Desarrollo de un Modelo de Estimación para la Prevención de Incendios Forestales

Luis Daniel Fuentes Licero
Dpto. de ingeniería de sistemas
Universidad del Norte
Barranquilla, Colombia
licerol@uninorte.edu.co

Edilberto Mario Rodríguez Fontalvo
Dpto. de ingeniería de sistemas
Universidad del Norte
Barranquilla, Colombia
edilbertof@uninorte.edu.co

Juan David Anzola Quiroga
Dpto. de ingeniería de sistemas
Universidad del Norte
Barranquilla, Colombia
jdanzola@uninorte.edu.co

Tutor de Proyecto: PhD. Eduardo Enrique Zurek Varela

***Abstract*— This project implements and evaluates predictive models for forest fire estimation using advanced machine learning techniques. A Random Forest model and a Bayesian model (Gaussian Naive Bayes) were compared, integrating historical, meteorological, topographical and socioeconomic data. CRISP-DM methodology and K-Fold cross-validation ensured the robustness and generalization of the models. The results show that the Random Forest model, with an accuracy of 83% was better than the Gaussian model with an accuracy of 64%. The most influential variables were the brightness temperature, the radiant power of the fire and the size of the fire scan. This work demonstrates the effectiveness of machine learning in wildfire prediction and provides valuable insights for future research and prevention.**

Resumen—Este proyecto implementa y evalúa modelos predictivos para la estimación de incendios forestales utilizando técnicas avanzadas de aprendizaje automático. Se compararon un modelo de Random Forest y uno Bayesiano (Gaussian Naive Bayes), integrando datos históricos, meteorológicos, topográficos y socioeconómicos. La metodología CRISP-DM y la validación cruzada K-Fold aseguraron la robustez y generalización de los modelos. Los resultados muestran que el modelo de Random Forest, con una precisión del 83% fue mejor que el Gaussiano frente al 64%. Las variables más influyentes fueron la temperatura de brillo, la potencia radiativa del fuego y el tamaño del escaneo del fuego. Este trabajo demuestra la eficacia del aprendizaje automático en la predicción de incendios forestales y proporciona insights valiosos para la investigación y prevención futuras.

Keywords*—*Predicción de Incendios Forestales, Aprendizaje Automático, Random Forest, Modelo Bayesiano, Validación Cruzada K-Fold, CRISP-DM, Datos Meteorológicos, Datos Topográficos, Datos Socioeconómicos, Prevención de Incendios

I. INTRODUCCIÓN

Los incendios forestales son una creciente preocupación mundial por su impacto negativo en el medio ambiente, la economía y la seguridad de las comunidades, requiriendo enfoques innovadores para su gestión y prevención (Hidayanto et al., 2021). Esta creciente amenaza, presionada por el cambio climático y el aumento de la intervención humana en áreas naturales, subraya la necesidad de adoptar enfoques innovadores como técnicas de análisis de datos y modelado para anticipar y mitigar los efectos de estos eventos para su prevención y gestión (Chew et al., 2020). El avance de tecnologías emergentes como Big Data, Machine Learning e Inteligencia Artificial (IA) ofrece nuevas posibilidades para desarrollar modelos analíticos avanzados para mejorar la eficacia de las medidas preventivas y la respuesta ante incendios forestales (García & González, 2012). Estos modelos permiten el análisis integral de datos históricos, meteorológicos, topográficos y socioeconómicos, optimizando así la toma de decisiones y la asignación de recursos para una gestión más efectiva y proactiva de los incendios forestales.

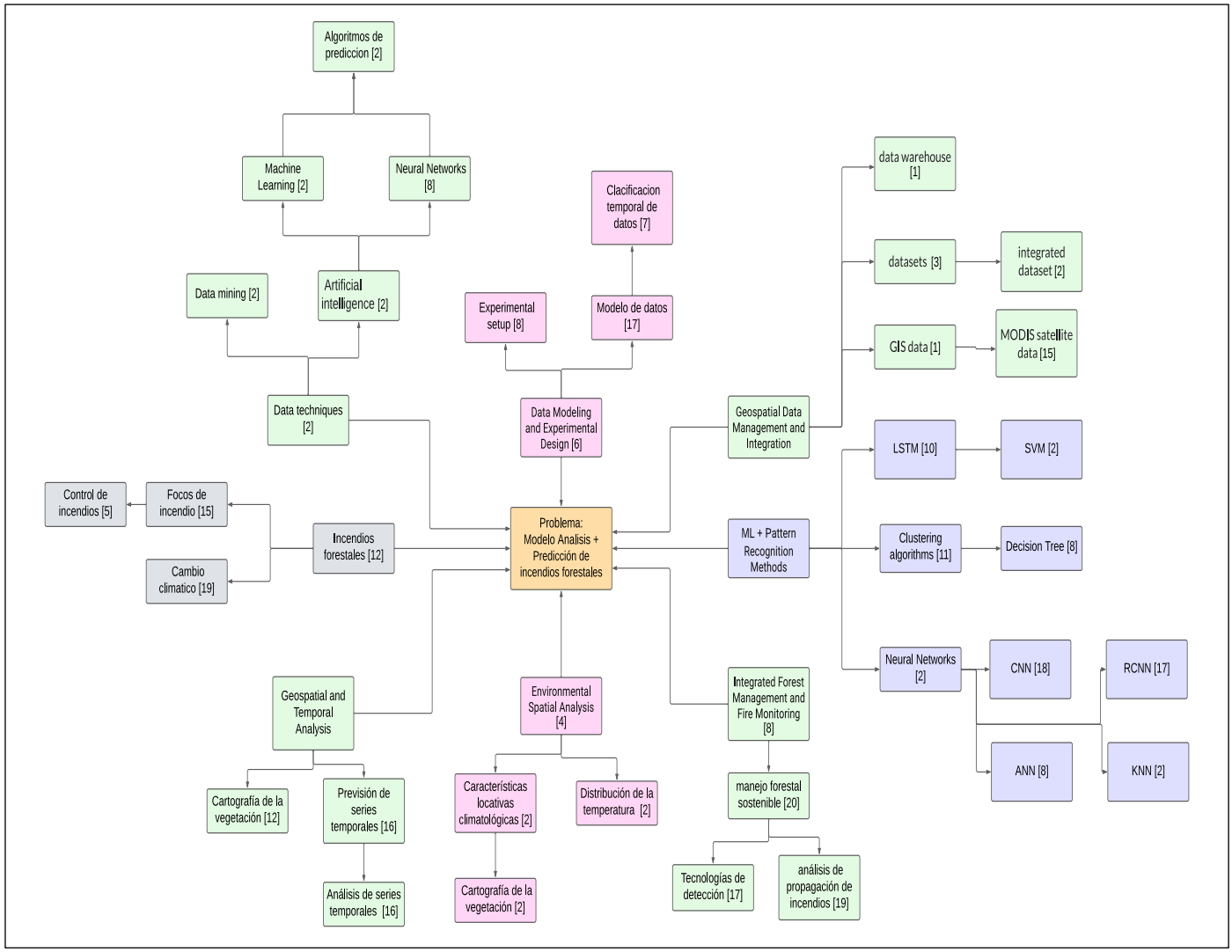
Investigaciones recientes destacan la importancia de estas tecnologías en el ámbito de los incendios forestales. Por ejemplo, el estudio de García y González (2012) explora la determinación de la tendencia espacial de los puntos de calor como estrategia para monitorear y prevenir efectivamente los incendios forestales. Además, la investigación de Ospino Rivera (2022) sobre modelos inteligentes para estimar áreas quemadas enfatiza la relevancia de las herramientas computacionales en la gestión de estos desastres. Del mismo modo, el trabajo de Castelli, Vanneschi y Popovič (2015) resalta la necesidad de estimar la progresión y el área afectada por los incendios, demostrando el valor de la predicción temprana para salvar vidas y proteger recursos naturales.

