PROYECTO FINAL DE INGENIERÍA

AQUA: DESARROLLO DE UN MODELO DE MACHINE LEARNING PARA PREVENIR INCENDIOS FORESTALES EN PINAMAR

Martínez Saucedo, Ana Carolina – LU1078889

Ingeniería Informática

Tutor:
Inchausti, Pablo Ezequiel, UADE

1 de noviembre de 2021

UNIVERSIDAD ARGENTINA DE LA EMPRESA

FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS

Agradecimientos

En primer lugar quiero agradecer profundamente a mi tutor Mg. Pablo E. Inchausti, por su constante acompañamiento y excelente predisposición a lo largo del desarrollo del trabajo, y por todas las oportunidades que me brindó para potenciar el proyecto y crecer profesionalmente.

De igual manera quiero agradecer al Dr. Marcos A. Saucedo, a Patricio Silva, Matías García, Lucas León y a todos los bomberos de la Asociación de Bomberos de Voluntarios de Pinamar, ya que sin su ayuda, predisposición y paciencia para responder consultas no hubiese sido posible arribar a los resultados de este proyecto.

Por último, quiero agradecer a toda mi familia, mi principal motor y fuente de motivación para dar lo mejor de mí, por siempre acompañarme y apoyarme. Y a Dios, por siempre guiarme en el camino.

Resumen

En los últimos años, la severidad de los incendios forestales ha llegado a niveles preocupantes tanto a nivel internacional como nacional. No obstante, gracias al avance de la tecnología es posible predecir la ocurrencia y magnitud de estos a través de modelos de Machine Learning especialmente desarrollados para tal fin. En línea con diversas investigaciones realizadas en materia de predicción espaciotemporal de incendios forestales, en el presente trabajo el objetivo fue desarrollar un modelo de Machine Learning que contribuya a la prevención de incendios forestales en el Partido de Pinamar. Para ello, se desarrolló un *pipeline* que genera el *dataset* de incendios forestales específico a la zona y se entrenaron diversos modelos predictivos, alcanzando una sensibilidad del 88.4% para predecir la ocurrencia de incendios forestales en Pinamar. Asimismo, se desarrolló una aplicación web para que bomberos y autoridades gubernamentales locales puedan interpretar fácilmente las predicciones de incendio y en consecuencia tomar decisiones para prevenirlos. De esta manera, a través de este trabajo se sentaron las bases necesarias para poder ampliar el área de predicción de incendios forestales a localidades vecinas.

Palabras clave: *machine learning, aprendizaje supervisado, incendios forestales, medioambiente.*

Abstract

In recent years, the severity of forest fires has reached worrying levels both internationally and nationally. Nonetheless, thanks to the advancement of technology it is now possible to predict their occurrence and magnitude through Machine Learning models specially developed for this purpose. In line with various investigations carried out in the field of spatiotemporal prediction of forest fires, in the present work the objective was to develop a Machine Learning model that contributes to the prevention of forest fires in the Pinamar district. Consequently, a data pipeline which generates the dataset of forest fires was developed and several predictive models were trained, achieving an 88.4% recall for forest fires occurrence prediction in Pinamar. Likewise, a web application was developed so that firefighters and local government authorities can easily interpret fire predictions and consequently make decisions to prevent them. In this way, through this work the necessary foundations were laid to be able to expand the forest fire prediction area to neighboring towns.

*Keywords****: machine learning, supervised learning, forest fires, environment.***

Contenidos

1. Introducción

Por definición un incendio forestal es un incendio no controlado en zonas cubiertas por vegetación y originado por causas naturales o humanas. Asimismo, las consecuencias ambientales, económicas y sociales que estos provocan en el mundo (Armando González-Cabán, 2013) 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.

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 (Argentina.gob.ar, 2021a), como para paliar las consecuencias económicas en las diversas regiones donde se han producido (Argentina.gob.ar, 2021b). Tan solo en el 2020 en Argentina se quemaron más de 1,1 millones de hectáreas a causa de incendios forestales (Servicio Nacional de Manejo del Fuego, 2020). Además, se registraron en el pasado año más de 74.113 focos activos, una cifra récord que representa un incremento del 251,9% con respecto al año anterior (Instituto Nacional de Pesquisas Espaciais, 2021).

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% 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 (Argentina.gob.ar, 2018b), la magnitud y desarrollo de los mismos dependen en gran medida de las condiciones climáticas específicas al momento y lugar donde se desarrollan (National Geographic, 2019). En efecto, el Servicio Nacional de Manejo del Fuego elabora diariamente mapas de peligro de incendios 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 (Argentina.gob.ar, 2018d).

1.1. Objetivos y alcance

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, el presente trabajo tiene como objetivo desarrollar un modelo predictivo de Machine Learning que contribuya a la prevención de incendios forestales centrando su atención a la ciudad de Pinamar, provincia de Buenos Aires, durante el año 2021. Para cumplir con este objetivo se plantean como objetivos específicos:

- Entrevistar a diversos especialistas del área con el fin de validar la solución y obtener información para abordar el desarrollo del modelo predictivo. Los objetivos deben estar redactados con un solo verbo en infinitivo
- Recopilar y procesar los datos de incendios históricos, climáticos y satelitales. con el que se construya el pipeline de datos
- Desarrollar un modelo de Machine Learning para prevenir incendios forestales.
- Desarrollar una aplicación web para visualizar las predicciones de incendio.

Los objetivos aquí planteados se integraron a un proyecto de investigación de mayor alcance, denominado *"A21T03 – Aplicaciones de Machine Learning para mejorar el uso de Recursos Naturales"* que, gracias al Instituto de Tecnología (INTEC) de la Universidad Argentina de la Empresa (UADE), posibilita la extensión y continuación del presente trabajo.

1.2. Estructura del documento

El presente trabajo está estructurado en distintas secciones. En la primera, *Antecedentes*, se describen los trabajos e investigaciones existentes con respecto a modelos de predicción de incendios forestales, analizando los avances realizados en el ámbito de estudio. Para comprender estos avances en el *Marco Teórico* se introducen los conceptos y la base teórica necesarios. A continuación, en el apartado *Estado del Arte* se describen los trabajos, modelos, investigaciones y proyectos desarrollados con el objetivo de determinar los aportes que el modelo propuesto realiza en la materia. Por último, en *User Research* se presentan los resultados de entrevistas realizadas a diversos especialistas del área con el propósito de validar la solución propuesta. En la segunda sección, *Descripción,* se detallan las características técnicas de este trabajo, profundizando en la arquitectura general de la solución, la justificación

a distintas decisiones de desarrollo y en cada uno de los componentes que conforman AQUA. En la posterior sección de *Metodología de Desarrollo* se explica la forma en que se desarrolló el presente trabajo, explicitando las herramientas utilizadas y el marco de desarrollo adoptado para cada componente del trabajo. Posteriormente se presentan los resultados de pruebas realizadas a la aplicación de visualización de predicciones y de los distintos modelos desarrollados y entrenados en la sección *Pruebas realizadas*. A continuación, se describe el análisis económico y financiero llevado a cabo para determinar la rentabilidad del proyecto. En la sección *Discusión* se comentan las dificultades encontradas al desarrollar el trabajo, y finalmente en la última sección *Conclusión* se recapitula acerca del trabajo realizado y el futuro de AQUA como proyecto.

2. Antecedentes

La predicción de incendios forestales a través de modelos de Machine Learning ha sido motivo de estudio desde hace ya tres décadas (Vega Garcia et al., 1996). Sin embargo, con el auge actual en los campos de Inteligencia Artificial y Big Data cada vez más investigadores han presentado nuevos modelos con resultados prometedores, realizando grandes aportes en la materia.

2.1. Marco Teórico

En esta sección se desarrollan en primer lugar los conceptos más importantes de Machine Learning y algunos algoritmos que fueron utilizados por investigadores y desarrolladores para crear modelos de predicción de incendios forestales. En segundo lugar, se definen conceptos relacionados con los incendios y los distintos factores ambientales que influyen en el origen y desarrollo de un incendio forestal.

2.1.1. Machine Learning

El aprendizaje automático o Machine Learning (ML) es una rama de la Inteligencia Artificial definida como "*el conjunto de métodos que pueden detectar patrones en los datos de forma automática y luego utilizar estos patrones descubiertos para predecir datos futuros o realizar otros tipos de decisiones bajo condiciones inciertas"* (Murphy, 2012). A su vez, (Russell et al., 2009a) describe el propósito del ML como "*adaptarse a nuevas circunstancias y detectar y extrapolar patrones"*.

Efectivamente el potencial de ML radica en la capacidad que una computadora tiene para aprender de los datos que le son brindados sin necesidad de haber sido programadas explícitamente para tal fin. En particular se distinguen tres tipos de aprendizaje:

- Supervisado: en este caso se aprende la función que mejor ajuste las entradas o características (*features*) de un conjunto de ejemplos o instancias a la salida esperada.
- No supervisado: consiste en aprender patrones a partir de las entradas provistas sin tener las salidas esperadas.
- Por refuerzo: el aprendizaje se da a través de la interacción con el entorno, obteniendo o no una recompensa al realizar acciones.

Tal como se mencionó anteriormente, los datos son el pilar fundamental de ML y la base sobre la cual un algoritmo aprende. Al proveer a un algoritmo de ML de un conjunto de datos obtenemos un modelo que se entrena con el objetivo de realizar predicciones con datos nuevos. El conjunto de datos que se utiliza durante el entrenamiento se conoce como conjunto de entrenamiento, y es un subconjunto del *dataset* completo que se cuenta para trabajar con el problema. Los datos restantes se agrupan en el conjunto de datos de prueba, utilizado para probar el rendimiento del modelo una vez entrenado.

Al obtener un modelo entrenado debemos evaluar qué tan acertadas son las predicciones que devuelve. Cuando estas predicciones son erróneas se puede estar en dos estados: el de *overfitting* o el de *underfitting*. El primero se da cuando el modelo se ajusta en demasía a las particularidades del conjunto de datos de entrenamiento, por lo que las predicciones sobre nuevos datos son incorrectas. En cambio, cuando un modelo se encuentra en estadio de *underfitting* no logra capturar los patrones subyacentes a los datos de entrenamiento, por lo que realiza predicciones incorrectas en tanto el conjunto de datos de entrenamiento como en el de prueba.

Consecuentemente el objetivo de un modelo de ML es poder generalizar bien a partir de los datos de entrenamiento. Esto es, obtener predicciones acertadas a partir de datos que el modelo no haya visto previamente. Por esta razón un modelo generalizable es aquel que tiene un buen rendimiento en los conjuntos de datos de entrenamiento y prueba, y es la característica que se busca obtener al desarrollarlos.

2.1.1.1. Aprendizaje supervisado

En línea a lo descrito anteriormente, los algoritmos supervisados son aquellos a los que se les provee de tanto las variables de entrada como del resultado esperado para que el algoritmo aprenda a partir de ellos. En particular, los algoritmos supervisados aprenden una función $y = f(x)$ tal que ajuste de mejor manera las variables de entrada x con la variable de salida y. Así, al obtener nuevos datos se puede predecir la salida (Brownlee, 2016d). Más específicamente, los algoritmos de aprendizaje supervisado buscan la función h (la hipótesis) que mejor aproxime a la función verdadera f (Russell et al., 2009b).

A su vez, los problemas de aprendizaje supervisado pueden clasificarse según la salida obtenida en dos tipos:

- Clasificación: la variable de salida es una categoría. Por ejemplo, en el caso de un problema de predicción de incendios la salida de un algoritmo de clasificación puede ser “incendio” o “no incendio”.
- Regresión: la variable de salida es un número. Si se toma como ejemplo la predicción de la cantidad de hectáreas quemadas por un incendio, la salida de un algoritmo de regresión es el número de hectáreas propiamente dicho.

A continuación, se enumeran algunos de los algoritmos supervisados más comunes utilizados para problemas de predicción de ocurrencia de incendios forestales (clasificación) y área quemada (regresión). Todos estos algoritmos pueden ser ajustados en el proceso de entrenamiento mediante hiperparámetros o configuraciones que son dados para controlar el proceso y mejorar el rendimiento del modelo.

Regresión logística

El objetivo de la regresión logística es el de encontrar un modelo que explique la relación entre la variable dependiente (salida) y las variables independientes (entradas) (Hosmer et al., 2004). Este es uno de los algoritmos de clasificación supervisados más sencillos y utilizados en ML. Su nombre proviene de la función que utiliza: la función logística, también conocida como función sigmoide (1).

$$Sigmoide = \sigma(x) = \frac{1}{1+e^{-x}} \quad (1)$$

La función sigmoide da como resultado un número acotado entre 0 y 1, por lo que el resultado puede ser interpretado como la probabilidad de que la entrada (X) pertenezca a una clase $(Y = 1)$, o más formalmente (2) (Brownlee, 2016b).

$$P(X) = P(Y=1|X) \quad (2)$$

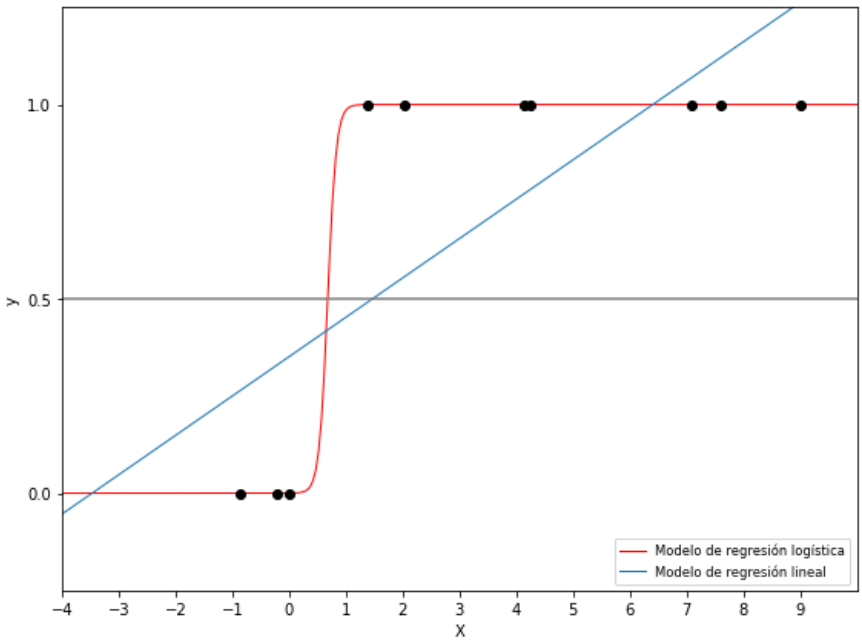

Figura 1: Comparación entre modelos de regresión logística y regresión lineal entrenados con datos ficticios.

Por esa razón, es común utilizar el algoritmo de regresión logística para problemas de clasificación binaria, en donde el objetivo es predecir una de dos clases posibles (como por ejemplo “incendio” y “no incendio”).

Los modelos de regresión logística son lineales, por lo que se obtienen buenos resultados cuando los datos son linealmente separables o pueden separarse por una frontera de decisión. Sin embargo, para modelos más complejos en donde hay múltiples fronteras de decisión los modelos de regresión logística no pueden capturar la complejidad de las relaciones en los datos.

Máquinas de Vectores Soporte

Las Máquinas de Vectores Soporte (o SVM por sus siglas en inglés *Support Vector Machines*) pueden utilizarse para problemas de clasificación y regresión. El objetivo es determinar los puntos (o vectores de soporte) que separan al máximo dos clases, aunque se puede generalizar a múltiples clases. Para ello cada ejemplo se representa como un vector en un espacio n-dimensional, donde n es la cantidad de atributos o características. Los vectores de soporte definen de esta forma un hiperplano que separa los datos linealmente (Luger, 2008).

En caso de que los datos no sean linealmente separables en una dimensión, estos se transforman para poder trabajar en un espacio de mayor dimensionalidad en donde encontrar el hiperplano sea más fácil. Para ello se recurre a funciones núcleo o *kernel*. Entre las funciones núcleo o kernel más comunes se encuentran:

- Kernel lineal (3):

$$k\left(x_i, x_j\right) = < x_i . x_j > \quad (3)$$

- Kernel RBF (*Radial Basis Function)* (4):

$$k\left(x_i, x_j\right) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right), \sigma \in (0, +\infty) \quad (4)$$

- Kernel polinómico (5):

$$k\left(x_i, x_j\right) = (< x_i . x_j > +1\ )^r, r > 1, r \in \mathbb{Z}^+ \quad (5)$$

(Liu et al., 2014)

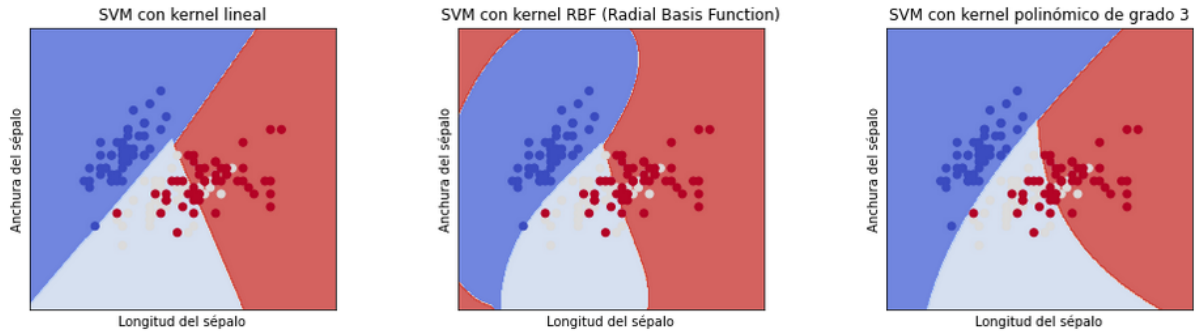

Figura 2: Comparación de las fronteras de decisión generadas a través de distintas funciones núcleo para clasificar las especies de la flor Iris (Fisher, 1988) según la anchura y longitud del sépalo.

Las SVM tienen la ventaja de ser efectivas en espacios de mayor dimensionalidad y capaces de capturar relaciones complejas y no lineales en los datos. Gracias a las funciones núcleo es posible encontrar un separador aún cuando los datos no son linealmente separables en una dimensión dada. No obstante, la elección de la función núcleo es fundamental para obtener buenos resultados, y el tiempo de entrenamiento es alto cuando el conjunto de datos de entrenamiento es grande (Mohamed, 2017).

Árboles de decisión

Los árboles de decisión reciben un conjunto de características y devuelven una decisión prevista como valor de salida para las entradas dadas. De la misma manera que SVM, los árboles de decisión también pueden utilizarse para tareas de clasificación y regresión.

Un árbol de decisión se representa mediante un árbol binario donde cada nodo del árbol representa un atributo de entrada. En caso de que la variable de entrada sea numérica, el nodo también es un punto de división sobre esa variable. Por otro lado, las hojas del árbol representan las salidas (Brownlee, 2016a), que pueden ser tanto etiquetas para problemas de clasificación como un valor continuo para problemas de regresión.

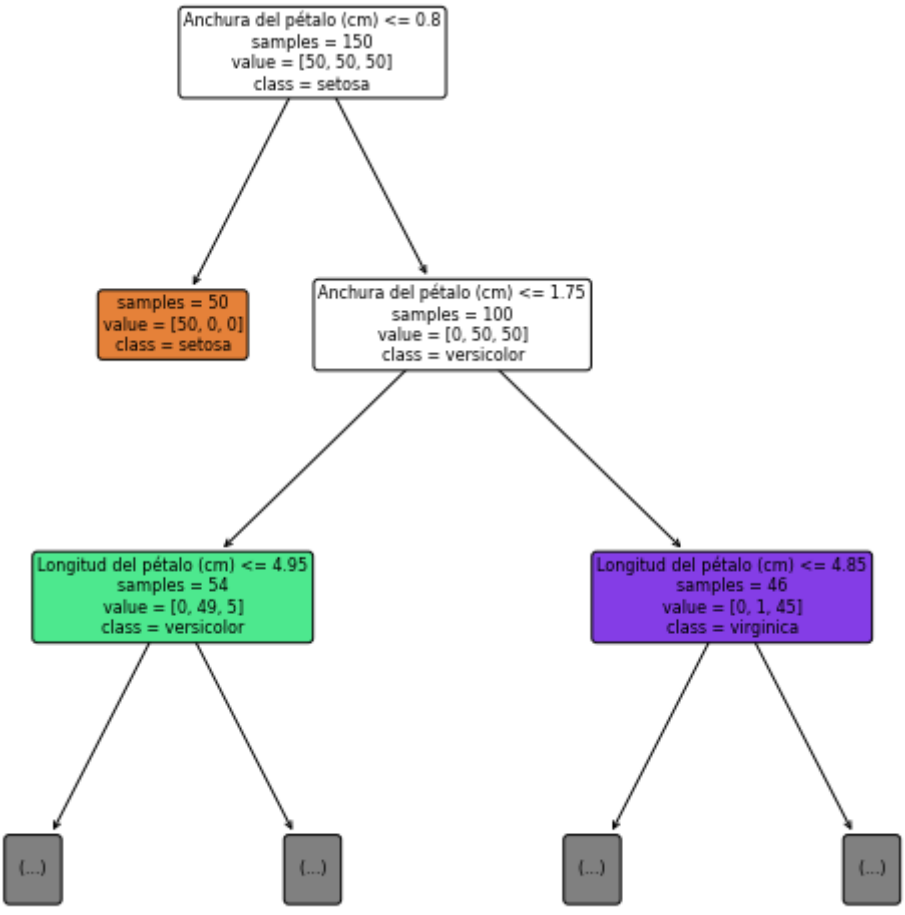

Figura 3: Representación de un árbol de decisión para clasificar las especies de la flor Iris según las longitudes del sépalo y pétalo, y las anchuras del pétalo y sépalo (Fisher, 1988). A fines visuales se omitieron los nodos de profundad mayor a 2.

Para crear el árbol de decisión se crean particiones sobre los datos de entrada y así encontrar las relaciones que existen entre las entradas y salidas de la forma más precisa posible. Sin embargo, se busca crear el árbol de menor tamaño para evitar que se ajuste demasiado a los datos de entrenamiento y no generalice bien a datos nuevos. Esto es posible mediante técnicas como la poda del árbol, la fijación de la profundidad del árbol y/o de la cantidad de nodos máxima.

Como se observa en la figura 3, una ventaja de los árboles de decisión es su fácil interpretación. Además, los árboles de decisión pueden trabajar con datos faltantes y *outliers* (valores extremos, distintos del resto de los valores en el conjunto de datos). Como desventajas se pueden enumerar el costo computacional de entrenamiento y la susceptibilidad de sobreajustarse a los datos del conjunto de entrenamiento, aunque como se mencionó anteriormente existen técnicas para evitarlo.

Métodos de ensamble

Los métodos de ensamble de algoritmos de ML combinan las predicciones de múltiples modelos con el fin de obtener predicciones más precisas. Entre las técnicas de ensamble más comunes se encuentran:

- Bagging: se construyen independientemente varios estimadores (ya sean clasificadores o regresores) utilizando el mismo algoritmo y distintos subconjuntos del conjunto de datos de entrenamiento. Una vez entrenados, las predicciones de cada modelo se promedian (en problemas de regresión) o votan (en problemas de clasificación). Por ejemplo, los bosques aleatorios o *Random Forests* son un ensamble de distintos árboles de decisión.
- Boosting: se desarrollan de manera secuencial varios *weak learners* o modelos sencillos, buscando corregir en cada paso los errores del modelo predecesor. Un ejemplo de esta técnica es el algoritmo *Gradient Boosted Trees*, en el que varios árboles de decisión individuales simples y con pocas ramificaciones se ajustan secuencialmente.

Redes Neuronales Artificiales

Tal como su nombre indica, las Redes Neuronales Artificiales (o ANN por sus siglas en inglés Artificial Neural Networks) simulan el comportamiento de una red neuronal biológica. Además, pueden utilizarse tanto en problemas de clasificación como de regresión.

Una ANN está formada por nodos o unidades (*units*), cada una compuesta por:

- Señales de entrada (X_j): se tratan de datos que provienen del entorno o de la activación de otros nodos.
- Pesos (W_j): determinan cuán fuerte es la conexión entre nodos y el signo de la conexión.
- Nivel de activación ($\sum W_j . X_j$): es la sumatoria de las entradas ponderadas por los pesos. Es decir, representa la fuerza acumulada de cada señal de entrada.
- Bias (b): es una constante que ajusta el resultado de la sumatoria de las entradas ponderadas por los pesos.
- Función de transferencia (f): calcula la salida del nodo determinando cuán superior o inferior es el nivel de activación con respecto a un valor de umbral establecido. Esta función es análoga a "activar" o "desactivar" el nodo cuando las entradas son correctas o incorrectas respectivamente. Dos ejemplos de funciones de transferencia son la función sigmoide (1) y la función umbral (6).

$$f(x) = \begin{cases} 0 \; si \; x \; < 0 \\ 1 \; si \; x \; \geq 0 \end{cases} \tag{6}$$

(Luger, 2008).

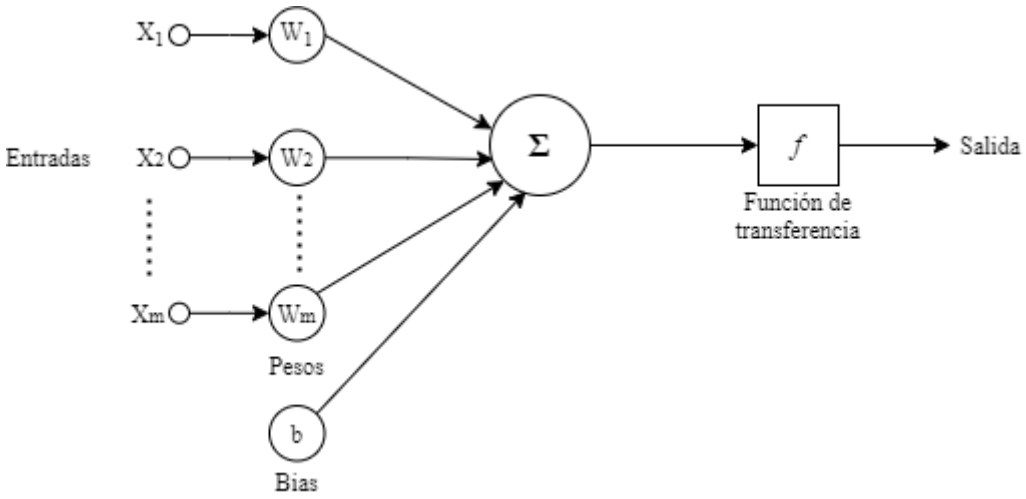

Figura 4: Estructura de un nodo o unidad (Haykin, 2008).

A su vez los nodos de una ANN pueden organizarse en capas, entre las que se distinguen:

- Capa de entrada: se encuentran los nodos que proveen los datos iniciales a la red neuronal.
- Capas ocultas: están conformadas por nodos ocultos. A mayor cantidad de capas ocultas más amplio es el espacio de hipótesis que puede representar la red y, por ende, más complejas son las relaciones que la ANN puede modelar.
- Capa de salida: contiene los nodos de salida. Dependiendo del tipo de problema (regresión, clasificación binaria, clasificación multiclase) es la cantidad de nodos presentes en esta capa.

(Haykin, 2008).

Existen diversas arquitecturas de ANN, siendo las principales las redes con alimentación hacia adelante o *feedforward neural networks,* y las redes recurrentes o *recurrent neural networks* (RNN). En las primeras los datos se pasan a través de las capas hasta llegar a la salida, mientras que en las segundas las salidas pueden ser sus propias entradas, formando así un ciclo (Russell et al., 2009b).

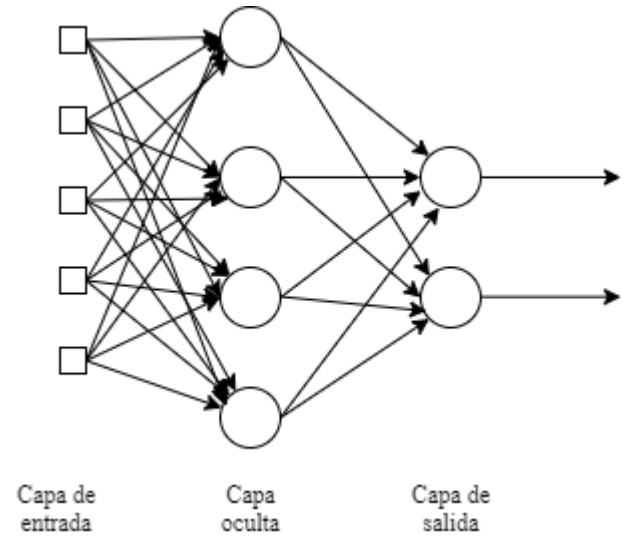

Figura 5: Feedforward Neural Network totalmente conectada (cada nodo en cada capa está conectado con todos los nodos de la capa adyacente), conformada por una capa de entrada, una capa oculta y una capa de salida (Haykin, 2008).

