ISSN 1870-4069

Propuesta de sistema computacional para incendios forestales y clasificación de días con riesgo a incendio usando temperatura y humedad en una reserva natural, caso de estudio

Carlos Sebastián Maya-Rojina, Luis Pastor Sánchez-Fernández, Mario Eduardo Rivero-Ángeles

Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, México

{cmayar2022,lsanchez,erivero}@cic.ipn.mx

Resumen. Se propone un sistema computacional para prevención, evaluación de riesgo y monitoreo de incendios forestales integrando diferentes herramientas tales como simuladores, redes de sensores e índices de vulnerabilidad, además, se profundiza en el estudio de caso sobre clasificación de días en los que hubo vulnerabilidad a incendios en una reserva ecológica utilizando datos meteorológicos de temperatura y humedad y los incendios registrados los últimos 22 años (2001-2023), aplicando y comparando entre sí diferentes algoritmos de clasificación, métodos de selección de instancias y de atributos. Se propone el agregado de índices de vulnerabilidad como entrada al clasificador, la implementación de métodos de balanceo y de selección de instancias, ampliar los experimentos usando los clasificadores con buenos resultados y descartar los que tuvieron un bajo desempeño.

Palabras clave: Incendios forestales, Estudio de caso de clasificación, vulnerabilidad a incendios.

Proposal for a Computer System for Forest Fires and Classification of Days with Fire Risk Using Temperature and Humidity in a Nature Reserve, Case Study

Abstract. A computational system for prevention, risk assessment and monitoring of forest fires is proposed, integrating different tools such as simulators, sensor networks and vulnerability indexes. In addition, a case study on the classification of days in which there was vulnerability to fires in an ecological reserve using meteorological data of temperature and humidity and the fires recorded in the last 22 years (2001-2023) is studied in depth, applying and comparing different classification algorithms, instance and attribute selection methods. It is proposed to add vulnerability indices as input to the classifier, implement balancing and instance selection methods, extend the experiments

using the classifiers with good results and to discard the ones that performed poorly.

Keywords: Forest fires, classification case study, vulnerability to fires.

1. Introducción

Los incendios forestales suelen tener consecuencias graves para el ecosistema, tales como pérdida de flora, fauna y de patrimonio biocultural, erosión del suelo, liberación de niveles altos de CO2 a la atmósfera, deforestación, fragmentación de ecosistemas entre otros [5, 6]. Según la CONAFOR, uno de los factores principales en la propagación de incendios es la limitada capacidad de respuesta para su tratamiento [7]. Por lo tanto, es importante el desarrollo de herramientas tecnológicas para su detección temprana, monitoreo y prevención.

En este trabajo se plantea en lo general un sistema computacional para atender el problema de los incendios forestales y se detallan los resultados de un primer acercamiento a la clasificación de los días propensos a incendios, tomando como caso de estudio una reserva natural y utilizando datos meteorológicos de temperatura y humedad, así como los días en los que sucedieron incendios de los últimos 22 años, haciendo uso de diferentes preprocesamientos y técnicas de clasificación, y se hace una comparación entre ellas respecto a su Sensibilidad, Precisión, Medida F1 y Especificidad. El sistema computacional propuesto así como los resultados pueden contribuir al desarrollo de un sistema de prevención, alerta temprana y monitoreo implementado en la reserva.

2. Descripción general del sistema computacional propuesto

Se plantea el desarrollo e integración de diferentes herramientas computacionales para generar un ecosistema alrededor de los incendios forestales: un simulador de incendios, la simulación de una red de sensores interaccionando con el simulador, la clasificación de días vulnerables y la creación de un índice de vulnerabilidad, el componente a explayar en el artículo es únicamente el primer acercamiento a la clasificación, el resto de herramientas se describen en su primera etapa de desarrollo y su objetivo final, esto para enmarcar el trabajo de los clasificadores dentro de un conjunto.

2.1. Simulador de incendios

Utilizando autómatas celulares, teniendo como características principales de cada celda: la cantidad de combustible, vulnerabilidad actual a un fuego, probabilidad de dispersión de un fuego a celdas vecinas (afectada por la inclinación y por la cantidad de viento) y siguiendo reglas de dispersión conocidas de los incendios forestales, capaz de reproducir el comportamiento de un incendio.

Actualmente su enfoque de desarrollo es a proporcionar un entorno de simulación para la realización de pruebas de la dinámica de una red de sensores con los incendios

forestales. Sin embargo, en el futuro los valores deben ser "mapeados" a la realidad o comparados con simuladores espaciales existentes, para su validación (tarea compleja por la escases de datos de la forma a través del tiempo de incendios forestales).