Estadísticas recientes indican un aumento en la frecuencia e intensidad de los incendios forestales a nivel global, lo que agudiza la urgencia de implementar soluciones tecnológicas avanzadas. Nuestro estudio propone el desarrollo de un modelo analítico que integra no solo datos históricos y meteorológicos, sino también información topográfica y socioeconómica, para estimar las áreas de mayor riesgo de incendios forestales. Además, busca mejorar el proceso de evaluación y respuesta ante estos eventos. Así mismo, la combinación de datos satelitales, sistemas de información geográfica y modelos de simulación ha demostrado ser una herramienta poderosa para comprender la dinámica de los incendios forestales (Hidayanto et al., 2021). Por lo tanto, esta estrategia se alinea con investigaciones previas que han demostrado la efectividad de modelos basados en Machine Learning para estimar la propagación de incendios forestales y optimizar la asignación de recursos (Chew et al., 2020). Este documento detalla el diseño e implementación de un sistema inteligente basado en técnicas de IA para la estimación y monitoreo de incendios forestales, destacando los métodos utilizados, los resultados obtenidos y las implicaciones prácticas de nuestra investigación. La adopción de metodologías como CRISP-DM refuerza la robustez y aplicabilidad de nuestras soluciones tecnológicas, posicionándolas como un pilar fundamental en la prevención y gestión de incendios forestales (Chew, Ooi & Pang, 2020).

II. DESCRIPCIÓN DEL PROBLEMA

Los incendios forestales representan una amenaza grave para los ecosistemas, causados por la crisis del cambio climático y la actividad humana. Su frecuencia ha aumentado en los últimos años, causando efectos devastadores en la biodiversidad, el suelo y la economía (Sharma et al, 2022). Nuestra propuesta se basa en el desarrollo de un modelo predictivo basado en datos históricos y variables relevantes, utilizando técnicas avanzadas de análisis de datos y aprendizaje automático para estimar la probabilidad de ocurrencia de incendios forestales en áreas específicas, lo que proporcionaría información valiosa para tomar decisiones informadas y mitigar los impactos negativos de los incendios forestales en el medio ambiente y la sociedad.

(Ver grafica 1)



Graf. 1 Mapa de problema planteado (Ver)

III. JUSTIFICACIÓN

La prevención y gestión eficaz de incendios forestales es un desafío global que demanda soluciones innovadoras, especialmente en un contexto donde el cambio climático y las actividades humanas incrementan el riesgo y la frecuencia de estos desastres. Este estudio aborda la estimación de incendios forestales, un componente crucial para la asignación eficiente de recursos y la mitigación de daños. Aunque nuestro enfoque se centra en un contexto local, las metodologías y tecnologías que exploramos tienen una aplicación y relevancia globales. Investigaciones anteriores han demostrado el potencial de la inteligencia artificial en la predicción de incendios, como el algoritmo de predicción de riesgo basado en SVM presentado por Sakr et al. (2010), y el uso de RNN para estimar la propagación del fuego, con una precisión notable de 94.77%, según Natekar et al. (2021). Estos avances subrayan el valor de las técnicas computacionales avanzadas en la mejora de la precisión predictiva y, por ende, en la protección de vidas y ecosistemas.

Adicionalmente, el análisis de susceptibilidad al fuego en áreas propensas a incendios utilizando modelos de aprendizaje automático, como en el estudio realizado en Pulang Pisau Regency, Indonesia, por Hidayanto et al. (2021), resalta la eficacia de estas herramientas en la predicción de incendios forestales. La relevancia de este trabajo se amplifica al considerar eventos locales recientes, que han demostrado la vulnerabilidad de nuestras comunidades y ecosistemas ante incendios forestales, resaltando la necesidad urgente de soluciones tecnológicas avanzadas para su prevención y manejo.

Por tanto, este estudio no solo se basa en una sólida investigación previa, sino que también aspira a expandir el conocimiento y las aplicaciones de herramientas tecnológicas en la lucha contra los incendios forestales. Mediante la integración de datos multidisciplinarios y el desarrollo de modelos analíticos y predictivos avanzados, buscamos proporcionar medios más efectivos y proactivos para la gestión de incendios forestales, con el potencial de salvar vidas humanas, preservar propiedades y proteger ecosistemas vulnerables a nivel local y global.

IV. Objetivo general

Desarrollar un modelo analítico que estime potenciales focos de incendios forestales, integrando datos multidisciplinarios y utilizando técnicas avanzadas de aprendizaje automático para mejorar la prevención y la gestión efectiva de estos eventos.

V. Objetivos específicos

- Consolidar y preprocesar conjuntos de datos históricos, meteorológicos, topográficos y socioeconómicos para análisis predictivo. Apoyado en los hallazgos de Garcia y González (2012).

- Analizar el conjunto de datos e identificar patrones de cambio climático y registros históricos en las últimas décadas para entender su impacto y descubrir las variables que más afectan y generan incendios forestales tomando en cuenta estudios previos como el de Priya y Vani (2023).

- Explorar diferentes técnicas de minería de datos, para analizar la estructura forestal y ajustar las fuentes de datos para mejorar la precisión del modelo, inspirados en el trabajo de Stojanova et al. (2006).

- Investigar y aplicar algoritmos de aprendizaje automático como random forest para mapear y estimar focos de incendios forestales en Colombia a partir de los datos recogidos, como lo demuestran Rosa et al. (2022).

- Implementar métodos de validación cruzada para evaluar la precisión y confiabilidad del modelo, con el objetivo de alcanzar una tasa de error mínima

- Desarrollar graficas interactivas que presenten de manera clara las estimaciones del modelo, permitiendo a los usuarios interpretar fácilmente los datos y las predicciones.

VI. Metodología de desarrollo

Para llevar a cabo el análisis de los incendios forestales, se adoptó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). El uso de esta metodología nos permitió una planificación organizada y efectiva de todas las etapas del proceso de análisis de datos, desde la comprensión del problema hasta el despliegue de los resultados.

Comprensión del problema: En esta fase, se llevó a cabo una revisión exhaustiva de la literatura, en el cual se identificaron y definieron los objetivos del proyecto, centrándose en comprender la problemática de los incendios forestales, sus impactos ambientales y socioeconómicos, así como los requisitos específicos de la investigación.

Comprensión de los datos: Durante esta etapa, se recopilaron y evaluaron los datos relevantes relacionados con los incendios forestales, incluidos registros históricos de incendios, datos climáticos y características del paisaje. Se realizó un análisis exploratorio de los datos para comprender su calidad, completitud y distribución, identificando posibles fuentes de errores que podrían afectar los resultados del análisis.

Preparación de los datos: En esta fase, se llevó a cabo la limpieza y preparación de los datos para su posterior análisis. Se realizaron tareas de limpieza, transformación y selección de variables, así como la integración de múltiples conjuntos de datos.

Modelado: Durante esta etapa, se construyeron y evaluaron modelos predictivos para estimar la ocurrencia y el comportamiento de los incendios forestales. Se realizaron experimentos rigurosos para comparar y evaluar la eficacia de diferentes modelos en función de métricas de rendimiento relevantes.

