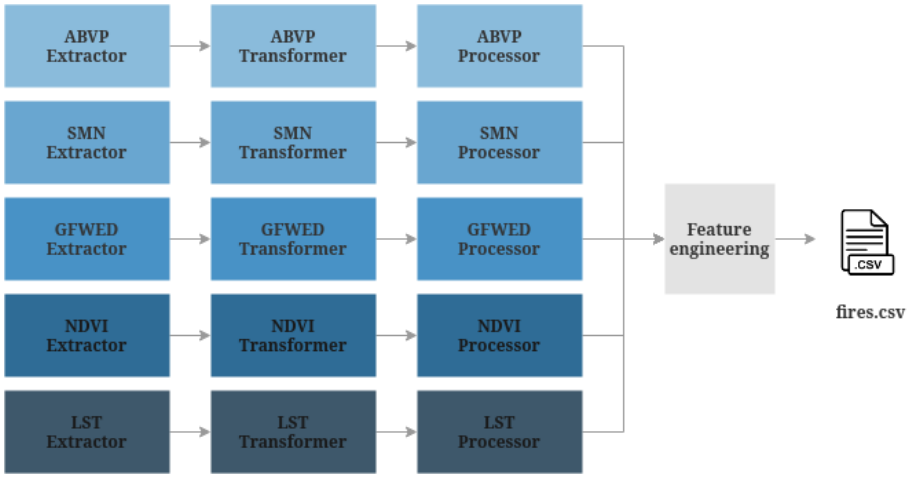

Fig. 1. Arquitectura conceptual del *pipeline* de datos.

dirección y velocidad del viento, las precipitaciones de las últimas 6 y 12 horas, el porcentaje de humedad y la temperatura.

En cuanto a los datos topográficos y de combustible, se recurrió a imágenes obtenidas de satélites de observación terrestre operados por la Administración Nacional de Aeronáutica y Espacio (NASA). A diferencia de las fuentes anteriores, los datos crudos provenientes de NASA abarcan escalas geográficas considerablemente mayores. En la mayoría de los casos se proveen archivos que contienen datos de todo el planeta, por lo que para el rango de años inicialmente considerado los datos crudos tienen un tamaño de más de 300 GB (*gigabytes*).

Tal como se puede apreciar en la Tabla 2, el *pipeline* de datos debe extraer de distintas fuentes datos heterogéneos que se almacenan para que en posteriores etapas se transformen a un formato que facilite su manipulación.

3.2 Transformación

Una vez extraídos los datos de las distintas fuentes, en la etapa de transformación se tiene como objetivo unificar la diversidad de formatos a un solo tipo (archivos CSV) y acotar la región geográfica a las coordenadas de interés (56° 57' 0" O, 37° 12' 0" S, 56° 48' 0" O, 37° 3' 0" S), reduciendo así el tamaño de los archivos y convirtiéndolos a un formato que facilite la manipulación de los mismos en posteriores etapas. Entre las tareas específicas que se llevan a cabo se destacan:

- Manipulación y conversión de archivos HDF utilizando H5py[1].
- Filtrado de datasets multidimensionales por coordenadas mediante xarray[2].

[1] `https://www.h5py.org/`
[2] `https://xarray.dev/`

Descripción	**Fuente**	**Resolución espacial**	**Resolución temporal**	**Formato**
Incendios forestales (Día, Mes, Día no laboral, Hora, Latitud, Longitud)	ABVP	N/A	Horaria	Hoja de cálculo de Microsoft Excel
Meteorología (Temperatura, Humedad, Viento, Precipitaciones)	SMN	N/A	Horaria	Hoja de cálculo de Microsoft Excel
Temperatura de superficie de suelo (LST)	NASA / MYD11C1 v006 [18]	0.05° × 0.05°	Diaria	Hierarchical Data Format
Índice de Vegetación de Diferencia Normalizada (NDVI)	NASA / MYD13Q1 v006 [5]	250m × 250m	Quincenal	NetCDF
Índice meteorológico de peligro de incendio e índices derivados (DC, DMC, FFMC, ISI, BUI, FWI)	NASA / GFWED GEOS-5 - GPM Late v5 [6]	0.1° × 0.1°	Diaria	NetCDF

Table 2. Origen y formato de los datos extraídos en el *pipeline* de datos.

– Manipulación y conversión de archivos XLSX con la librería Pandas[3].
– Creación y manipulación de matrices multidimensionales con NumPy[4].
– Geocodificación de direcciones mediante GeoPy[5].