El desarrollo de esta herramienta podría funcionar, además del entorno para la simulación de una red de sensores, como asistencia en el combate a incendios, tanto de manera preventiva (identificar zonas clave para la instalación de tanques de agua, mayor frecuencia en la vigilancia) como durante el incendio (indicando al cuerpo de bomberos un plan de acción óptimo para la reducción de daños). Se encontraría limitada por la cantidad de datos disponibles para su validación, y ya validada, por los datos disponibles del área a implementarse, así como por la naturaleza caótica de los incendios.

2.2. Simulador de red de sensores

Su objetivo inicial es evaluar diferentes esquemas de comunicación (que determinan las probabilidades de envío de mensaje y de despertar, como las funciones de probabilidad y el algoritmo que sigue cada nodo) implementados en la red, en relación a su capacidad de monitorear un incendio, así como de disminuir su consumo energético (dado que los nodos estarán instalados en áreas sin un acceso constante a fuentes de energía, es necesario que sean óptimos energéticamente hablando), permitiendo generar y corroborar un modelo matemático de cada esquema.

Esta sección del sistema computacional también permite evaluar diferentes configuraciones de la red, sus posicionamientos adecuados para un mejor monitoreo y capacidad de detección, así como los primeros pasos para la simulación de técnicas de enrutamiento de información (en el que un nodo no tiene alcance al nodo objetivo por la distancia y necesita recurrir a nodos intermedios para hacer llegar el mensaje).

2.3. Clasificación de días vulnerables a incendios

Se ejecutaron pruebas con diferentes clasificadores, targets y preprocesamientos. Esta parte del modelo computacional será explayado en las siguientes secciones.

2.4. Índice de vulnerabilidad

Implementando un sistema difuso para obtener un índice de vulnerabilidad hora a hora, utilizando los datos disponibles de temperatura y humedad, su desarrollo permitirá indicar al personal de reserva una primera alerta y debería ser implementado como atributo de apoyo en la clasificación, así como sentar las bases de un sistema difuso más integral que tome en cuenta características propias de la vegetación y factores humanos tales como la cantidad de interacción de una zona de la reserva con personas, el momento del ciclo escolar (que define la cantidad de población estudiantil que accede a C.U.), etc.

3. Estado del arte - clasificadores

Alonso-Fontenla et. al. 2003 [3] Presentan un sistema integrando redes neuronales para la predicción de riesgo a incendio, un sistema que apoye en el combate a incendios y un sistema de asistencia en la planeación de la recuperación de áreas quemadas. La sección de predicción de riesgo se enfoca a Galicia (comunidad autónoma de España, 36,000km2, utilizando los datos de incendios de 13 años (1988-2001) y las variables de temperatura del mismo día y de días previos, humedad diaria, precipitación y el histórico de incendios, con un total de 13,906 muestras de incendio y 125,156 de no incendio, utilizaron una estrategia de crecimiento de redes neuronales, en la que se agregan capas y neuronas de manera progresiva hasta alcanzar un óptimo, obtuvieron una Exactitud de 78.9%, una Sensibilidad de 90.4% y una especificidad de 67.4%.

Sakr-Elhajj et. al, 2011 [1] Propusieron un sistema de predicción de incendios forestales orientado a zonas con poca infraestructura, utilizando Redes Neuronales Artificiales y Máquinas de Soporte Vectorial utilizando las variables del día anterior al incendio de 8 años (2000-2008): mínima y máxima temperatura del día, promedio de humedad y de velocidad del viento, radiación solar total y el acumulado del nivel de precipitación empezando por el 1ro de Octubre, para predecir la cantidad de incendios que habrían en el Líbano cada día para la temporada de incendios, con clasificadores entrenados para cada mes, alcanzando un accuracy de 94.21%. Sin embargo, no se mencionan otras medidas de evaluación (el accuracy podría ser de 95% si solo hubo un 5% de días con incendios y se clasifican todos como sin incendios), ni la proporción de días con o sin incendio.

Chang-Zhu et. al. 2013 [2] Utilizaron regresión logística para predicción de incendios de 28 años (1980-2009) utilizando puntos de incendios (2,359) en una provincia de China (454,800km2) y separándolos de puntos del mismo día en otra locación (con una proporción aproximada de 1.5 días sin incendio por cada día con incendio en el muestreo), utilizando mapas de variables de densidad de población, elevación, inclinación, distancia a caminos, promedio anual de temperatura y precipitación, promedio diario de velocidad de viento, temperatura y humedad, precipitación, mínima temperatura y humedad, máxima temperatura y tiempo con radiación solar, y el tipo de vegetación, una precisión de 51.6%, sensibilidad de 72.7% y especificidad de 0.78%.