El proceso de aprendizaje en una ANN consiste en comparar las salidas de la red con el valor de la salida esperado para calcular una medida de error (como la pérdida o función de costo) y en consecuencia ajustar los pesos de los nodos para minimizar el error obtenido. Específicamente se calcula el gradiente de toda la red para determinar la magnitud del cambio que debe aplicarse en el valor de cada peso y así actualizarlos. Este proceso se repite hasta obtener los resultados deseados (Hecht-Nielsen, 1989).

Como principal ventaja de las ANN se encuentra la capacidad que tienen para resolver tanto problemas lineales como no lineales y detectar relaciones complejas entre las variables dependiente e independientes. Aun así no existen lineamientos específicos acerca de cuál es la arquitectura e hiperparámetros óptimos, sino que requieren ser establecidos mediante la prueba y error (Mohamed, 2017).

2.1.1.2. Evaluación de modelos

Una vez entrenado un modelo existen varias métricas para establecer cuán capaz es de proveer predicciones correctas con datos nuevos. No obstante, las métricas que se pueden utilizar dependen del tipo de problema que el modelo resuelva.

En los problemas de clasificación donde se cuenta con una clase positiva y una clase negativa se pueden distinguir las siguientes métricas:

- Exactitud (*accuracy*): determina el rendimiento general del modelo, calculando la proporción de predicciones correctas sobre el total de predicciones.
- Precisión (*precision*): indica cuál es el porcentaje de predicciones efectivamente positivas sobre el total de predicciones positivas.
- Sensibilidad (*recall*): mide cuán correcta es la identificación de verdaderos positivos por parte del modelo.
- Valor F (*F1 Score*): es una métrica utilizada cuando las clases están desbalanceadas (es decir, cuando la distribución de clases en el conjunto de datos es desigual). Es el promedio ponderado de la sensibilidad y precisión.

Asimismo existen representaciones gráficas para entender el rendimiento de un modelo. La Característica Operativa del Receptor (o ROC por sus siglas en inglés *Receiver Operating Curve*) representa la sensibilidad frente a la especificidad conforme se varía el umbral de discriminación. Por otro lado, también se puede representar el Área bajo la curva Característica Operativa del Receptor o AUC (*Area Under the receiver operating Curve*) (Kulkarni et al., 2020). En este caso, cuanto mayor es el AUC, más capaz es el modelo de distinguir entre clases positivas y negativas.

Por otro lado, en los problemas de regresión el rendimiento de un modelo se mide determinando cuán cercanas fueron las predicciones al valor (numérico) esperado. Las métricas básicas son:

- Error cuadrático medio (*Mean Squared Error* o MSE) (7).

$$MSE = \frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n} \quad (7)$$

- Raíz del error cuadrático medio (*Root Mean Squared Error* o RMSE) (8).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n}} \quad (8)$$

- Error absoluto medio (*Mean Absolute Error* o MAE) (9).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\|p_i - y_i\| \quad (9)$$

(Brownlee, 2016c).

2.1.2. Incendios forestales

Formalmente un incendio forestal se define como *"fuegos no controlados (sean de origen natural o antrópico) que ocurren en ecosistemas terrestres, y que se propagan por la vegetación, sea esta el tipo que sea (bosque, sabana, matorral, pastizal, humedal, turbera, etc.)* (Pausas, 2020).

En Argentina, los incendios forestales originados naturalmente son causados en su mayoría por la caída de rayos durante tormentas eléctricas en las que no llueve, aunque en algunas regiones de la Patagonia la actividad volcánica también es otro origen natural. No obstante, estos incendios representan solo el 5% del total: el 95% restante son originados por causas antrópicas, como el establecimiento de campamentos y fogatas, deportes, instalaciones eléctricas deficientes, accidentes o negligencia. (Argentina.gob.ar, 2018b).

A fin de poder comprender cuáles son los factores influyentes en tanto la ocurrencia como desarrollo de incendios forestales se debe explicitar cuáles son las etapas que todo incendio forestal atraviesa. En primer lugar se encuentra la etapa de precalentamiento, en donde se eleva la temperatura del combustible (vegetación en el caso de los incendios forestales) hasta llegar a los 200°C y evaporar los compuestos volátiles de las resinas. Cuando la temperatura asciende los 300°C o 400°C el material combustible se incendia, y al alcanzar los 600°C la combustión continúa incluso si se retirara la fuente de calor. Una vez que el combustible se agota se hacen visibles las cenizas, marcando el final del incendio (Argentina.gob.ar, 2018a).

Como se puede evidenciar, el combustible es un factor importante en el desarrollo de un incendio forestal. No obstante, existen otros elementos determinantes que se deben tomar en cuenta a la hora de estudiar el origen y finalización de un incendio. A continuación se describen en detalle estos factores.

2.1.2.1. Factores influyentes

Al iniciarse un incendio existen factores como el área, la velocidad de propagación de las llamas, y la longitud y altura de las llamas que modifican el comportamiento del mismo, lo que permite saber a los bomberos 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.

Meteorología

Se trata del componente que más varía en el tiempo y espacio. El clima es el factor que determina en qué épocas del año los incendios forestales son más propensos de producirse, ya que afecta a la salud de la vegetación. Entre las variables meteorológicas que influyen en la probabilidad de ocurrencia de incendios forestales se encuentran:

- La temperatura, que afecta la velocidad en que los combustibles se secan. Además, a mayor temperatura los incendios son más difíciles de controlar.
- La humedad relativa, que se la relaciona con la humedad de los combustibles, por lo que una baja humedad relativa ambiente aumenta la probabilidad de ocurrencia de incendios.
- El viento, que determina la dirección, intensidad y velocidad en que el fuego se propaga.
- La precipitación, cuya distribución temporal y cantidad determinan el agua que la vegetación tiene disponible para desarrollarse. Por esta razón la probabilidad de que se produzcan incendios luego de sequías aumenta.
- Las nubes, que originan tormentas y debajo las cuales es común que se produzcan fuertes ráfagas de viento que intensifiquen el fuego.

(Argentina.gob.ar, 2018c) (Moscovich et al., 2014).

Topografía

Los factores topográficos varían de forma espacial y pueden considerarse constantes en el tiempo. No obstante, estos factores afectan de diversas maneras la probabilidad y comportamiento de los incendios como se detalla a continuación:

- Altura: las condiciones meteorológicas y el tipo de suelo se ven modificados conforme la altura aumenta. Además, el tipo de combustible que se encuentra disponible varía en las distintas alturas.
- Exposición: la exposición al sol dispone la distribución y condiciones en que se encuentra la vegetación en un momento dado.
- Pendiente: la inclinación del terreno determina a qué velocidad y en qué dirección se propagarán las llamas durante un incendio.
- Relieve: modifica la dirección y velocidad de los vientos, afectando consecuentemente el comportamiento del fuego durante un incendio.

(Argentina.gob.ar, 2018c) (Moscovich et al., 2014).

Combustible

El combustible en un incendio forestal es la vegetación, la cual varía tanto temporal como espacialmente. Las características de los combustibles que se consideran influyentes en los incendios forestales son:

- La ubicación, forma y tamaño de los combustibles: estos factores determinan la facilidad de ignición. Por ejemplo, las ramas finas se encienden con mayor facilidad que troncos de gran tamaño.
- La continuidad: es la distancia que existe entre la vegetación. La propagación del fuego es más sencilla cuando la distancia entre combustibles es menor, de otra manera sería necesaria la presencia de vientos fuertes para propiciar la transmisión.
- La compactación: es la relación entre la carga de combustible por superficie y la altura del combustible. Esta determina la superficie de combustible expuesta al fuego.
- La carga y densidad: se refiere al peso de los combustibles en un área medida en kilogramos por metros cuadrados. La densidad de un combustible determina cuánto calor puede absorber antes de encenderse.
- La composición química y el contenido de humedad: la presencia de resinas, aceites y otras sustancias volátiles provoca que el fuego se propague rápida e intensamente.

(Argentina.gob.ar, 2018c) (Moscovich et al., 2014).

En el partido de Pinamar los pinos son una de las especies vegetales predominantes (Respirá Pinamar, 2020). Junto con otras coníferas, los pinos contienen resina que los hacen más inflamables que otras especies. Además, la resistencia al paso del fuego en coníferas es menor (Dirección General de Protección Civil y Emergencias - Ministerio del Interior - España, 2021).

2.1.2.2. Índices de peligro de incendio

Considerando todos los factores que influyen en el inicio y desarrollo de incendios forestales, resulta necesaria la previsión de las condiciones favorables de ocurrencia para poder prevenirlos y combatirlos en caso de ocurrir. 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 Incendios o FWI (por sus siglas en inglés *Forest fire Weather Index*) (National Wildfire Coordinating Group, 2021). 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 (Waidelich et al., 2019).

Para determinar el riesgo de incendio el FWI se vale de seis componentes relacionados jerárquicamente. 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).

(National Wildfire Coordinating Group, 2021) (Villers-Ruiz et al., 2012).

Una vez calculados 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.

(National Wildfire Coordinating Group, 2021) (Villers-Ruiz et al., 2012).

Por último, estos dos últimos índices se combinan linealmente para obtener el FWI, que indica 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.

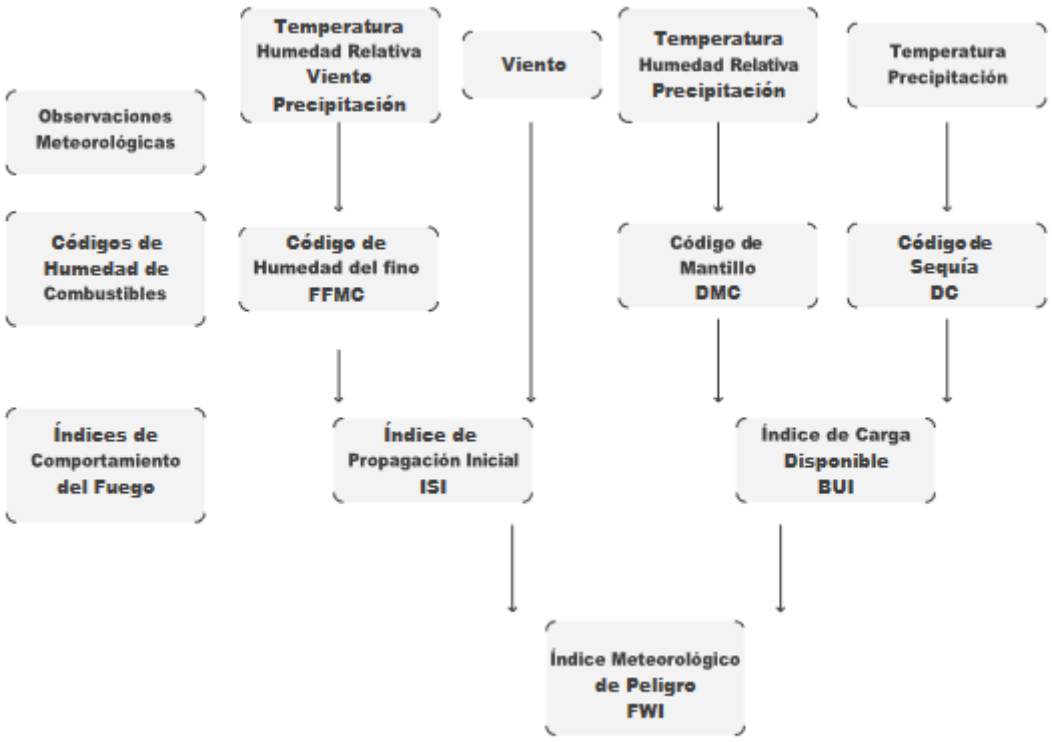

Figura 6: Estructura del FWI (Secretaría de Ambiente y Desarrollo Sustentable, 2021).

2.2. Estado del Arte

Los avances tecnológicos de las últimas décadas han contribuido a que investigadores de todo el mundo trabajen en pos de encontrar soluciones a problemas que el hombre y su entorno sufren diariamente. Los incendios forestales no han sido la excepción, y a causa del cambio climático estos han despertado mayor interés debido a que el mismo ha provocado un aumento en la severidad, impacto y superficie final quemada por los incendios (Flannigan et al., 2000).

Entre los avances tecnológicos más importantes en relación con el estudio de incendios forestales se puede mencionar la inclusión de sensores especializados a satélites que permiten, entre otras tareas, detectar focos de incendios, monitorear los cambios en la distribución y salud de la vegetación, determinar la temperatura de la superficie y estudiar el cambio climático a escala planetaria. Como los datos recopilados por estos satélites son de libre acceso, el interés de investigadores y estudiantes por utilizarlos para resolver problemas concretos a través de modelos de ML ha crecido. (Jain et al., 2020)

Como se describió anteriormente, la ocurrencia de incendios forestales depende de muchos factores interrelacionados que están resumidos en el "Triángulo de Comportamiento del Fuego" (Boulandier et al., 2001). Por esta razón, los trabajos que tienen como objetivo de estudio a los incendios forestales se enfocan en al menos un elemento del triángulo, y los modelos que se han desarrollado giran principalmente en torno a las siguientes áreas de estudio:

- Predicción espacial y temporal de incendios.
- Detección de incendios forestales.
- Predicción de área quemada por incendios.
- Detección del área quemada a causa de incendios.
- Simulación de propagación de fuego una vez iniciado el incendio.

Considerando esta clasificación, los modelos desarrollados en el contexto del proyecto AQUA se encuentran enmarcados en las áreas de predicción espacial-temporal de incendios forestales y el área quemada por los mismos.

Con el fin de comparar los resultados obtenidos en la materia de incendios forestales se realiza un compendio de los trabajos y productos desarrollados internacional, nacional y localmente.

2.2.1. A nivel internacional

En Estados Unidos la empresa de la industria geoespacial Sanborn ofrece un software de gestión de incendios forestales a agencias gubernamentales locales, estatales y federales. Uno de los módulos denominado WFRAS incluye el análisis de riesgo de incendio basado en datos de sensores remotos propietarios, imágenes y datos satelitales. El producto está disponible para cualquier estado continental (esto es, todos los estados de Estados Unidos excepto Alaska, Hawái y Puerto Rico) ya que ése es el alcance de su flota de aeronaves para recolectar imágenes áreas (Sanborn, 2021).

Por su parte, la empresa española TecnoSylva cuenta en su cartera de productos con Wildfire Analyst, un software que simula el comportamiento de incendios forestales en tiempo real y establece perímetros en zonas donde el peligro de incendios forestales es mayor. Para esto se vale de datos meteorológicos del Servicio Meteorológico Nacional de Estados Unidos, aunque también ofrece la integración con estaciones meteorológicas locales (Tecnosylva, 2017).

Más allá de las soluciones comerciales mencionadas, (Rodrigues et al., 2014) compararon la precisión de distintos modelos de ML para predecir la probabilidad de incendios forestales en la región peninsular de España a través de dos clases: baja probabilidad y alta probabilidad. Con tal fin se utilizaron como variables explicativas aquellas que caracterizan la ecología (áreas protegidas), infraestructura (líneas de tensión, vías de tren), demografía y actividades humanas (urbanas y rurales) de una región particular del país. Tal como se realizó en la mayoría de los modelos de predicción de incendios que tienen en cuenta el factor espacial, se dividió el área de interés en una cuadrícula de una dimensión preestablecida con el objetivo de trabajar con celdas de cierta superficie. De los modelos desarrollados y evaluados, *Random Forest* y *Boosted Regression Trees* obtuvieron la precisión más alta con valores de 74.6% y 73% respectivamente.

Siguiendo una línea similar, en el vecino país de Brasil investigadores desarrollaron dos modelos para predecir el riesgo de ocurrencia de incendios forestales en el Distrito Federal utilizando redes neuronales y regresión logística. En este caso se utilizaron varias variables topográficas además de registros históricos, como la distancia de los puntos de incendio a caminos, la densidad poblacional de una zona particular, la disponibilidad de agua superficial, la cobertura del suelo y el Índice de Vegetación Senescente de Diferencia

Normalizada o NDSVI (por sus siglas en inglés *Normalized Difference Senescent Vegetation Index*). A través del modelo de regresión logística los investigadores obtuvieron una precisión del 67.8%, mientras que con la red neuronal la precisión obtenida fue de un 66.55% (Bem et al., 2018).

Por otro lado, una investigación reciente consistió en utilizar métodos de ensamble de modelos para predecir el área que un incendio forestal quemará (Xie et al., 2019). En este caso se utilizó el conocido y ampliamente utilizado *dataset* de incendios ocurridos en el parque Montesinho, Portugal, entre los años 2000 y 2003 (Cortez et al., 2007). El mencionado conjunto de datos contiene además las condiciones meteorológicas del momento en que dichos incendios sucedieron, lo que les permitió a su vez incluir los diversos índices intermedios que componen el FWI. Entre los métodos de ensamble evaluados, aquel con el que se obtuvieron los mejores resultados fue el de *Extreme Gradient Boosting*, con una precisión de 72.3%.

Otro desarrollo interesante en cuanto a las variables explicativas utilizadas es CityGuard, una herramienta que permite visualizar los resultados del modelo de predicción de incendios denominado NeuroFire. Este modelo fue desarrollado para pronosticar la probabilidad mensual de incendios en la ciudad de Zhengzhou, China. Para ello se utilizaron variables que consideren los efectos temporales (registros de incendios históricos, meteorología) y espaciales (población del área, puntos de interés, actividades humanas) influyentes en la probabilidad de incendios. El modelo está compuesto por dos partes: la clasificación temporal de incendios a través de una red neuronal GRU-CRF (*Gated Recurrent Unit – Conditional Random Field* o unidad recurrente cerrada combinada con un campo aleatorio condicional), y la predicción espacial de peligro de incendio mediante una función de pérdida inspirada en BPR (*Bayesian Personalized Ranking* o ranking bayesiano personalizado). NeuroFire alcanzó una sensibilidad de 55.8% y un valor de AUC de 0.763 (Wang, Qianru et al., 2019).

Además de los mencionados productos de software y modelos que buscan predecir incendios forestales existen aplicaciones para dispositivos móviles que determinan el peligro de incendio dadas ciertas condiciones climáticas ingresadas por el usuario. Entre ellas se encuentran Fire weather calculator, Calc-FDI y Wildfire Analyst Pocket. Estas aplicaciones son gratuitas y se basan en índices de peligro de incendio calculados tales como el FWI, o los obtenidos del Sistema Nacional de Evaluación de Peligro de Incendios de Estados Unidos (o NFDRS por sus siglas en inglés *National Fire Danger Rating System*).

2.2.2. A nivel nacional

El número de *papers*, investigaciones y proyectos realizados en el país con relación a la predicción de incendios forestales es considerablemente menor que el descripto a nivel internacional e incluso regional. Sin embargo, (Cardenas et al., 2016) desarrollaron un sistema de predicción de incendios forestales en la provincia de Córdoba utilizando redes neuronales y SVM. En este caso hicieron uso de datos meteorológicos recopilados a través de un módulo perteneciente al sistema, en donde los bomberos pueden ingresar las condiciones climáticas en las que ocurrieron los incendios. También se utilizaron registros históricos de incendios, y en la fase final del proyecto se comenzaron a emplear datos provenientes de imágenes satelitales. De esta manera obtuvieron una precisión preliminar del 61.6% al predecir la ocurrencia de incendios forestales.

2.2.3. A nivel local

En la actualidad no existe registro de la existencia de un modelo de predicción de incendios forestales en la Costa Atlántica argentina ni tampoco alguno específicamente desarrollado para Pinamar.

Los Bomberos Voluntarios de Pinamar cuentan desde el año 2019 con un sistema de gestión administrativa buscando migrar progresivamente los registros en soporte papel. De esta manera pueden obtener estadísticas e información relevante sobre los incendios y gestionar digitalmente todos los documentos que deben generar.

Por esta razón, como futura línea de trabajo se proyecta integrar este sistema administrativo con el sistema de visualización de riesgo de incendios forestales y el modelo propuesto de predicción, con el objetivo de ofrecer una solución integral a los bomberos que les permitan tomar decisiones con la mayor cantidad de información posible, ahorrando tiempo y dinero.

2.2.4. Conclusión

En todo el mundo se han desarrollado numerosos modelos y productos para predecir incendios forestales utilizando distintos algoritmos y aplicando distintos enfoques tal como se evidencia en la Tabla I.

TABLA I: Comparación entre modelos existentes y el propuesto.

Modelo	**Dimensiones consideradas**	**Escala temporal**	**Escala espacial**
(Rodrigues et al., 2014)	Topografía Demografía	Dimensión temporal no considerada	Nacional
(Bem et al., 2018)	Topografía Combustible	Dimensión temporal no considerada	Regional
(Xie et al., 2019)	Meteorología Combustible	Diaria	Regional
(Wang, Qianru et al., 2019)	Meteorología Topografía Demografía	Mensual	Regional
(Cardenas et al., 2016)	Meteorología Combustible	Diaria	Regional
Propuesto	Meteorología Topografía Combustible	Diaria	Local

Tan solo considerando modelos de predicción de incendios, hasta fines del año 2020 existían casi 300 *papers* en los que investigadores han desarrollado y comparado distintos modelos predictivos obteniendo diversos resultados. (Jain et al., 2020).

Internacionalmente existen varias soluciones comerciales orientadas a organizaciones de mediana a gran envergadura, ya que el costo de las mismas es alto. No obstante, estas soluciones brindan predicciones especializadas al área de interés valiéndose de varias fuentes de datos, como imágenes aéreas, satelitales, datos meteorológicos recopilados por redes de estaciones meteorológicas e incluso registros históricos de incendios locales, regionales y nacionales.

A nivel nacional el software creado por (Cardenas et al., 2016) es, hasta la fecha de redacción del presente trabajo, el único que permite visualizar la ubicación en donde la probabilidad de ocurrencia de un incendio forestal es mayor. Sin embargo, el proyecto de investigación se encontraba aún en desarrollo.

Teniendo en cuenta el estado del arte, el valor del presente trabajo radica en el enfoque local que se adopta, permitiendo capturar las características meteorológicas y topográficas particulares al área de estudio considerada (inicialmente el Partido de Pinamar) así también como la variación temporal en la ocurrencia de incendios forestales en la zona.

2.3. User Research

Para validar la problemática que se busca solucionar en el contexto del proyecto AQUA y completar uno de los objetivos específicos del presente Proyecto Final de Ingeniería se entrevistaron a diversos especialistas de cada área relevante al desarrollo. A continuación, se detallan los principales resultados de las mismas.

2.3.1. Entrevista a Matías García (Asociación Bomberos Voluntarios de Pinamar)

Con el objetivo de entender con mayor profundidad el dominio y las problemáticas de un cuartel de bomberos se entrevistó a Matías García, ayudante principal de bomberos de la Asociación de Bomberos Voluntarios de Pinamar. En particular se hizo foco al trabajo que se realiza en el cuartel antes, durante y después de un incendio forestal con el fin de comprender cómo un modelo de predicción de incendios forestales puede ayudar en su labor diaria.

Para entender cómo trabajan los bomberos a la hora de combatir incendios es importante listar las actividades previas que se realizan y las personas involucradas en cada tarea. Internamente un cuartel está organizado en secciones o áreas, cada una a cargo de un jefe o encargado. Algunas de estas secciones son automotores, personal, capacitación, materiales y equipos.

Dentro del cuartel de bomberos, García resalta la importancia del rol del cuartelero. Este se encarga, entre otras actividades, de recibir llamadas de emergencia las 24 horas del día, registrar los datos que la persona que se comunicó haya provisto, y accionar la sirena para avisar sobre la emergencia a bomberos y a la población en general. En el caso de los Bomberos Voluntarios de Pinamar, el tiempo que transcurre entre que el cuartelero activa la alarma y el cuerpo de bomberos se dirige al incendio en cuestión es de 2 a 3 minutos en promedio. Esta rápida respuesta es posible gracias al trabajo semanal que se realizan en las denominadas guardias, donde los móviles -autobombas como comúnmente se conocen-,

materiales y equipos se mantienen y dejan listos para responder a cualquier emergencia en el menor tiempo posible.

Si bien el tiempo que transcurre entre la alarma y la salida de bomberos a la emergencia es muy corto, en condiciones desfavorables el fuego en incendios forestales puede avanzar a razón de 5 metros por minuto. Además, la persona que se comunica al cuartel para informar sobre el incendio puede haberlo hecho muchos minutos después de que efectivamente haya comenzado el siniestro, por lo que en esos casos los bomberos se encuentran con incendios forestales ya avanzados. Controlarlos se torna más difícil conforme pasa el tiempo, y hay situaciones donde es necesario pedir refuerzos.

Para combatir incendios forestales los bomberos se valen de muchos materiales. En particular, se preparan unidades forestales que están equipadas con tanques de agua y una motobomba que expulsa el agua por las mangueras al incendio. Estas unidades poseen también mangueras de distinto diámetro con sus respectivas boquillas (lanzas, pico o puntas de manguera). Además, las unidades forestales disponen de materiales de zapa, como palas, rastrillos, látigos, hachas y picos. El uso de los materiales mencionados depende del tipo de combustible del fuego.

En cuanto a móviles, la Asociación de Bomberos Voluntarios de Pinamar cuenta con 4 *unimogs* o autobombas forestales especialmente preparadas para movilizarse en terrenos irregulares y de difícil circulación, un móvil de ataque rápido (se trata de una camioneta con un pequeño tanque de agua) y camionetas para traslado de material o apoyo logístico.

Tanto los móviles como materiales requieren de mantenimiento periódico, y dependiendo de la magnitud del incendio forestal el desgaste de los equipos es mayor o menor. Este mantenimiento no sólo lleva tiempo (semanalmente cada bombero dedica alrededor de 10 horas a tareas administrativas y de mantenimiento) sino también que incurre en dinero. Si bien la sección automotores está conformada en su mayoría por mecánicos, -ahorrándose de esta manera el gasto de dinero en mano de obra externa-, la Asociación debe pagar combustible, aceites y repuestos, entre otros insumos. En muchas ocasiones se pierden o rompen materiales al combatir incendios forestales de gran magnitud, por lo que deben ser repuestos. En este punto García destaca que todos los materiales y móviles requeridos tienen precios fijados en dólares, por lo que cada vez se torna más difícil reponer materiales y adquirir equipamiento más avanzado. Tal es así que, en el incendio forestal de mayor magnitud del partido ocurrido en Valeria del Mar en diciembre de 2016, el gobierno nacional tuvo que subsidiar a los Bomberos

Voluntarios de Pinamar para reponer un *unimog* que había sido perdido en combate a causa del estado avanzado del incendio. La suma de la partida transferida en ese año ascendió a más de 55000 dólares.

Los bomberos poseen conocimiento empírico sobre las condiciones en las que existen mayores probabilidades de incendios forestales, y coinciden en que el clima es un factor importante para determinar el riesgo de incendio forestal. En efecto, utilizan la denominada Regla del 30 o Factor 30-30-30. Esta regla se cumple los días en los que la temperatura es igual o superior a los 30 grados centígrados, la humedad relativa del ambiente es menor al 30% y la velocidad del viento es igual o superior a los 30 kilómetros por hora. Esta combinación de condiciones meteorológicas es muy común en verano, donde además la afluencia de turistas a la ciudad es mayor. Este último punto es sumamente relevante, ya que según García el 99% de los incendios forestales en Pinamar son causados por el hombre (ya sea intencional o accidentalmente). Esto implica que a mayor cantidad de personas visitando bosques para realizar actividades tales como campamentos y fogatas, mayor es la probabilidad de que se produzcan incendios forestales bajo condiciones climáticas desfavorables.

Por estas razones, los bomberos realizan campañas de prevención de incendios a través de diferentes medios de comunicación, como publicaciones en sus redes sociales y entrega de revistas y folletos en instituciones públicas y privadas. Además, los días en que la Regla del 30 se cumple y el peligro de incendios forestales es mayor, informan a la población sobre el riesgo a través de sus redes sociales. Para alertar a turistas y habitantes también se valen del mapa de peligro de incendios (Figura 7) elaborado diariamente por el Servicio Nacional de Manejo del Fuego utilizando el FWI.

En esta sección hacer un resumen de la información mas importante de la entrevista y hacer referencia al instrumento anexo utilizado para recopilar la información.