En lo que respecta a registros de incendio, en esta etapa se descartan aquellos registros que tengan atributos faltantes que son requeridos para el entrenamiento de algoritmos de ML, como la ubicación o la fecha de ocurrencia. A su vez, se eliminan los registros que corresponden a incendios ocurridos en ubicaciones que no pertenecen al Partido de Pinamar.

3.3 Preprocesamiento

En la etapa de procesamiento se busca, en primer lugar, reunir todos los atributos para caracterizar cada registro de incendio forestal según su ubicación y fecha de ocurrencia. En segundo lugar, en este paso se establece la granularidad geográfica con la que se trabajarán los datos reunidos y transformados en etapas anteriores.

En varios trabajos realizados en el ámbito de predicción de incendios forestales el área de interés es descrita mediante una grilla en donde se establecen regiones de determinado tamaño, generalmente con una escala de kilómetros [13, 19, 4]. Esta grilla es necesaria para traducir las coordenadas geográficas de los incendios históricos en coordenadas (X, Y) que sean más fáciles de interpretar por un modelo de ML. Para el área de interés de este trabajo el tamaño de esta grilla fue de 250 metros $\times$ 250 metros como se puede apreciar en la Fig. 2, ya que es la resolución mínima disponible en los atributos (NDVI) y el área es representativa considerando la superficie y las características topográficas del Partido de Pinamar.

3.4 Análsis de datos

Si bien esta etapa requiere de un análisis manual de los datos procesados hasta este punto, el análisis de los datos es fundamental para entender la distribución de los datos, sus relaciones y encontrar patrones subyacentes. Consecuentemente se elaboraron en la aplicación Jupyter[6] distintos gráficos para adquirir entendimiento cualitativo de los incendios forestales ocurridos entre los años 2015 y 2019. Gracias a estas visualizaciones (Fig. 3 y Fig. 4) se llegaron a las siguientes conclusiones:

– Durante las primaveras y los veranos los incendios forestales son más peligrosos ya que la superficie quemada es mayor.
– Desde las 8 de la mañana hasta las 3 de la tarde los incendios ocurridos han quemado más hectáreas.

[3] `https://pandas.pydata.org/`
[4] `https://numpy.org/`
[5] `https://geopy.readthedocs.io/en/stable/`
[6] `https://jupyter.org/`

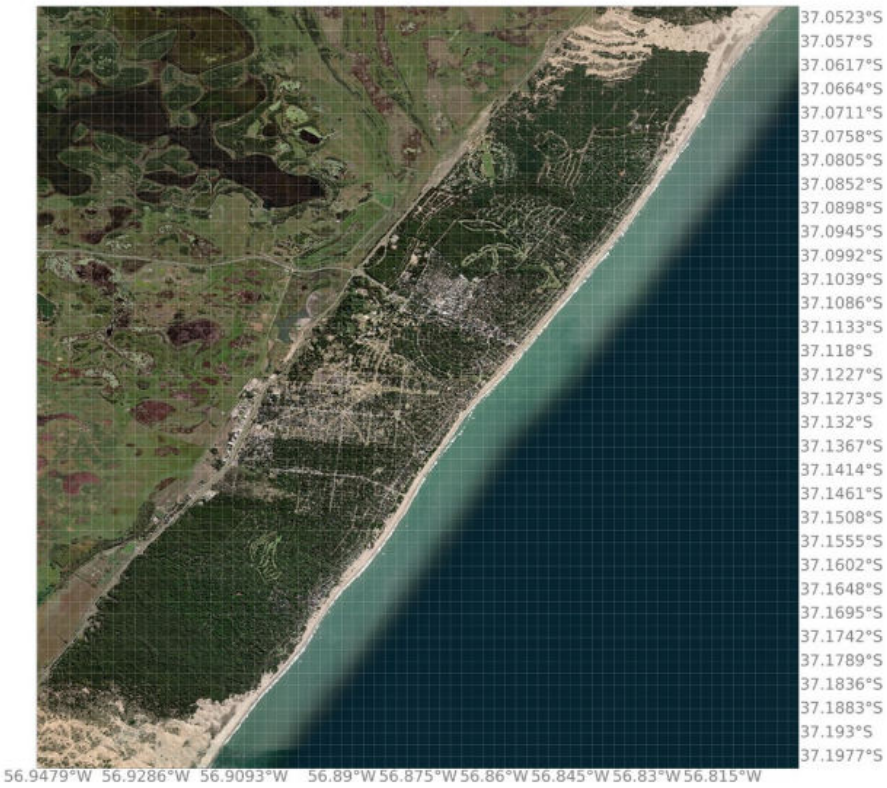

Fig. 2. Grilla establecida para el área de interés.