Oulad-Mousannif et. al. 2019 [4] utilizaron técnicas de data mining, redes neuronales y máquinas de soporte vectorial, para clasificar entre regiones con incendio y regiones sin incendio empleando datos de la densidad de vegetación, temperatura de superficie y anomalía térmica, todos provenientes de satélites con una resolución espacial de 1km y diaria de 1 día, en un área de 20,000 km^2 en el centro de Canadá, con incendios de un año (2013-2014), con 386 totales y 418 muestras agregadas de no incendio (que pertenecen a datos de la zona donde hubo un incendio, de otro día) ; alcanzando un 97% de precisión, con 98% de sensibilidad.

En nuestra propuesta nos enfocamos a un área mucho más específica (1.245 km2), implicando una menor cantidad de datos de incendios, utilizando únicamente datos de temperatura y humedad, una expansión de los atributos y selección de los mismos, enfocando el proyecto a la predicción y a medidas preventivas de incendios de áreas relativamente pequeñas, evaluando el sistema de una manera más cercana a la implementación.

4. Materiales y métodos

4.1. Descripción de los datos utilizados para las pruebas de clasificación

Se utilizaron dos fuentes de datos para esta sección del sistema computacional. La primera, de datos meteorológicos que fue obtenida del Programa de Estaciones Meteorológicas del Bachillerato Universitario (PEMBU), de la estación meteorológica del Plantel Sur del Colegio de Ciencias y Humanidades (CCH-Sur), que es de libre acceso. La segunda, del histórico de incendios, proporcionada por la Reserva Ecológica del Pedregal de San Ángel (REPSA) como parte de la colaboración académica realizada para este proyecto.

Los datos meteorológicos consisten en Fecha y hora, y los valores de Temperatura, Humedad Relativa, Rapidez del viento sostenido, Dirección del viento sostenido, Rapidez de racha, Dirección de racha, Presión barométrica, Precipitación, Radiación Solar, Índice UV y Dosis UV censados por la estación cada media hora desde enero del 2001 hasta enero del 2023 siendo un total de 387169 muestras.

Los datos de incendios consisten en Número de Identificación del incendio, Fecha, Periodo, Año, Mes, Hora, Tipo de siniestro, X,Y, Ubicación, Acción, Foto y Superficie en m^2, Superficie en ha, Zona, Clave, Elementos Bomberos, Duración (hrs.), Temperatura ambiental, Humedad ambiental, Temperatura de la superficie del fuego, Factor de inicio, Fuente de la información, Número de Identificación del reporte de incendio, Fecha de levantamiento, Técnica, Plano cartográfico y la Persona que Elaboró, registrados por el personal de la Secretaria Ejecutiva de la REPSA (SEREPSA) desde enero del 2001 hasta marzo del 2023, con un total de 322 incendios reportados.

4.2. Preprocesamiento

De la fuente de datos de incendios se empleó únicamente la información del día del incendio, ya que el resto de columnas no se consideran útiles por el momento para la clasificación, o se encuentran demasiado incompletas para su uso (Superficie del incendio). De la fuente de datos meteorológicos se utilizaron solamente los valores de temperatura y humedad relativa, puesto que son los datos más completos y de sensores económicos (pensando en una futura implementación en hardware de la red de sensores), sus valores fueron normalizados a un intervalo de 0 a 1 y las instancias con valores nulos fueron eliminadas del dataset.

Expansión de atributos De la fuente de información original de los datos meteorológicos se obtuvieron los siguientes valores tanto para temperatura como para humedad:

- Promedio de todos los datos.
- Valor más alto y más bajo.
- Promedio de los 5, 10 y 20 valores más altos y más bajos.

Estos fueron extraídos del día anterior al incendio y de los 5, 10 y 20 días precedentes al mismo. De lo anterior se obtuvieron un total de 72 atributos para la clasificación.

Selección de atributos Se implementaron tres métodos de selección de atributos.

Por rankeo de atributos La selección se realizó tomando los primeros 10 atributos, dándoles puntuaciones a cada uno respecto a diferentes métricas, entre ellas:

- Ganancia de Información y Radio de Ganancia de Información - Calculan la ganancia de información de cada característica respecto a la variable objetivo usando la entropía condicional.
- X^2 - Determina si existe o no una relación significativa entre dos variables.
- ReliefF - Para cada instancia se mide la distancia a la instancia más cercana de la misma clase y a la más cercana perteneciente a otra clase, y actualiza el peso de cada atributo en función de las distancias obtenidas.