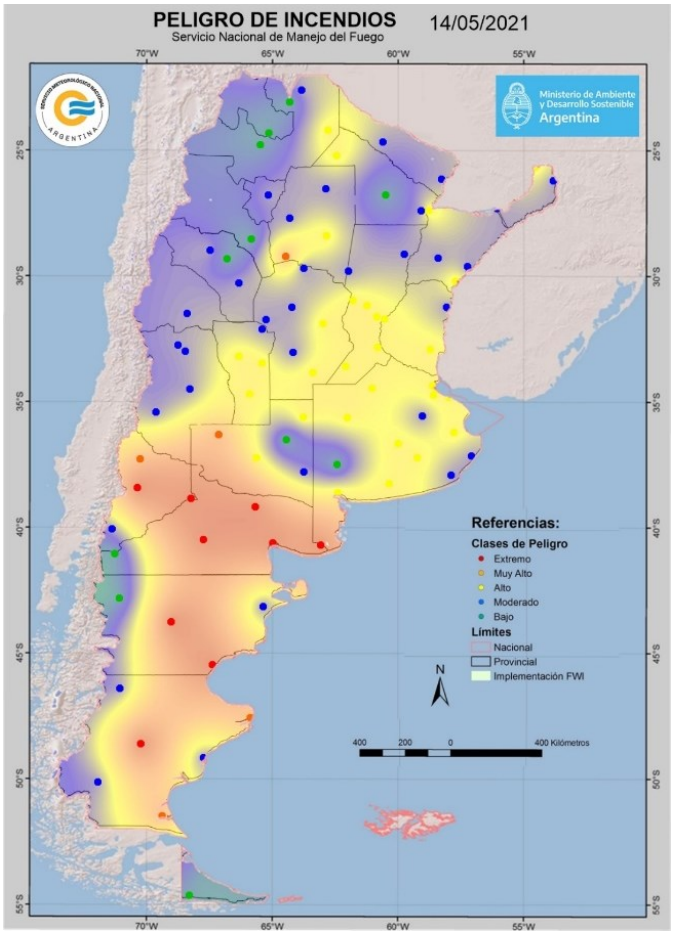

Figura 7: Mapa de peligro de incendio a escala nacional del día 14 de mayo de 2021. (Servicio Meteorológico Nacional, 2021)

Según afirmó García, los bomberos han tenido que combatir incendios forestales de menor magnitud en muchas ubicaciones del Partido de Pinamar. Sin embargo, gracias a su experiencia conocen cuáles son las zonas en donde se producen los incendios forestales de mayor magnitud. Tanto García como el bombero Lucas León aseveraron que la Reserva Natural de Cariló, la Reserva Forestal de Valeria del Mar, el Camino de los Pioneros y la zona norte de Pinamar (también denominada La Frontera) son las ubicaciones donde los incendios forestales han sido más importantes. Efectivamente en estas ubicaciones los bosques son más frondosos, por lo que esperan que, bajo condiciones desfavorables, los incendios sean peligrosos.

2.3.2. Entrevista a Marcos Saucedo (Servicio Meteorológico Nacional)

Dado que uno de los pilares del modelo de predicción propuesto se basa en la relación que existe entre la probabilidad de ignición, propagación e impacto del fuego y las

condiciones meteorológicas del medio en donde se producen se realizó una entrevista a Marcos Saucedo, director de la Dirección de Pronósticos del Tiempo y Avisos del Servicio Meteorológico Nacional (SMN).

Los objetivos de la entrevista fueron:

- Determinar cuáles son dichas condiciones meteorológicas, teniendo en cuenta la topografía y características particulares de la Costa Atlántica argentina en términos climáticos.
- Evaluar la calidad de los datos climáticos provenientes de estaciones meteorológicas automáticas instaladas en Pinamar.

Tal como se evidenció en los modelos de predicción existentes detallados en la sección *Estado del Arte*, Saucedo afirmó la estrecha relación existente entre las condiciones climáticas y la facilidad de propagación de las llamas durante un incendio. Efectivamente las variables que influencian en mayor medida son la temperatura, humedad, velocidad y dirección del viento, y la precipitación acumulada durante las horas anteriores a dicho incendio.

No obstante, Saucedo aclaró que en la Costa Atlántica argentina las probabilidades que se den condiciones climáticas que provoquen incendios de forma natural son muy bajas. No es el caso de la Patagonia, por ejemplo, donde se producen tormentas eléctricas sin lluvia o tormentas secas en las que los rayos que caen en la superficie pueden iniciar incendios forestales.

Además, manifestó la relación que existe entre la variación anual de las condiciones climáticas y la vegetación, donde las abundantes o escasas precipitaciones durante un año influencian la superficie verde total del siguiente. En otras palabras, si durante un año las precipitaciones fueron abundantes, es altamente probable que se observe un crecimiento importante de la vegetación al año siguiente. Esto implica una mayor densidad de combustible, por lo que en caso de producirse un incendio la superficie susceptible de ser quemada es mayor.

Un índice que es muy útil para observar la salud y existencia de vegetación en una zona dada es el Índice de Vegetación de Diferencia Normalizada o NDVI (por sus siglas en inglés *Normalized Difference Vegetation Index)*. En particular Saucedo recomendó observar las series temporales correspondientes al clima y NDVI para comprender cómo estas variables están relacionadas, y prestar atención a cómo varía este índice a través de los años para entender la evolución de la cantidad de superficie cubierta por vegetación, superficie susceptible de ser quemada en un incendio forestal.

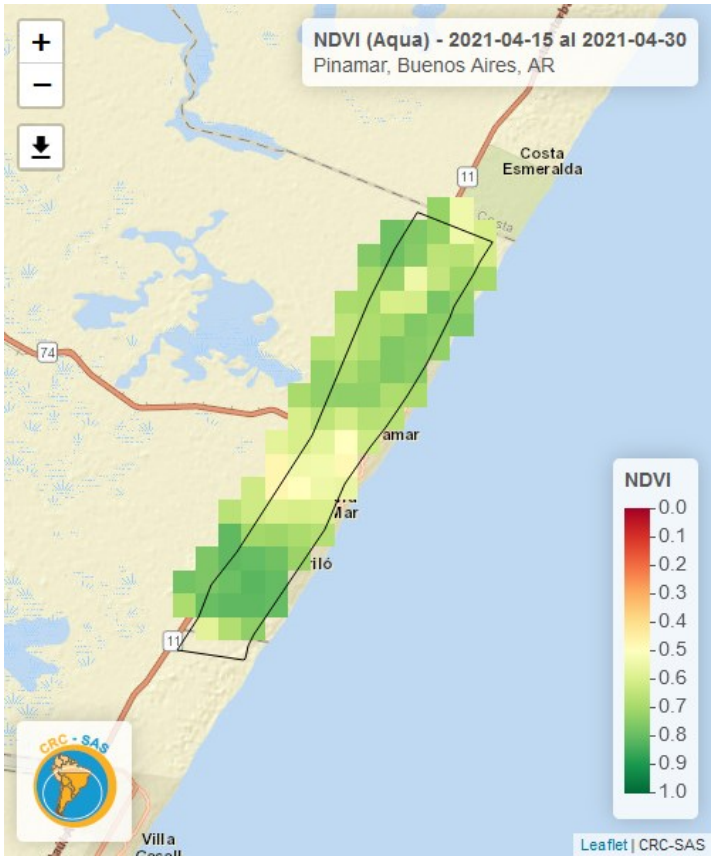

Figura 8: Valores de NDVI en el Partido de Pinamar de la segunda quincena del mes de abril de 2021, obtenidos del satélite de observación terrestre Aqua operado por la NASA (Administración Nacional de Aeronáutica y Espacio o *National Aeronautics and Space Administration* por sus siglas en inglés). Las tonalidades verdes indican condiciones de máximo verdor. (Centro Regional de Climas do Sul da América do Sul, 2021)

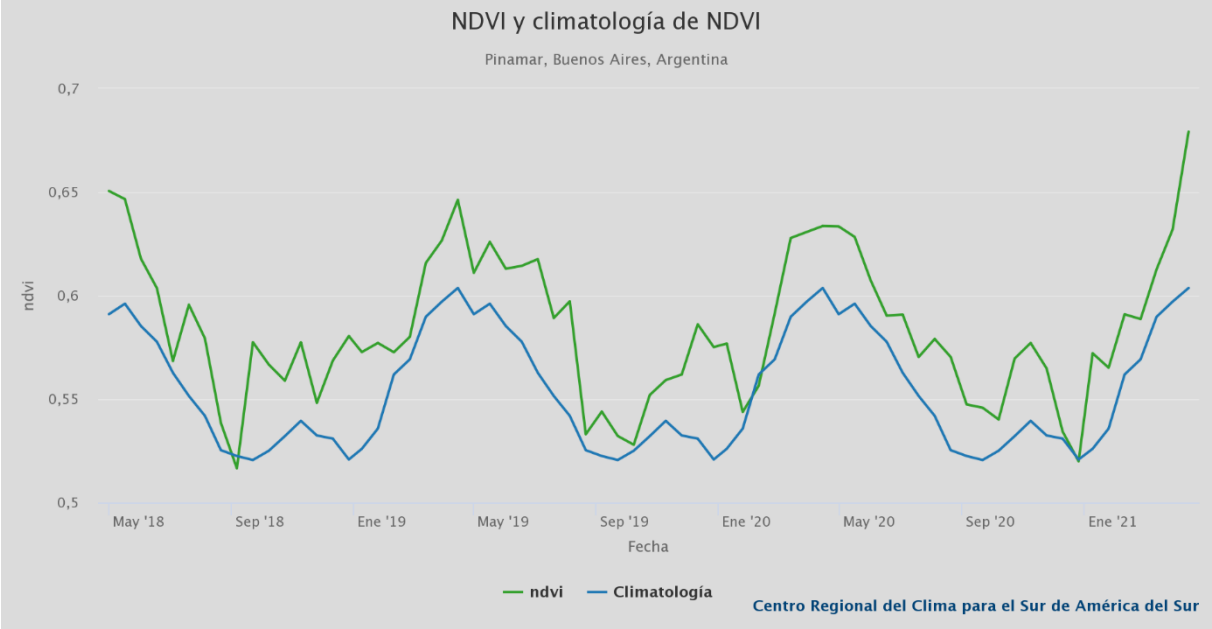

Figura 9: Serie temporal de los últimos 36 meses del promedio espacial de NDVI y valores climatológicos correspondientes a Pinamar. (Centro Regional de Climas do Sul da América do Sul, 2021)

En cuanto a la calidad de los datos climáticos provistos por estaciones meteorológicas automáticas en Pinamar, Saucedo recomendó utilizar los datos provenientes de la estación meteorológica instalada y supervisada por el SMN en la ciudad vecina de Villa Gesell. Las razones radican en que las estaciones automáticas no son controladas por un especialista para detectar anormalidades en los datos recopilados (que en general se producen cuando el instrumental de la estación falla) ni mantenidas periódicamente. Por otro lado, las estaciones sinópticas de superficie cumplen con condiciones establecidas por la Organización Meteorológica Mundial, por lo que se puede asegurar la confiabilidad de las mediciones.

Al consultar sobre la factibilidad del uso de los datos climáticos de Villa Gesell para entrenar un modelo de ML cuya área de interés se encuentra a 25 kilómetros de distancia, Saucedo confirmó dicha factibilidad justificando la similitud de condiciones climáticas en el área circundante. A su vez, afirmó que los datos de la estación de Villa Gesell son válidos para un área 40 kilómetros a la redonda gracias a la topografía particular de la zona costera. Esta afirmación confirma la posibilidad de ampliación del área de predicción de incendios a una escala regional en un futuro *release*.

Sin embargo, se remarcó que, dado que las estaciones del SMN son supervisadas por personal del Servicio, no se registran datos en el horario nocturno (de 22 horas hasta las 8 horas) ya que al no tratarse de una zona donde se necesiten pronósticos del tiempo las 24 horas (como sí ocurre en las zonas cercanas a aeropuertos) no hay personal contratado en dicho turno. Para calcular datos aproximados en esa franja de tiempo, Saucedo recomendó promediar los valores climáticos máximos y mínimos del día.

2.3.3. Conclusión

Todos los entrevistados confirmaron la relación que existe entre las condiciones climáticas y la probabilidad de incendios forestales. Sin embargo, coincidieron en que tanto el FWI como los datos climáticos son globales a toda la extensión del partido, por lo que a través de ellos únicamente no es posible inferir patrones o determinar ubicaciones específicas en donde el peligro de incendio es mayor o menor. Si bien los bomberos locales cuentan con conocimiento empírico sobre las zonas donde la susceptibilidad de producirse incendios forestales es mayor, no han determinado aún una relación clara entre los incendios que se han producido en estas ubicaciones y las condiciones climáticas que se dieron en ese momento.

No obstante, se pudo detectar una relación entre los lugares donde los bomberos afirman que los incendios que se producen son más importantes y la variación climática anual que Saucedo recomendó analizar. Esta relación se refleja en el NDVI, que permite determinar zonas con mayor superficie cubierta por vegetación y, en consecuencia, zonas más susceptibles de ocurrencia de incendios de gran magnitud.

En conclusión, determinando la relación geoespacial y temporal de los incendios forestales y las condiciones meteorológicas de un momento dado, los bomberos podrían anticiparse a aquellos de mayor magnitud, prepararse con más información sobre cómo evolucionará un incendio e incluso prevenirlos activamente, ahorrando tiempo y recursos valiosos.

3. Descripción

El objetivo principal de AQUA es de proveer a bomberos, entidades gubernamentales y a la población de Pinamar predicciones de incendios forestales de forma rápida y comprensible a fin de que los distintos organismos involucrados en la toma de decisiones cuenten con más información a la hora de gestionar recursos para la prevención de incendios forestales. Para cumplir dicho objetivo se desarrollaron los siguientes componentes que conforman el núcleo de la solución, a saber:

- *Pipeline* de recolección, transformación y procesamiento de datos.
- Entrenamiento de modelos predictivos.
- Software de visualización de predicciones.

Consecuentemente, en esta sección se especifican los atributos de calidad, la arquitectura de la solución y los detalles de cada componente que la conforman, así también como las etapas que guiaron su desarrollo.

3.1. Atributos de calidad

Los atributos de calidad permiten medir diversas propiedades de un sistema para determinar el grado en el que dicho sistema cumple con los objetivos y necesidades de distintos *stakeholders* (Bass et al., 2012). Por esta razón, y con el fin de justificar las decisiones arquitectónicas y guiar el desarrollo de la solución, se han definido atributos de calidad para AQUA en el marco de la disponibilidad, modificabilidad y usabilidad tal como se detalla a continuación.

3.1.1. Disponibilidad

La disponibilidad de un sistema se refiere al nivel de operabilidad de este, junto con la habilidad que tiene para recuperarse de fallas de modo tal que el tiempo fuera de servicio no exceda un umbral considerado como aceptable para un intervalo de tiempo (Bass et al., 2012). En este sentido, con AQUA se busca que en caso de que el Servicio de Predicciones no responda peticiones de los clientes se redirija el tráfico a un servidor pasivo para que bomberos y entidades gubernamentales puedan consultar las predicciones en todo momento. La definición de este atributo de calidad se detalla en la Tabla II.

TABLA II: Atributo de calidad de disponibilidad.

Elemento	Descripción
Origen del estímulo	El Servicio de Predicciones no responde peticiones de clientes.
Estímulo	Petición de predicción de incendios forestales a resolver.
Ambiente	Operación normal.
Componentes	Servicio de Predicciones, API Gateway.
Respuesta	Fallo registrado y notificado, tráfico redirigido a servidor pasivo, predicciones realizadas y enviadas a clientes.
Medida de respuesta	Tráfico redirigido a servidor pasivo en menos de 3 segundos, 100% de peticiones realizadas y enviadas.

3.1.2. Modificabilidad

El eje de este atributo de calidad está conformado por el costo y riesgo de realizar cambios sobre un sistema. Por ello es fundamental controlar la complejidad del proceso de modificación a través de la definición de módulos con bajo acoplamiento y alta cohesión, de forma tal que cada uno sea independiente de los cambios que se efectúan sobre otros módulos y a su vez realice únicamente la tarea para la cual fue diseñado (Bass et al., 2012). En este aspecto, la incorporación a AQUA de nuevas localidades o áreas de interés en donde predecir incendios forestales no debe afectar las funcionalidades ya implementadas, tal como se detalla en la Tabla III.

TABLA III: Atributo de calidad de modificabilidad.

Elemento	Descripción
Origen del estímulo	Usuario final.
Estímulo	Petición para añadir nuevas ubicaciones a la predicción de incendios forestales.
Ambiente	Etapa de operación.
Componentes	*Pipeline* de datos, Entrenamiento de modelos de Machine Learning.
Respuesta	Ubicaciones añadidas y consideradas en el entrenamiento de modelos predictivos, modelo productivo desplegado.

Medida de respuesta	Ubicaciones añadidas y consideradas en el entrenamiento de modelos predictivos en menos de una semana, modelo productivo desplegado en un lapso menor a 2 semanas.

3.1.3. Usabilidad

La usabilidad de un sistema alude al grado de facilidad en que un usuario puede llevar a cabo determinadas tareas dentro del mismo. Algunas áreas de la usabilidad incluyen la incorporación de tutoriales sobre el uso del sistema para que el usuario se familiarice con las funcionalidades, el desarrollo de interfaces de usuario que minimicen la probabilidad e impacto de que el usuario cometa un error, y la adaptación de la interfaz para satisfacer las necesidades del usuario (Bass et al., 2012). En consecuencia, el software de visualización de AQUA debe ser comprensible por las distintas entidades que harán uso de las predicciones de incendios forestales para tomar o no medidas preventivas tal como se detalla en la Tabla IV.

TABLA IV: Atributo de calidad de usabilidad.

Elemento	**Descripción**
Origen del estímulo	Usuario final.
Estímulo	Usuario consulta las predicciones de incendios forestales para una zona dada.
Ambiente	Estado normal, en tiempo de ejecución.
Componentes	Software de visualización de predicciones.
Respuesta	Se presentan las predicciones de manera familiar para el usuario, presentando también un cuadro de ayuda.
Medida de respuesta	El usuario consulta las predicciones e interactúa con el sistema en un lapso de 2 segundos, o recurre al cuadro de ayuda y efectúa la consulta en un lapso menor a 30 segundos.

3.2. Arquitectura conceptual

Dentro de la arquitectura conceptual de AQUA detallada en la Figura 10 se pueden distinguir dos ramas: la rama de laboratorio y la de producción. Por un lado, la rama de laboratorio está conformada por el *pipeline* de datos cuya salida -el *dataset*- es utilizada por el

componente encargado de entrenar los distintos modelos de ML. Por otro lado, en la rama de producción los modelos entrenados en la rama de laboratorio se despliegan a la Nube para que el Servicio de Predicciones pueda consumir las predicciones.

La principal distinción entre ambas ramas radica en el objetivo con el que fueron desarrolladas: la rama de laboratorio está diseñada para la experimentación con datos y modelos, mientras que la rama de producción tiene como objetivo principal poner a disposición las predicciones de los modelos productivos a los componentes que las requieran. Además, en el marco del presente trabajo la rama de laboratorio se trabajó de forma *offline*, esto es, sin haber contratado a un proveedor de servicios en la Nube. No obstante, los componentes de esta rama (el *pipeline* de datos y entrenador de modelos) se encuentran empaquetados en contenedores Docker, lo que permite en futuras iteraciones ser desplegados fácilmente para trabajar en la Nube.

Teniendo en cuenta que los bomberos cuentan con un sistema de administración y gestión documental (Martínez Saucedo et al., 2021), se planteó una arquitectura de microservicios para la rama de producción en vistas de una futura integración con dicho sistema. Entre las ventajas de adoptar una arquitectura de microservicios en este contexto se encuentran la posibilidad de utilizar tecnologías especializadas para cada problema a resolver (en este caso utilizando *frameworks* y lenguajes de programación utilizados para ML), la capacidad de aislar fallos en determinados servicios sin afectar el funcionamiento de todo el sistema, y la facilidad para escalar aquellos servicios que lo necesiten (Newman, 2015).

En consideración con el alcance definido para el presente trabajo se definieron dos microservicios: el Servicio de Predicciones, encargado de proveer las predicciones a clientes a través de una API REST; y el Servicio Histórico, que consulta una base de datos NoSQL para poner a disposición la información acerca de los incendios forestales ocurridos para que, junto con las predicciones, los usuarios finales que interactúan con la aplicación de visualización puedan interpretar y contrastar las predicciones con la historia de incendios forestales en Pinamar. Ambos microservicios son consumidos por los clientes a través de un API Gateway encargado de enrutar las peticiones de los clientes al servicio correspondiente, traducir protocolos y administrar *cross-cutting concerns* como la gestión de logs, autorización y autenticación.

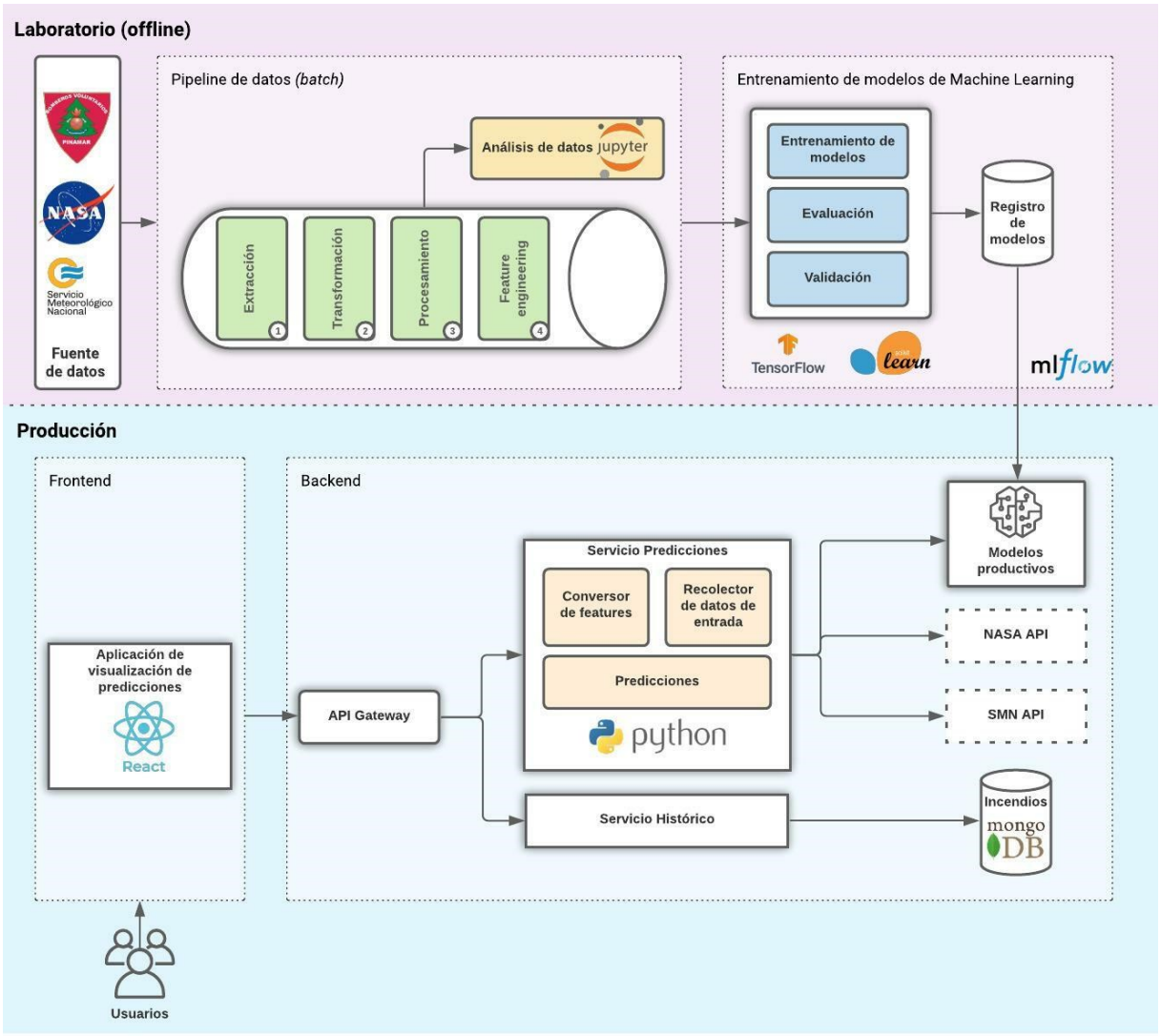

Figura 10: Arquitectura conceptual de la solución: *pipeline* de datos, entramiento de modelos y aplicación de visualización.

En las siguientes secciones se detallan en profundidad las características de cada uno de los componentes de la arquitectura propuesta, así también como sus responsabilidades y dependencias con otros componentes.

3.3. Pipeline de recolección de datos

El objetivo de este componente es recopilar los datos provenientes de distintas fuentes para transformarlos y procesarlos a fin de obtener el *dataset* final con el que se entrenarán los diversos modelos predictivos de incendios forestales. Para esto se desarrolló un *pipeline* automatizado en el que cada paso (Figura 11) se ejecuta 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. Como se contempla ampliar la zona de predicción a localidades aledañas al Partido de Pinamar, el *pipeline* admite configurar diversos parámetros (ver Anexo B) a través de archivos de configuración externos que faciliten la selección de la extensión geográfica y temporal de los datos.

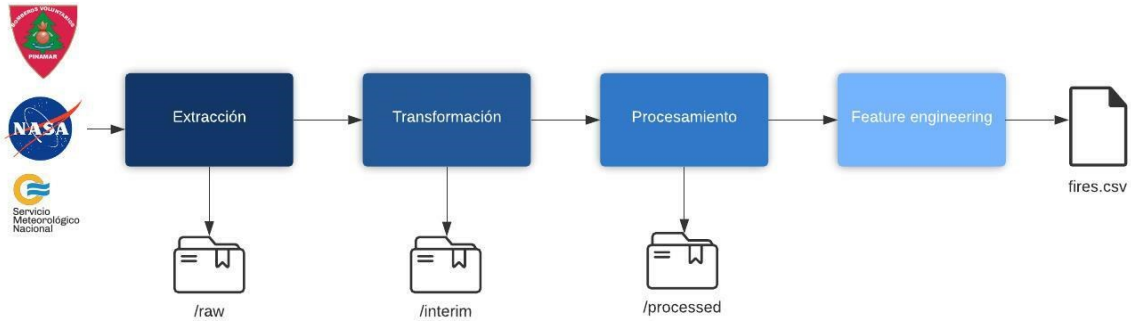

Figura 11: Arquitectura del *pipeline* de datos.

En particular, el *pipeline* se encarga de reunir y crear los atributos descritos en la Tabla V, que tiene en cuenta los factores influyentes durante los incendios forestales y son descritos en el mencionado Triangulo de Comportamiento del Fuego (TCF).

TABLA V. Atributos recopilados y generados por el *pipeline* de datos.

Atributo	**Descripción**	**Valores válidos**	**Unidad de medida**	**Factor en TCF**
Día	Día del mes.	1 - 31	N/A	N/A
Mes	Mes del año.	1 - 12	N/A	N/A
Día no laboral	Feriados, sábados y domingos.	0 - 1	N/A	N/A
Hora	Hora del incendio.	0 - 23	N/A	N/A
X	Valor en eje X (Longitud).	0 - 70	N/A	N/A
Y	Valor en eje Y (Latitud).	0 - 64	N/A	N/A
LST	Temperatura de Superficie del Suelo.	150 - 1310.7	K (Kelvin)	Topografía

NDVI	Índice de Vegetación de Diferencia Normalizada.	-0.2 - 1	N/A	Combustible
Elevación	Elevación.	Escala abierta	m (Metros)	Topografía
DC	Código de sequía.	0 - 1000	N/A	Combustible
DMC	Código de humedad del mantillo.	Escala abierta	N/A	Combustible Meteorología
FFMC	Código de humedad de combustibles finos.	0 - 101	N/A	Combustible Meteorología
ISI	Velocidad de propagación del incendio.	Escala abierta	N/A	Combustible Meteorología
BUI	Combustible disponible.	Escala abierta	N/A	Combustible Meteorología
FWI	Índice meteorológico de peligro de incendio.	Escala abierta	N/A	Combustible Meteorología
Temperatura	Temperatura (dato horario).	Escala abierta	°C (Grados centígrados)	Meteorología
Humedad	Humedad relativa (dato horario).	0 - 100	% (Porcentaje)	Meteorología
Viento	Velocidad del viento (dato horario).	0 - 100	km/h (Kilómetros por hora)	Meteorología
Precipitaciones	Precipitaciones (últimas 24 horas).	0 - 100	mm (Milímetros)	Meteorología

A continuación se describen las responsabilidades y tareas llevadas a cabo por cada componente del *pipeline*, junto con las decisiones técnicas que debieron tomarse frente a distintos desafíos que se presentaron en la etapa de desarrollo.