– La mayor cantidad de incendios forestales ocurrieron durante la tarde (12 del mediodía a 6 de la tarde).
– El 93% de los incendios forestales han quemado menos de 1 hectárea.
– La mayoría de los incendios forestales se produjeron en las localidades de Ostende y Valeria del Mar.

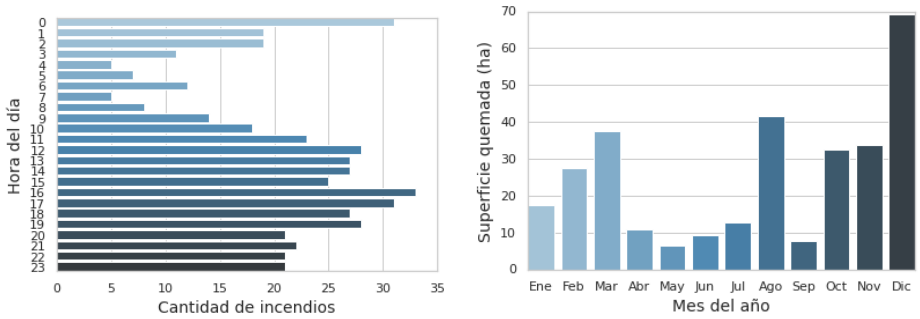

Fig. 3. Cantidad de incendios ocurridos según la hora del día (izquierda) y magnitud de incendios según el mes del año (derecha).

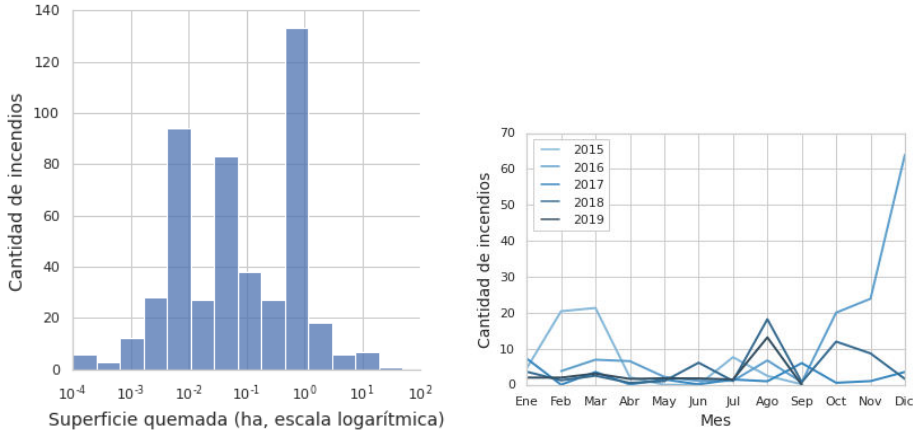

Fig. 4. Cantidad de incendios según magnitud (izquierda) y según el mes y año (derecha).

Por otro lado, el *dataset* que se genera en esta etapa permite confirmar ciertas tendencias comúnmente observadas por bomberos de todo el país en lo que respecta a la relación entre las condiciones meteorológicas (Fig. 5) y de combustible

(Fig. 6) con la magnitud de los incendios forestales. Particularmente, los incendios forestales de mayor magnitud de Pinamar (superficie quemada mayor a las 5 hectáreas) han ocurrido en días secos donde no se registraron precipitaciones en las últimas 24 horas. Si bien se espera que los días secos y calurosos también sean propicios para el desarrollo de incendios más devastadores, a partir del *dataset* se puede inferir que la humedad y temperatura no inciden proporcionalmente con la superficie final quemada por incendios forestales.

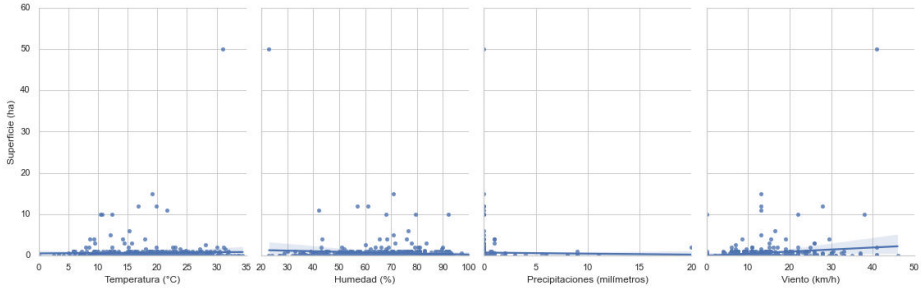

Fig. 5. Relación entre la superficie quemada en hectáreas por incendios forestales y distintas variables climáticas.