Evaluación: En esta fase, se evaluaron y validaron los modelos desarrollados utilizando conjuntos de datos de prueba independientes. Se llevaron a cabo análisis de sensibilidad y robustez para evaluar la estabilidad y generalización de los modelos en diferentes condiciones.

Despliegue: Finalmente, en esta etapa, se implementaron los modelos validados en un entorno operativo para su uso en la predicción y gestión de incendios forestales. Se desarrollaron herramientas y aplicaciones basadas en los modelos para proporcionar información útil y oportuna a los gestores de emergencias y tomadores de decisiones.

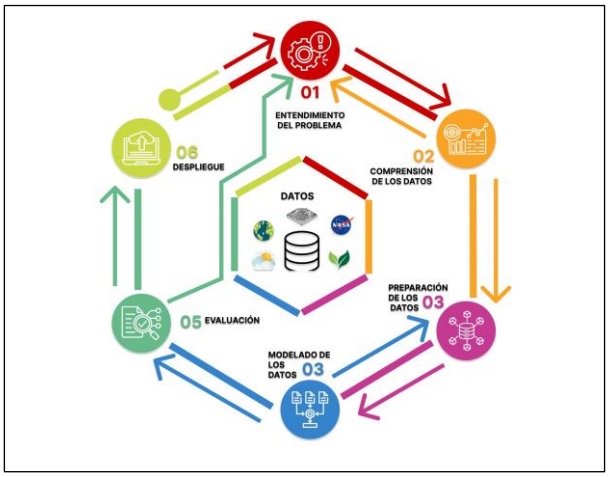


Fig. 2 Diagrama de metodología

VII. REVISIÓN SISTEMÁTICA DE LA LITERATURA

Tabla I
Artículos Encontrados

Fuente	Keywords	Resultados sin filtro	Tipo de articulo: paper de investigación	Tipo de articulo: conferencia
ACM DL	Forest Fire detection AI	495,743	290,541	61440
IEEE Xplore	Forest Fire detection AI	1375	1320	49

Tabla II
Artículos Útiles

Fuente	Aptos leer abstract	Artículos descartados	Aptos para revisión completa	Artículos seleccionados
ACM DL	28	22	6	5
IEEE Xplore	41	15	26	13

VIII. MARCO TEÓRICO

Los incendios forestales representan una amenaza creciente para los ecosistemas naturales en todo el mundo, con consecuencias devastadoras para la biodiversidad y el medio ambiente (Ospino Rivera, 2022). Estos eventos son provocados por una variedad de factores, incluyendo condiciones climáticas extremas, actividades humanas y cambios en el uso del suelo, lo que los convierte en un problema multidimensional (Carta et al., 2023). Además de la pérdida directa de vegetación, los incendios forestales emiten grandes cantidades de gases de efecto invernadero, contribuyendo así al cambio climático y exacerbando sus efectos (Chen et al., 2012). La frecuencia e intensidad de los incendios forestales están aumentando en muchas partes del mundo debido al cambio climático y a la actividad humana, lo que subraya la necesidad urgente de abordar esta crisis (Sakr et al., 2010).

Los avances en la inteligencia artificial y la minería de datos ofrecen nuevas oportunidades para abordar la predicción y detección de incendios forestales (Stojanova et al., 2006). Algoritmos como los Bosques Aleatorios (RF), las Máquinas de Vectores de Soporte (SVC) y la Clasificación por Refuerzo de Gradiente (GBC) han demostrado ser efectivos en la modelización de la susceptibilidad a los incendios forestales (Hidayanto et al., 2021). La aplicación de técnicas de aprendizaje automático, como la regresión logística y los árboles de decisión, ha permitido mejorar la precisión en la predicción de la ocurrencia de incendios (Chew et al., 2020). Además, la fusión de datos de diferentes fuentes, como imágenes satelitales y datos meteorológicos, ha proporcionado una visión más completa y precisa de la dinámica de los incendios forestales (Pal, 2022).

En última instancia, la combinación de enfoques tradicionales y tecnologías emergentes puede conducir a una mejor comprensión y gestión de los incendios forestales (Morales et al., 2014). La integración de datos geoespaciales con modelos de inteligencia artificial puede proporcionar herramientas poderosas para prevenir y combatir los incendios forestales (Stojanova et al., 2006). Sin embargo, es crucial seguir investigando y desarrollando nuevas técnicas y metodologías para hacer frente a este desafío en constante evolución (Preeti et al., 2021). Al aprovechar el potencial de la inteligencia artificial y la ciencia de datos, podemos avanzar hacia un futuro en el que los incendios forestales sean menos frecuentes y destructivos, protegiendo así nuestros preciados recursos naturales y ecosistemas (Chen et al., 2012).

IX. MARCO CONCEPTUAL

Incendios Forestales: Son fuegos no controlados que ocurren en áreas de vegetación y pueden ser causados por factores naturales como rayos, o por actividades humanas.

Inteligencia Artificial (IA): Tecnología que permite que las máquinas imiten las capacidades cognitivas humanas, fundamental para el desarrollo de nuestro modelo predictivo.

Aprendizaje Automático (Machine Learning): Un subcampo de la IA que se centra en la creación de sistemas capaces de aprender de los datos y mejorar con la experiencia, esencial para el análisis predictivo de incendios.

Datos Multidisciplinarios: Conjuntos de datos que incluyen información meteorológica, topográfica, histórica y socioeconómica, utilizados para alimentar y entrenar el modelo predictivo.

Topografía: Estudio de las características físicas de la superficie de la tierra y su descripción detallada, incluyendo elevación, relieve y pendiente, que influyen en el riesgo y comportamiento de los incendios.

Condiciones Meteorológicas: Factores como temperatura, humedad, precipitación y viento que afectan directamente la ignición, propagación y severidad de los incendios forestales.

Minería de Datos: Refiere al proceso de descubrir patrones, correlaciones y anomalías en grandes conjuntos de datos utilizando algoritmos y técnicas estadísticas.

Random Forest: Un ensamble de árboles de decisión para clasificación y regresión que mejora la precisión predictiva mediante la reducción del sobreajuste.

Decision Tree: Un modelo de predicción que usa una estructura de árbol para tomar decisiones, basándose en la segmentación de los datos y las reglas de decisión inferidas.

Regresión Logística: Un modelo estadístico que se utiliza para predecir la probabilidad de una variable dependiente categórica, especialmente útil para la clasificación binaria.

Support Vector Machines (SVM): Una técnica de aprendizaje supervisado eficaz para clasificación y regresión, caracterizada por la búsqueda del hiperplano que mejor divide un conjunto de datos en clases.

Redes Neuronales Artificiales: Modelos computacionales inspirados en el cerebro humano que son capaces de aprender y hacer predicciones o tomar decisiones basándose en los datos de entrada.

Datos Geoespaciales: Información que tiene una ubicación geográfica implícita, como la latitud y longitud, y que se usa a menudo para toma de decisiones basadas en la localización.

CRISP-DM (Cross-Industry Standard Process for Data Mining): Un proceso estándar de la industria que guía a los analistas de datos en la minería de datos y en la construcción de modelos predictivos.

Algoritmos de Clustering: Métodos utilizados para agrupar un conjunto de objetos de tal manera que los objetos en el mismo grupo (o cluster) son más similares entre sí que con los de otros grupos.

Validación Cruzada: Una técnica para evaluar la capacidad predictiva de un modelo estadístico y su rendimiento en un conjunto de datos independiente, a menudo usada para proteger contra el sobreajuste.