3.3.1. Extracción

Para desarrollar el modelo predictivo de incendios forestales en el Partido de Pinamar se recopilaron datos de incendios forestales históricos, datos climáticos recolectados por la estación meteorológica de Villa Gesell y datos ambientales provenientes de satélites de observación terrestre entre los años 2015 y 2019 inclusive. La razón por la cual se eligieron dichos años radica en la disponibilidad de registros de incendios por parte de la Asociación de Bomberos Voluntarios de Pinamar (ABVP), ya que el registro en soporte digital de los mismos comenzó a realizarse en el año 2015. Además, se excluyó el año 2020 ya que por la pandemia de COVID-19 los bomberos no han podido acercarse al cuartel para cumplir tareas administrativas y en consecuencia generar los partes correspondientes a los incendios de dicho año. Asimismo, según indicó el ayudante principal de bomberos, Matías García, el año 2020 fue atípico en cuanto a incendios, ya que durante los meses en que rigió el decreto presidencial de Aislamiento Social Preventivo y Obligatorio (ASPO) (Boletín Oficial de la República Argentina, 2020) no hubo incendios forestales.

Con el objetivo de obtener información histórica de incendios forestales en el Partido de Pinamar se recopilaron diversas planillas de incendios en formato Excel donde se incluyen datos como la fecha y hora, duración, superficie quemada, móviles y bomberos involucrados en su extinción. Dado que dichas planillas no incluyen la ubicación de los incendios, se procesaron manualmente los partes en formato papel que son almacenados en los archivos de la Asociación de Bomberos Voluntarios de Pinamar para poder completar la información faltante.

En cuanto a los datos climáticos, tal como recomendó Marcos Saucedo se utilizaron los recopilados por la estación meteorológica de la vecina ciudad de Villa Gesell. Para ello se realizó un pedido de información meteorológica 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 de diversas estaciones meteorológicas a partir del 26 de noviembre de 2017, aunque en el presente trabajo se desarrolló un componente para consultar la API en vistas de ampliar el rango de años en futuras iteraciones.

Tal como se evidencia en la Tabla VI, la diversidad de orígenes, resoluciones espaciales y formatos de datos dificulta el trabajo con modelos de ML. Por estas razones, en las

etapas posteriores del *pipeline* estos datos crudos se transforman y procesan para trabajar con ellos de forma unificada.

TABLA VI: Origen y formato de los datos extraídos en el *pipeline* de datos.

Descripción	**Origen / Dataset**	**Resolución espacial**	**Generación de los datos**	**Cobertura geográfica**	**Formato**
Incendios forestales (Día, Mes, Día no laboral, Hora, X, Y)	ABVP / Incendios 2015-2019.	N/A.	Horaria.	Partido de Pinamar.	.xlsx (Hoja de cálculo de Microsoft Excel).
Meteorología (Temperatura, Humedad, Viento, Precipitaciones)	SMN.	N/A.	Horaria.	Partido de Pinamar. Partido de Villa Gesell. Partido de General Juan Madariaga. Partido de la Costa.	.xlsx (Hoja de cálculo de Microsoft Excel).
Temperatura de superficie de suelo (LST)	NASA / MYD11C1 v006 (Wan, Zhengming et al., 2015)	0.05° × 0.05°.	Diaria.	Global.	.hdf (Hierarchical Data Format)**.**
Índice de Vegetación de Diferencia	NASA / MYD13Q1 v006	250 metros × 250 metros.	Quincenal.	Partido de Pinamar.	.nc (Archivo NetCDF).

Normalizada (NDVI)	(Didan, Kamel, 2015)				
Índice meteorológico de peligro de incendio e índices derivados (DC, DMC, FFMC, ISI, BUI, FWI)	NASA / GFWED GEOS-5 - GPM Late v5 (Field et al., 2015)	0.1° × 0.1°.	Diaria.	60S – 60N.	.nc (Archivo NetCDF).

3.3.2. 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 (Tabla VII). Para ello se utilizaron las librerías H5py (manipulación y conversión de archivos HDF), xarray (filtrado de *datasets* multidimensionales por coordenadas), Pandas (manipulación y conversión de archivos XLSX), NumPy (creación y manipulación de matrices multidimensionales) y GeoPy (geocodificación de direcciones).

No obstante, los siguientes *datasets* no necesitaron transformaciones por las razones a continuación descritas:

- Meteorología: el CIM proveyó el *dataset* correspondiente a los datos climáticos de la ciudad de Villa Gesell (aplicables a todo el Partido de Pinamar) en un formato tabular fácil de manipular.
- Índice de Vegetación de Diferencia Normalizada (NDVI): la API provista por la NASA para este *dataset* permite el filtrado por coordenadas geográficas al momento de descargarlo, por lo que en esta etapa del *pipeline* no fue necesario acotar la región geográfica.

TABLA VII: Transformaciones aplicadas sobre los distintos *datasets*.

Origen / Dataset	Transformación
ABVP / Incendios 2015-2019	Geocodificación de ubicaciones de incendios consultando a Google Geocoding API. Filtrado de incendios combatidos fuera del Partido de Pinamar. Filtrado de incendios con datos faltantes (fecha, ubicación). Adición de columna "Día no laboral" (fines de semana y feriados).
NASA / MYD13Q1 v006 (Didan, Kamel, 2015)	Filtrado de datos a coordenadas de interés (37°2'S 56° 95' O, 37°05'S 56°8'O).
NASA / GFWED MERRA2 - GPM Late v5 (Field et al., 2015)	Filtrado de datos a coordenadas de interés (37°2'S 56° 95' O, 37°05'S 56°8'O).

3.3.3. Procesamiento

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.

Tal como se describió en la sección de Antecedentes, en la mayoría de los 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). 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. En el marco del proyecto AQUA, el tamaño de esta grilla es de 250 metros × 250 metros (Figura 12), ya que es la resolución mínima disponible en los atributos (NDVI) y el tamaño es representativo considerando la superficie (63 kilómetros cuadrados) y las características topográficas del Partido de Pinamar.

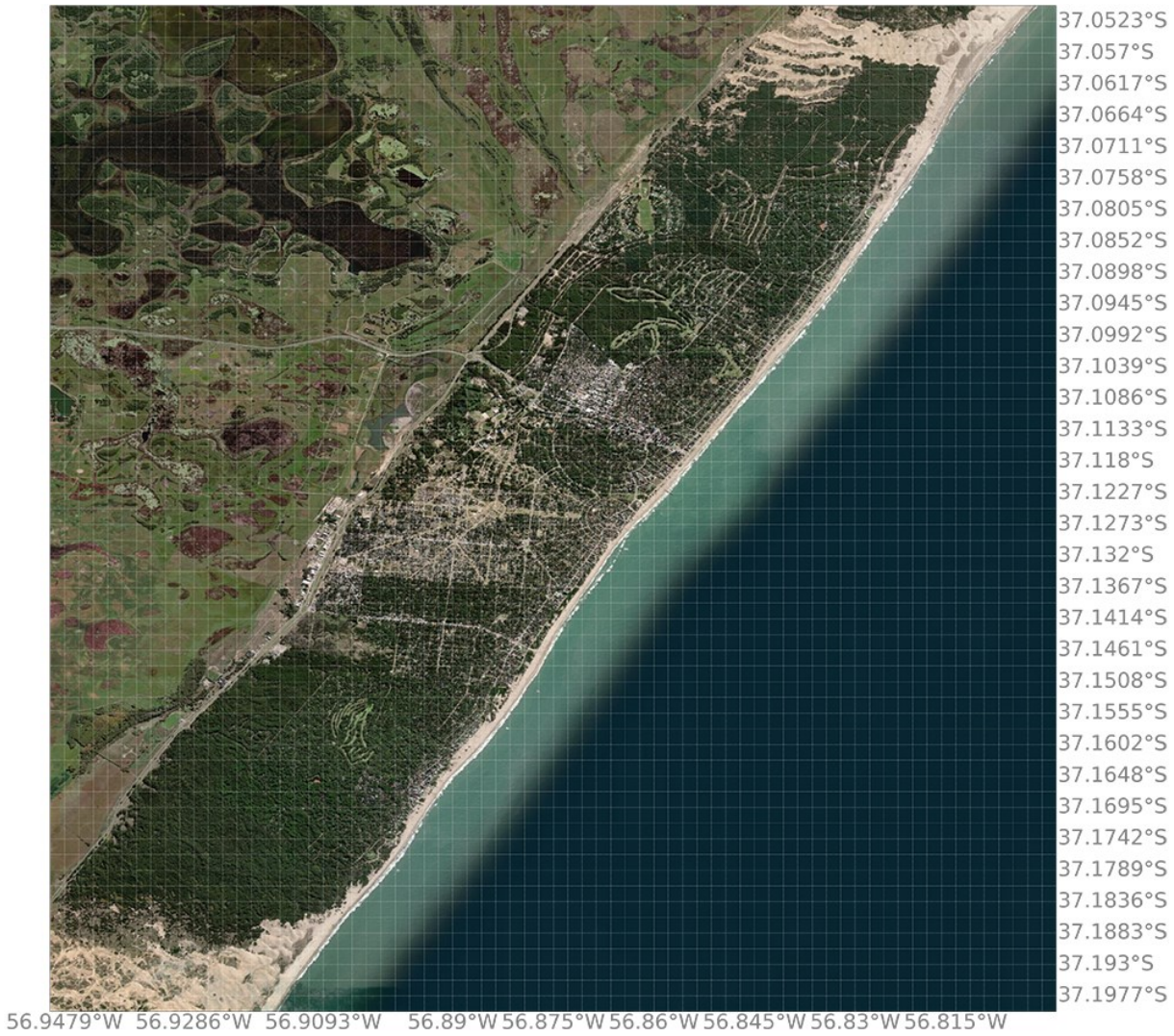

Figura 12: Grilla establecida para el área de interés.

Otra tarea importante que se lleva a cabo en esta fase es la de remuestrear el *dataset* de incendios forestales. Como el objetivo del modelo predictivo es determinar para cada coordenada (X, Y) si ocurrirá o no un incendio forestal dadas ciertas condiciones meteorológicas, topográficas y ambientales, el problema a resolver por el modelo es de clasificación. Los modelos de clasificación tienen buenos resultados cuando la distribución de la variable objetivo (en este caso, 0=No o 1=Sí) está balanceada, esto es, cuando la proporción de casos en donde se produjo un incendio forestal y en donde no es semejante. No obstante, cuando se cuentan con más registros de una determinada clase por sobre otra, el modelo puede ser influenciado por la clase mayoritaria, ignorando completamente la minoritaria. Esta influencia se hace notoria al evaluar los resultados del modelo, en donde la incapacidad para

predecir la clase minoritaria se hace evidente al obtener el porcentaje de predicción para dicha clase (Yap et al., 2014).

Como los datos que se cuentan corresponden enteramente a la clase 1 (Incendio), en esta etapa se generó aleatoriamente una cantidad proporcionada de registros de "No Incendio" en concordancia con algoritmos propuestos en la literatura por autores como (Stojanova et al., 2012). 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. En particular, el algoritmo empleado para la generación de puntos de clase 0 se detalla en el Anexo C.

3.3.4. Análisis de datos

La etapa de 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 distintos gráficos para adquirir entendimiento cualitativo de los incendios forestales ocurridos entre los años 2015 y 2019. Gracias a estas visualizaciones (Anexo D) 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.
- 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.

Posterior al análisis de las distintas visualizaciones se elaboró la matriz de correlación entre variables (Figura 13) para entender cómo éstas se relacionan entre sí utilizando el método estadístico de Spearman, que es el apropiado cuando la distribución de los datos no es normal. Los valores de correlación se pueden interpretar de la siguiente manera:

- 0 – 0.25: relación escasa o nula.
- 0.26 – 0.5: relación débil.
- 0.51 – 0.75: relación moderada a fuerte.

- 0.76 – 1: relación fuerte a perfecta.

(Martínez Ortega et al., 2009).

La matriz de correlación muestra que, si bien las variables están relacionadas, la mayoría tienen una relación débil entre sí (valores menores a 0.5). No obstante, la correlación entre variables no implica causalidad, por lo que no se puede concluir que una variable tenga un efecto sobre otra.

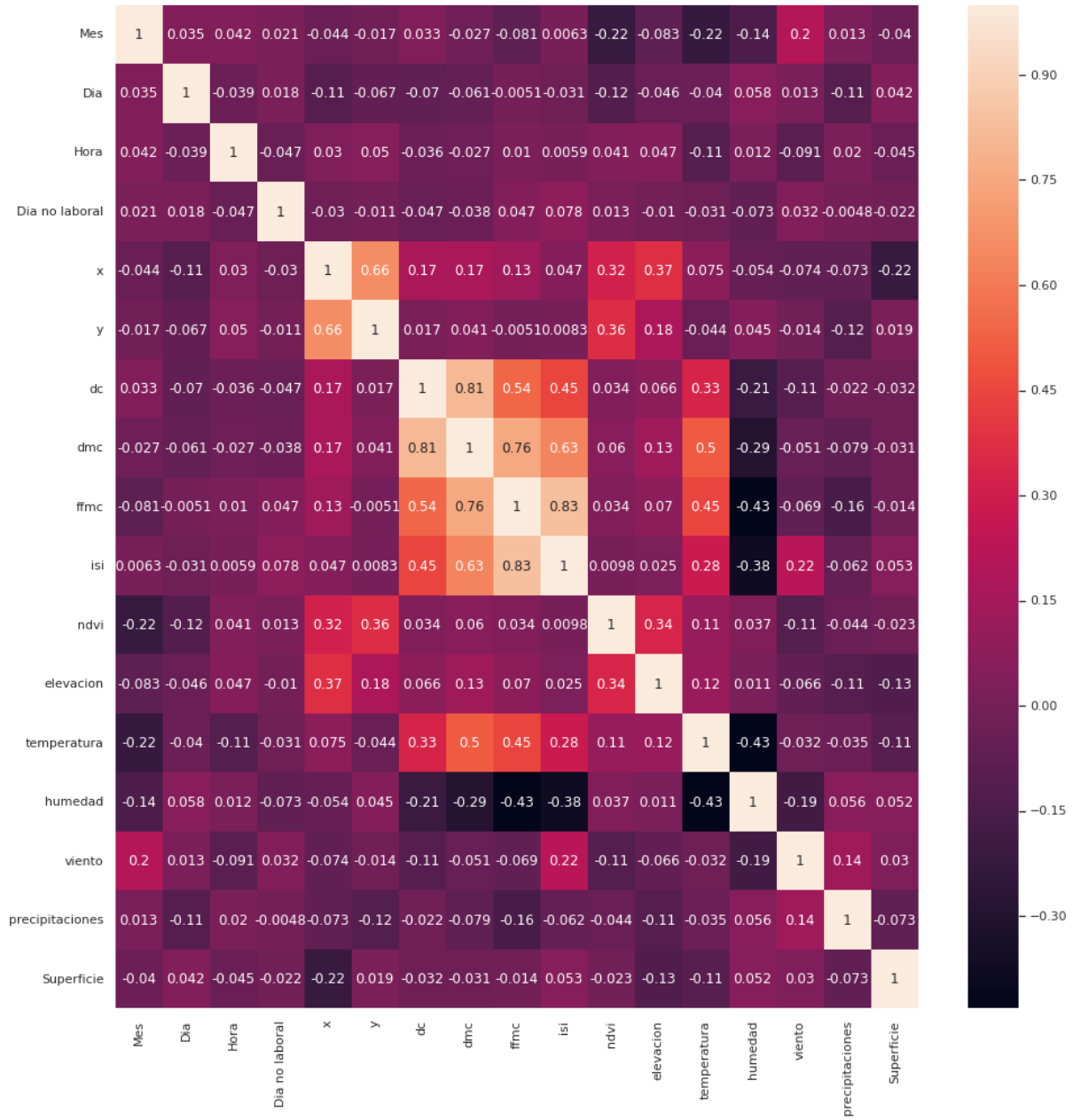

Figura 13: Matriz de correlación entre las variables. Valores cercanos a 1 o -1 indican una relación fuerte entre variables.

En línea con la elaboración de la matriz de correlación y con el objetivo de detectar si existe colinealidad entre las variables (esto es, que dos variables independientes tengan una alta correlación) se llevó a cabo el método de factor de inflación de la varianza (o VIF por sus siglas en inglés *Variance Inflation Index*). De esta forma se busca evitar que los modelos de regresión no sean capaces de distinguir los efectos que cada variable independiente provoca sobre la variable dependiente. En general, un valor de VIF mayor o igual a 5 indica un problema de multicolinealidad (Ramasubramanian et al., 2016). No obstante, en el *dataset* de incendios sólo se descartaron las variables BUI, LST y FWI ya que sus VIF son mayores a 300. La razón por la cual estas variables adquieren valores tan altos radica en que BUI y FWI son producto de la combinación lineal entre otros índices, mientras que LST está estrechamente relacionada con la temperatura que miden las estaciones sinópticas de superficie. Los valores de VIF obtenidos por cada atributo se encuentran detallados en el Anexo E.

3.3.5. Feature engineering

Como consecuencia del conocimiento adquirido sobre el *dataset* en la etapa de Análisis de Datos, en la etapa de *feature engineering* o ingeniería de características 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 VIII. De esta manera, 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.

TABLA VIII: Agrupamiento de atributos según el tipo de variable.

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.

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 (Zheng et al., 2018). La elección de este método por sobre otros tales como *Integer Encoding* está justificada por el hecho de que *One-Hot-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.

x	y	dia no labora	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

Figura 14: *Dataset* final de incendios forestales. Para una mejor visualización se omitieron las columnas generadas al aplicar la técnica de *One-Hot-Encoding*.

3.4.Entrenamiento de modelos

La predicción de incendios forestales a través de modelos de ML ha sido estudiada en los últimos años desde distintas perspectivas. En determinados trabajos se ha buscado predecir si dadas ciertas condiciones ambientales, meteorológicas o demográficas se producirá un incendio o no (problema de clasificación binaria), mientras que en otros el objetivo

ha sido predecir la superficie que podría ser quemada en caso de que efectivamente se desarrolle un incendio forestal (problema de regresión).

La experiencia manifestada en la literatura y en el presente trabajo demuestran que la distribución de la magnitud de incendios forestales es altamente despareja: según el conocimiento adquirido en la etapa de Análisis de Datos, la mayoría de los incendios forestales que se producen a lo largo del tiempo son pequeños (menores a una hectárea en el caso de Pinamar), aunque representan el 37% de la superficie quemada históricamente. Por el contrario, el 61% de la superficie quemada entre los años 2015-2019 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.

Con el fin de abordar este problema, y en línea con trabajos como el de (Wang, Sally et al., 2019), el enfoque adoptado con AQUA es el de dividir en dos pasos la predicción de incendios forestales (Figura 15): en primer lugar, se predice por cada coordenada de la grilla si se producirá o no un incendio forestal (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), se predicen las hectáreas que podrían quemarse (regresión). En el caso particular de la predicción de hectáreas quemadas se han abordado enfoques alternativos a la regresión lineal en vistas de obtener mejores resultados, como la regresión por cuantiles y la predicción de clases de incendio forestal según su magnitud en hectáreas.

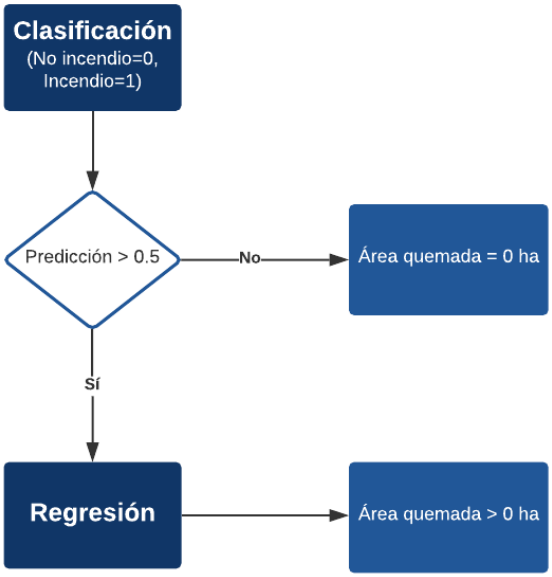

Figura 15: Diagrama de los pasos en la fase de entrenamiento de modelos.

Una de las tareas previas al entrenamiento de modelos consiste en preparar los datos, debido a que el rendimiento de determinados algoritmos como SVM y redes neuronales artificiales convergen más rápido a una solución cuando los datos tienen valores más pequeños o se asemejan a una distribución normal. En consecuencia, se aplicó la técnica de escalada *MinMax* (10) que transforma todas las características para que sus valores se encuentren en el rango [0, 1]. De esta manera, la distribución original de las características no se ve alterada y la distancia relativa entre los valores se mantiene. (Esposito et al., 2020).

$$X_{escalado} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (10)$$

En la instancia anterior al entrenamiento de modelos se dividió el *dataset* original en los conjuntos de entrenamiento (75% de los registros) y prueba (25% de los registros) para trabajar con los distintos algoritmos de ML. En el caso particular de los algoritmos de regresión se aplicó previamente una transformación logarítmica sobre la variable objetivo (Superficie) para reducir la asimetría de la distribución tal como se puede apreciar en la Figura 16.

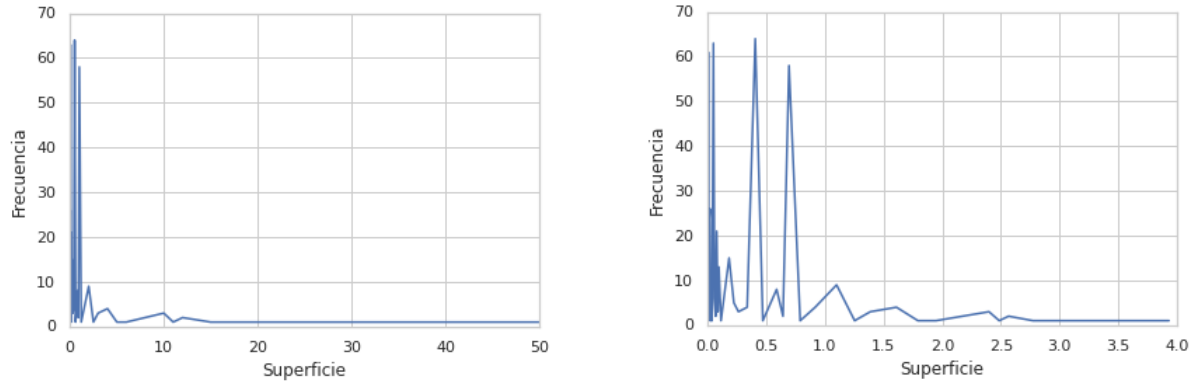

Figura 16. Distribución de la variable “Superficie” antes y después de su transformación.

El proceso de entrenamiento de modelos está compuesto por una serie de pasos comunes a todos los algoritmos utilizados, y por cada uno se efectuaron las siguientes tareas:

- Optimización de hiperparámetros aplicando la técnica de validación cruzada, utilizando 5 particiones sobre el conjunto de datos de entrenamiento.
- Entrenamiento del modelo con el conjunto de datos de entrenamiento utilizando las librerías y frameworks Scikit-learn, Tensorflow y Keras.

- Evaluación de modelos según tipo de problema con el conjunto de datos de pruebas.
- Almacenamiento del modelo entrenado al registro de modelos.

Puesto que se han evaluado diversos algoritmos para resolver tanto problemas de clasificación como regresión, se utilizó la herramienta MLflow para gestionar los modelos entrenados, sus versiones, hiperparámetros y métricas de evaluación. De igual manera, MLflow permite desplegar estos modelos localmente para realizar tanto pruebas de concepto como demostraciones en las etapas tempranas del entrenamiento.

A continuación se detallan los hiperparámetros optimizados para cada algoritmo, junto con los enfoques alternativos considerados a la hora de predecir la superficie que un incendio forestal puede quemar.

3.4.1. Modelos de predicción de incendios forestales

La predicción de ocurrencia o no de incendios forestales es un problema de clasificación binaria, en donde la variable objetivo puede tomar el valor 0 para indicar la no ocurrencia de un incendio forestal y el valor 1 para la ocurrencia. Por esta razón se consideraron algoritmos de clasificación para el entrenamiento de los modelos tal como se puede comprobar en la Tabla IX.

TABLA IX: Estrategia de optimización de hiperparámetros para algoritmos de clasificación.

Algoritmo	**Tipo de búsqueda de parámetros**	**Hiperparámetros evaluados**
Red neuronal artificial	En grilla	Cantidad de nodos por capa: [(50, 50, 50), (50, 100, 50), (50, 25, 10), (100,)]. Función de activación: [ReLu, Tangente Hiperbólica]. Optimizador de pesos: [Gradiente Estocástico Descendiente, Adam]. Alpha: [0.05, 0.001, 0.0001, 0.00001]. Tasa de aprendizaje: [Constante, Adaptativa].
Árbol de decisión	En grilla	Profundidad máxima del árbol: [10, 6, 3, Ilimitada].

		Cantidad máxima de características a considerar: [33, 26, 20, 10, 5]. Número mínimo de ejemplos para ser un nodo hoja: [1, 3, 5, 9, 10]. Criterio de Ganancia de Información: [Gini, Entropía].
Gradient Boosting	Aleatoria	Función de pérdida: [Deviance]. Tasa de aprendizaje: [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2]. Número mínimo de ejemplos para dividir un nodo interno: [0.1, 0.14, 0.17, 0.21, 0.25, 0.28, 0.32, 0.35, 0.39, 0.43, 0.46, 0.5]. Número mínimo de ejemplos para ser un nodo hoja: [0.1, 0.14, 0.17, 0.21, 0.25, 0.28, 0.32, 0.35, 0.39, 0.43, 0.46, 0.5]. Profundidad máxima de cada clasificador: [3, 5, 8]. Cantidad máxima de características a considerar: [$log_{2\ (características)}$,$\sqrt{características}$]. Función de calidad de división: [Friedman, MSE]. Fracción de los ejemplos para entrenar clasificadores individuales: [0.5, 0.62, 0.8, 0.85, 0.9, 0.95, 1.0]. Cantidad de etapas de *boosting*: [10, 50, 100, 200, 500, 1000].
Regresión Logística	En grilla	C: valores de 0.00002 a 1 espaciados linealmente de a 100 valores.
Random Forest	En grilla	Bootstrap: [Sí]. Profundidad máxima de cada clasificador: [80, 90, 100, 110, 120, 150]. Cantidad máxima de características a considerar: [4, 6, 12, 24, 33]. Número mínimo de ejemplos para dividir un nodo interno: [8, 10, 12].

		Número mínimo de ejemplos para ser un nodo hoja: [3, 4, 5]. Número de árboles en el bosque: [100, 200, 500, 1000].
SVM	En grilla	C: valores de 0.1 a 10000 espaciados linealmente de a 10 valores. Gamma: [1, 0.1, 0.001, 0.0001]. Kernel: [Linear, Polinómico, RBF].

3.4.2. Modelos de predicción de superficie quemada

La predicción de la superficie quemada por un incendio forestal es inherentemente un problema de regresión, ya que el objetivo radica en obtener un número que represente esa superficie. No obstante, además de plantear este problema como uno de regresión tradicional, se consideró la regresión por cuantiles para predecir la superficie quemada debido a que ha sido utilizada en trabajos como los de (Rijal, 2018) y (Wang, Sally et al., 2019) y presenta las siguientes ventajas:

- Se modela la distribución condicional de la variable objetivo.
- No se asume la distribución de la variable objetivo.
- No es sensible a *outliers.*

(Rodriguez, 2017).

Los distintos algoritmos de regresión utilizados junto con los hiperparámetros optimizados se detallan en la Tabla VI. En el caso particular de los modelos de redes neuronales artificiales se realizó una integración con la herramienta Wandb durante el entrenamiento, ya que permite visualizar la importancia de los hiperparámetros ajustados, el rendimiento de los modelos conforme se ajusta dichos hiperparámetros y la arquitectura de las redes entrenadas.

TABLA X: Estrategia de optimización de hiperparámetros para algoritmos de regresión.

Algoritmo	**Tipo de búsqueda de parámetros**	**Hiperparámetros evaluados**
Red neuronal artificial	Aleatoria	Tasa de aprendizaje: [0.001, 0.0001, 0.00001, 0.000001, 0.0000001]. Dilución: [0.15, 0.2, 0.25, 0.3, 0.4, 0.5].

		Cantidad de nodos en capa oculta n°1: [4, 9, 14, 18, 25]. Cantidad de nodos en capa oculta n°2: [4, 9, 14, 18, 25]. Función de activación: [ReLu, Tangente Hiperbólica, Sigmoide].
Árbol de decisión	En grilla	Profundidad máxima del árbol: [10, 6, 3, Ilimitada]. Cantidad máxima de características a considerar: [33, 20, 10, 5]. Número mínimo de ejemplos para ser un nodo hoja: [1, 3, 5, 9, 10]. Criterio de Ganancia de Información: [MAE, Friedman].
Gradient Boosting (Cuantil 0.05)	En grilla	Cuantil: 0.05. Función de pérdida: Cuantil. Tasa de aprendizaje: [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2]. Cantidad de etapas de *boosting*: [10, 50, 100, 200, 500, 1000]. Profundidad máxima de cada regresor: [2, 5, 10, 15, 20, 80, 100]. Número mínimo de ejemplos para ser un nodo hoja: [1, 5, 10, 20, 30, 50]. Número mínimo de ejemplos para dividir un nodo interno: [2, 5, 10, 20, 30, 50].
Gradient Boosting (Cuantil 0.95)	En grilla	Cuantil: 0.95. Función de pérdida: Cuantil. Tasa de aprendizaje: [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2].