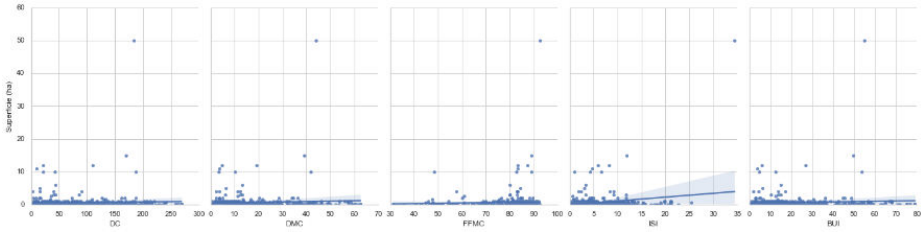

Fig. 6. Relación entre la superficie quemada en hectáreas por incendios forestales y los componentes del índice meteorológico de peligro de incendio.

3.5 Feature engineering

Como consecuencia del conocimiento adquirido sobre el *dataset* en la etapa de Análisis de Datos, en la etapa de *feature engineering* se determinan aquellos atributos que serán relevantes a la hora de predecir incendios, se aplican transformaciones sobre ellos y se crean nuevos atributos a partir de los existentes. En particular, para facilitar el trabajo con los distintos atributos los mismos se agruparon para representar el tipo de variable que describen tal como se especifica en la Tabla 3. De esta manera, el *pipeline* admite parámetros que permitan

seleccionar los atributos a incluir en el *dataset* final. En consecuencia, las tareas de entrenar y evaluar modelos predictivos intercambiando diferentes tipos de variables se hacen de manera más ordenada y trazable a lo largo de las distintas ejecuciones.

Tipo de variable	**Atributos**
Temporal	Mes, Día, Hora, Día no laboral
Espacial	X, Y
Topográfica	Elevación
Combustible	DC, DMC, FFMC, ISI, NDVI
Meteorológica	Temperatura, Humedad, Viento, Precipitaciones

Table 3. Agrupamiento de atributos según el tipo de variable.

Con el objetivo de llevar a cabo las tareas descriptas para esta etapa se utilizó una fracción del *dataset* creado (30%) que fue separado para no ser utilizado al entrenar los modelos predictivos. De esta manera se busca evitar un estado de *overfitting* o sobreajuste en los modelos entrenados, dando lugar a predicciones poco precisas y estimaciones excesivamente optimistas [11].

Habiendo identificado y seleccionado el conjunto de características con el que en la posterior etapa se entrenarán los modelos, se creó una nueva característica correspondiente al día de la semana en que se produjeron los incendios. Asimismo, se agruparon las horas en que se desarrollaron incendios por momentos del día (mañana, tarde y noche), ya que al analizar los datos se detectó una variación en la cantidad de incendios que se producen cada 8 horas.

En lo que respecta a transformación de características, se aplicó la técnica de *One-Hot-Encoding* a las variables categóricas Momento del día, Día y Mes para convertirlas en numéricas, creando un vector de características por categoría donde cada uno contiene bits que representan la pertenencia o no a dicha categoría [20]. La elección de este método por sobre otros tales como *Integer Encoding* está justificada por el hecho de que *OneHot-Encoding* no representa relaciones ordinales entre categorías que puedan ser incorrectamente modeladas por un modelo de ML.

Por último, en esta etapa se establecieron valores por defecto a los registros de incendio con una o más características faltantes. En el caso particular del *dataset* de incendios forestales del Partido de Pinamar la única característica faltante en algunos registros es el NDVI. Dado que este índice representa el índice de verdor o salud de la vegetación en un par de coordenadas dadas, la ausencia de este valor implica también ausencia de vegetación. Por esta razón, la transformación llevada a cabo consistió en convertir los valores nulos por ceros.