PCA Es una técnica estadística multivariable que reduce la dimensionalidad de un dataset, transformando el conjunto de variables correlacionadas entre sí a un conjunto de variables no correlacionadas entre sí llamadas componentes principales, estas variables principales están ordenadas de forma tal que las primeras mantienen la mayoría de la variación presente en el dataset original [10]. De ella se seleccionaron los primeros 10 atributos.

Manual Se eligieron los 11 atributos que se consideraron con mayor injerencia en la vulnerabilidad a incendios en la zona. Los atributos seleccionados fueron:

- Promedio de temperatura del día anterior.
- Promedio de temperatura de los 5 días anteriores.
- Temperatura más alta del día anterior.
- Promedio de 5 temperaturas más altas del día anterior.
- Temperatura más alta de los 5 días anteriores.
- Promedio de 5 temperaturas más altas de los 5 días anteriores.
- Humedad más alta del día anterior.
- Promedio de 5 humedades más altas del día anterior.
- Promedio de 5 humedades más bajas del día anterior.
- Promedio de 5 humedades más bajas de los 5 días anteriores.
- Promedio de 10 humedades más bajas de los 5 días anteriores.

4.3. Algoritmos de selección de instancias

Eliminación de meses con bajo porcentaje de incendios Se eliminaron los meses que contenían menos del 3 y del 5% de incendios de los últimos 22 años.

Target Se utilizaron dos targets diferentes. El primero, en el que los días anteriores a cuando hubo un incendio tienen un valor de 1 y el resto de días tienen un 0; el segundo, en el que se aplicó una ventana en la que el día anterior, el mismo día y un día posterior al incendio tienen un valor de 1 y el resto de días tienen un 0, pensando en que la reserva era más vulnerable a incendios el día mismo del incendio, pero que esta vulnerabilidad existía un día anterior y un día posterior (no se genera ni se disipa de manera inmediata), este segundo target fue utilizado únicamente para cuando se implementó una selección de instancias en los datos.

4.4. Clasificadores supervisados utilizados

Naive Bayes Basado en la regla de Bayes, supone que los atributos son independientes entre sí, simplificando el sistema [9], construye reglas que permiten clases predictivas para nuevos datos, es fácil de implementar e interpretar y puede aplicarse sin complicación a datasets extensos, sin embargo, requiere computar una gran cantidad de probabilidades condicionales.

Redes Neuronales Artificiales Del campo de la Inteligencia Artificial, están inspiradas en el funcionamiento biológico del sistema nervioso, se conectan por ”capas” de varias neuronas y se utiliza un proceso de aprendizaje con base en ejemplos para optimizar la función de clasificación [11], son ampliamente utilizadas para problemas de clasificación e inferencia, guardan información a lo largo de toda la red y trabajan bien con información incompleta, sin embargo requieren procesos paralelos para mejorar los tiempos de entrenamiento y no poseen interpretabilidad de los resultados.

CN2 Inductor de Reglas Induce reglas “Si... entonces...” usando entropía o error laplaciano como su heurística de búsqueda [12], funciona bien para conjuntos de datos con ruido independiente de su extensión, sin embargo, tiene problemas para representar la interacción entre variables y con datasets desbalanceados.

Máquina de Soporte Vectorial Proviene de los modelos lineales, se transforma el espacio de entrada haciendo uso de combinación no lineal y generando un límite de decisión [9], requiere pocos datos para su entrenamiento, sin embargo, son lentas en su entrenamiento y en implementación, dependiendo de la extensión de los datos, pueden requerir grandes cantidades de memoria disponible.

5. Resultados y discusión

5.1. Diseño experimental

Los experimentos se realizaron combinando la selección de instancias, la elección de atributos y el target. Dando un total de 80 experimentos.

Siendo las 3 formas de selección de instancias:

- Sin selección de instancias.
- Seleccionando únicamente los meses con más del 3% de incendios.
- Seleccionando únicamente los meses con más del 5% de incendios.

Las 4 formas de selección de atributos:

- Por rankeo, los 10 primeros.
- PCA, los 10 primeros.
- Manual, los 11 considerados más adecuados.
- Sin selección.

Los 4 clasificadores:

- Naive Bayes.
- Redes Neuronales.
- Inductor de reglas CN2.

Tabla 1. Tabla de los resultados utilizando los datos sin selección de instancias y el target puntual.