		Cantidad de etapas de *boosting*: [10, 50, 100, 200, 500, 1000]. Profundidad máxima de cada regresor: [2, 5, 10, 15, 20, 80, 100]. Número mínimo de ejemplos para ser un nodo hoja: [1, 5, 10, 20, 30, 50]. Número mínimo de ejemplos para dividir un nodo interno: [2, 5, 10, 20, 30, 50].
Random Forest	En grilla	Bootstrap: [Sí]. Profundidad máxima de cada regresor: [80, 90, 100, 110, 120, 150]. Cantidad máxima de características a considerar: [4, 6, 12, 24, 33]. Número mínimo de ejemplos para dividir un nodo interno: [8, 10, 12]. Número mínimo de ejemplos para ser un nodo hoja: [3, 4, 5]. Número de árboles en el bosque: [100, 200, 500, 1000].
SVM	En grilla	C: valores de 0.1 a 10000 espaciados linealmente de a 10 valores. Gamma: [1, 0.1, 0.001, 0.0001]. Kernel: [Linear, Polinómico, RBF].

De forma alternativa al abordaje del problema de predicción del área quemada como uno de regresión, se desarrolló una estrategia para predecir clases de incendios forestales según el área quemada. En consecuencia se trata de un problema de clasificación multiclase, por lo que se emplearon los mismos algoritmos utilizados para predecir la ocurrencia o no de incendios forestales. Para ello se establecieron las siguientes categorías para la variable objetivo "Superficie":

- Clase 0: Superficie quemada < 0.05 hectáreas.

- Clase 1: Superficie quemada ≥0.05 hectáreas y < 0.5 hectáreas.
- Clase 2: Superficie quemada ≥ 0.5 hectáreas y < 1 hectárea.
- Clase 3: Superficie quemada > 1 hectárea.

3.5. Visualización de predicciones

Como parte de los objetivos del presente trabajo se encuentra el desarrollo de una aplicación de visualización en donde las predicciones de incendio devueltas por los modelos de ML puedan ser fácilmente interpretadas por los distintos actores que las utilizarán para tomar decisiones. En correspondencia con la arquitectura conceptual planteada, este objetivo es cumplido a través de los componentes que conforman la rama de producción, a su vez dividida en *backend* y *frontend.*

En lo que respecta a componentes que conforman el *backend* de AQUA y proveen los datos necesarios para visualizar las predicciones se pueden distinguir la base de datos documental de incendios históricos, los modelos de ML productivos, y los microservicios responsables de obtener las predicciones de dichos modelos junto con los incendios históricos. Por otro lado, en el primer *release* de AQUA el *frontend* está conformado por una aplicación web, aunque fácilmente pueden agregarse clientes tales como sistemas de terceros o clientes móviles que deseen conocer las predicciones de incendio.

En las siguientes secciones se detallan los componentes que forman parte de la rama productiva de la arquitectura, segmentando el sistema de visualización de predicciones desde el punto de vista lógico: presentación, lógica de negocio y persistencia.

3.5.1. Presentación

La presentación es lo que permite a usuarios finales interactuar con el dominio del negocio. En este caso, la presentación de AQUA consiste en una aplicación web en donde las predicciones de incendio forestal puedan visualizarse fácil y rápidamente mediante mapas de calor, sumado a una capa histórica que permita consolidar lo ocurrido en términos de incendios forestales (pasado) con lo predicho (futuro).

La arquitectura implementada para la aplicación web es Redux, una variación de la arquitectura Flux. De acuerdo con (Garreau et al., 2018) Flux fue concebida como una alternativa al patrón MVC (Modelo Vista Controlador) que proponía un flujo de información bidireccional a través de los modelos, vistas y controladores, pero que resultaba difícil de seguir

en aplicaciones de gran envergadura. En su lugar, Flux propone un flujo unidireccional a través de los siguientes componentes:

- Acciones: inician el cambio de estado de la aplicación. Las mismas son generadas por la interacción del usuario o por un evento del servidor.
- Despachador: un único despachador recibe las acciones de la aplicación para dirigirlas al almacén de estado correspondiente.
- Almacenes de estado: en Flux los datos correspondientes a cada parte del dominio de la aplicación están almacenados en su propio almacén. Al recibir acciones del despachador, el almacén actualiza el estado correspondiente.
- Vistas: son aquellas que se suscriben a las actualizaciones de los almacenes para mostrar datos.

A diferencia de Flux, en Redux existe un único almacén de estado inmutable de la aplicación, y consecuentemente no cuenta con un despachador. Como el estado es inmutable, existen funciones llamadas reductoras (o *reducers*) que especifican cómo se debe transformar el siguiente estado de la aplicación para devolverlo tal como se especifica en la Figura 17. Las ventajas de adoptar esta arquitectura para la aplicación web son la predictibilidad del estado de la aplicación (sobre todo cuando la aplicación crece en complejidad), la facilidad para realizar pruebas y escalar la aplicación.

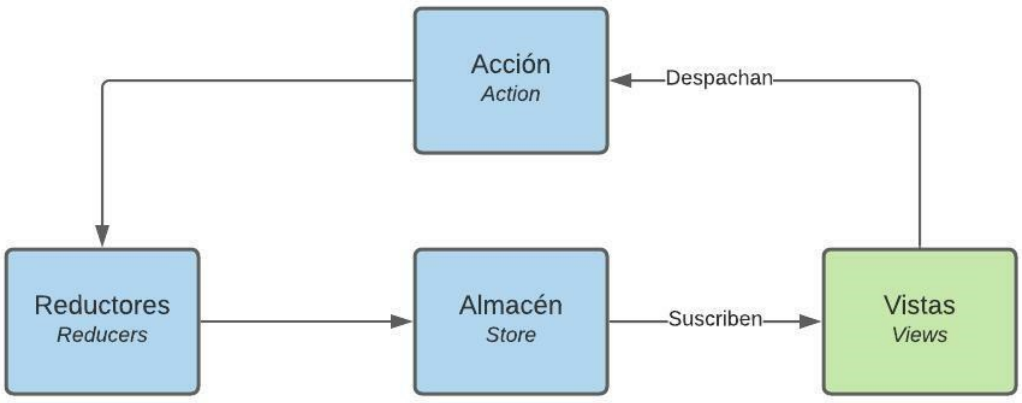

Figura 17: Componentes y flujo de la información en la arquitectura Redux.

En lo que respecta a la interfaz gráfica de la aplicación de visualización de predicciones, se priorizó la facilidad de uso e interpretación de predicciones. Para ello, cuando el usuario desplaza el cursor sobre el mapa las predicciones de incendio se presentan en la barra inferior tal como indica la Figura 18. Por otro lado, con el fin de facilitar la interacción con la aplicación, las capas de predicciones e incendios históricos se pueden habilitar o deshabilitar según la preferencia del usuario. A su vez los incendios forestales pueden filtrarse por fecha con el fin de contrastar el comportamiento de los incendios forestales en el pasado con el que se predice a futuro (Figura 19).

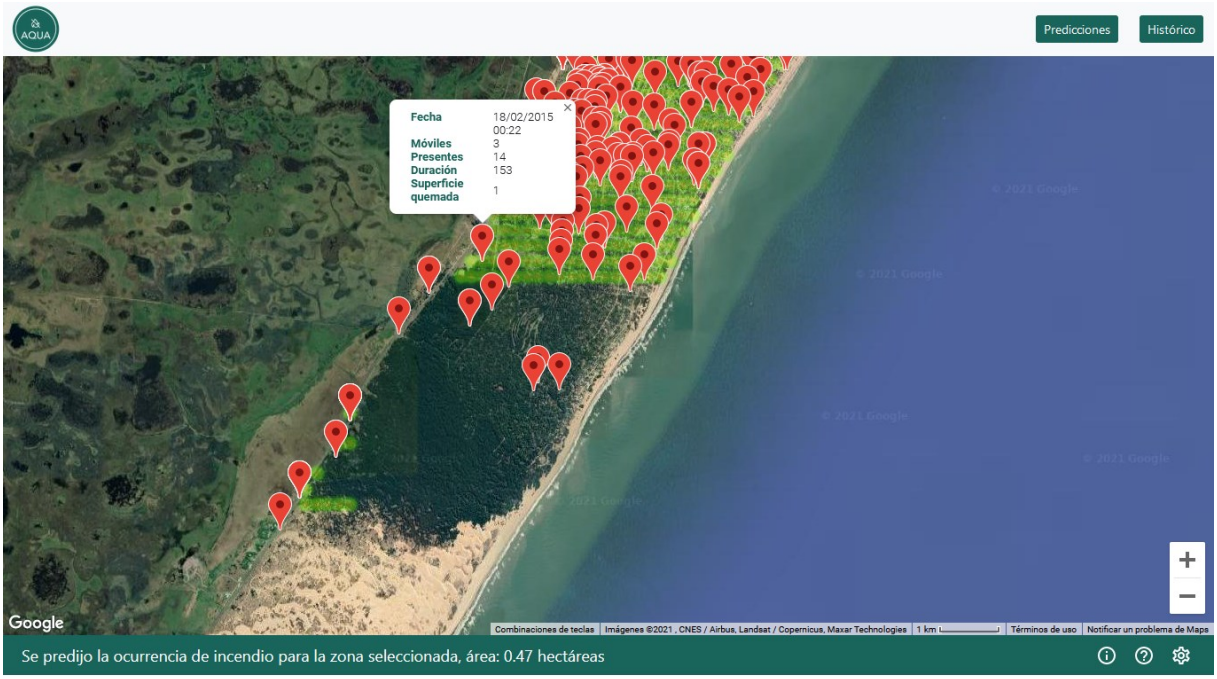

Figura 18: Aplicación de visualización de predicciones con capa histórica activada.

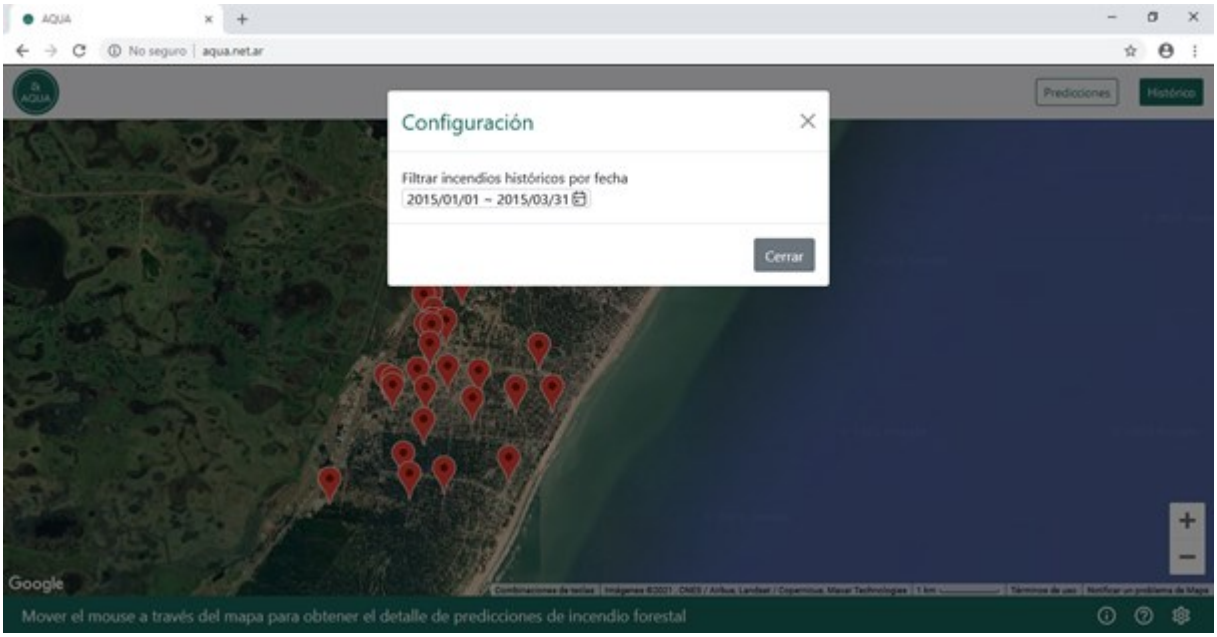

Figura 19: Aplicación de visualización de predicciones con capa histórica activada, capa de predicciones desactivada e incendios históricos filtrados (primer trimestre del año 2015).

3.5.2. Lógica de negocio

La lógica de negocio de AQUA abarca tanto la obtención de predicciones de incendios forestales como de incendios forestales ocurridos en la misma área de interés. Consecuentemente se desarrollaron dos microservicios que aborden estas tareas de forma independiente, considerando las distintas necesidades de tecnologías, escalamiento e integración que presentan ambos servicios. Los recursos que ponen a disposición estos microservicios se detallan en el Anexo G, y la secuencia lógica para tanto obtener las predicciones de incendio como los incendios históricos se describen en el Anexo H.

3.5.2.1. Servicio de predicciones

El servicio de predicciones es el encargado de obtener las predicciones de incendio para una fecha, horario y par de coordenadas solicitados por los clientes. Para ello, este servicio se vale de los siguientes componentes:

- Recolector de datos de entrada (a través del componente *pipeline* de datos).
- Conversor de *features*.
- Modelos productivos de ML.

Dado que la responsabilidad de este servicio está fuertemente atada a modelos de ML, APIs de terceros y módulos externos, se adoptó una arquitectura hexagonal (Figura 20) que permita independizar la lógica de obtención de predicciones de cambios en componentes externos a dicha lógica. Esto es posible ya que el grafo de dependencias externas siempre apunta hacia el interior del hexágono -el dominio-, desligando al mismo del conocimiento acerca de dependencias externas y así lograr un bajo acoplamiento. Otra ventaja que presenta la arquitectura hexagonal es la facilidad de realizar pruebas sobre el dominio, permitiéndose abstraer de la forma en que se comunica con componentes externos.

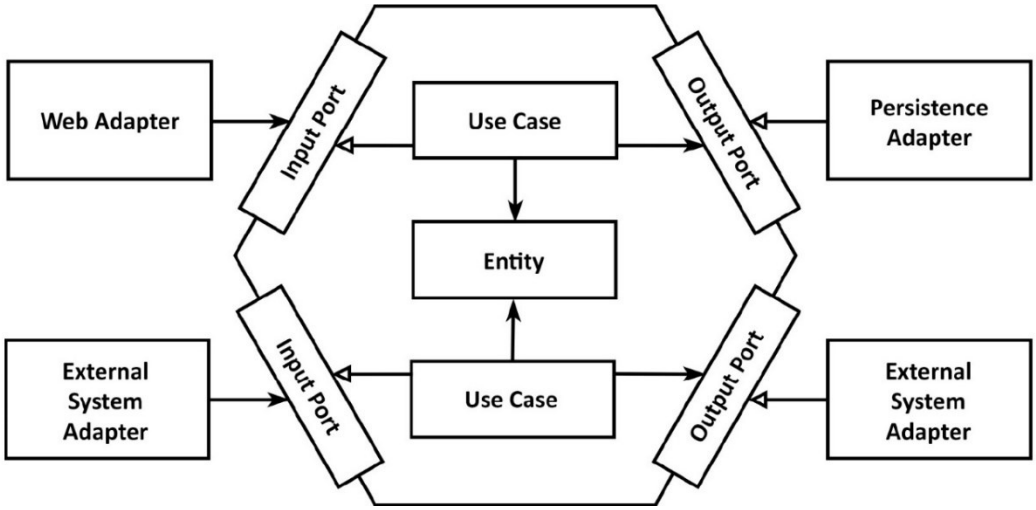

Figura 20: Diagrama conceptual de la arquitectura hexagonal (Hombergs, 2019). El centro del hexágono está conformado por la lógica del dominio, la cual está rodeada puertos y adaptadores de forma tal que el núcleo lógico de la aplicación no se vea afectado por cambios en APIs externas, fuentes de datos o bases de datos.

La arquitectura hexagonal también es conocida bajo el nombre de arquitectura de puertos y adaptadores. Estos conceptos están relacionados a la forma en que la información fluye a través de la arquitectura: los eventos que provienen del ambiente externo llegan a puertos, y luego adaptadores específicos a la tecnología convierten estos eventos en llamadas a la aplicación (Cockburn, 2005). Estos puertos y adaptadores son los que permiten aislar el núcleo de la aplicación, y la implementación de la arquitectura hexagonal del servicio de predicciones se detalla en la Figura 21.

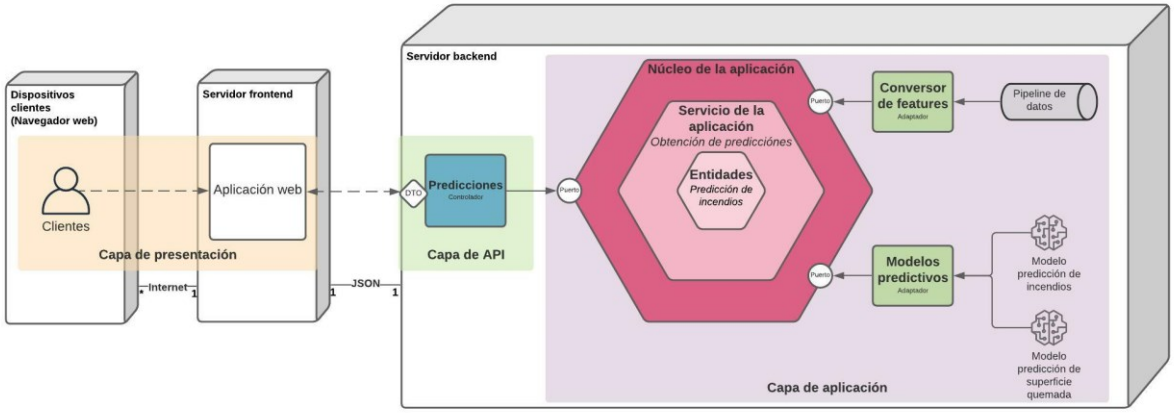

Figura 21: Arquitectura hexagonal del servicio de predicciones (vista lógica y de despliegue combinadas).

3.5.2.2. Servicio histórico

El servicio histórico es responsable de proveer información sobre los incendios que se produjeron en el Partido de Pinamar de modo tal que los usuarios finales de AQUA cuenten con una capa de información extra que les permita contraponer las predicciones con la susceptibilidad de incendios histórica en una zona dada. Para este fin, el servicio obtiene de una base de datos documental (MongoDB) los registros almacenados. Dada la sencillez de la responsabilidad de este servicio, se decidió utilizar la arquitectura en capas horizontales, cada una con un rol específico dentro de la aplicación.

La principal ventaja de la separación de la aplicación en capas es la consecuente separación de responsabilidades, lo que significa que los componentes de una capa sólo se enfocan en tareas pertinentes a la capa en la que se encuentran. Por otro lado, esta arquitectura permite modificar capas individualmente sin afectar el funcionamiento de las otras (Richards, 2015). En particular, para este servicio se definieron cuatro capas (Figura 22) tal como se detalla a continuación:

- Datos: esta capa centraliza el acceso a los datos, particularmente a la base de datos MongoDB. Además, se implementa el patrón *Repository* que permite abstraer el origen de datos subyacente e independizar al dominio de este (Fowler, 2002).
- Dominio: incluye las entidades y modelos del negocio, en particular se modela la ocurrencia de incendios forestales.

- Presentación: en ella se incluyen controladores HTTP que se encargan de procesar las peticiones que realizan los clientes y enviarlas al núcleo de la aplicación para que se realicen las acciones requeridas.
- Servicio: esta capa actúa como una fachada de la lógica de negocio, donde se abstraen y centralizan las reglas del negocio. Si bien en este primer *release* las operaciones que realiza esta capa son básicas (cargar incendios históricos y listarlos), en futuros *releases* se integrará el sistema de gestión documental actualmente disponible en el cuartel de bomberos de Pinamar (Martínez Saucedo et al., 2021) con este servicio. Consecuentemente se prevé que la lógica de negocio se complejice a raíz de las futuras integraciones.

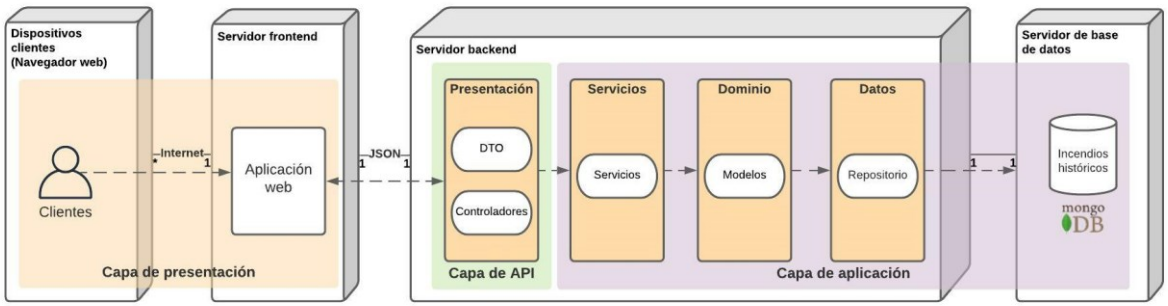

Figura 22: Arquitectura en capas del servicio histórico (vista lógica y de despliegue combinadas).

3.5.3. Persistencia

A lo largo del presente proyecto se han optado por distintas estrategias de persistencia según las necesidades de cada componente. En particular, los componentes de la rama de laboratorio trabajan con archivos CSV para generar los *datasets* utilizados tanto para entrenar los modelos de ML como para obtener las predicciones de incendios forestales en producción. Por otro lado, los modelos de clasificación y regresión son almacenados en archivos *pickle* (.pkl) y *hierarchical data format* (.h5) respectivamente. En el caso de la rama productiva, además de consumir archivos generados por la rama de laboratorio, se utilizaron bases de datos no relacionales para almacenar los datos sobre incendios históricos y las respuestas de la API de AQUA en caché. Esto se debió a las ventajas que presentan este tipo de

bases de datos por sobre las relacionales: pueden escalar horizontal y verticalmente, son distribuidas por diseño, tienen esquemas flexibles y soportan grandes volúmenes de datos.

Los incendios forestales históricos almacenados en una base de datos no relacional y documental fueron inicialmente extraídos de planillas que elaboran internamente los bomberos voluntarios de Pinamar. De todos los datos recopilados en estas planillas se filtraron los atributos que son considerados importantes por bomberos a la hora de contrastar las predicciones de incendios con los incendios forestales ocurridos, a saber:

- Fecha y hora en la que se produjo el incendio forestal.
- Móviles utilizados en combate.
- Cantidad de bomberos presentes en combate.
- Tiempo insumido por los bomberos para combatir el incendio forestal (minutos).
- Área quemada por el incendio (hectáreas).
- Latitud correspondiente a la ubicación donde se produjo el incendio forestal.
- Longitud correspondiente a la ubicación donde se produjo el incendio forestal.

3.6.Despliegue

En lo que respecta al despliegue de la solución, se eligió al proveedor de servicios en la Nube Amazon Web Services (AWS) ya que cuenta con los servicios requeridos por AQUA, además de proveer un nivel de servicio gratuitito para ciertos productos. Tal como se puede observar en la Figura 23, dentro de la Nube de Amazon se definieron a su vez redes privadas a través del servicio *Virtual Private Cloud* (VPC) para poder monitorear el tráfico y restringir el acceso a direcciones no autorizadas.

Con el objetivo de tener una baja latencia se propone trabajar en la región de Sudamérica, particularmente Brasil (sa-east-1). Por otro lado, para contar con una alta disponibilidad se definieron dos zonas de disponibilidad (sa-east-1a y sa-east-1b), de forma tal que el balanceador de carga de aplicación (*Application Load Balancer*) enrute las peticiones a ambas zonas o una de ellas dependiendo de la salud del servicio objetivo. En este caso, los servicios objetivos corresponden a contenedores y servicios de *Elastic Container Service*

(ECS). Asimismo, se contempla el uso del servicio *Auto Scaling*, que permite escalar recursos de forma rápida en caso de que algún servicio lo necesite en un momento dado.

En cuanto a la infraestructura definida para desplegar los distintos componentes que conforman la arquitectura de AQUA se definió que los microservicios de Predicciones e Histórico se encuentren alojados en contenedores Docker dentro de ECS, mientras que la base de datos MongoDB de incendios históricos se despliegue en un servicio web EC2 (*Elastic Compute Cloud*). De esta manera se aprovechan los beneficios que ofrece ECS en relación con la implementación, administración y escalado de recursos. Por otra parte, se planteó desplegar los modelos predictivos en *SageMaker*, ya que ofrece un servicio administrado (*Endpoints*) que permite realizar predicciones a través de una API REST y MLflow cuenta con conectores para automatizar la puesta en producción. Además, *SageMaker* ofrece la posibilidad de escalar las instancias automáticamente, asegurando en consecuencia alta disponibilidad para realizar predicciones.

El *frontend* de AQUA está conformado por la aplicación de visualización de predicciones, desarrollada en React y desplegada a través de *Amplify*. *Amplify* consiste en un conjunto de herramientas y servicios que facilitan el despliegue de aplicaciones web gracias a la integración que cuenta con diversos *frameworks* web, entre los que se encuentra React. Los usuarios finales interactuarán con esta aplicación web que a su vez enviará las peticiones a *API Gateway,* servicio administrado que es la puerta de entrada a los servicios *backend*. Este servicio ofrece una baja latencia y almacenamiento en *caché*, lo que mejora la velocidad de respuesta.

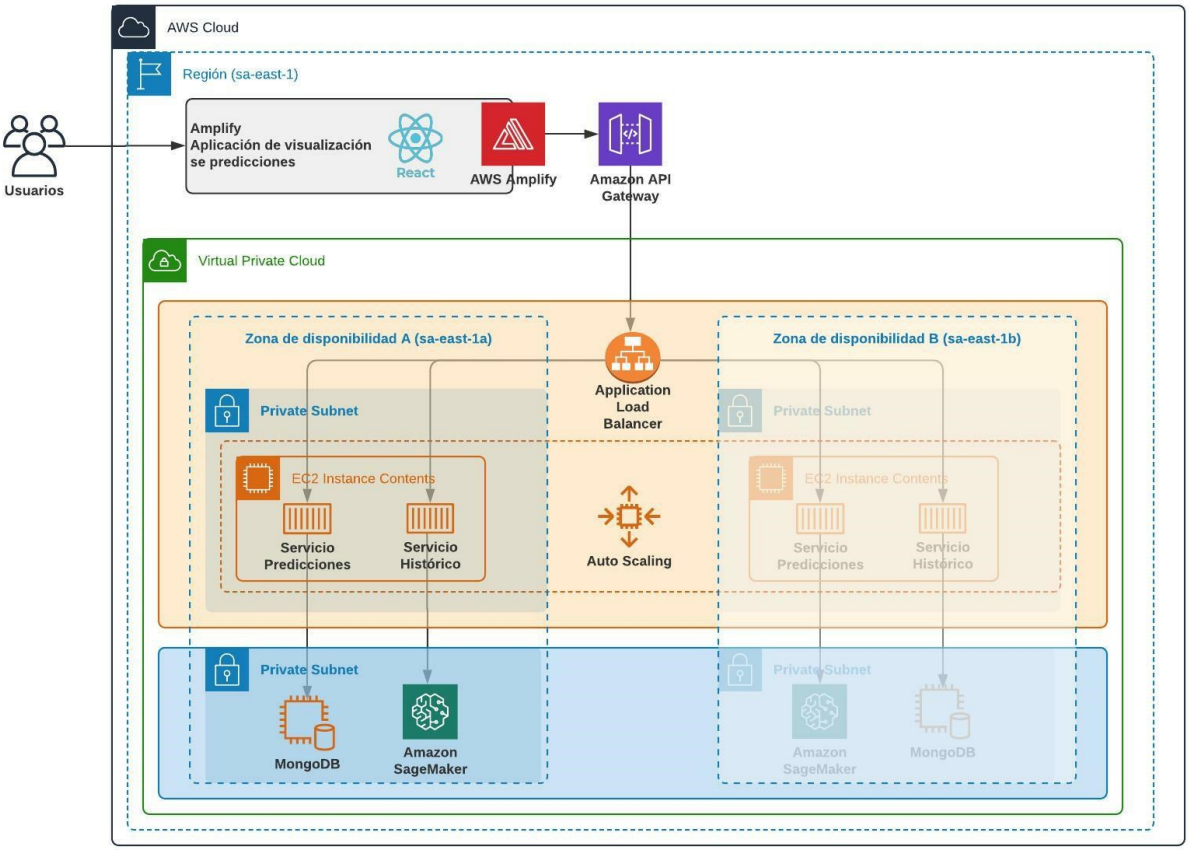

Figura 23: Arquitectura de AQUA en tiempo de despliegue.