4 Resultados

En el marco del presente trabajo se configuró el *pipeline* para obtener como resultado el *dataset* de incendios forestales (Fig. 7) para el área descrita por las coordenadas 56° 57' 0" O, 37° 12' 0" S, 56° 48' 0" O, 37° 3' 0" S y desde el 01/01/2015 hasta el 01/01/2020. El mismo requirió procesar 302.95 GB de datos crudos para un área de 68 kilómetros cuadrados.

x	y	dia no labori	temperatura	humedad	viento	precipitacion	lst	elevacion	dc	dmc	ffmc	isi	bui	fwi	ndvi	Superficie
26	35	0	26.55	75	12.5	0	300.26	13.99965	44.479923	13.657989	86.80505	3.6837015	15.45328	5.1182256	0.54089999	0.8
24	36	0	24.8	83	6	0	300.26	8.74283981	44.479923	13.657989	86.80505	3.6837015	15.45328	5.1182256	0.39430001	1.2
31	40	0	21.8	77	10	0	303.88	19.1897507	30.081812	10.990948	86.73585	5.2334514	11.488267	6.1005793	0.64099997	0.05
42	34	1	22.1	91.5	9.5	6	305.32	10.6517363	7.44247	3.9052541	57.890266	1.6193064	3.7804859	0.62112325	0.38209999	0.08
32	40	1	22.1	91.5	9.5	6	305.32	17.995533	7.44247	3.9052541	57.890266	1.6193064	3.7804859	0.62112325	0.6049	0.002
31	40	1	21.9	91	13	0	305.32	19.1897507	7.44247	3.9052541	57.890266	1.6193064	3.7804859	0.62112325	0.64099997	0.06
32	40	1	23.4	74	24	0	304.38	17.995533	22.360628	6.599125	83.01819	4.0297456	7.594776	3.7070892	0.6049	0.08
39	41	1	22.5	71	22	0	304.38	17.2795448	22.360628	6.599125	83.01819	4.0297456	7.594776	3.7070892	0.44479999	0.08
41	48	0	24.45	67.5	8.5	0	308.16	12.0616665	84.7967	13.161542	88.287346	11.574877	18.964317	15.333416	0.63330001	0.05
46	46	0	26.7	56	11	0	308.16	18.9806347	84.7967	13.161542	88.287346	11.574877	18.964317	15.333416	0.5643	0.08
10	18	1	21.55	67	6.5	0.4	304	12.3574419	22.52976	7.004245	83.42087	2.551946	7.882242	2.0920477	0.69620001	0.01
46	44	1	24.2	73.5	9.5	0	304.44	17.4103794	29.552475	9.361469	85.96932	4.645797	10.448441	5.1564493	0.63810003	0.03
51	47	0	25.7	83	15	0	300.26	11.165144	45.72546	13.804427	86.80631	3.6843605	15.733826	5.1751933	0.72060001	0.007
34	31	0	28.9	67	19	0	300.26	12.5392952	44.479923	13.657989	86.80505	3.6837015	15.45328	5.1182256	0.55760002	0.008
31	40	0	24.5	77	26	0	300.26	19.1897507	41.40625	9.6695175	75.069855	1.7914311	12.210375	1.7131221	0.65240002	0.006
25	32	0	24.3	89	13	0	300.26	14.7812281	41.40625	9.6695175	75.069855	1.7914311	12.210375	1.7131221	0.47	0.9
31	40	0	21.6	74	10	0	300.9	19.1897507	56.12544	15.008098	88.16141	6.940873	17.989853	9.933572	0.65240002	0.09
24	36	0	21.6	74	10	0	300.9	8.74283981	56.12544	15.008098	88.16141	6.940873	17.989853	9.933572	0.39430001	0.034
24	36	0	21.6	74	10	0	300.9	8.74283981	56.12544	15.008098	88.16141	6.940873	17.989853	9.933572	0.39430001	0.09

Fig. 7. *Dataset* final de incendios forestales creado por el *pipeline* de datos. Para una mejor visualización se omitieron las columnas generadas al aplicar la técnica de *One-Hot-Encoding.*

El *dataset* generado está conformado por 597 registros de incendios forestales ocurridos en el Partido de Pinamar durante los años 2015 – 2019 caracterizados por las condiciones climáticas, topográficas y de combustible del día y hora de ocurrencia. Cabe destacar que con la arquitectura propuesta agregar nuevas fuentes de datos -y, por ende, nuevos atributos- en futuras versiones es una tarea sencilla. Esto es posible ya que por cada atributo se cuenta con un extractor, transformador y procesador independiente que es referenciado en tiempo de ejecución tal como se indica en el Algoritmo 1.