Herramientas de visualización de datos: Esenciales para la fácil interpretación de los datos multidisciplinarios, permiten transformar los datos en gráficos, tablas y diagramas comprensibles.

X. Arquitectura lógica de la solución

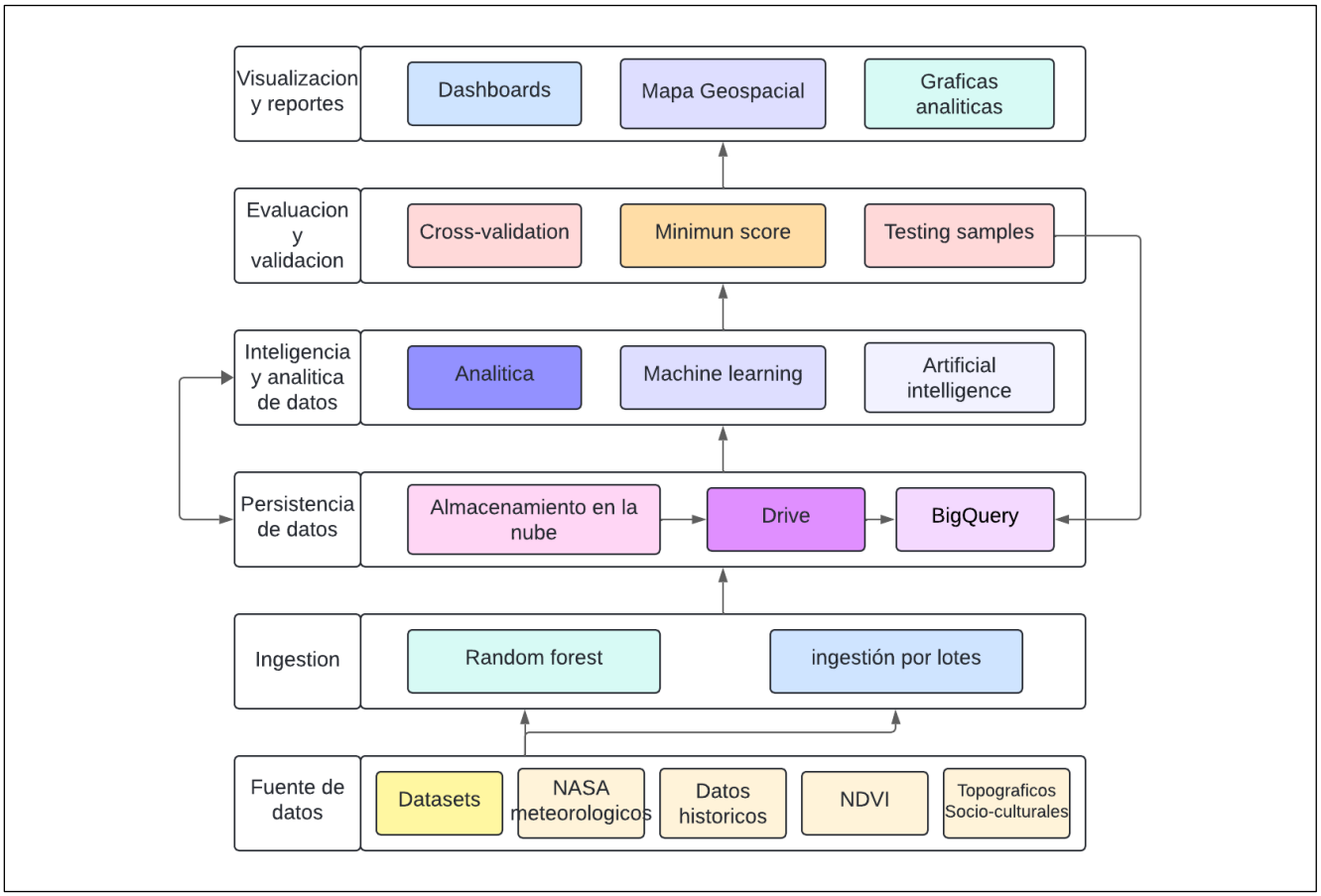
Nuestra arquitectura lógica articula un ecosistema de capas interconectadas que orquestan desde la recopilación intuitiva de datos a través de la interfaz de usuario hasta la presentación inteligible de predicciones de incendios forestales. La integración y procesamiento de datos diversificados alimenta el núcleo de modelado predictivo, donde algoritmos de aprendizaje automático refinan continuamente las estimaciones de riesgo. Los datos son gestionados de manera segura y eficiente, asegurando la integridad y disponibilidad para análisis subsiguientes. Visualizaciones dinámicas y accesibles cierran el ciclo, transformando la información procesada en herramientas prácticas para la toma de decisiones estratégicas en la prevención de incendios forestales.

(Ver grafica 3)