4. Metodología de Desarrollo

A lo largo del desarrollo de los componentes que conforman AQUA se adoptaron distintas metodologías de desarrollo según los estándares establecidos en la industria. Si bien el enfoque a lo largo del avance del presente trabajo fue mayoritariamente ágil en miras de desarrollar software de forma iterativa e incremental tal como se enuncia en los valores del *Agile Manifesto* (Beck et al., 2001), durante la fase de entrenamiento de modelos se flexibilizó la metodología adoptada debido a la naturaleza inherentemente experimental del desarrollo de modelos de ML. En efecto, el proceso de ML puede describirse a través de siete pasos:

1. Recolección de datos.
2. Preparación de datos.
3. Elección de algoritmos de ML.
4. Entrenamiento de modelos.
5. Evaluación de modelos.
6. Optimización de hiperparámetros.
7. Evaluación de modelos.

(Guo, 2017).

En los proyectos de ML la entrada de un componente depende en gran medida de la salida del otro. Efectivamente para comenzar a desarrollar y entrenar modelos es necesario contar con los datos que son base para las predicciones. Por esta razón, para abordar esta dependencia se dividieron las etapas de ML en subproyectos, de forma tal de poder trabajar con un enfoque ágil en cada uno de ellos.

Los primeros dos pasos o etapas fueron comprendidas en el desarrollo del *pipeline* de datos, subproyecto en donde la metodología adoptada fue Scrum y cuyos fundamentos pueden consultarse en el Anexo F. La adopción de esta metodología permitió trabajar en *sprints* de una semana de forma tal de reducir la incertidumbre relacionada a la recolección y manipulación de datos geográficos provenientes de distintas fuentes.

Las últimas cinco etapas fueron englobadas en el subproyecto de desarrollo y entrenamiento de modelos predictivos. Si bien también se trabajó con Scrum en este subproyecto, la dificultad de estimar el tiempo para realizar las tareas previstas y la técnica predominante a la hora de entrenar modelos de la prueba y error llevó a incumplir en la mayoría de los *sprints* los plazos previstos para las tareas del *sprint backlog.* No obstante, la índole

flexible de las metodologías ágiles permitió corregir el curso del desarrollo al adquirir nuevos conocimientos y llegar a nuevas conclusiones con los modelos desarrollados.

En lo que concierne a herramientas utilizadas para el desarrollo, en la Tabla XVI se detalla para cada subproyecto o componente de la solución cuáles son y con qué fin han sido seleccionadas, y las Figuras 24 y 25 explicitan algunas de las herramientas utilizadas para el entrenamiento de modelos predictivos.

TABLA XVI: Lenguajes de programación, frameworks, herramientas y librerías utilizadas en cada subproyecto de la solución.

Subproyecto	**Lenguajes de programación, frameworks, herramientas y librerías**
Pipeline de datos	Lenguaje de programación: Python 3.7.3. Contenerización: Docker. Gestor de paquetes: Conda. Análisis y visualización de datos: Jupyter lab. Conversor de archivos HDF4 a HDF5: h4toh5. Manipulación de archivos HDF5: H5py. Manipulación y análisis de datos: • Pandas. • NumPy. Generación de gráficos: Matplotlib. Procesamiento y visualización de datos geoespaciales: • Cartopy. • Seaborn. Manipulación de *datasets* multidimensionales: • Xarray. • Rioxarray (*datasets* geoespaciales). Manipulación de archivos XLSX: Openpyxl. Manipulación de archivos NC: NetCDF4. Geocodificación de direcciones: GeoPy.
Entrenamiento de modelos	Lenguaje de programación: Python 3.7.3 Contenerización: Docker.

	Gestor de paquetes: Conda. Plataforma de ciclo de vida del proyecto: MLflow. Manipulación y análisis de datos: • Pandas. • NumPy. Librerías de machine Learning: • Scikit-Learn • TensorFlow • Keras Administración y visualización de modelos: Wandb.
Visualización de predicciones	**Backend** Lenguaje de programación: Python 3.7.3. Contenerización: Docker. Gestor de paquetes: Conda. Framework web: Flask. Base de datos: MongoDB. Caché: Redis. Serialización de objetos: Marshmallow. Inyección de dependencias / inversión de control: Inject. Manipulación de datos: • Pandas. • NumPy. Librerías de machine Learning: • Scikit-Learn • TensorFlow • Keras • Cloudpickle **Frontend** Lenguaje de programación: Contenerización: Docker. Gestor de paquetes: npm.

	Framework: React. Manejo de estado: Redux Librerías: • Moment • Bootstrap • React MultiDate Picker • React Google Maps

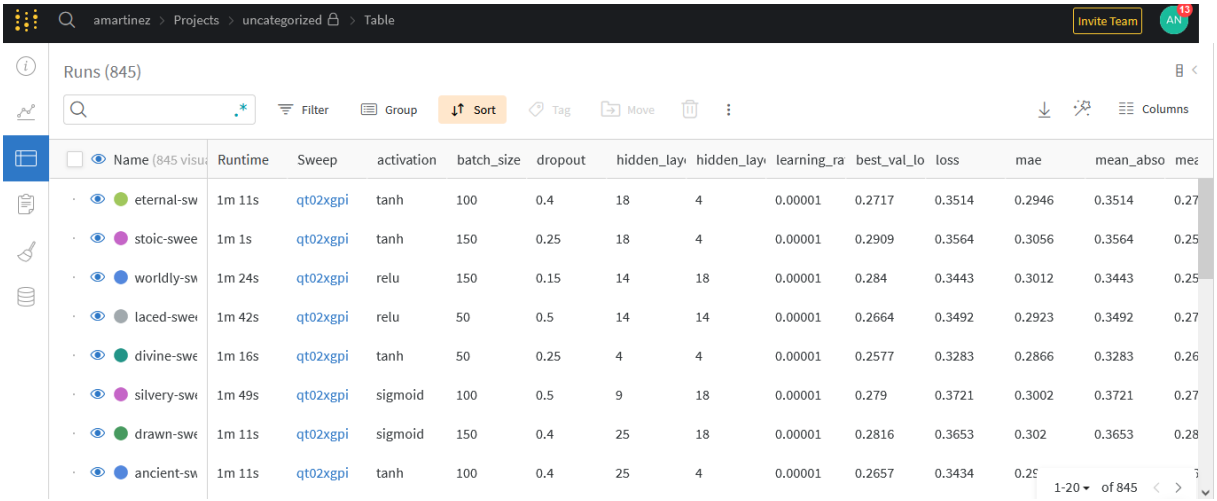

Figura 24: Entrenamiento de redes neuronales con Wandb.

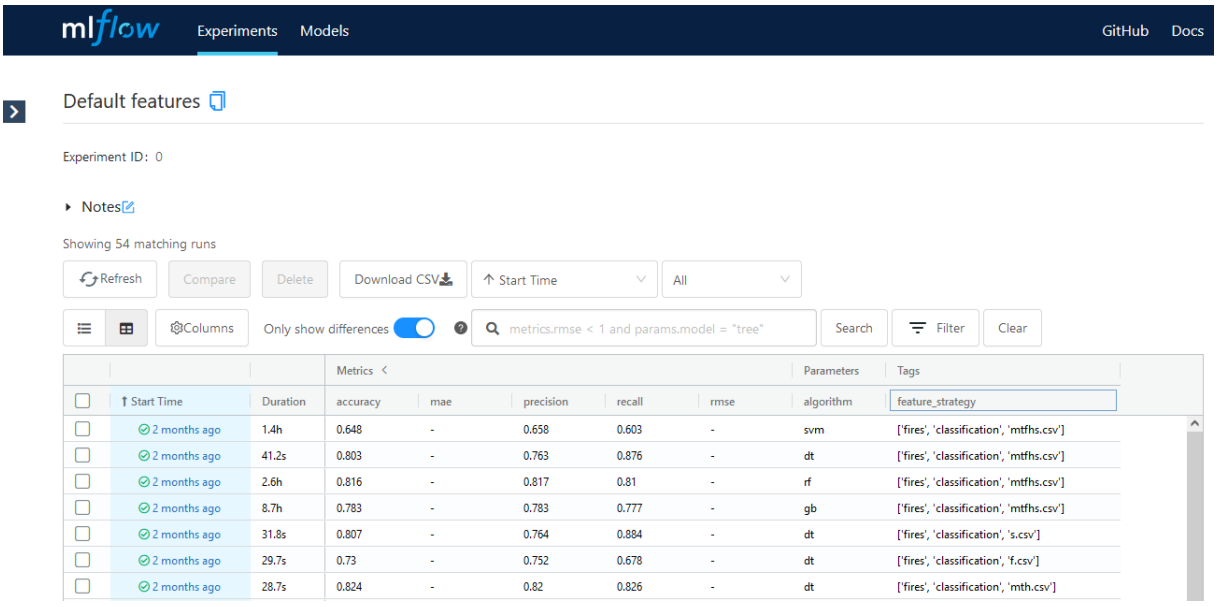

Figura 25: Entrenamiento de modelos de Machine Learning con MLflow.

5. Pruebas realizadas

Con el objetivo de determinar los mejores modelos para predecir tanto la ocurrencia como superficie quemada por los incendios forestales se calcularon diversas métricas considerando el conjunto de datos de pruebas, el cual no fue incluido durante el entrenamiento de los modelos. Esta etapa de evaluación es fundamental para establecer cuáles son los modelos que mejores resultados han brindado y por consiguiente serán desplegados para brindar las predicciones que los clientes consultarán. A su vez se realizó una prueba con usuarios finales sobre la visualización de predicciones de incendios forestales, principal funcionalidad de AQUA. Es importante visualizar las pantallas de prueba de la Aplicación AQUA donde se registre su funcionalidad y alcance?

5.1. Modelos entrenados

Durante el desarrollo y entrenamiento de los modelos de clasificación y regresión se optimizaron los hiperparámetros con el fin de obtener los modelos que mejores resultados provean. No obstante, la forma de evaluarlos depende del tipo de problema que el modelo resuelva: para el caso de los modelos de predicción de incendios forestales (clasificación binaria) se calculó la exactitud, precisión y sensibilidad; mientras que para los modelos de predicción de superficie quemada se calcularon las métricas de MAE y RMSE. En el caso particular del modelo de regresión por cuantiles se estableció como medida de evaluación la cobertura, que determina el porcentaje de valores que efectivamente se encuentra incluido en el cuantil establecido (Frank et al., 2016).

En cuanto a la predicción de incendios forestales, el modelo con mayor exactitud es el de Random Forest con un 82.4%. Coincidentemente este modelo fue también el que obtuvo el valor de precisión más alto (82%). No obstante, el modelo con mayor sensibilidad es el de Árbol de Decisión con un 88.4%, convirtiéndolo en aquel que mejor distingue la ocurrencia de incendios tal como se puede observar en la Tabla XVII.

TABLA XVII: Resultados de la evaluación de modelos de predicción de incendios forestales.

Modelo	**Mejores hiperparámetros**	**Resultados**		
		Exactitud	**Precisión**	**Sensibilidad**
Red neuronal artificial	Cantidad de nodos por capa: (50, 100, 50).	74.2%	71.6%	79.3%

	Función de activación: Tangente Hiperbólica. Optimizador de pesos: Adam. Alpha: 0.0001. Tasa de aprendizaje: Constante.			
SVM	C: 5555.6. Gamma: 1. Kernel: Linear.	61.5%	62.4%	56.2%
Gradient Boosting	Función de pérdida: Deviance. Tasa de aprendizaje: 0.15. Número mínimo de ejemplos para dividir un nodo interno: 0.28. Número mínimo de ejemplos para ser un nodo hoja: 0.1. Profundidad máxima de cada clasificador: 8. Cantidad máxima de características a considerar: $\sqrt{características}$. Función de calidad de división: MSE. Fracción de los ejemplos para entrenar clasificadores individuales: 0.62. Cantidad de etapas de *boosting*: 100.	77.9%	77.7%	77.7%
Regresión Logística	C: 0.878	62.3%	63.6%	56.2%
Árboles de decisión	Profundidad máxima del árbol: 3. Cantidad máxima de características a considerar: 59.	80.7%	76.4%	**88.4%**

	Número mínimo de ejemplos para ser un nodo hoja: 5. Criterio de Ganancia de Información: Entropía.			
Random Forest	Bootstrap: Sí. Profundidad máxima de cada clasificador: 80. Cantidad máxima de características a considerar: 36. Número mínimo de ejemplos para dividir un nodo interno: 12. Número mínimo de ejemplos para ser un nodo hoja: 4. Número de árboles en el bosque: 200.	**82.4%**	**82%**	82.6%

En lo que respecta a modelos de predicción de superficie quemada, los mejores modelos en términos de valores de MAE y RMSE son redes neuronales artificiales. Particularmente, el MAE más bajo corresponde a un valor de 0.255 mientras que el mejor RMSE es de 0.178. No obstante, la arquitectura e hiperparámetros de ambas redes difieren entre sí tal como se puede apreciar en la Tabla XVIII.

TABLA XVIII: Resultados de la evaluación de modelos de predicción de superficie quemada.

Modelo	**Mejores hiperparámetros**	**Resultados**	
		MAE	**RMSE**
Red neuronal artificial (Modelo 1)	Tasa de aprendizaje: 0.0001. Dilución: 0.4. Cantidad de nodos en capa oculta n°1: 18. Cantidad de nodos en capa oculta n°2: 4.	0.295	**0.178**

	Función de activación: Tangente Hiperbólica.		
Red neuronal artificial (Modelo 2)	Tasa de aprendizaje: 0.0001. Dilución: 0.5 Cantidad de nodos en capa oculta n°1: 9. Cantidad de nodos en capa oculta n°2: 18. Función de activación: ReLu.	**0.255**	0.215
SVM	C: 4444.5 Gamma: 0.001 Kernel: RBF	0.274	0.438
Árboles de decisión	Profundidad máxima del árbol: 3. Cantidad máxima de características a considerar: 33. Número mínimo de ejemplos para ser un nodo hoja: 5. Criterio de Ganancia de Información: Friedman.	0.294	0.46
Random Forest	Bootstrap: Sí. Profundidad máxima de cada regresor: 80. Cantidad máxima de características a considerar: 24. Número mínimo de ejemplos para dividir un nodo interno: 12. Número mínimo de ejemplos para ser un nodo hoja: 3. Número de árboles en el bosque: 100.	0.299	0.448

En lo que concierne al enfoque de predicción de clases de incendios forestales, los resultados obtenidos no fueron buenos: tal como se puede observar en la Tabla XIX en ningún modelo se logró superar una exactitud del 40%, y este rendimiento pobre puede ser atribuido a la cantidad desbalanceada de registros por clase. En efecto, ningún modelo tuvo la capacidad de predecir incendios pertenecientes a la Clase 2 o a la Clase 3. Coincidentemente, el rendimiento es mejor para los registros pertenecientes a la Clase 0 o 1, que corresponde a los incendios de menor magnitud y más frecuentes en el *dataset* de incendios. Si bien se consideró disminuir la cantidad de clases para que cada una contenga una cantidad balanceada de registros, no se encontró un conjunto de clases que abarque una cantidad uniforme de registros y sea representativa para poder tomar decisiones al respecto. En otras palabras, de haber considerado una división en dos clases (incendios de menos de 1 hectárea e incendios de más de 1 hectárea) las predicciones de los modelos no proveerían suficiente información para que los bomberos tomen decisiones concluyentes acerca de la gestión de sus recursos.

TABLA XIX: Resultados de la evaluación de modelos de predicción de superficie quemada por clases.

Modelo	**Mejores hiperparámetros**	**Resultados**
		Exactitud
SVM	C: 0.1. Gamma: 1. Kernel: RBF.	31.4%
Gradient Boosting	Función de pérdida: Deviance. Tasa de aprendizaje: 0.2. Número mínimo de ejemplos para dividir un nodo interno: 0.32. Número mínimo de ejemplos para ser un nodo hoja: 0.39. Profundidad máxima de cada clasificador: 3. Cantidad máxima de características a considerar: $\sqrt{características}$. Función de calidad de división: Friedman.	29.8%

	Fracción de los ejemplos para entrenar clasificadores individuales: 0.95. Cantidad de etapas de *boosting*: 1000.	
Regresión Logística	C: 0.061	34.7%
Árboles de decisión	Profundidad máxima del árbol: 6. Cantidad máxima de características a considerar: 10. Número mínimo de ejemplos para ser un nodo hoja: 10. Criterio de Ganancia de Información: Gini.	33.8%
Random Forest	Bootstrap: Sí. Profundidad máxima de cada clasificador: 80. Cantidad máxima de características a considerar: 6. Número mínimo de ejemplos para dividir un nodo interno: 10. Número mínimo de ejemplos para ser un nodo hoja: 4. Número de árboles en el bosque: 100.	**36.4%**

Con respecto al modelo de regresión por cuantiles, la cobertura de este fue de un 90.9%, siendo el mismo un valor cercano al intervalo de confianza esperado con los cuantiles (95% - 5% = 90%), por lo que se puede concluir que el resultado de este modelo ha sido satisfactorio.

En conclusión, el mejor resultado para predecir la ocurrencia de incendios forestales en términos de sensibilidad se obtuvo a través del modelo de Árbol de Decisión. En cuanto a la predicción de área a quemar se obtuvieron dos modelos de redes neuronales artificiales cuyos rendimientos en términos de MAE y RMSE difieren. No obstante, RMSE es una métrica que pondera los errores de predicción altos, lo que permite dar cuenta de cuán grandes son los errores que el modelo comete. Debido a que en el presente trabajo se busca que las predicciones de superficie quemada no sobreestimen ni subestimen el valor final con el cual los bomberos y entidades toman decisiones, el mejor modelo corresponde a aquel con menor RMSE (Modelo 1).

5.2. Visualización de predicciones

Con el fin de validar que la aplicación web de visualización de predicciones de incendios forestales cumpla con el atributo de usabilidad previamente definido y la funcionalidad sea la esperada por los usuarios finales, se llevó a cabo una prueba en un entorno de desarrollo con bomberos del cuartel de Pinamar según el caso de prueba definido en la Tabla XX.

TABLA XX: Caso de prueba para la funcionalidad "*Visualización de predicciones de incendios forestales*".

Identificador del caso de prueba	CP01
Descripción	Visualización de mapa de predicciones de incendios forestales e incendios históricos.
Pasos	Ingresar a la aplicación de AQUA desde un navegador web.
Datos	date=01-01-2021 hour=21-05 latitude_1=-37.2 longitude_1=-56.95 latitude_2=-37.05 longitude_2=-56.8
Resultados esperados	Se muestra en pantalla el mapa centrado en el par de coordenadas establecido, la fecha ingresada por el usuario y marcadores en el mapa indicando las ubicaciones de incendios forestales.
Resultado de la prueba	Satisfactorio.

Si bien el resultado de la prueba de la funcionalidad fue exitoso, el usuario final propuso como mejora la posibilidad de ocultar los marcadores correspondientes a los incendios históricos para facilitar la interpretación de los resultados. En consecuencia se desarrolló un componente que permita activar o desactivar las capas de información disponibles: predicciones e incendios históricos.

Cuantos usuarios realizaron las pruebas?. Donde se muestran las pantallas de los resultados que generan esta conclusión?

6. Análisis económico

Con el objetivo de determinar la viabilidad económico-financiera de AQUA se llevaron a cabo diversos análisis considerando el público objetivo de la solución, los gastos asociados y los ingresos proyectados. En las subsiguientes secciones de detallan estos análisis y los escenarios evaluados para establecer la rentabilidad del proyecto.

6.1. Modelo de negocio

Si bien la solución desarrollada en el marco del presente proyecto está, al momento de redacción de este documento, apuntada a una región geográfica y mercado acotados (Partido de Pinamar), la misma puede adaptarse fácilmente a otras regiones del país con una mínima configuración de parámetros para la extracción de datos crudos. Por esta razón, cualquier región en que la vegetación (bosques, pastizales, reservas naturales) tenga un valor económico, AQUA permite proteger los recursos vegetales gracias a las predicciones rápidas y de valor para el área geográfica de interés. Estas predicciones son de suma importancia para personas u organismos privados y estatales que necesitan de información para la toma de decisiones.

Dentro del público objetivo de AQUA se pueden distinguir distintos segmentos de clientes según el alcance de las decisiones que puedan tomar contando con las predicciones de incendios forestales para una zona dada:

- Cuarteles de bomberos: las predicciones les permiten alocar eficientemente recursos humanos y materiales con antelación para poder controlar de manera más rápida incendios de mayor magnitud.
- Entidades gubernamentales: el Estado cuenta con mecanismos de prevención para la protección de bosques y reservas naturales, por lo que a través de AQUA puede gestionar patrullas y medidas de prevención y concientización en conjunto con bomberos locales.
- Sector agrícola-ganadero: las quemas controladas que se llevan a cabo en campos y pastizales permiten eliminar residuos vegetales remanentes, y siempre se realizan bajo ciertas condiciones climáticas para evitar perder el control del incendio.
- Turistas y habitantes de la zona: las fogatas, campings y demás actividades recreativas y deportivas en zonas boscosas pueden evitarse

cuando AQUA alerta sobre condiciones propicias para incendios forestales de gran magnitud.

Considerando el público objetivo, se propone un modelo de negocio de suscripción mensual para cuarteles de bomberos, entidades gubernamentales y el sector agrícola-ganadero. Este modelo es flexible en términos de ingresos periódicos, permitiendo cubrir los costos de mantenimiento de la infraestructura de AQUA con tarifas accesibles para cuarteles de bomberos. El presupuesto estimado del proyecto para el primer año se detalla en la Tabla XI, mientras que el detalle de los servicios y proveedores contratados para la infraestructura se especifica en la Tabla XII.

TABLA XI: Recursos del proyecto AQUA.

Insumo	**Descripción**	**Cantidad**	**Costo**
Recursos Humanos	Desarrollador Frontend.	25 horas	USD 500
	Desarrollador Backend.	40 horas	USD 600
	Project Manager.	40 horas	USD 600
	Ingeniero de Machine Learning.	195 horas	USD 2925
Computadora	Lenovo IdeaPad S340.	1 unidad	USD 600
Infraestructura	AWS.	N /A	USD 257
	Impuesto país (30%).	N /A	USD 77
Presupuesto total: USD 5559			

TABLA XII: Costo mensual de mantenimiento de la infraestructura de AQUA, región América del Sur (AWS Pricing calculator).

Servicio	**Descripción**	**Costo mensual**
Amazon API Gateway	Unidades de solicitudes de la API HTTP (miles), Tamaño promedio de cada solicitud (256 KB), Unidades de solicitud de la API REST (millones), Tamaño de memoria caché (GB) (Ninguno), Unidades de mensaje WebSocket (miles), Tamaño promedio del mensaje (32 KB), Solicitudes (10000 por mes)	USD 15,9

Amazon EC2	Sistema operativo (Linux), Cantidad (2), Estrategia de precios (Instancias bajo demanda), Cantidad de almacenamiento (15 GB), Tipo de instancia (t4g.large)	USD 44,94
Application Load Balancer	Número de balanceadores de carga de aplicaciones (1)	USD 24,93
Presupuesto mensual de mantenimiento (más impuestos): USD 111		

Algunos servicios que utiliza AQUA no se incluyeron en el presupuesto ya que el tráfico previsto para los primeros 5 años está incluido en las capas gratuitas que ofrecen AWS y Google. Particularmente es el caso de los productos AWS Amplify, Amazon SageMaker Endpoints, Google Maps Platform (Maps y Geocoding) y Google Elevation API.

Considerando los costos mensuales de mantenimiento se fijó el precio de suscripción para cuarteles de bomberos y el sector agrícola-ganadero en USD 10 mensuales, siendo este un costo accesible para tanto organizaciones sin fines de lucro como propietarios de campos locales y tomando en cuenta la frecuencia de uso mensual que se estima que realicen de AQUA.

6.2. Análisis financiero

Con el fin de evaluar financieramente AQUA al momento presente se utilizaron diversos métodos de valoración de inversiones: el VAN (Valor Actual Neto), la TIR (Tasa Interna de Rentabilidad) y el *pay back* o tiempo de retorno de inversión. Para ello, se definieron los siguientes valores para las variables que a continuación se detallan:

- I_o (inversión inicial): USD 5559.
- n (períodos de tiempo): 5 años.
- i (costo de recursos o rendimiento mínimo aceptable): 1,75 % (TEA para constituir un plazo fijo en dólares en el Banco Nación).

Además de los valores detallados, para evaluar financieramente AQUA es necesario definir el flujo de fondos neto (FFN) que se estima para cada periodo de tiempo. En consecuencia, se definieron tres escenarios financieros para calcular el VAN, la TIR y el *pay back*: optimista, neutro y pesimista. El detalle de la cantidad de cuarteles de bomberos y campos a los cuales se estiman brindar servicios en el horizonte temporal y escenarios definidos junto con los cálculos de cada indicador financiero se describe en el Anexo I.

6.2.1. VAN

El Valor Actual Neto o VAN (11) permite obtener el rendimiento actualizado de los flujos originados por una inversión, es decir, la rentabilidad que se obtendría al invertir en el proyecto. Cuando el VAN toma un valor positivo entonces el proyecto es rentable. En contrapartida, un VAN negativo significa que la inversión es mayor al flujo de fondos neto y el proyecto no es rentable.

$$VAN = -I_0 + \sum_1^n \frac{FFN}{(1+i)^n} \quad (11)$$

El VAN obtenido para distintos escenarios en el marco del proyecto AQUA se detalla en la Tabla XIII.

TABLA XIII: VAN del proyecto.

Escenario	VAN
Optimista	$1566
Neutro	-$1351
Pesimista	-$4271

6.2.2. TIR

La tasa interna de retorno o TIR es un indicador estrechamente relacionado con el VAN ya que se calcula igualando el VAN a cero para despejar i, obteniendo así un valor porcentual que permite comparar proyectos para determinar con cuál se lograría un mayor aumento patrimonial de la inversión. En el caso de AQUA la TIR obtenida se describe en la Tabla XIV.

TABLA XIV: TIR del proyecto.

Escenario	TIR
Optimista	8%
Neutro	-4%
Pesimista	-21%

6.2.3. Pay back

El *pay back* es el tiempo que es necesario que transcurra para que los flujos de fondos netos sean iguales al monto de la inversión inicial. En otras palabras, el *pay back* indica cuánto tiempo demora recuperar el capital invertido. Este valor en conjunto con el VAN y la TIR brinda más información acerca del riesgo asociado a invertir en un proyecto. Tomando en consideración los distintos escenarios propuestos, el *pay back* de AQUA se detalla en la Tabla XV.

TABLA XV: *Pay back* del proyecto en años.

Escenario	***Pay back* (años)**
Optimista	4
Neutro	6
Pesimista	9

6.3.Conclusiones

De los escenarios planteados, AQUA es rentable sólo en el optimista en un horizonte de 5 años. Esto se debe principalmente a que las estimaciones fueron realizadas con la premisa de no contar con un plan de marketing enfocado en captar más clientes en distintos puntos del país de forma rápida en lugar de concentrarse únicamente en la Provincia de Buenos Aires. Por esta razón el precio mensual de suscripción se estableció en un valor accesible para tanto el sector público como privado, aunque incrementando el mismo a USD 20 el proyecto se volvería rentable en todos los escenarios.

7. Discusión

Dentro de un proyecto de ML, las etapas de recolección y preparación de datos son fundamentales. En efecto, los datos son la materia prima de cualquier modelo de ML y determinan en gran medida el rendimiento que los modelos tendrán. A partir de esta afirmación radica la importancia de que los datos utilizados para entrenar modelos sean de calidad y describan la realidad de la forma más fiel posible.

En este sentido, las primeras dificultades surgidas durante el desarrollo de AQUA estuvieron relacionadas al proceso manual de recolección de los datos de incendios de Pinamar y la incompletitud de datos en los registros de incendios de los meses correspondientes a octubre, noviembre y diciembre de 2015. Por esta razón, de los 750 registros de incendios que los bomberos locales han provisto para los años 2015 – 2019 se han podido utilizar 597 para la fase de entrenamiento de modelos, posteriormente a la limpieza y procesamiento de los datos crudos. Esto pone en relieve la importancia de que los bomberos cuenten con un sistema de gestión documental, como el presentado en (Martínez Saucedo et al., 2021), que les permita trabajar con registros digitales fácilmente exportables en lugar de utilizar planillas en papel. De esta manera la calidad de todos los datos se vería asegurada, y el proceso de extracción y transformación de datos de incendios se tornaría menos complejo.