Datos sin Selección de Instancias, Target puntual					
Método de selección de atributos	Matriz de confusión/ Medidas de desempeño	Clasificador			
		Naive Bayes	Red Neuronal	CN2	SVM
Top 10	TN	3973	5147	5090	4667
	FP	1222	48	105	528
	FN	31	124	129	90
	TP	105	12	7	46
	Sensibilidad	0.7721	**0.0882**	0.0515	**0.3382**
	Especificidad	0.7648	***0.9908***	0.9798	0.8984
	Precisión	**0.0791**	***0.2000***	0.0625	**0.0801**
	F1	***0.1435***	**0.1224**	0.0565	**0.1296**
PCA	TN	3234	5044	4960	4939
	FP	1871	61	145	166
	FN	13	129	120	122
	TP	123	7	16	14
	Sensibilidad	***0.9044***	0.0515	**0.1176**	0.1029
	Especificidad	0.6335	0.9881	0.9716	**0.9675**
	Precisión	0.0617	0.1029	0.0994	0.0778
	F1	0.1155	0.0686	**0.1077**	0.0886
Manual	TN	5134	6203	6160	5656
	FP	1141	72	115	619
	FN	100	177	175	150
	TP	87	10	12	37
	Sensibilidad	0.4652	0.0535	0.0642	0.1979
	Especificidad	**0.8182**	0.9885	**0.9817**	0.9014
	Precisión	0.0708	0.1220	**0.0945**	0.0564
	F1	0.1230	0.0743	0.0764	0.0878
Sin Selección	TN	3234	5044	4960	4939
	FP	1871	61	145	166
	FN	13	129	120	122
	TP	123	7	16	14
	Sensibilidad	***0.9044***	0.0515	**0.1176**	0.1029
	Especificidad	0.6335	0.9881	0.9716	**0.9675**
	Precisión	0.0617	0.1029	0.0994	0.0778
	F1	0.1155	0.0686	**0.1077**	0.0886

- Máquina de Soporte Vectorial.

Los 2 tipos de target

- Día puntual.

Tabla 2. Tabla de los resultados utilizando sin datos de meses con > 3% de incendios y el target puntual.

Datos Selección de Meses > 3%, Target puntual					
Método de selección de atributos	Matriz de confusión/ Medidas de desempeño	Clasificador			
		Naive Bayes	Red Neuronal	CN2	SVM
Top 10	TN	2738	3587	3498	2825
	FP	905	56	145	818
	FN	61	137	142	91
	TP	87	11	6	57
	Sensibilidad	0.5878	0.0743	0.0405	**0.3851**
	Especificidad	0.7516	0.9846	0.9602	0.7755
	Precisión	**0.0877**	0.1642	0.0397	0.0651
	F1	***0.1526***	0.1023	0.0401	**0.1114**
PCA	TN	2151	3281	3208	3230
	FP	1192	62	135	113
	FN	29	119	114	125
	TP	105	15	20	9
	Sensibilidad	***0.7836***	**0.1119**	**0.1493**	0.0672
	Especificidad	0.6434	0.9815	0.9596	**0.9662**
	Precisión	0.0810	***0.1948***	**0.1290**	**0.0738**
	F1	0.1468	**0.1422**	**0.1384**	0.0703
Manual	TN	3407	4023	3957	3447
	FP	671	55	121	631
	FN	122	178	168	158
	TP	63	7	17	27
	Sensibilidad	0.3405	0.0378	0.0919	0.1459
	Especificidad	**0.8355**	***0.9865***	**0.9703**	0.8453
	Precisión	0.0858	0.1129	0.1232	0.0410
	F1	0.1371	0.0567	0.1053	0.0641
Sin Selección	TN	2151	3281	3208	3230
	FP	1192	62	135	113
	FN	29	119	114	125
	TP	105	15	20	9
	Sensibilidad	***0.7836***	**0.1119**	**0.1493**	0.0672
	Especificidad	0.6434	0.9815	0.9596	**0.9662**
	Precisión	0.0810	***0.1948***	**0.1290**	**0.0738**
	F1	0.1468	**0.1422**	**0.1384**	0.0703

- Ventana.

5.2. Parámetros utilizados en cada clasificador

- Naive Bayes.
  - Sin parámetros.
  - Red neuronal Perceptrón Multicapa, con 3 capas escondidas de 50, 100 y 20 neuronas.
  - 3 capas ocultas.

Tabla 3. Tabla de los resultados utilizando sin datos de meses con >3% de incendios y el target de ventana.