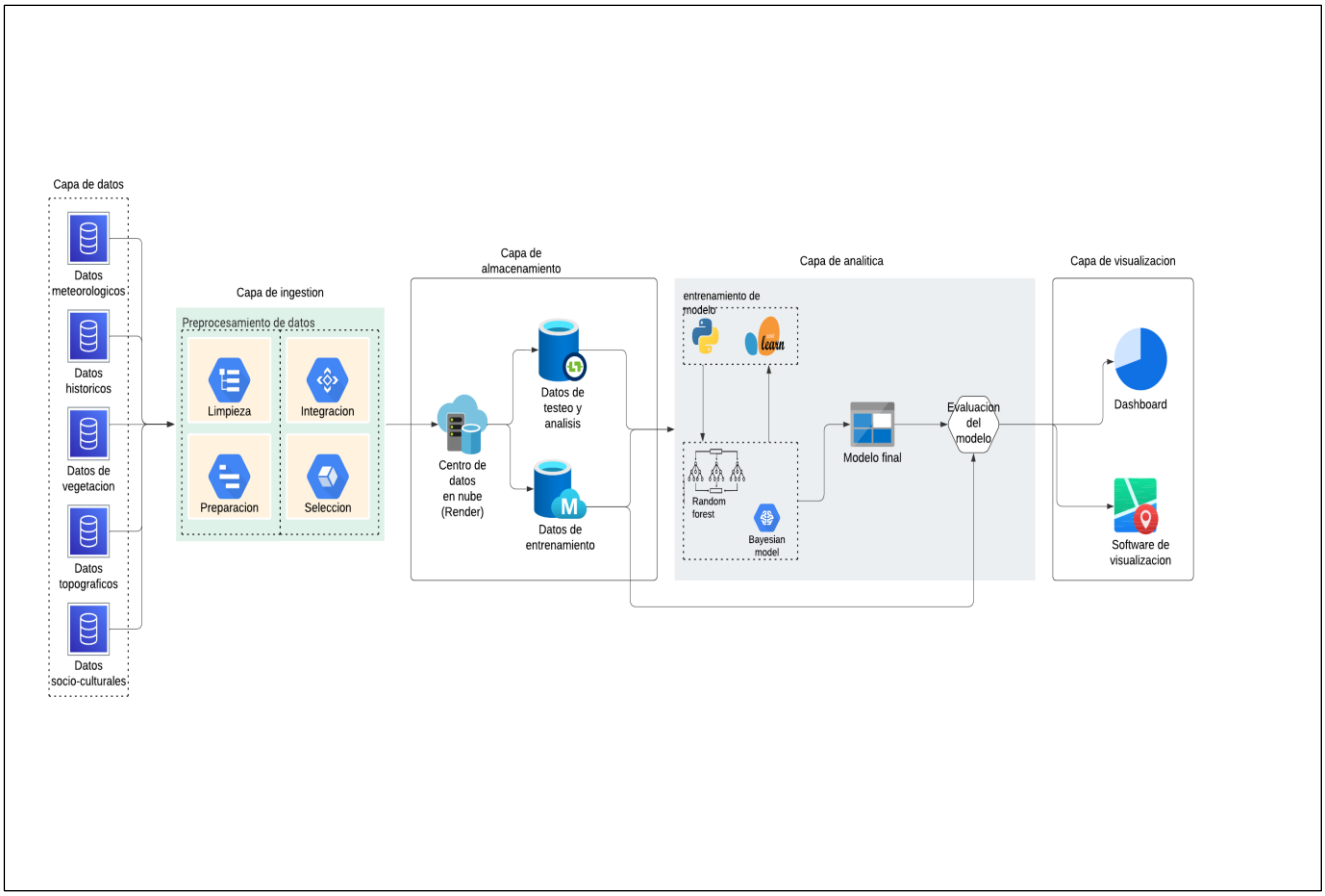
XI. Arquitectura física de la solución

La arquitectura física de esta solución está estructurada con una serie de capas interconectadas, lo que permite un procesamiento y análisis eficientes de los datos. En su núcleo se encuentra la capa de datos, que sirve de repositorio fuente para diversas fuentes de datos relacionados con los incendios forestales. Estos datos se introducen sin problemas en el sistema a través de la capa de ingestión, donde los procesos de integración armonizan flujos de datos dispares. Posteriormente, la capa de almacenamiento, que utiliza el almacén de datos de BigQuery, actúa como un amortiguador para acomodar las fluctuaciones en el volumen y la velocidad de los datos, garantizando un flujo de datos fluido dentro del sistema. Dentro de la capa analítica, se desarrollan y evalúan modelos para estimar la ocurrencia de incendios forestales y su comportamiento, aprovechando el rico conjunto de datos almacenados en BigQuery, aquí evaluaremos distintos tipos de modelos y adquiriremos información valiosa como las variables más significativas y así ajustar el modelo para aumentar su precisión. Por último, la capa de visualización sigue siendo fundamental, ya que proporciona a los usuarios finales los datos y gráficas, esto con el fin de explorar e interpretar los resultados de los análisis, contribuyendo a la toma de decisiones informadas sobre la mitigación y gestión del riesgo de incendios.

(Ver grafica 4)



Graf. 3 Diagrama de arquitectura de la solución (Ver)



Graf. 4 Diagrama de arquitectura de la solución (Ver)

XII. PROTOTIPO DE LA SOLUCIÓN

A. Descripción de algunas variables

Variable	Descripción	Posibles valores
latitude	Especifica la posición norte-sur de un punto en la superficie de la Tierra.	Numérica
longitude	Especifica la posición este-oeste de un punto en la superficie de la Tierra.	Numérica
population_density	Cantidad de personas que viven en un 1km^2.	Numérica
land_cover_type	Clasifica la superficie terrestre según su uso o vegetación	Terra Firma, Wetland, Cropland, Built-up, Not registered, Open surface water, Ocean, Snow/ice
wind_speed	Medida de la rapidez del viento en una determinada área	Numérica
vapor_pressure_deficit	Diferencia entre la cantidad de vapor de agua presente en el aire y la cantidad que el aire puede retener cuando está saturado	Numérica
vapor_pressure	Medida de la presión ejercida por el vapor de agua en la atmósfera.	Numérica
soil_moisture	Cantidad de agua presente en el suelo	Numérica
Reference_evapotranspiration	Cantidad de agua que se evapora y transpira de una superficie de referencia.	Numérica
climate_water_deficit	Cantidad por la cual la demanda de agua excede la oferta durante un período	Numérica
brightness_temperature	Medida de la radiación emitida por la superficie terrestre en diferentes longitudes de onda	Numérica
scan_fire_size	Extensión del fuego detectada durante un escaneo de satélite	Numérica
confidence	Probabilidad de que un punto detectado sea realmente un incendio	l, n, h
fire_radiative_power	Cantidad de energía radiada por un incendio	Numérica
fire_type	Una clasificación del tipo de incendio	presumed vegetation fire, other static land source, offshore, active volcano
ndvi	Índice que mide la salud y la densidad de la vegetación usando imágenes de satélite	Numérica
seasons	Divisiones del año basadas en patrones climáticos	dry season, rainy season

B. Descripción código y técnicas de análisis

Como se mencionó anteriormente, este proyecto tiene un enfoque multidisciplinario, por lo que se mezclaron cinco conjuntos de datos para crear el conjunto de datos final, que sería el objeto de estudio para desarrollar el modelo objetivo.

1. Preparación preliminar

Debido a que algunos datasets presentaban sub-datasets, se creó un script que descarga y une de forma automatizada los conjuntos de datos. El script descarga el conjunto de datos históricos de incendios forestales en Colombia. Posteriormente, descarga el conjunto de datos del NDVI de Colombia, que no proporciona las coordenadas geográficas sino el código por departamento. Por lo tanto, fue necesario utilizar la librería geopy para hallar las coordenadas geográficas de cada dato. También se descarga el conjunto de datos de las normales climatológicas utilizando un script que proporciona el dataset de TERRACLIMATE, empleando las coordenadas geográficas del dataset de incendios forestales.

Además, se descarga el conjunto de datos de cobertura terrestre, el cual presentaba algunos datos nulos, y se aplicaron métodos de interpolación y extrapolación para rellenarlos. Finalmente, se descarga el conjunto de datos de densidad de población, que presentaba las mismas condiciones que el conjunto de datos anterior mencionado.

Además, se realizaron las siguientes tareas: Se cambiaron los nombres de algunas columnas para hacerlas más intuitivas. Debido a la presencia de muchas variables categóricas ordinales, se utilizó la técnica de label encoding. Se realizó una matriz de correlación para revisar si había variables altamente correlacionadas. Si existían, se mantenía solo una de las variables para disminuir el número de variables, estableciendo un umbral de correlación de 0.85. Basándonos

en la literatura, se eligieron dos técnicas para evaluar cuál tenía un mejor desempeño.

2. Creación del modelo

Para la creación del modelo se utilizaron dos técnicas, ambas utilizando la biblioteca de Python scikit-learn que implementa métodos de aprendizaje supervisado para la clasificación: sklearn.ensemble.RandomForestClassifier y sklearn.naive_bayes.GaussianNB.

3. Algoritmo de RandomForestClassifier

El RandomForestClassifier es un algoritmo de clasificación que utiliza múltiples árboles de decisión independientes. Emplea bagging (Bootstrap Aggregating) y selección aleatoria de características para construir árboles robustos, reduciendo la correlación entre ellos. Cada árbol se entrena con subconjuntos de datos creados mediante muestreo con reemplazo.

Los parámetros clave para ejecutar el RandomForestClassifier incluyen:

- n_estimators: Número de árboles en el bosque. Para este proyecto, se usaron 120 árboles.
- max_depth: Profundidad máxima de cada árbol. Se estableció en 10 para limitar el sobreajuste.
- random_state: Semilla aleatoria para garantizar la reproducibilidad, establecida en 42.
- class_weight: Balanceo de clases para manejar datos desequilibrados, configurado como "balanced".

Este modelo es especialmente bueno para conjuntos de datos grandes y complejos con muchas características, ya que maneja bien la multicolinealidad y reduce el riesgo de sobreajuste. La salida del RandomForestClassifier es la clase predicha para cada observación, junto con las probabilidades asociadas para cada clase.