No obstante el tamaño del *dataset* de incendios obtenido, el rendimiento de los modelos de clasificación y regresión ha sido muy bueno en comparación a los resultados que se han obtenido a través de otros modelos de predicción de incendios forestales desarrollados a nivel nacional e internacional. En este aspecto resulta fundamental destacar el rol del preprocesamiento de datos, el proceso de análisis exploratorio de los datos de incendios y la etapa de *feature engineering* para obtener un *dataset* que sea de la mayor calidad posible. Asimismo, el *user research* y la investigación llevada a cabo para comprender el estado del arte permitieron analizar los distintos enfoques utilizados en los modelos de predicción de incendios forestales y en consecuencia adoptar el más apto para el contexto de AQUA.

En miras de una integración con sistemas de gestión documental en cuarteles de bomberos para un futuro *release* de AQUA, es importante considerar el aspecto de seguridad de la solución. Si bien la información de incendios forestales es de público acceso y no incluye datos sensibles, la autorización y autenticación de usuarios será una tarea prioritaria en la

próxima iteración ya que los sistemas de gestión documental con los cuales se integrará la solución administran información sensible que debe ser resguardada de accesos no autorizados.

Por último, este tipo de modelos predictivos de incendios forestales pueden aplicarse de forma alternativa en el sector agrícola-ganadero. Las quemas controladas que se llevan a cabo con fines productivos y en consideración con los procesos naturales deben realizarse bajo ciertas condiciones climáticas para evitar perder el control de estas quemas. Por esta razón se propone como futura línea de investigación a desarrollar dentro del proyecto *"A21T03 – Aplicaciones de Machine Learning para mejorar el uso de Recursos Naturales"* la adaptación y aplicación de AQUA en este ámbito.

Además del proceso de recolección de datos, encontró alguna limitación, dificultad o diferencia en los resultados, al momento de realizar las pruebas respecto al alcance planteado inicialmente para el primer reléase de la App?

8. Conclusiones

En las últimas décadas la severidad y magnitud de los incendios forestales ha alcanzado niveles preocupantes en todo el mundo. Conforme han transcurrido los años y la tecnología ha avanzado, más investigadores se han abocado en el desarrollo de modelos predictivos que contribuyan a la prevención de incendios forestales para paliar las pérdidas económicas, ecológicas y sociales que los mismos generan. No obstante, a nivel nacional las investigaciones realizadas en la materia son escasas pese a que Argentina ha alcanzado cifras récord de focos de incendio en el último año.

En consecuencia, el presente trabajo ha buscado contribuir a la prevención de incendios forestales a través del desarrollo de un modelo de predicción de incendios forestales enfocado en el Partido de Pinamar. Los objetivos definidos para el alcance se han cumplido, obteniendo como productos finales los modelos predictivos propiamente dichos (contando con una sensibilidad del 88.4% y un RMSE del 0.178), y un software para visualizar las predicciones de incendio. En este sentido, las entrevistas tempranas con especialistas resultaron de vital importancia para validar la solución. A su vez, del desarrollo de AQUA también se ha desprendido el *pipeline* de datos para recolectar datos meteorológicos y de incendios, siendo el mismo configurable a distintas ubicaciones geográficas y permitiendo generar el *dataset* específico a la zona de interés posteriormente utilizado para entrenar los modelos predictivos.

Aunque AQUA debe seguir perfeccionándose, mediante el proyecto se han sentado las bases necesarias para continuar con el desarrollo e investigación de modelos que predigan incendios forestales en el país utilizando datos específicos a cada región. Esto último marca un hito importante, ya que al momento de redacción del presente documento los escasos modelos predictivos desarrollados en Argentina se han valido de datos de incendios de otros países para poder entrenar los modelos que se han desarrollado.

Es importante remarcar que en Argentina los bomberos son voluntarios y trabajan ad honorem. Este aspecto estuvo presente a lo largo del desarrollo de AQUA de forma tal que el desarrollo requiera la menor inversión posible, priorizando el uso de librerías *open source* y la extracción de datos provenientes de fuentes abiertas. Gracias a estas decisiones se logró obtener un producto que contribuya al trabajo diario de los bomberos en todo el país a un bajo costo, de modo tal que puedan enfocar sus esfuerzos a la tarea que mejor saben hacer que es proteger a las personas.

Las conclusiones deben estar alineadas con los objetivos específicos y con las conclusiones de cada sección desarrollada (estado del arte, user research, pruebas realizadas y análisis económico).

Indicar que normas utilizo para detallar la bibliografía

9. Bibliografía

ARGENTINA.GOB.AR, 2018a. ¿Cómo se mantiene un fuego? *Argentina.gob.ar* [en línea]. [Consulta: 24 mayo 2021]. Disponible en: https://www.argentina.gob.ar/ambiente/fuego/conocemas/combustion.

ARGENTINA.GOB.AR, 2018b. ¿Cómo se originan los incendios? *Argentina.gob.ar* [en línea]. [Consulta: 26 abril 2021]. Disponible en: https://www.argentina.gob.ar/ambiente/fuego/conocemas/origen.

ARGENTINA.GOB.AR, 2018c. ¿Cuáles son las variables y qué factores las afectan? *Argentina.gob.ar* [en línea]. [Consulta: 24 mayo 2021]. Disponible en: https://www.argentina.gob.ar/ambiente/fuego/conocemas/variables.

ARGENTINA.GOB.AR, 2018d. Mapa de peligro de incendio. [en línea]. [Consulta: 28 marzo 2021]. Disponible en: https://www.argentina.gob.ar/ambiente/fuego/alertatemprana/indices.

ARGENTINA.GOB.AR, 2021a. El Gobierno nacional lanzó una campaña de prevención de incendios forestales. *Argentina.gob.ar* [en línea]. [Consulta: 8 mayo 2021]. Disponible en: https://www.argentina.gob.ar/noticias/el-gobierno-nacional-lanzo-una-campana-de-prevencion-de-incendios-forestales.

ARGENTINA.GOB.AR, 2021b. Obtener financiamiento para MiPyMEs afectadas por los incendios forestales del área cordillerana en Chubut. *Argentina.gob.ar* [en línea]. [Consulta: 8 mayo 2021]. Disponible en: https://www.argentina.gob.ar/obtener-financiamiento-para-mipymes-afectadas-por-los-incendios-forestales-del-area-cordillerana-en.

ARMANDO GONZÁLEZ-CABÁN, 2013. 17 The Economic Dimension of Wildland Fires. , pp. 9.

BASS, L., CLEMENTS, P. y KAZMAN, R., 2012. *Software Architecture in Practice*. S.l.: Addison Wesley. ISBN 978-0-321-81573-6.

BECK, K., BEEDLE, M., BENNEKUM, A. van, COCKBURN, A., CUNNINGHAM, W., FOWLER, M., GRENNING, J., HIGHSMITH, J., HUNT, A., JEFFRIES, R., KERN, J., MARICK, B., Robert C., MELLOR, S., SCHWABER, K., SUTHERLAND, J. y THOMAS, D., 2001. Principles behind the Agile Manifesto. [en línea]. [Consulta: 6 julio 2021]. Disponible en: https://agilemanifesto.org/principles.html.

BEM, P.P. de, JÚNIOR, O.A. de C., MATRICARDI, E.A.T., GUIMARÃES, R.F., GOMES, R.A.T., BEM, P.P. de, JÚNIOR, O.A. de C., MATRICARDI, E.A.T., GUIMARÃES, R.F. y GOMES, R.A.T., 2018. Predicting wildfire vulnerability using logistic regression and artificial neural networks: a case study in Brazil's Federal District. *International Journal of Wildland Fire*, vol. 28, no. 1, pp. 35-45. ISSN 1448-5516, 1448-5516. DOI 10.1071/WF18018.

BOLETÍN OFICIAL DE LA REPÚBLICA ARGENTINA, 2020. *AISLAMIENTO SOCIAL PREVENTIVO Y OBLIGATORIO* [en línea]. 19 marzo 2020. S.l.: s.n. 297/2020. Disponible en: https://www.boletinoficial.gob.ar/detalleAviso/primera/227042/20200320.

BOULANDIER, J.J., ESPARZA, F., GARAYOA, J., ORTA, C. y ANITUA, P., 2001. *Comportamiento del fuego forestal*. 11 abril 2001. S.l.: s.n.

BROWNLEE, J., 2016a. Classification and Regression Trees. En: Google-Books-ID: PCJnAQAACAAJ, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. S.l.: Machine Learning Mastery, pp. 72-74.

BROWNLEE, J., 2016b. Logistic Regression. En: Google-Books-ID: PCJnAQAACAAJ, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. S.l.: Machine Learning Mastery, pp. 51-54.

BROWNLEE, J., 2016c. Simple Linear Regression. En: Google-Books-ID: PCJnAQAACAAJ, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. S.l.: Machine Learning Mastery, pp. 43-45.

BROWNLEE, J., 2016d. Supervised, Unsupervised and Semi-Supervised Learning. En: Google-Books-ID: PCJnAQAACAAJ, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. S.l.: Machine Learning Mastery, pp. 16-18.

CARDENAS, M., CASTILLO, J., MEDEL, R., CASCO, O., NAVARRO, M., GUTIERREZ, S. y CURTI, A., 2016. Sistema de predicción de incendios forestales para la provincia de Córdoba. .

CENTRO REGIONAL DE CLIMAS DO SUL DA AMÉRICA DO SUL, 2021. Índices de vegetación. [en línea]. [Consulta: 9 mayo 2021]. Disponible en: https://www.crc-sas.org/es/aplicaciones.php.

COCKBURN, A., 2005. Hexagonal architecture. *Alistair Cockburn* [en línea]. [Consulta: 28 septiembre 2021]. Disponible en: https://alistair.cockburn.us/hexagonal-architecture/.

CORTEZ, P. y MORAIS, A. de J.R., 2007. A data mining approach to predict forest fires using meteorological data. *Associação Portuguesa para a Inteligência Artificial (APPIA)*, pp. 512-523. ISSN 13 978-989-95618-0-9.

DIDAN, KAMEL, 2015. *MYD13Q1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V006* [en línea]. 2015. S.l.: NASA EOSDIS Land Processes DAAC. [Consulta: 23 julio 2021]. Disponible en: https://lpdaac.usgs.gov/products/myd13q1v006/.

DIRECCIÓN GENERAL DE PROTECCIÓN CIVIL Y EMERGENCIAS - MINISTERIO DEL INTERIOR - ESPAÑA, 2021. Incendios. [en línea]. [Consulta: 24 mayo 2021]. Disponible en:

https://www.proteccioncivil.es/catalogo/carpeta02/carpeta24/vademecum17/vdm010.htm.

ESPOSITO, D. y ESPOSITO, F., 2020. Metrics of Machine Learning. En: Google-Books-ID: VjDNDwAAQBAJ, *Introducing Machine Learning*. S.l.: Microsoft Press, pp. 285-288. ISBN 978-0-13-558838-3.

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. y WANG, X., 2015. Development of a Global Fire Weather Database. *Natural Hazards and Earth System Sciences*, vol. 15, no. 6, pp. 1407-1423. ISSN 1561-8633. DOI 10.5194/nhess-15-1407-2015.

FISHER, R.A., 1988. UCI Machine Learning Repository: Iris Data Set. [en línea]. [Consulta: 24 mayo 2021]. Disponible en: https://archive.ics.uci.edu/ml/datasets/iris.

FLANNIGAN, M.D., STOCKS, B.J. y WOTTON, B.M., 2000. Climate change and forest fires. *Science of The Total Environment*, vol. 262, no. 3, pp. 221-229. ISSN 0048-9697. DOI 10.1016/S0048-9697(00)00524-6.

FOWLER, M., 2002. Object-Relational Metadata Mapping Patterns. *Patterns of Enterprise Application Architecture*. USA: Addison-Wesley Longman Publishing Co., Inc., pp. 322-326. ISBN 978-0-321-12742-6.

FRANK, H., E, H. y RALF, M., 2016. Method. En: Google-Books-ID: sUiGDQAAQBAJ, *Proceedings. 26. Workshop Computational Intelligence, Dortmund, 24. - 25. November 2016*. S.l.: KIT Scientific Publishing, pp. 15-17. ISBN 978-3-7315-0588-4.

GARREAU, M. y FAUROT, W., 2018. 1. Introducing Redux. *Redux in Action*. Shelter Island, New York: s.n., ISBN 978-1-61729-497-6.

GUO, Y., 2017. The 7 Steps of Machine Learning. [en línea]. [Consulta: 6 julio 2021]. Disponible en: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

HAYKIN, S., 2008. Introduction. *Neural Networks and Learning Machines*. New York: s.n., pp. 21-24. ISBN 978-0-13-147139-9.

HECHT-NIELSEN, 1989. Theory of the backpropagation neural network. *International 1989 Joint Conference on Neural Networks*. S.l.: s.n., pp. 593-605 vol.1. DOI 10.1109/IJCNN.1989.118638.

HOMBERGS, T., 2019. *Get Your Hands Dirty on Clean Architecture: A hands-on guide to creating clean web applications with code examples in Java*. S.l.: s.n. ISBN 978-1-83921-196-6.

HOSMER, D.W. y LEMESHOW, S., 2004. Introduction to the Logistic Regression Model. En: Google-Books-ID: Po0RLQ7USIMC, *Applied Logistic Regression*. S.l.: John Wiley & Sons, pp. 1. ISBN 978-0-471-65402-5.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS, 2021. Monitoramento dos Focos Ativos por País - Programa Queimadas - INPE. [en línea]. [Consulta: 8 mayo 2021]. Disponible en: https://queimadas.dgi.inpe.br/queimadas/portal-static/estatisticas_paises/.

JAIN, P., COOGAN, S.C.P., SUBRAMANIAN, S.G., CROWLEY, M., TAYLOR, S. y FLANNIGAN, M.D., 2020. A review of machine learning applications in wildfire science and management. En: arXiv: 2003.00646, *Environmental Reviews*, vol. 28, no. 4, pp. 478-505. ISSN 1181-8700, 1208-6053. DOI 10.1139/er-2020-0019.

KULKARNI, A., CHONG, D. y BATARSEH, F.A., 2020. 5 - Foundations of data imbalance and solutions for a data democracy. En: F.A. BATARSEH y R. YANG (eds.), *Data Democracy* [en línea]. S.l.: Academic Press, pp. 83-106. [Consulta: 24 mayo 2021]. ISBN 978-0-12-818366-3. Disponible en: https://www.sciencedirect.com/science/article/pii/B9780128183663000058.

LIU, Z. y XU, H., 2014. Kernel Parameter Selection for Support Vector Machine Classification. *Journal of Algorithms & Computational Technology*, vol. 8, pp. 163-178. DOI 10.1260/1748-3018.8.2.163.

LUGER, G., 2008. Machine Learning: Connectionist. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Boston: s.n., pp. 482-484. ISBN 978-0-321-54589-3.

MARTÍNEZ ORTEGA, R.M., TUYA PENDÁS, L.C., MARTÍNEZ ORTEGA, M., PÉREZ ABREU, A. y CÁNOVAS, A.M., 2009. EL COEFICIENTE DE CORRELACION DE LOS RANGOS DE SPEARMAN CARACTERIZACION. *Revista Habanera de* Ciencias Médicas, vol. 8, no. 2, pp. 0-0. ISSN 1729-519X.

MARTÍNEZ SAUCEDO, A.C., RÍOS, B., CONNELL, F. y PERROTTA, J., 2021. AQUA: Sistema de administración y gestión documental de emergencias para cuarteles de bomberos. [en línea]. 15° Simposio de Informática en el Estado (SIE) - 50 JAIIO. [Consulta: 31 octubre 2021]. Disponible en: https://www.youtube.com/watch?v=zpZI8KZ4S5Y&t=5000s.

MOHAMED, A., 2017. Comparative Study of Four Supervised Machine Learning Techniques for Classification. , vol. 7.

MOSCOVICH, F.A., IVANDIC, F. y BESOLD, L.C., 2014. Manual de combate de incendios forestales y manejo de fuego. (Nivel Inicial). En: Accepted: 2019-09-04T16:53:44Z [en línea]. info:ar-repo/semantics/libro. S.l.: Ediciones INTA. [Consulta: 24 mayo 2021]. Disponible en: http://repositorio.inta.gob.ar:80/handle/20.500.12123/5780.

MURPHY, K.P., 2012. Introduction. *Machine Learning: A Probabilistic Perspective*. S.l.: The MIT Press, pp. 1-2. ISBN 978-0-262-01802-9.

NATIONAL GEOGRAPHIC, 2019. Wildfires. *National Geographic Society* [en línea]. [Consulta: 26 abril 2021]. Disponible en: http://www.nationalgeographic.org/encyclopedia/wildfires/.

NATIONAL WILDFIRE COORDINATING GROUP, 2021. Fire Weather Index (FWI) System. [en línea]. [Consulta: 24 mayo 2021]. Disponible en: https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system.

NEWMAN, S., 2015. Microservices. *Building Microservices: Designing Fine-Grained Systems*. Beijing Sebastopol, CA: s.n., pp. 4-11. ISBN 978-1-4919-5035-7.

PAUSAS, J.G., 2020. Un mundo inflamable. En: Google-Books-ID: i1jZDwAAQBAJ, *Incendios forestales*. S.l.: Los Libros De La Catarata, pp. 4-5. ISBN 978-84-9097-895-5.

RAMASUBRAMANIAN, K. y SINGH, A., 2016. Machine Learning theory and practices. En: Google-Books-ID: jYrJDQAAQBAJ, *Machine Learning Using R*. S.l.: Apress, pp. 254-256. ISBN 978-1-4842-2334-5.

RESPIRÁ PINAMAR, 2020. Naturaleza. [en línea]. [Consulta: 27 marzo 2021]. Disponible en: https://www.respirapinamar.com.ar/naturaleza/.

RICHARDS, M., 2015. Layered Architecture. *Software Architecture Patterns* [en línea]. S.l.: s.n., [Consulta: 29 septiembre 2021]. ISBN 9781491924242. Disponible en: https://www.oreilly.com/library/view/software-architecture-patterns/9781491971437/ch01.html.

RIJAL, B., 2018. Quantile regression: An alternative approach to modelling forest area burned by individual fires. *International Journal of Wildland Fire*, vol. 27. DOI 10.1071/WF17120.

RODRIGUES, M. y DE LA RIVA, J., 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, vol. 57, pp. 192-201. ISSN 1364-8152. DOI 10.1016/j.envsoft.2014.03.003.

RODRIGUEZ, R.N., 2017. Five Things You Should Know about Quantile Regression. [en línea]. [Consulta: 1 agosto 2021]. Disponible en: https://www.semanticscholar.org/paper/Five-Things-You-Should-Know-about-Quantile-Rodriguez/7e824688e34f6010ee203c86643b17ccbd9acf90.

RUSSELL, S. y NORVIG, P., 2009a. Introduction. *Artificial Intelligence: A Modern Approach*. 3rd. USA: Prentice Hall Press, pp. 2-3. ISBN 978-0-13-604259-4.

RUSSELL, S. y NORVIG, P., 2009b. Learning from Examples. *Artificial Intelligence: A Modern Approach*. 3rd. USA: Prentice Hall Press, pp. 693-697, 727-732. ISBN 978-0-13-604259-4.

SANBORN, 2021. Wildfire Risk Management Software | Wildfire Map. [en línea]. [Consulta: 15 mayo 2021]. Disponible en: https://www.sanborn.com/wfras-software/.

SCRUM.ORG, 2020. The Scrum Framework Poster. *Scrum.org* [en línea]. [Consulta: 17 agosto 2021]. Disponible en: https://www.scrum.org/resources/scrum-framework-poster.

SECRETARÍA DE AMBIENTE Y DESARROLLO SUSTENTABLE, 2021. *Índice de peligro de incendios forestales* [en línea]. 2021. S.l.: s.n. Disponible en: https://www.smn.gob.ar/sites/default/files/mapasdepeligro.pdf.

SERVICIO METEOROLÓGICO NACIONAL, 2021. Índices de peligro de incendio. [en línea]. [Consulta: 14 mayo 2021]. Disponible en: https://www.smn.gob.ar/indices_peligro_fuego.

SERVICIO NACIONAL DE MANEJO DEL FUEGO, 2020. Reporte de incendios 2020. [en línea]. S.l.: Disponible en: https://www.argentina.gob.ar/sites/default/files/31-dic-reporte_incendios_.pdf.

STOJANOVA, D., KOBLER, A., OGRINC, P., ŽENKO, B. y DŽEROSKI, S., 2012. Estimating the risk of fire outbreaks in the natural environment. *Data Mining and Knowledge Discovery*, vol. 24, no. 2, pp. 411-442. ISSN 1573-756X. DOI 10.1007/s10618-011-0213-2.

TECNOSYLVA, 2017. Wildfire Analyst | Tecnosylva. *Wildfire Analyst | Tecnosylva* [en línea]. [Consulta: 15 mayo 2021]. Disponible en: https://tecnosylva.es/wildfire-analyst.

VEGA GARCIA, LEE B.S, WOODARD P.M, y TITUS S.J, 1996. Applying neural network technology to human-caused wildfire occurrence prediction. *AI applications* [en línea], [Consulta: 16 mayo 2021]. ISSN 1051-8266. Disponible en: https://agris.fao.org/agris-search/search.do?recordID=US19970095050.

VILLERS-RUIZ, L., CHUVIECO, E. y AGUADO, I., 2012. Aplicación del índice meteorológico de incendios canadiense en un Parque Nacional del centro de México. *Revista mexicana de ciencias forestales*, vol. 3, no. 11, pp. 25-40. ISSN 2007-1132.

WAIDELICH, S., ZIMMERMAN, V., LANERI, K. y DENHAM, M.M., 2019. Fire Weather Index assessment and visualization. *XXV Congreso Argentino de Ciencias de la Computación (CACIC) (Universidad Nacional de Río Cuarto, Córdoba, 14 al 18 de octubre de 2019)* [en línea]. S.l.: s.n., [Consulta: 25 mayo 2021]. ISBN 978-987-688-377-1. Disponible en: http://sedici.unlp.edu.ar/handle/10915/90905.

WAN, ZHENGMING, HOOK, SIMON y HULLEY, GLYNN, 2015. *MYD11C1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006* [en línea]. 2015. S.l.: NASA EOSDIS Land Processes DAAC. [Consulta: 23 julio 2021]. Disponible en: https://lpdaac.usgs.gov/products/myd11c1v006/.

WANG, Qianru, ZHANG, J., GUO, B., HAO, Z., ZHOU, Y., SUN, J., YU, Z. y ZHENG, Y., 2019. CityGuard: Citywide Fire Risk Forecasting Using A Machine Learning

Approach. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, pp. 1-21. DOI 10.1145/3369814.

WANG, Sally y WANG, Y., 2019. *Predicting wildfire burned area in South Central US using integrated machine learning techniques*. S.l.: s.n.

XIE, Y. y PENG, M., 2019. Forest fire forecasting using ensemble learning approaches. *Neural Computing and Applications*, vol. 31, no. 9, pp. 4541-4550. ISSN 1433-3058. DOI 10.1007/s00521-018-3515-0.

YAP, B.W., RANI, K.A., RAHMAN, H.A.A., FONG, S., KHAIRUDIN, Z. y ABDULLAH, N.N., 2014. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. En: T. HERAWAN, M.M. DERIS y J. ABAWAJY (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*. Singapore: Springer, pp. 13-22. ISBN 978-981-4585-18-7. DOI 10.1007/978-981-4585-18-7_2.

ZHENG, A. y CASARI, A., 2018. Categorical Variables. En: Google-Books-ID: sthSDwAAQBAJ, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. S.l.: O'Reilly Media, Inc., pp. 77-81. ISBN 978-1-4919-5319-8.

ANEXO A. Glosario

Algoritmo: conjunto de pasos u operaciones para hallar la solución a un problema.

API: (*Application Programming Interfaces)*: conjunto definiciones y protocolos para el desarrollo e integración de aplicaciones para que las funcionalidades definidas por un sistema puedan ser utilizadas por otro.

Árbol binario: estructura de datos no lineal compuesta por nodos o elementos que se estructuran de forma jerárquica, en donde cada nodo puede tener dos nodos hijos como máximo.

Big Data: conjunto grande de datos que crece a un ritmo rápido, por lo que requiere un manejo especializado para procesarlos.

Bootstrap: método de remuestreo en el que se usan muestras con reemplazo al azar, o puesto en otras palabras, las observaciones utilizadas para la muestra pueden volver a ser utilizadas.

Espacio de hipótesis: conjunto de todas las hipótesis (funciones) posibles que un modelo puede devolver.

Espacio dimensional: espacio generado por la cantidad de variables de entrada o características del conjunto de datos, en donde cada una representa una dimensión.

Frontera de decisión: recta (en caso de tener dos variables de entrada) o hiperplano (cuando hay más de dos variables de entrada) que separa una clase de otra, permitiendo clasificar datos.

Función de costo: función que devuelve el error (representado numéricamente) entre los resultados predichos y los reales.

Función hipótesis (*h*): función que mejor describe la variable objetivo en función de las variables de entrada o características.

Función verdadera: función desconocida que mapea las variables de entrada con la variable de salida.

Hiperparámetros: conjunto de configuraciones externas al modelo que se establecen antes de comenzar con el proceso de aprendizaje y no pueden ser determinadas a partir de los datos.

Hiperplano: subespacio plano y afín de dimensiones p-1. Por ejemplo, un hiperplano en $\mathbb{R}$ es un punto, un hiperplano en el plano es una recta y un hiperplano en el espacio es un plano.

Hoja (de un árbol binario): nodo que no contiene referencias a nodos hijo.

Inteligencia Artificial: habilidad de una computadora para realizar tareas asociadas a seres inteligentes o imitar el comportamiento inteligente de los seres humanos.

Linealmente separable: condición de existencia de un elemento (recta, plano, etc.) de una dimensión menor que separe los grupos que se encuentran en una dimensión dada.

Nodo: elemento de un árbol que contiene un dato.

Regresión por cuantiles: método estadístico que modela la relación entre variables independientes y los cuantiles de una variable dependiente.

Relación ordinal: conjunto de variables en donde cada una tiene un orden particular.

Serie temporal: conjunto de observaciones ordenado cronológicamente, en donde se representa el valor que adquiere una variable a lo largo del tiempo.

Validación cruzada: técnica para medir el rendimiento de distintos modelos dividiendo el conjunto de datos de entrenamiento en bloques iguales, donde en cada entrenamiento el conjunto de datos de validación cambia.

ANEXO B. Parámetros configurables del pipeline de datos

El *pipeline* de datos está preparado para extraer datos de cualquier rango de fechas y coordenadas geográficas. No obstante, estas configuraciones son válidas especialmente para los atributos extraídos de fuentes como la NASA, que trabaja con datos a escala global. Por otro lado, el filtrado de fechas es lo que permite al momento de redacción del presente trabajo obtener los datos correspondientes a los años 2015-2019, limitación impuesta por la falta de registros de incendios previos o posteriores a dichos años. Los parámetros que se pueden ajustar se detallan en la Tabla XXI.

TABLA XXI: Parámetros configurables del *pipeline* de datos.

Nombre del parámetro	**Descripción**
yearRange	Es una lista con el rango de años que se desea trabajar. Formato: [desde, hasta].
from	Fecha *desde*. Inicio del rango de años, especificando día y mes. Formato: "%d-%m-%Y".
to	Fecha *hasta*. Fin del rango de años, especificando día y mes. Formato: "%d-%m-%Y".
coordinates	Es una lista con las coordenadas que se desean trabajar. Formato: [longitud_1, latitud_1, longitud_2, latitud_2].
boundingBox	Es un objeto que al igual que coordinates especifica las coordenadas del área del interés. Formato: {lower_left_longitude, lower_left_latitude, upper_right_longitude, upper_right_latitude}.
polygon	Es un listado de coordenadas que demarcan el área de interés válida para generar puntos de "no incendio" o ejemplos negativos. *Nota: dado que coordinates y boundingBox actualmente son cuadrados, polygon excluye coordenadas inválidas como las correspondientes al mar.*

ANEXO C. Generación de puntos de "No Incendio"

Con el objetivo de balancear el *dataset* en términos de ejemplos positivos (Incendio) y negativos (No incendio), se implementó el siguiente algoritmo para generar los ejemplos negativos ya que ha sido utilizado en diversos trabajos que han estudiado la predicción de incendios forestales.

Por cada incendio ocurrido repetir los pasos i, ii y iii:

i. Seleccionar todos los incendios ocurridos en un lapso de ± 3 días.
ii. Por cada incendio seleccionado en i, crear una región de 1.5 kilómetros a la redonda.
iii. Generar un par de coordenadas aleatorias que se encuentren dentro del área exterior a la unión de las regiones generadas en el punto ii.