Datos Selección de Meses >3%, Target de ventana					
Método de selección de atributos	Matriz de confusión/ Medidas de desempeño	Clasificador			
		Naive Bayes	Red Neuronal	CN2	SVM
Top 10	TN	2728	3699	3443	2872
	FP	1065	94	350	921
	FN	156	417	367	280
	TP	305	44	94	181
	Sensibilidad	0.6616	0.0954	0.2039	**0.3926**
	Especificidad	0.7192	***0.9752***	0.9077	0.7572
	Precisión	**0.2226**	0.3188	0.2117	**0.1642**
	F1	**0.3332**	0.1469	0.2077	**0.2316**
PCA	TN	2059	3014	2970	2529
	FP	1088	133	177	618
	FN	90	160	210	216
	TP	240	170	120	114
	Sensibilidad	***0.7273***	**0.5152**	**0.3636**	0.3455
	Especificidad	0.6543	0.9577	**0.9438**	**0.8036**
	Precisión	0.1807	***0.5611***	**0.4040**	0.1557
	F1	0.2895	***0.5371***	**0.3828**	0.2147
Manual	TN	2897	3628	3525	2302
	FP	903	172	275	1498
	FN	242	383	377	287
	TP	221	80	86	176
	Sensibilidad	0.4773	0.1728	0.1857	0.3801
	Especificidad	**0.7624**	0.9547	0.9276	0.6058
	Precisión	0.1966	0.3175	0.2382	0.1051
	F1	0.2785	0.2238	0.2087	0.1647
Sin Selección	TN	2059	3014	2970	2529
	FP	1088	133	177	618
	FN	90	160	210	216
	TP	240	170	120	114
	Sensibilidad	***0.7273***	**0.5152**	**0.3636**	0.3455
	Especificidad	0.6543	0.9577	**0.9438**	**0.8036**
	Precisión	0.1807	***0.5611***	**0.4040**	0.1557
	F1	0.2895	***0.5371***	**0.3828**	0.2147

Tabla 4. Tabla de los resultados utilizando sin datos de meses con >5% de incendios y el target puntual.

Datos Selección de Meses >5%, Target puntual					
Método de selección de parámetros	Matriz de confusión/ Medidas de desempeño	Clasificador			
		Naive Bayes	Red Neuronal	CN2	SVM
Top 10	TN	1890	2366	2316	2435
	FP	545	69	119	0
	FN	58	115	108	124
	TP	66	9	16	0
	Sensibilidad	0.5323	0.0726	0.1290	0.0000
	Especificidad	0.7762	0.9717	0.9511	1.0000
	Precisión	**0.1080**	0.1154	0.1185	NaN
	F1	***0.1796***	0.0891	0.1236	0.0000
PCA	TN	1588	2371	2322	2432
	FP	847	64	113	3
	FN	45	106	107	124
	TP	79	18	17	0
	Sensibilidad	***0.6371***	**0.1452**	**0.1371**	0.0000
	Especificidad	0.6522	0.9737	0.9536	***0.9988***
	Precisión	0.0853	***0.2195***	**0.1308**	0.0000
	F1	0.1505	**0.1748**	**0.1339**	0.0000
Manual	TN	2551	2968	2921	2425
	FP	486	69	116	612
	FN	120	161	153	133
	TP	49	8	16	36
	Sensibilidad	0.2899	0.0473	0.0947	0.2130
	Especificidad	**0.8400**	**0.9773**	**0.9618**	0.7985
	Precisión	0.0916	0.1039	0.1212	0.0556
	F1	0.1392	0.0650	0.1063	0.0881
Sin Selección	TN	1588	2371	2322	2432
	FP	847	64	113	3
	FN	45	106	107	124
	TP	79	18	17	0
	Sensibilidad	0.6371	**0.1452**	**0.1371**	0.0000
	Especificidad	0.6522	0.9737	0.9536	***0.9988***
	Precisión	0.0853	***0.2195***	**0.1308**	0.0000
	F1	0.1505	**0.1748**	**0.1339**	0.0000

- 50,100 y 20 neuronas.

– CN2 Inductor de reglas.

- Regla de ordenamiento= Ordenada.
- Algoritmo de cobertura: exclusivo.
- Medida de evaluación: entropía.
- Ancho de beam: 5.
- Cobertura mínima de la regla: 1.
- Longitud máxima de la regla: 5.

Tabla 5. Tabla de los resultados utilizando datos de meses con >5% de incendios y el target de ventana.