4. Algoritmo de GaussianNB

El GaussianNB es una implementación del algoritmo Naive Bayes para datos continuos, que asume que las características siguen una distribución gaussiana (normal). Es un clasificador probabilístico que calcula la probabilidad posterior de cada clase dada una entrada, utilizando el teorema de Bayes y asumiendo la independencia condicional entre las características.

El GaussianNB se comporta mejor con conjuntos de datos donde las características siguen una distribución normal y hay independencia entre ellas. Es rápido y eficiente para conjuntos de datos grandes y de alta dimensionalidad.

Para nuestro conjunto de datos, GaussianNB es adecuado porque muchas de las variables meteorológicas y topográficas pueden aproximarse a distribuciones normales. Además, su simplicidad y eficiencia lo hacen útil para un procesamiento rápido.

La salida de GaussianNB es la clase predicha para cada observación, junto con las probabilidades asociadas para cada clase. Esto permite no solo obtener la predicción más probable, sino también medir la confianza en cada predicción.

C. Descripción código y técnicas de análisis

```
def create_models(df, target):
    X = df.drop(columns=[target]).values
    y = df[target].values

    random_forest_model = RandomForestClassifier(n_estimators=120,
max_depth=10, class_weight="balanced", random_state=42)
    random_forest_info = get_avg_training(X, y, random_forest_model)

    bayesian_model = GaussianNB()
    bayesian_info = get_avg_training(X, y, bayesian_model)

    return random_forest_model, random_forest_info, bayesian_model,
bayesian_info
```

La anterior función crea los 2 modelos con sus respectivos parámetros para hacer el entrenamiento. Finalmente, se retornan los modelos y su respectiva información.

```
def get_info_model(y_test, y_pred, class_names, average='micro'):
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, average=average)
    recall = recall_score(y_test, y_pred, average=average)
    f1 = f1_score(y_test, y_pred, average=average)
    cm = confusion_matrix(y_test, y_pred, labels=class_names)

    return accuracy, precision, recall, f1, cm
```

La anterior función halla las medidas para poder evaluar que tan correctamente se realizó el modelo

```
def get_avg_training(X, y, model):
    avg_scores = []
    cm_total = np.zeros((3, 3))
    class_names = ['l', 'n', 'h']
    kf = KFold(n_splits=5, shuffle=True, random_state=42)

    for train_index, test_index in kf.split(X):
        X_train, X_test = X[train_index], X[test_index]
        y_train, y_test = y[train_index], y[test_index]

        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
```

```
        accuracy, precision, recall, f1, cm = get_info_model(y_test,
y_pred, class_names)
        avg_scores.append([accuracy, precision, recall, f1])
        cm_total += cm



    avg_scores = np.array(avg_scores)
    accuracy_avg = np.mean(avg_scores[:, 0])
    precision_avg = np.mean(avg_scores[:, 1])
    recall_avg = np.mean(avg_scores[:, 2])
    f1_avg = np.mean(avg_scores[:, 3])
    cm_total_df = pd.DataFrame(cm_total, index=class_names,
columns=class_names)

    return accuracy_avg, precision_avg, recall_avg, f1_avg,
cm_total_df
```

En esta sección de código se usa validación cruzada con Kfold para entrenar el modelo 5 veces. Para cada iteración se hallan el accuracy, precision, recall y f1 score, para así hallar un promedio de esas medidas. Además, se halla la matriz de confusión.

D. Metricas

Modelo	Accuracy	Precision	Recall
Random Forest	0.7525005076318212	0.6255724412219686	0.7939872565110171
GaussianNB	0.6458054910823028	0.4342015699602187 5	0.5314414491598796

Revisando las métricas obtenidas, podemos concluir que:

- El modelo tiene un buen rendimiento general.
- La accuracy refleja bien el F1 score, sugiriendo que la proporción de clases no está afectando desproporcionadamente el rendimiento del modelo.
- El modelo está funcionando bien tanto en términos de identificar correctamente las instancias positivas como en evitar falsos positivos.

E. Ejemplos

```
# Medellin
ff_value = {
    'latitude': 6.25184,
    'longitude': -75.56359,
    'population_density': 19134.373047,
    'land_cover_type': 'Terra Firma',
    'land_cover_subtype': 'Tree cover',
    'vegetation_percent': '35% short vegetation cover',
    'wind_speed': 1.3,
    'vapor_pressure_deficit': 0.82,
    'vapor_pressure': 1.913,
    'minimum_temperature': 289.45,
    'snow_water_equivalent': 0.0,
    'surface_shortwave_radiation': 207.8,
    'soil_moisture': 36,
    'runoff': 108.4,
    'precipitation_accumulation': 221.7,
    'Reference_evapotranspiration': 113.3,
    'climate_water_deficit': 0.0,
    'palmer_drought_severity_index': -3.4,
    'brightness_temperature': 321.6,
    'scan_fire_size': 1.3,
    'fire_radiative_power': 9.0,
    'daynight': 'D',
    'fire_type': 'presumed vegetation fire',
    'n_pixels_ndvi': 15.0,
    'ndvi': 0.7695,
    'ndvi_anomaly_percent': 100.7927,
    'year': 2023,
    'seasons': 'dry season',
    'month': 4,
    'day': 2
}
```

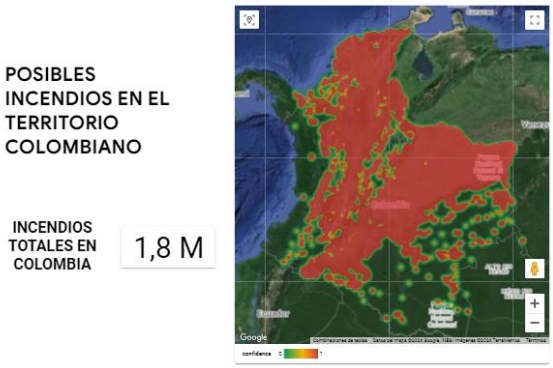
```
mp_values = np.array([value for _, value in ff_value.items()]).res-
hape(1, -1)
value_predicted = random_forest_model.predict(mp_values)[0]
```

```
if value_predicted == "l":
    print("0 - 30% de probabilidad de que ocurra un incendio")
elif value_predicted == "n":
    print("30 - 80% de probabilidad de que ocurra un incendio")
elif value_predicted == "h":
    print("80 - 100% de probabilidad de que ocurra un incendio")
```

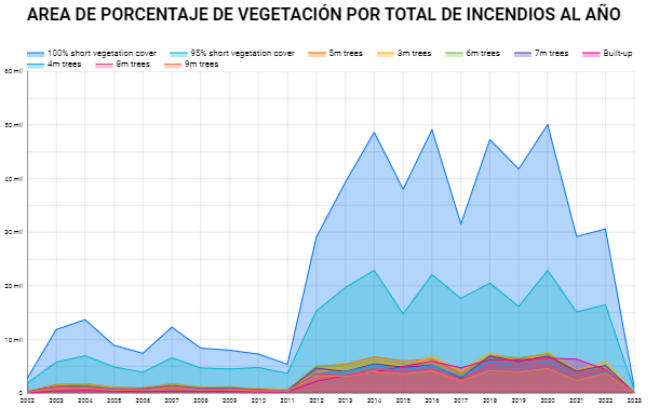
30 - 80% de probabilidad de que ocurra un incendio

En este ejemplo se usa el modelo de RandomForestClassifier para estimar la probabilidad de que ocurra un incendio en Medellín

F. Resultados



El uso de mapas de calor y puntos de datos ayuda a identificar las áreas más afectadas y facilita la planificación de medidas preventivas y de respuesta. La cifra de 1.8 millones sugiere el total acumulado de incidentes registrados.