A partir del dataset obtenido, fue posible desarrollar modelos que permiten predecir la ocurrencia (problema de clasificación binaria) y magnitud (problema de regresión) de incendios forestales en la zona utilizando como variable objetivo la superficie quemada en hectáreas. La adopción de este enfoque fue necesaria ya que a partir del análisis de datos se pudo concluir que la distribución de la magnitud de incendios forestales es altamente despareja (Fig. 3 y Fig. 4): la mayoría de los incendios forestales ocurridos en Pinamar han sido pequeños, aunque representan el 37% de la superficie quemada históricamente. Por el contrario, el 61% de la superficie quemada fue producida por el 6% del total de incendios forestales. Esta disparidad afecta directamente el rendimiento de modelos que, en este tipo de distribuciones, se verán influenciados por lo que se refleja en la mayoría de los casos, ya sea la cantidad de incendios o el promedio de superficie quemada.

La metodología para predicir incendios forestales consistió, en primer lugar, en predecir por cada coordenada de la grilla si se producirá o no un incendio fore-

Algorithm 1 *Pipeline* de datos de incendios forestales

```
dataset ← initializeDataset()
task ← getParam("task")  ▷ Valores posibles: "extract", "transform", "process", "engineer", "all"
coordinates ← getParam("coordinates")  ▷ Formato: [longitud_1, latitud_1, longitud_2, latitud_2]
from ← getParam("from")  ▷ Fecha desde. Formato: %d-%m-%Y
to ← getParam("to")  ▷ Fecha hasta. Formato: %d-%m-%Y
if task = "all" then
    dataSources ← getParam("dataSources")  ▷ Valores posibles: "lst", "ndvi", "gfwed", "smn", "abvp"
    for dataSource in dataSources do
        extractor ← getExtractor(dataSource)
        transformer ← getTransformer(dataSource)
        processor ← getProcessor(dataSource)
        extractor.extract(from, to, coordinates)
        transformer.transform(coordinates)
        processor.process(dataset)
    end for
    joinAttributes(dataset)
    engineer(dataset)
    export(dataset)
end if
```

stal (clasificación binaria); y en segundo lugar, para aquellas coordenadas en las que se haya predicho la ocurrencia de un incendio forestal (con una probabilidad mayor a 0.5), predecir las hectáreas que podrían quemarse (regresión). A partir de esta división, dependiendo del tipo de problema los siguientes pasos difieren. Por un lado, para el problema de clasificación se reemplazan los registros cuyo valor de superficie quemada es mayor a 0 hectáreas por 1 (indicando ocurrencia de incendio). Luego, como los datos que se cuentan corresponden enteramente a la clase 1 (Incendio), se genera aleatoriamente una cantidad proporcionada de registros de la clase 0 (No Incendio) (Algoritmo 2) en concordancia con algoritmos propuestos en la literatura [14]. El objetivo de este tipo de algoritmos es que las ubicaciones de los puntos en donde se produjeron incendios y donde no estén espacial y temporalmente relacionados con los atributos.

Por otro lado, para abordar el problema de regresión se aplica una transformación logarítmica sobre la variable objetivo (superficie quemada) para reducir la asimetría de la distribución. A continuación, en ambos problemas se optimizan los hiperparámetros de los distintos algoritmos seleccionados utilizando la técnica de búsqueda exhaustiva, y se entrenan modelos aplicando la técnica de validación cruzada utilizando 5 particiones sobre el conjunto de datos de entrenamiento. Por último, se evalúan los modelos entrenados a través de las métricas correspondientes al tipo de problema

Entre los algoritmos utilizados se entrenaron aquellos que mostraron ser efectivos en la literatura [10] como árboles de decisión, máquinas de vectores soporte

Algorithm 2 Generación de ejemplos negativos ("no incendio")

$dias \leftarrow 3$
$kilometros \leftarrow 1.5$
$P \leftarrow \varnothing$ ▷ Puntos de "no incendio"
for $puntoDeIncendio$ **in** $incendios$ **do**
  $fechaIncendio \leftarrow incendio.fecha$
  $incendiosRecientes \leftarrow seleccionarIncendiosEnLapso(dias, fechaIncendio)$
  $R \leftarrow \varnothing$ ▷ Región donde se produjeron incendios
  for $incendioReciente$ **in** $incendiosRecientes$ **do**
    $coordenadas \leftarrow incendioReciente.coordenadas$
    $R \leftarrow R \cup crearRegionEn(kilometros, coordenadas)$
  end for
  $P \leftarrow P \cup seleccionarCoodenadasEnArea(R')$
end for