Datos Selección de Meses >5%, Target de ventana					
Método de selección de parámetros	Matriz de confusión/ Medidas de desempeño	Clasificador			
		Naive Bayes	Red Neuronal	CN2	SVM
Top 10	TN	2052	2559	2469	1919
	FP	721	214	304	854
	FN	188	324	334	222
	TP	231	95	85	197
	Sensibilidad	0.5513	0.2267	0.2029	0.4702
	Especificidad	0.7400	0.9228	0.8904	**0.6920**
	Precisión	**0.2426**	0.3074	0.2185	**0.1874**
	F1	**0.3370**	0.2610	0.2104	**0.2680**
PCA	TN	1495	2138	2062	1418
	FP	761	118	194	838
	FN	111	132	190	129
	TP	192	171	113	174
	Sensibilidad	***0.6337***	**0.5644**	**0.3729**	**0.5743**
	Especificidad	0.6627	***0.9477***	**0.9140**	0.6285
	Precisión	0.2015	***0.5917***	**0.3681**	0.1719
	F1	0.3057	***0.5777***	**0.3705**	0.2646
Manual	TN	2120	2553	2516	1902
	FP	665	232	269	883
	FN	240	326	332	269
	TP	181	95	89	152
	Sensibilidad	0.4299	0.2257	0.2114	0.3610
	Especificidad	**0.7612**	0.9167	0.9034	0.6829
	Precisión	0.2139	0.2905	0.2486	0.1469
	F1	0.2857	0.2540	0.2285	0.2088
Sin Selección	TN	1495	2138	2062	1418
	FP	761	118	194	838
	FN	111	132	190	129
	TP	192	171	113	174
	Sensibilidad	**0.6337**	**0.5644**	**0.3729**	**0.5743**
	Especificidad	0.6627	***0.9477***	**0.9140**	0.6285
	Precisión	0.2015	***0.5917***	**0.3681**	0.1719
	F1	0.3057	***0.5777***	**0.3705**	0.2646

- Máquina de Soporte Vectorial.
    - Costo (C)= 1.
    - Epsilon de regresión de pérdida: 0.1.
    - Kernel: RBF.
    - Tolerancia numérica: 0.0010.
    - Límite de iteraciones: 100.

Esquema de validación. Se usó un esquema de validación cruzada estratificada con 5 pliegues.

Medidas de desempeño seleccionadas Se utilizaron las medidas de desempeño de: Sensibilidad, que el % de días con vulnerabilidad clasificados correctamente del total de vulnerables (indica la capacidad de identificar los días vulnerables dentro del total de días vulnerables); Precisión, que representa el % dentro de los clasificados con vulnerabilidad, los días que realmente son vulnerables acorde al target utilizado, entre más alto sea se evitarán falsas alarmas al cuerpo de bomberos; la Medida F1, que combina estas dos medidas de desempeño; y finalmente, como medida menos importante, se seleccionó la Especificidad, que indica el porcentaje de días clasificados como sin vulnerabilidad dentro del total de días sin vulnerabilidad.

5.3. Estudio comparativo

Se obtuvieron las siguientes tablas, agrupadas por la selección de instancias. Para cada valor de Sensibilidad, Especificidad, Precisión y F1, se indica en **negritas** el valor más alto de cada clasificador por cada selección de instancias, en *cursiva* el mejor valor por cada selección de instancias, mientras que subrayado el mejor valor global.

5.4. Resultados generales

Podemos observar que los resultados de los experimentos sin selección de atributos y con el uso de PCA son prácticamente los mismos.

La combinación de selección de instancias/atributos y clasificador con mejores resultados en la precisión y la medida F1 fue la red neuronal aplicada a los datos con selección de los meses con >5% de incendios y el target de ventana, detectando un total de 56% de los días vulnerables con una precisión del 59%; el clasificador con mejor resultado en la medida de Sensibilidad fue el clasificador Naive Bayes, con PCA/sin selección de atributos y sin selección de instancias, sin embargo, cuenta con una precisión de 0.0616, lo que implica un alto porcentaje de falsas alarmas.

6. Conclusiones y trabajo a futuro

Uno de los factores negativos en los resultados fue el desbalance en el dataset, pues afecta directamente el desempeño de los clasificadores utilizados, así como el descarte prematuro de variables meteorológicas. Ante esto, se propone implementar maneras de equilibrar el dataset o aplicar algoritmos de selección y condensación de instancias únicamente a la clase mayoritaria.

Para la primera condensación de los datos (pasar de mediciones cada media hora a valores cada día), se propone agregar una división del día en dos, basado en la distribución de temperatura y humedad, día/noche y la integración del índice de vulnerabilidad del sistema difuso a las pruebas de clasificación.

Respecto al proceso de diseño de experimentos, se propone descartar los clasificadores con un mal desempeño general (SVM y CN2) y ampliar las pruebas con los clasificadores que dieron buenos resultados (Redes Neuronales, Naive Bayes).

Para los métodos de clasificación se propone probar otros algoritmos de clasificación, tales como redes neuronales tipo LSTM (con un método de aprendizaje

para datasets desbalanceados), sistemas neurodifusos u otros algoritmos más resilientes al desbalance de los datos.