La gráfica ilustra cómo diferentes tipos de vegetación y cobertura del suelo se ven afectados por incendios a lo largo de los años. Las áreas apiladas permiten visualizar la contribución relativa de cada tipo de vegetación a la superficie total afectada por incendios. Notablemente, hay un aumento significativo en la superficie afectada a partir de 2014, lo cual puede estar relacionado con cambios climáticos, políticas de manejo de tierras, o eventos naturales específicos.

La longitud de cada barra indica el número de incendios registrados para cada año, y los colores muestran la distribución entre los diferentes tipos de cobertura terrestre. El predominio del color azul (Terra Firma) sugiere que la mayoría de los incendios ocurren en tierras firmes. El aumento en la longitud de las barras a partir de 2014 indica un incremento en la incidencia de incendios, posiblemente debido a factores ambientales o humanos.

XIII. CONCLUSIONES

El proyecto de estimación de incendios forestales ha alcanzado importantes hitos en la implementación y evaluación de modelos de clasificación, utilizando técnicas avanzadas de análisis de datos y aprendizaje automático. La investigación se centró en desarrollar modelos que puedan estimar la probabilidad de incendios forestales en Colombia, integrando datos históricos, meteorológicos, topográficos y socioeconómicos. Los modelos desarrollados, Random Forest y Gaussian Naive Bayes, se entrenaron y evaluaron utilizando un enfoque de validación cruzada, proporcionando resultados significativos.

Desde el punto de vista teórico, este proyecto ha demostrado la relevancia y efectividad del uso de técnicas de aprendizaje automático para abordar problemas complejos como la predicción de incendios forestales. La elección de los modelos Random Forest y Gaussian Naive Bayes se basó en sus respectivas capacidades para manejar diferentes tipos de datos y capturar relaciones complejas en los mismos. Random Forest es particularmente útil para nuestro proyecto por su capacidad para manejar grandes conjuntos de datos con muchas características y es robusto frente a datos ruidosos y desbalanceados. El modelo Gaussian Naive Bayes, por otro lado, asume que las características siguen una distribución normal y es adecuado para datos continuos, proporcionando una interpretación probabilística útil en la toma de decisiones basada en riesgos. Su eficiencia y rendimiento computacional, así como su simplicidad, lo hacen rápido de entrenar, incluso en grandes volúmenes de datos como los que usamos para el modelo.

La metodología CRISP-DM proporcionó una estructura robusta para el desarrollo del modelo, asegurando una comprensión profunda del problema, una preparación adecuada de los datos y una evaluación rigurosa de los modelos. La validación cruzada K-Fold se utilizó para asegurar que los modelos no estuvieran sobreajustados y para proporcionar estimaciones confiables de su rendimiento en datos no vistos. Este enfoque divide los datos en múltiples subconjuntos (folds), entrena el modelo en algunos de ellos y lo valida en los restantes, lo que mejora la generalización del modelo y reduce la posibilidad de sobreajuste. Además, la importancia de una arquitectura lógica y física bien definida fue fundamental para la organización y gestión de los datos. La arquitectura lógica proporcionó una estructura clara para el procesamiento y almacenamiento de datos, mientras que la arquitectura física garantizó que los datos estuvieran accesibles y bien organizados en el sistema, permitiendo un flujo de trabajo eficiente y reproducible.

Metodológicamente, se integraron datos de múltiples fuentes y disciplinas. Los datos meteorológicos se obtuvieron del conjunto de datos TerraClimate, que proporciona información climática mensual de alta resolución temporal y espacial. Los datos topográficos y de cobertura del suelo provinieron del conjunto de datos globales de cambio de uso y cobertura de la tierra. Finalmente, los datos socioeconómicos, específicamente la densidad poblacional, fueron obtenidos del proyecto WorldPop, que ofrece estimaciones de densidad poblacional ajustadas a los totales oficiales de la ONU a una resolución de 1 km. Esta integración permitió un análisis más completo y detallado, mejorando la precisión y confiabilidad de los modelos predictivos.

Los resultados obtenidos en las métricas de evaluación para ambos modelos muestran que, para el modelo de Random Forest, la exactitud, precisión y recall son aproximadamente del 75%. En contraste, para el modelo Gaussian Naive Bayes, estos parámetros alcanzan el 64%. Estos resultados indican que el modelo Random Forest superó al modelo Gaussian Naive Bayes en todas las métricas evaluadas, sugiriendo que el modelo Random Forest es más efectivo para estimar la ocurrencia de incendios forestales en las condiciones estudiadas.

Una posible explicación de por qué el modelo Gaussian Naive Bayes tuvo un rendimiento inferior al modelo de Random Forest podría estar relacionada con las suposiciones de distribución normal de las características en el modelo Gaussian, que pueden no ajustarse bien a todas las variables en el conjunto de datos. Además, Random Forest es más robusto frente a la multicolinealidad y puede manejar mejor las relaciones no lineales entre las características y la variable objetivo.

Las variables más influyentes en la predicción de incendios forestales, según la importancia de características del modelo de Random Forest, fueron la temperatura de brillo, la potencia radiativa del fuego y el tamaño del escaneo del fuego. La temperatura de brillo (brightness_temperature), que tuvo la mayor influencia en el modelo, es un indicador directo de la intensidad del calor en una región específica, lo que permite al modelo identificar áreas con alta probabilidad de incendios. La potencia radiativa del fuego (fire_radiative_power) mide la energía emitida por el fuego y es fundamental para evaluar la magnitud y severidad de los incendios, permitiendo al modelo diferenciar entre pequeños brotes y grandes incendios. El tamaño del escaneo del fuego (scan_fire_size) refleja el área afectada por el incendio y es esencial para determinar la extensión del daño. Comprender la importancia de estas variables puede mejorar significativamente futuras investigaciones y políticas de prevención de incendios forestales. Al centrarse en estos factores clave, los investigadores y responsables de políticas pueden desarrollar estrategias más efectivas para monitorear y mitigar los riesgos de incendios forestales, optimizando los recursos y enfocándose en las áreas más vulnerables.

En términos de aplicación, los modelos desarrollados pueden ser herramientas valiosas para las autoridades y gestores de emergencias en la planificación y asignación de recursos para la prevención y mitigación de incendios forestales. Las visualizaciones interactivas y las herramientas desarrolladas permiten una interpretación clara y oportuna de los datos y predicciones, facilitando la toma de decisiones informadas.

REFERENCES

[1] Stojanova, D., Panov, P., Kobler, A., Džeroski, S., & Taškova, K. (2006, October). Learning to predict forest fires with different data mining techniques. In Conference on data mining and data warehouses (SiKDD 2006), Ljubljana, Slovenia (pp. 255-258).

[2] Sakr, G. E., Elhajj, I. H., Mitri, G., & Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction. In 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (pp. 1311-1316). Montreal, QC, Canada. doi: 10.1109/AIM.2010.5695809

[3] Chen, X. C., et al. (2012). A new data mining framework for forest fire mapping. In 2012 Conference on Intelligent Data Understanding (pp. 104-111). Boulder, CO, USA. doi: 10.1109/CIDU.2012.6382190

[4] García, M., & González, P. H. (2012). Determinación de la tendencia espacial de los puntos de calor como estrategia para monitorear los incendios forestales en Durango, México. Bosque (Valdivia), 33, 63-68. http://doi.org/10.4067/S0717-92002012000100007

[5] Morales, G. A., Morales, R. S., Valencia, C. F., & Akhavan-Tabatabaei, R. (2014). A forest fire propagation simulator for Bogotá. IEEE Press, 1505–1515.