(SVM) y redes neuronales artificiales, tratando cada problema de predicción por separado. Como resultado, mediante un árbol de decisión se obtuvo el mejor resultado para predecir la ocurrencia de incendios con una sensibilidad del 88.4% (Tabla 4). Por otro lado, a través de una red neuronal artificial que predice la magnitud de un incendio forestal en hectáreas se obtuvo un error cuadrático medio (RMSE) de 0.178 ha (Tabla 5). En este último caso se entrenaron dos redes neuronales con distintas arquitecturas: mientras la red n°1 tiene 18 nodos en la capa oculta n°1 y 4 en la n°2, la red n°2 posee 9 y 18 respectivamente. Por otro lado, la red n°1 utiliza como función de activación la tangente hiperbólica, mientras que la n°2 utiliza la ReLu.

Modelo	**Exactitud**	**Precisión**	**Sensibilidad**	**F1 score**
ANN	74.2%	71.6%	79.3%	75.3%
SVM	61.5%	62.4%	56.2%	59.1%
GB	77.9%	77.7%	77.7%	77.7%
LR	62.3%	63.6%	56.2%	59.7%
DT	80.7%	76.4%	**88.4%**	82%
RF	**82.4%**	**82%**	82.6%	**82.3%**

Table 4. Resultados de la evaluación de modelos de predicción de incendios forestales.

Habiendo entrenado estos modelos, se pudo analizar cuáles son las variables que demostraron ser importantes a la hora de explicar la ocurrencia de incendios forestales en Pinamar. Correspondientemente a la experiencia manifestada por bomberos locales, la ubicación es un factor importante ya que los incendios son más comunes en zonas de alta densidad poblacional. En efecto, la variable longitud tuvo una importancia del 83%. En segundo lugar, la variable elevación resultó en una importancia del 8%. Por último, el NDVI obtuvo una importancia del 6%. Estos resultados demuestran que el componente humano es el más

Modelo	MAE	RMSE
ANN (n°1)	0.295	**0.178**
ANN (n°2)	**0.255**	0.215
SVM	0.274	0.438
DT	0.294	0.46
RF	0.299	0.448

Table 5. Resultados de la evaluación de modelos de predicción de superficie quemada.

importante a la hora de predecir incendios en Pinamar, por lo que en futuras líneas de investigación variables de hábitat e infraestructura social podrían ser incorporadas en el *pipeline*.

5 Conclusiones

En el presente trabajo se desarrolló un sistema que permite extraer, transformar y procesar de distintas fuentes datos meteorológicos, topográficos y de combustible para crear un *dataset* de incendios forestales del Partido de Pinamar. Si bien el *pipeline* de datos fue concebido para un área geográficamente acotada, el mismo fue diseñado con miras a extender el área de interés a la región circundante, por lo que admite configuraciones que permitan abarcar áreas más extensas (modificando coordenadas geográficas) y lapsos de tiempo más amplios.

Gracias al *pipeline* fue posible desarrollar y entrenar distintos modelos de ML que predigan la ocurrencia y magnitud de incendios forestales en la zona, obteniendo resultados prometedores. Asimismo, la posibilidad de seleccionar un subconjunto de atributos permitió obtener distintos *datasets* que acompañen a la experimentación con varias combinaciones de atributos y algoritmos durante el entrenamiento de modelos.

Como futuras líneas de trabajo se planifica agregar al *pipeline* otros atributos meteorológicos y poblacionales que contribuyan a explicar la ocurrencia y magnitud de incendios forestales. En particular, para el Partido de Pinamar la dirección del viento permitiría estudiar si la predominancia de vientos del sudeste es un factor determinante durante el desarrollo de incendios forestales. Por otro lado, considerando que según los bomberos locales todos los incendios forestales del partido son causados por el hombre, la afluencia de turistas a la zona y la densidad poblacional son variables que aportarían información valiosa a las autoridades locales ya que sería posible inferir patrones de comportamiento estacionales que, sumados a las condiciones climáticas y ambientales, propician la ocurrencia de incendios forestales.

Agradecimientos. Los autores agradecen a la Universidad Argentina de la Empresa (UADE) y al Instituto de Tecnología (INTEC) por el apoyo brindado en el presente trabajo realizado en el marco del Proyecto Final de Ingeniería en Informática “AQUA: Desarrollo de un modelo de Machine Learning para

prevenir incendios forestales en Pinamar", articulado en el ACyT "Aplicaciones de Machine Learning para mejorar el uso de Recursos Naturales" (A21T03).