La base de datos de incendios está probablemente sesgada, ya que el sistema de monitoreo actual en la REPSA es deficiente (generalmente las alarmas se dan por peatones que marcan a emergencias, suben un Twitter etiquetando al perfil de la reserva o algún miembro de la SEREPSA pasó en auto por las cercanías y logró vislumbrar el humo), es probable que haya habido incendios en la reserva sin que el equipo responsable se percatara de, y por lo tanto no se registrara.

Así mismo, los datos de la estación meteorológica están incompletos, desde una ausencia de datos por algunas horas hasta semanas enteras sin información y se desconoce el mantenimiento dado a la estación, o si hubo un cambio del equipo instalado. De la misma manera, la información disponible sobre los días en los que hubo incendios no indica la totalidad de los días en los que la reserva era vulnerable, ya que pudo haber días donde hubo una alta vulnerabilidad, pero no se inició un incendio ya que no hubo el factor humano (caótico) iniciador del fuego.

Dada la actual situación del monitoreo de incendios, podría implementarse el clasificador que dio mejores resultados (red neuronal entrenada con el target de ventana) en la reserva ecológica para que pueda dar alertas tempranas al cuerpo de bomberos y puedan aumentar la frecuencia de vigilancia de la reserva en caso de alerta de vulnerabilidad alta; sin embargo, la implementación en otras reservas naturales debe tomar muchas consideraciones dada la situación específica de la REPSA, tanto por sus altos niveles de interacción humana (y distribuida en ciclos relacionados con los semestres escolares), como las características específicas de su ecosistema nacido sobre piedra volcánica (con capacidades de filtrado de agua, retención de humedad y temperatura y relieve específicas), basado principalmente en matorral xerófilo y con un % considerable de especies invasoras flamables [8], que si bien lo hace único, también estas características limitan las posibilidad de implementar el mismo sistema en otras reservas (sería necesario una recalibración del clasificador).

Referencias

1. Sakr, G. E., Elhajj, I. H., Mitri, G.: Efficient forest fire occurrence prediction for developing countries using two weather parameters. Engineering Applications of Artificial Intelligence, vol. 24, no. 5, pp. 888–894 (2011) doi: https://doi.org/10.1016/j.engappai.2011.02.017
2. Chang, Y., Zhu, Z., Bu, R., Chen, H., Feng, Y., Li, Y., Wang, Z.: Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. Landscape Ecology, vol. 28, pp. 1989–2004 (2013) doi: 10.1007/s10980-013-9935-4
3. Alonso-Betanzos, A., Fontenla-Romero, O., Guijarro-Berdiñas, B., Hernández Pereira, E., Andrade, M. I. P., Jiménez, E., Carballas, T.: An intelligent system for forest fire risk prediction and firefighting management in Galicia. Expert systems with applications, vol. 25, no. 4, pp. 545–554 (2003) doi: 10.1016/S0957-4174(03)00095-2
4. Sayad, Y. O., Mousannif, H., Al-Moatassime, H.: Predictive modeling of wildfires: A new dataset and machine learning approach. Fire safety journal, vol. 104, pp. 130-146 (2019) doi: 10.1016/j.firesaf.2019.01.006
5. Torres-Rojo, J. M.: Estudio de tendencias y perspectivas del sector forestal en América Latina al año 2020: informe nacional: México. 2nd Ed. México (2004)
6. Schwela, D. H., Goldammer, J. G., Morawska, L. H., Simpson, O.: Health guidelines for vegetation fire events: guideline document. pp. 219 (1999)

7. CONAFOR - Gerencia de manejo del fuego: Programa de manejo del fuego, 2019. 1st Ed. CONAFOR, México (2019)
8. SEREPSA- La reserva ecológica del Pedregal de San Ángel: Atlas de riesgos, UNAM México (2012)
9. Clasificación de patrones: Métodos supervisados, Porta Zamorano, Jordi, Universidad Autónoma de Madrid http://www.iula.upf.edu/materials/050418porta 4.pdf
10. Jollife, I. T.: Principal component analysis. 2nd Ed. Springer, New York – EUA (2002)
11. Flórez-López, R., Fernández-Fernández, J. M.: Las redes neuronales artificiales, fundamentos teóricos y aplicaciones prácticas. 2nd Ed. Netbiblo, España (2008)
12. Ceruto-Cordovés, T., Rosete-Suárez, A., Espín-Andrade, R.: Obtención de predicados difusos a partir de datos utilizando metaheurísticas. Revista Internacional de Investigación de Operaciones – RIIO, vol. 1, no. 1, pp. 29–37 (2010)