El resultado se puede observar en la Figura 26, donde los puntos rojos corresponden a incendios forestales y los azules a los puntos generados aleatoriamente para representar lugares donde no han ocurrido incendios forestales.

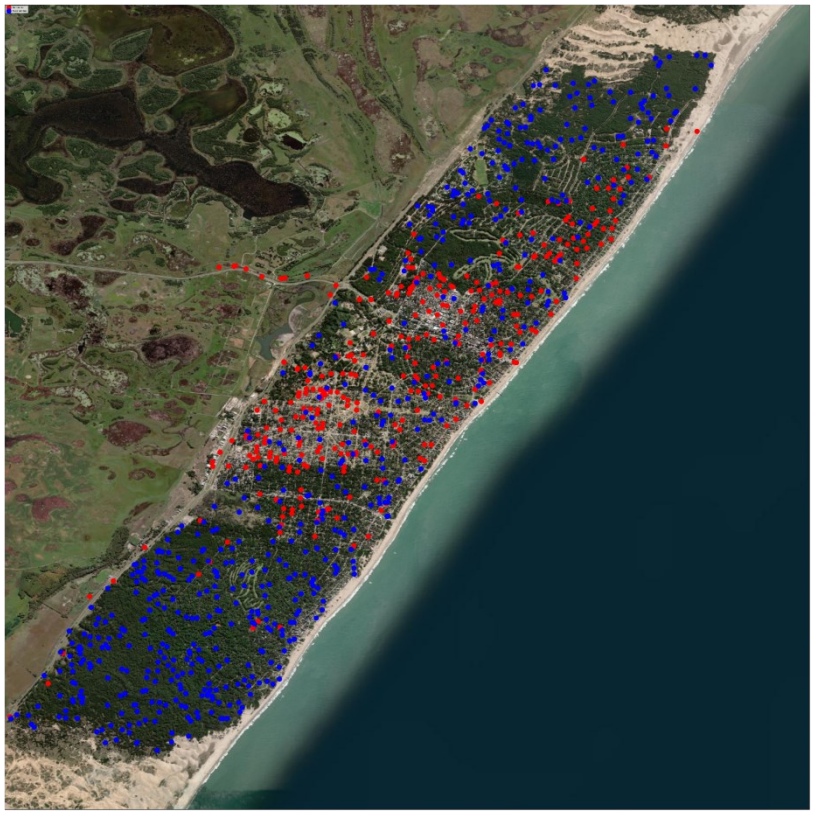

Figura 26: Ubicaciones de incendios forestales durante los años 2015 – 2019 (rojo) y sus correspondientes puntos de no incendio (azul).

ANEXO D. Análisis de datos

En miras de entender la distribución de los atributos que conforman el *dataset* de incendios forestales y las relaciones entre ellos y la variable objetivo (Superficie), se elaboraron los gráficos que se detallan a continuación.

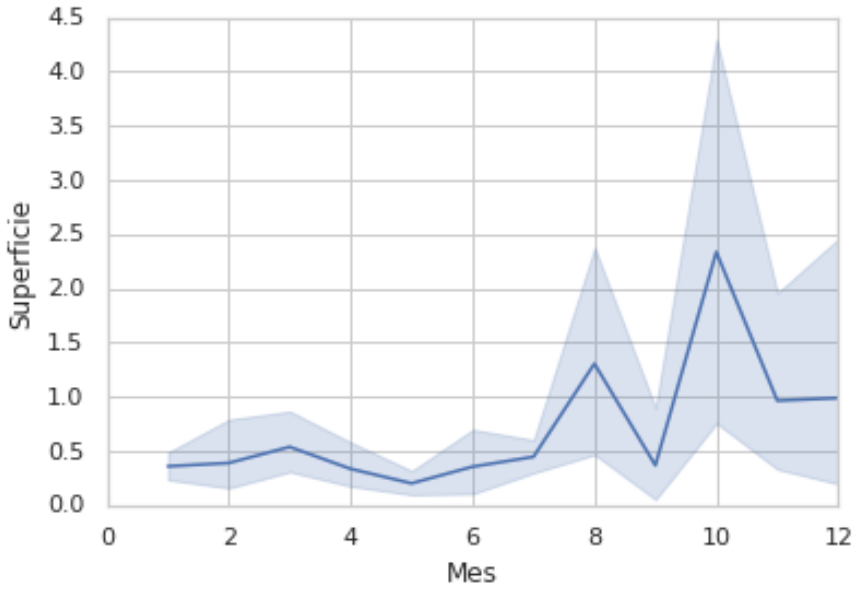

Figura 26: Relación entre superficie quemada por incendios forestales y el mes del año.

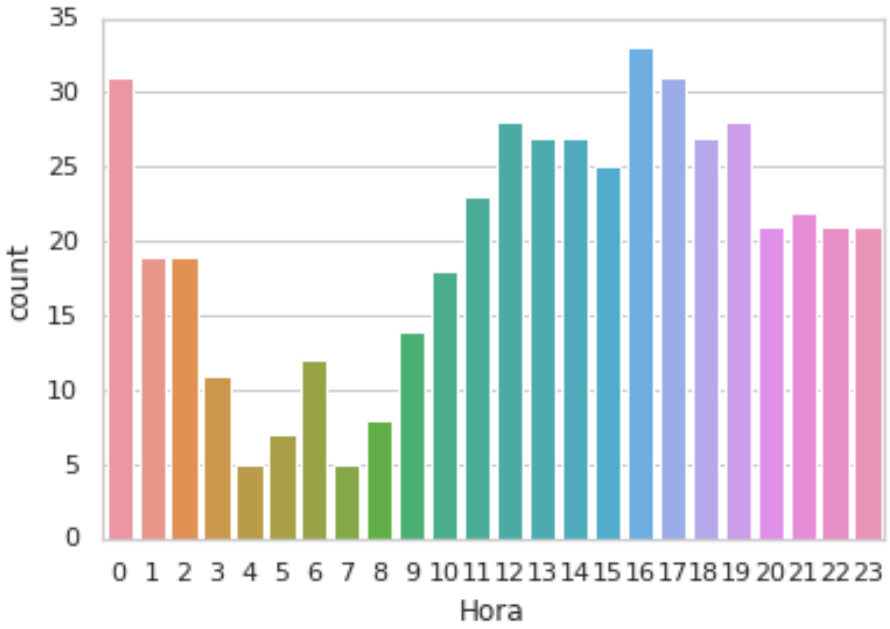

Figura 27: Relación entre cantidad de incendios forestales ocurridos y la hora del día.

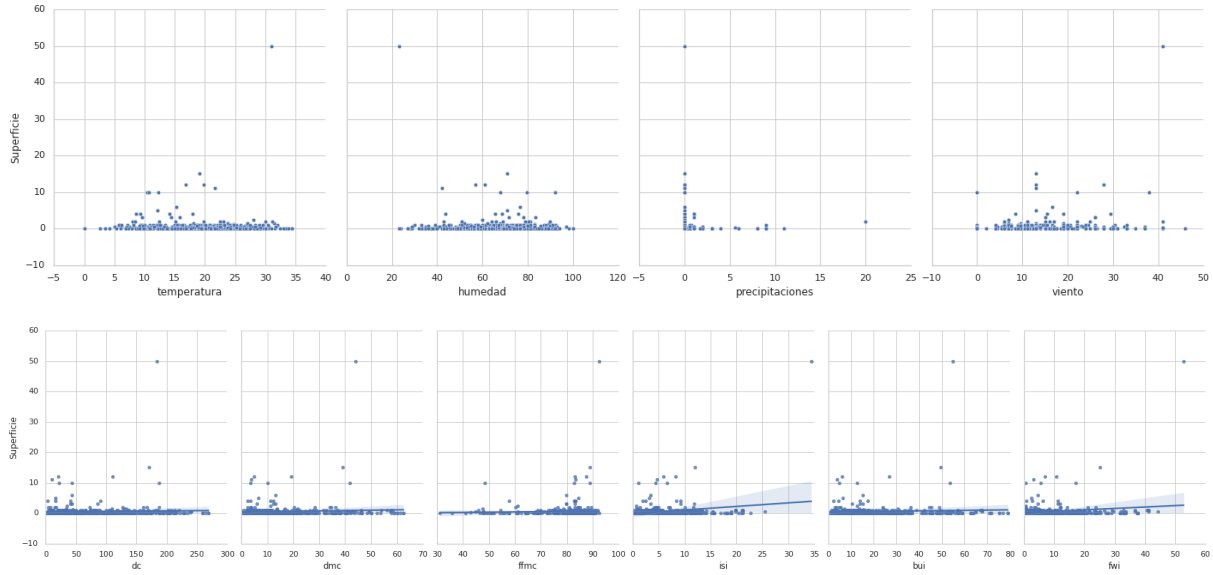

Figura 28: Relación entre variables meteorológicas y de combustible con la superficie afectada por incendios forestales.

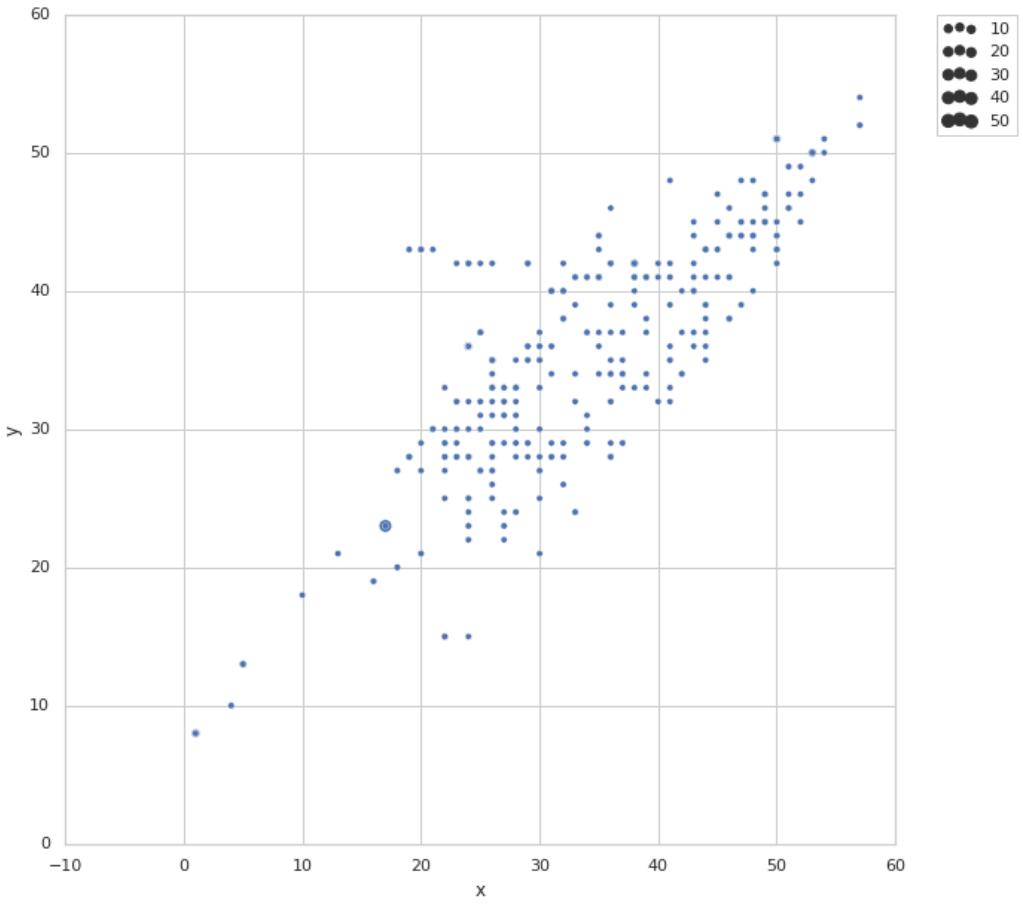

Figura 29: Ubicaciones de incendios forestales durante los años 2015 – 2019 según la grilla definida.

ANEXO E. Factores de inflación de la varianza

Con el objetivo de determinar si las variables independientes están altamente correlacionadas entre sí se llevó a cabo el cálculo de factores de inflación de la varianza (VIF), arrojando los resultados detallados en la Tabla XXII.

TABLA XXII: Valores de VIF para los atributos del *dataset* de incendios.

Atributo	**VIF**
Mes	3.81
Día	4.56
Hora	4.67
Día no laboral	1.52
X	30.28
Y	54.65
DC	10.19
DMC	12.12
FFMC	53.37
ISI	6.28
NDVI	16.9
Elevación	13.84
Temperatura	13.41
Humedad	17.32
Viento	5.39
Precipitaciones	1.11
Superficie	1.16

ANEXO F. Marco de trabajo Scrum

El marco de trabajo Scrum tiene como objetivo gestionar el desarrollo de un producto con especial énfasis en el valor que el mismo le provee al cliente. Para esto los proyectos se estructuran *sprints* o iteraciones de duración fija en las que se propone llevar a cabo una serie de tareas (contenidas en el *sprint backlog*) que puedan realizarse en dicho lapso. De esta forma, en cada iteración el producto se desarrolla gradual e incrementalmente y se valida con el cliente el valor que se agregó finalizado cada *sprint*.

En este marco de trabajo se distinguen varios artefactos que facilitan el trabajo a realizar: el *product backlog* es una lista priorizada de todos los requerimientos que se tienen sobre el producto, el *sprint backlog* consiste en ítems del *product backlog* que serán trabajados durante un *sprint*, y por último el incremento es el conjunto de ítems del *product backlog* que han sido desarrollados. Además, dentro de Scrum se contemplan tres roles: desarrolladores, (construyen el producto), *product owner* (prioriza el *product backlog*) y *Scrum master* (da soporte al equipo para aplicar la metodología Scrum en el desarrollo del producto).

Scrum cuenta a su vez con una serie de eventos que definen la forma en la que se trabaja bajo esta metodología, tal como se puede observar en la Figura 30. En primer lugar se encuentra el *sprint planning,* en donde el equipo define los objetivos del *sprint* y las tareas que pueden llevarse a cabo en el mismo. En segundo lugar, diariamente se lleva a cabo la *daily meeting,* una reunión donde los desarrolladores ponen en común el progreso para detectar la necesidad de ajustar o no el *sprint backlog.* En tercer lugar se realiza la *sprint review* o revisión del sprint en conjunto con el cliente para validar el incremento y determinar si deben realizar ajustes sobre lo trabajado. Por último, la reunión de *sprint retrospective* consiste en identificar, una vez finalizado un *sprint*, cursos de acción para mejorar la calidad y efectividad de la forma de trabajo en general para implementar en el próximo *sprint*, promoviendo así la mejora continua.

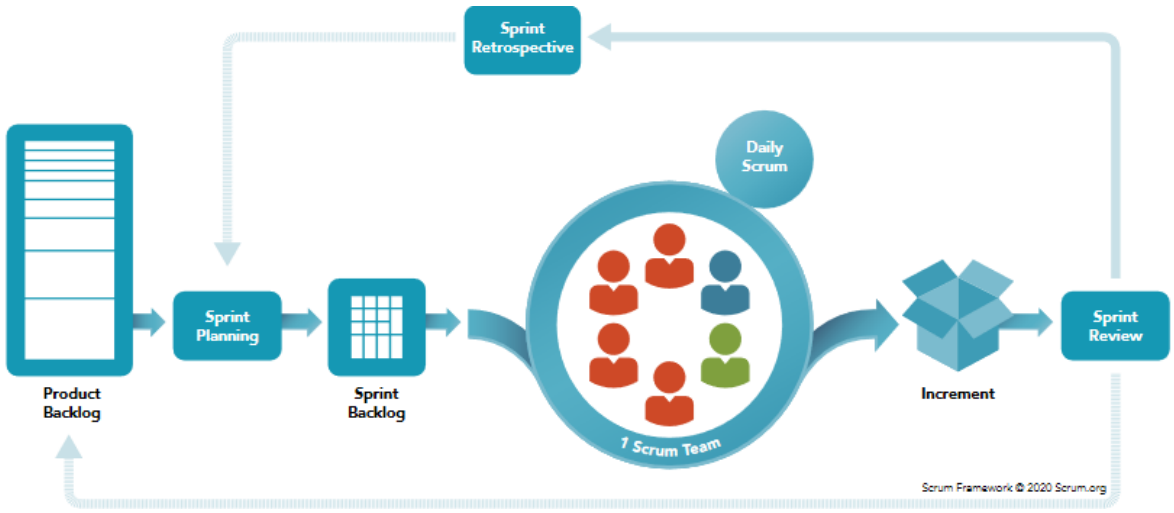

Figura 30: Marco de trabajo *Scrum* (Scrum.org, 2020).

ANEXO G. Documentación de la API de AQUA

En el primer *release* de AQUA se disponen los *endpoints* en la API que permiten obtener predicciones de incendios e información histórica de incendios forestales. Las respuestas de estos *endpoints* se estructuran según los siguientes objetos:

- *FirePrediction* (predicción de incendio para una celda de la grilla, o par de coordenadas del mapa). Atributos:
  - latitude: latitud correspondiente a la ubicación sobre la cual se predice el área a quemar por un incendio forestal.
  - longitude: longitud correspondiente a la ubicación sobre la cual se predice el área a quemar por un incendio forestal.
  - burntArea: área predicha a quemar por un incendio forestal. El valor 0 indica que se no predijo la ocurrencia de incendio forestal en las coordenadas dadas.
- *FireOccurrence* (incendio forestal histórico). Atributos:
  - date: fecha y hora en la que se produjo el incendio forestal.
  - holiday: indica si se trató de un feriado o fin de semana.
  - fleet: cantidad de móviles utilizados en combate.
  - men: cantidad de bomberos presentes en combate.
  - duration: tiempo insumido en minutos por los bomberos para combatir el incendio forestal.
  - area: área quemada por el incendio.
  - latitude: latitud correspondiente a la ubicación donde se produjo un incendio forestal.
  - longitude: longitud correspondiente a la ubicación donde se produjo un incendio forestal.

TABLA XXIII: Descripción de la API de AQUA.

Método	**URL**	**Parámetros**	**Respuesta**
GET	/api/v1.0/predictions	date (*String)*: fecha de predicción.	predictions: lista de objetos *FirePrediction*.

		hour (*String)*: hora de predicción. latitude_1 (*Float)*: primera latitud del área de predicción. latitude_2 (*Float)*: segunda latitud del área de predicción. longitude_1 (*Float)*: primera longitud del área de predicción. longitude_2 (*Float)*: segunda longitud del área de predicción.	metadata: objeto con los valores de exactitud, precisión, sensibilidad y RMSE de los modelos productivos.
GET	/api/v1.0/fires	from (*String)*: incendios ocurridos a partir de dicha fecha. to (*String)*: incendios ocurridos hasta dicha fecha. latitude_1 (*Float)*: primera latitud del área de interés. latitude_2 (*Float)*: segunda latitud del área de interés. longitude_1 (*Float)*: primera longitud del área de interés.	fires: lista de objetos *FireOccurrence.*

		longitude_2 (*Float)*: segunda longitud del área de interés.	
POST	/api/v1.0/fires	fires: lista de objetos *FireOccurrence*	N/A

ANEXO H. Diagramas de secuencia

En el presente apartado se detallan a través de diagramas de secuencia la interacción que se produce entre los distintos componentes de AQUA para obtener tanto las predicciones de incendio forestal como los incendios históricos.

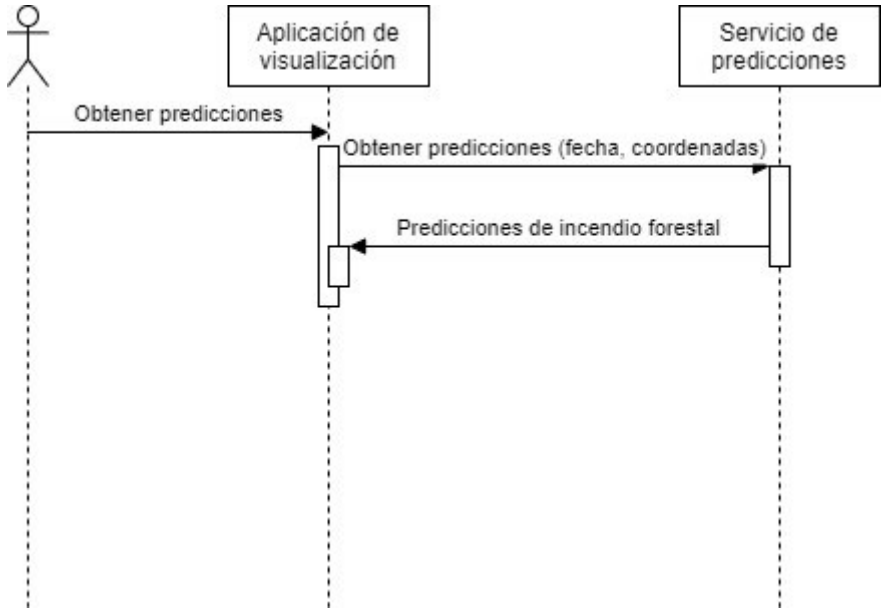

Figura 31: Diagrama de secuencia: Obtener predicciones (1).

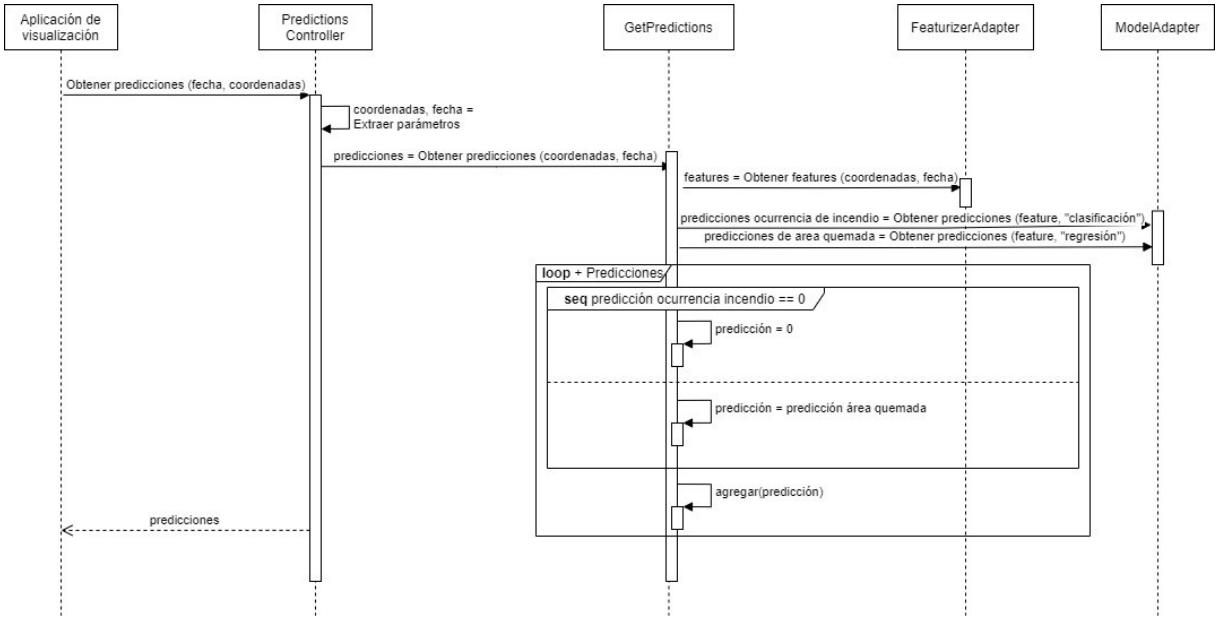

Figura 32: Diagrama de secuencia: Obtener predicciones (2).

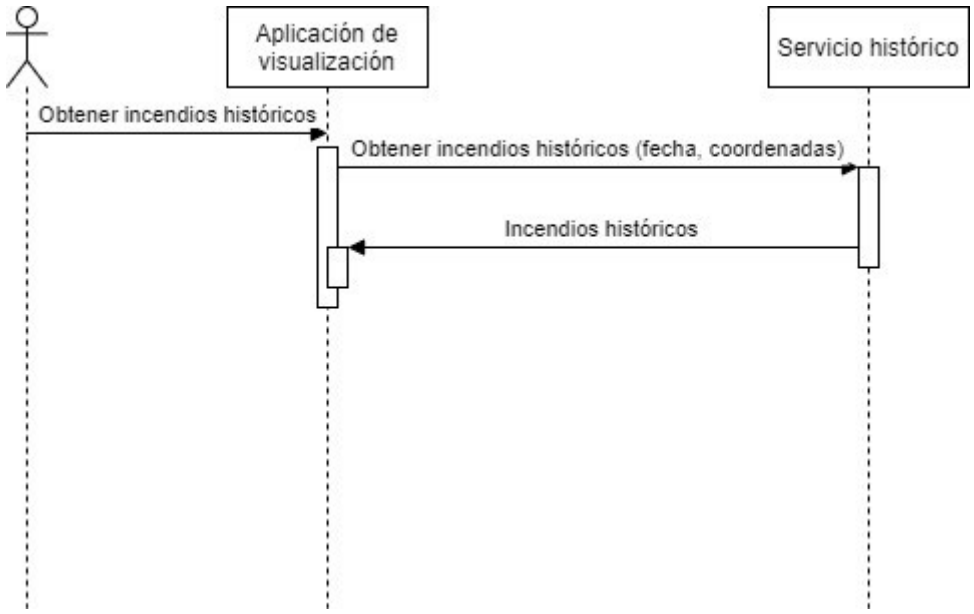

Figura 33: Diagrama de secuencia: Obtener incendios históricos (1).

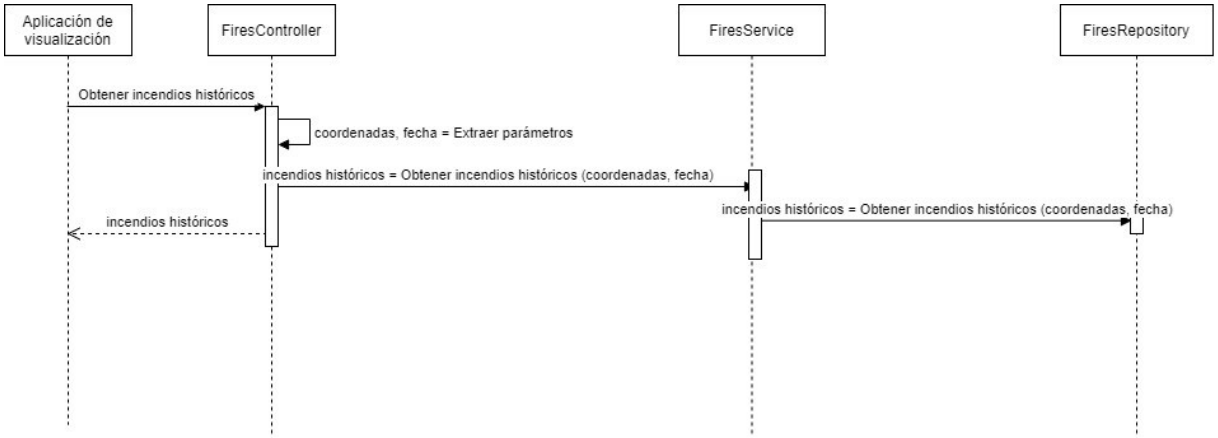

Figura 34: Diagrama de secuencia: Obtener incendios históricos (2).

ANEXO I. Análisis financiero

Con el fin de calcular el VAN, TIR y pay *back* de AQUA se definieron tres escenarios posibles en el horizonte de tiempo considerado para analizar financieramente el proyecto. A continuación se detallan los valores de referencia utilizados y los flujos de fondos netos obtenidos con estos valores.

TABLA XXIV: Estimación de cuarteles de bomberos que utilizarán AQUA bajo distintos escenarios en un horizonte de 5 años.

	1	**2**	**3**	**4**	**5**
Optimista	3	6	10	15	20
Neutro	2	4	8	10	12
Pesimista	1	2	5	6	8

TABLA XXV: Estimación de dueños de campos agrícolas-ganaderos que adquirirán AQUA bajo distintos escenarios en un horizonte de 5 años.

	1	**2**	**3**	**4**	**5**
Optimista	5	10	15	30	40
Neutro	3	8	12	25	30
Pesimista	2	5	10	15	20

TABLA XXVI: Estimación de flujos de fondos neto considerando distintos escenarios en un horizonte de 5 años (considerando impuestos del 35%).

FFN		**Período**					
		0	1	2	3	4	5
Escenario	Optimista	-$5559	-$242	$382	$1084	$2644	$3814
	Neutro	-$5559	-$476	$70	$694	$1864	$2410
	Pesimista	-$5559	-$632	-$320	$304	$772	$1318