References

1. Argentina.gob.ar: ¿Cómo se originan los incendios? (Mar 2018), https://www.argentina.gob.ar/ambiente/fuego/conocemas/origen
2. Argentina.gob.ar: Asistencia financiera a PYMEs y emprendedores afectados por los incendios de Chubut (Apr 2021), https://www.argentina.gob.ar/noticias/asistencia-financiera-pymes-y-emprendedores-afectados-por-los-incendios-de-chubut
3. Argentina.gob.ar: El Gobierno nacional lanzó una campaña de prevención de incendios forestales (Apr 2021), https://www.argentina.gob.ar/noticias/el-gobierno-nacional-lanzo-una-campana-de-prevencion-de-incendios-forestales
4. Cortez, P., Morais, A.: A Data Mining Approach to Predict Forest Fires using Meteorological Data (Jan 2007)
5. Didan, K.: MYD13Q1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 (2015). https://doi.org/10.5067/MODIS/MYD13Q1.006, https://lpdaac.usgs.gov/products/myd13q1v006/, type: dataset
6. Field, R.D., Spessa, A.C., Aziz, N.A., Camia, A., Cantin, A., Carr, R., de Groot, W.J., Dowdy, A.J., Flannigan, M.D., Manomaiphiboon, K., Pappenberger, F., Tanpipat, V., Wang, X.: Development of a Global Fire Weather Database. Natural Hazards and Earth System Sciences **15**(6), 1407–1423 (Jun 2015). https://doi.org/10.5194/nhess-15-1407-2015, https://nhess.copernicus.org/articles/15/1407/2015/, publisher: Copernicus GmbH
7. de Manejo del Fuego, S.N.: Reportes diarios del Servicio Nacional de Manejo del Fuego correspondientes a 2020. (Nov 2020), https://www.argentina.gob.ar/ambiente/fuego/reporte-2020
8. González-Cabán, A.: The Economic Dimension of Wildland Fires. pp. 229–237 (Jan 2013)
9. Group, N.W.C.: Fire Weather Index (FWI) System | NWCG, https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system
10. Jain, P., Coogan, S.C., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D.: A review of machine learning applications in wildfire science and management. Environmental Reviews **28**(4), 478–505 (2020). https://doi.org/10.1139/er-2020-0019, https://doi.org/10.1139/er-2020-0019
11. Johnson, M.K.a.K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. http://www.feat.engineering/
12. Queimadas, P.: Estatísticas dos estados e regiões - Programa Queimadas - INPE, https://queimadas.dgi.inpe.br/queimadas/portal-static/estatisticas_paises/
13. Rodrigues, M., de la Riva, J.: An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environmental Modelling & Software **57**, 192–201 (Jul 2014). https://doi.org/10.1016/j.envsoft.2014.03.003, https://www.sciencedirect.com/science/article/pii/S1364815214000814
14. Stojanova, D., Kobler, A., Ogrinc, P., Ženko, B., Džeroski, S.: Estimating the risk of fire outbreaks in the natural environment. Data Mining and Knowledge Discovery **24**(2), 411–442 (Mar 2012). https://doi.org/10.1007/s10618-011-0213-2, https://doi.org/10.1007/s10618-011-0213-2

15. Varol, T., Ertuğrul, M.: Determining the relationship the burned areas and the number of fires and drought index: sample of Antalya (Oct 2016)
16. Villers, L., Chuvieco, E., Aguado, I.: Aplicación del índice meteorológico de incendios canadiense en un Parque Nacional del centro de México. Revista Mexicana de Ciencias Forestales **3**, 25–40 (Jun 2012). https://doi.org/10.29298/rmcf.v3i11.515
17. Waidelich, S., Zimmerman, V., Laneri, K., Denham, M.M.: Fire Weather Index assessment and visualization (2019), http://sedici.unlp.edu.ar/handle/10915/90905
18. Wan, Z., Hook, S., Hulley, G.: MYD11C1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006 (2015). https://doi.org/10.5067/MODIS/MYD11C1.006, https://lpdaac.usgs.gov/products/myd11c1v006/, type: dataset
19. Xie, Y., Peng, M.: Forest fire forecasting using ensemble learning approaches. Neural Computing and Applications **31**(9), 4541–4550 (Sep 2019). https://doi.org/10.1007/s00521-018-3515-0, https://doi.org/10.1007/s00521-018-3515-0
20. Zheng, A., Casari, A.: Categorical Variables. In: Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, pp. 77–81. Beijing Boston Farnham Sebastopol Tokyo (2018)