[6] Castelli, M., Vanneschi, L., & Popovič, A. (2015). Predicting Burned Areas of Forest Fires: an Artificial Intelligence Approach. Fire Ecology, 11, 106-118. http://doi.org/10.4996/fireecology.1101106

[7] Chew, Y. J., Ooi, S. Y., & Pang, Y. H. (2020). Experimental Exploratory of Temporal Sampling Forest in Forest Fire Regression and Classification. In 2020 8th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). Yogyakarta, Indonesia. doi: 10.1109/ICoICT49345.2020.9166231

[8] Preeti, T., Kanakaraddi, S., Beelagi, A., Malagi, S., & Sudi, A. (2021). Forest Fire Prediction Using Machine Learning Techniques. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-6). Hubli, India. doi: 10.1109/CONIT51480.2021.9498448

[9] Hidayanto, N., Saputro, A. H., & Nuryanto, D. E. (2021). Peatland Data Fusion for Forest Fire Susceptibility Prediction Using Machine Learning. In 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 544-549). Yogyakarta, Indonesia. doi: 10.1109/ISRITI54043.2021.9702762

[10] Natekar, S., Patil, S., Nair, A., & Roychowdhury, S. (2021). Forest Fire Prediction using LSTM. In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1-5). Belagavi, India. doi: 10.1109/INCET51464.2021.9456113

[11] Li, Y., Zhang, S., & Fu, G. (2022). Forest fire modeling and analysis based on K-means clustering algorithm and time series forecasting. In 2022 2nd International Conference on Bioinformatics and Intelligent Computing (BIC 2022) (pp. 310–316). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3523286.3524560

[12] Ospino Rivera, J. L. (2022). Régimen de incendios y densidad de la vegetación (NDVI) en la subregión de los Montes de María con énfasis en el bosque seco tropical entre 2012-2022. Retrieved from http://hdl.handle.net/10584/11468

[13] Pal, R. (2022). Mazelink: Detecting and Predicting Forest Fires. In Proceedings of the 12th Indian Conference on Human-Computer Interaction (IndiaHCI '21) (pp. 80–83). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3506469.3506481

[14] Reddy, P. R., & Kalyanasundaram, P. (2022). Novel Detection of Forest Fire using Temperature and Carbon Dioxide Sensors with Improved Accuracy in Comparison between two Different Zones. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 524-527). London, United Kingdom. doi: 10.1109/ICIEM54221.2022.9853107

[15] Rosa, S. L., Abdul Kadir, E., Syukur, A., Irie, H., Wandri, R., & Evizal, M. F. (2022). Fire Hotspots Mapping and Forecasting in Indonesia Using Deep Learning Algorithm. In 2022 3rd International Conference on Electrical Engineering and Informatics (ICon EEI) (pp. 190-194). Pekanbaru, Indonesia. doi: 10.1109/IConEEI55709.2022.9972281

[16] Santos, B. Z., Araujo Soriano, B. M., Narciso, M. G., Silva, D. F., & Cerri, R. (2023). A New Time Series Framework for Forest Fire Risk Forecasting and Classification. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). Gold Coast, Australia. doi: 10.1109/IJCNN54540.2023.10191502

[17] Meena, U., Munjal, G., Sachdeva, S., Garg, P., Dagar, D., & Gangal, A. (2023). RCNN Architecture for Forest Fire Detection. In 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 699-704). Noida, India. doi: 10.1109/Confluence56041.2023.10048878

[18] Singh, J., Aarthi, M. S., & Idikkula, A. S. (2023). Convolutional Neural Networks for Early Detection of Forest Fires. In 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 777-780). Pudukkottai, India. doi: 10.1109/ICACRS58579.2023.10404568

[19] Priya, R. S., & Vani, K. (2023). Climate Change Forecast for Forest Fire Risk Prediction using Deep Learning. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1065-1070). Coimbatore, India. doi: 10.1109/ICACCS57279.2023.10112983

[20] Carta, F., Zidda, C., Putzu, M., Loru, D., Anedda, M., & Giusto, D. (2023). Advancements in forest fire prevention: A comprehensive survey. Sensors, 23(14), 66

SÍNTESIS DE LA TABLA RSL

Articulo	Autores	Año	Palabras clave	Fuente
Artificial intelligence for forest fire prediction	Sakr, G. E., Elhajj, I. H., Mitri, G., Wejinya, U. C.	2010	Support vector machines, Data mining, Prediction algorithms, Equations, Fires, Weather forecasting, Machine Learning, SVM, Forest Fire Prediction	IEEE Xplore
Predicting Burned Areas of Forest Fires: an Artificial Intelligence Approach	Castelli, M., Vanneschi, L., Popovič, A.	2015	climatic data, forest fires, genetic programming, Portugal, Semantics-	SpringerOpen
Learning to predict forest fires with different data mining techniques	Stojanova, D., Panov, P., Kobler, A., Džeroski, S., Taškova, K.	2006	Forest Fires, Data Mining Techniques, Predictive Models, GIS, ALADIN Weather Prediction Model, MODIS Satellite Data, Slovenia, Machine Learning, Ensemble Methods, Forest Structure	IEEE Xplore
Forest Fire Prediction using LSTM	Natekar, S., Patil, S., Nair, A., Roychowdhury, S.	2021	Recurrent neural networks, Biological system modeling, Time series analysis, Ecosystems, Fires, Forestry, Climate change, forest fire, LSTM, time series, RNN	IEEE Xplore
Forest fire modeling and analysis based on K-means clustering algorithm and time series forecasting	Li, Y., Zhang, S., Fu, G.	2022	Firest fire, clustering algorithm	ACM DL
Forest Fire Detection Method Based on Deep Learning	Wang, W., Huang, Q., Liu, H., Jia, Y., Chen, Q.	2022	Deep learning, Training, Sensitivity, Biological system modeling, Fires, Forestry, Real-time systems, deep learning, flame detection, forest fire, object detection	IEEE Xplore
Experimental Exploratory of Temporal Sampling Forest in Forest Fire Regression and Classification	Chew, Y. J., Ooi, S. Y., Pang, Y. H.	2020	Hidden Markov models, temporal sampling forest, random forest, temporal data classification, temporal regression analysis, forest fire	IEEE Xplore
Forest Wild Fire Detection using Deep Learning Approach	K, A. S., D, M. D., A. A., Mary, G., S, B. P., M, P. M., A. B., C.	2023	Wireless communication, Wireless sensor networks, Machine learning algorithms, Disasters, Forestry, Vegetation, Prediction algorithms, Forest Fire, WSN, Correlation, Sensors	IEEE Xplore
Climate Change Forecast for Forest Fire Risk Prediction using Deep Learning	Priya, R. S., Vani, K.	2023	Deep learning, Temperature distribution, Rain, Forestry, Global warming, Predictive models, Market research, Climate change, Machine learning, climate change, forest fire, greenhouse gas, temperature forecast	IEEE Xplore
Peatland Data Fusion for Forest Fire Susceptibility Prediction Using Machine Learning	Hidayanto, N., Saputro, A. H., Nuryanto, D. E.	2021	Radio frequency,Training,Seminars,Roads,Static VAr compensators,Vegetation mapping,Support vector machine classification,Peatlands,Machine learning,Forest fire,Susceptibility,Prediction	IEEE Xplore