UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS
Revista BOLETÍN DE CIENCIAS DE LA TIERRA

DETECCIÓN Y PRONÓSTICO DE INCENDIOS FORESTALES EN EL DEPARTAMENTO DEL VICHADA

DETECTION AND PROGNOSIS OF FOREST FIRES IN THE DEPARTMENT OF VICHADA

Sebastian Camilo Pachón García
Posgrado en Recursos Hidráulicos, Ingeniero Ambiental Universidad Nacional de Colombia sede Medellín,
scpachong@unal.edu.co

RESUMEN

La incidencia de los incendios forestales ha aumentado con los efectos del cambio climático a nivel mundial, debido a que las interacciones naturales que propician el incendio son complejas, predecir la ocurrencia de este fenómeno es complejo. Por ello, poder determinar el lugar y el momento de estos eventos es una tarea importante para la planeación del territorio y delimitación de zonas de riesgo. Estos análisis se han realizado bajo diversos modelos como físicos, dinámicos (interacción entre variables) y basados en machine learning también. Como la ocurrencia de estos eventos no es tan recurrente, el conjunto de datos utilizado para este análisis esta desequilibrado, lo que infiere en los resultados. Para este caso, se tomaron 9 variables, obtenidas de datos satelitales con una resolución temporal diaria, durante el 2015-2019, con un tamaño de pixel de 10km, evaluadas en 10 modelos supervisados de machine learning. Los resultados muestran que los modelos ensamblados como el Random Forest, el Xgboost y el Stocastic Gradient son los que tienen mejores métricas, presentando un recall y un precisión de más de 0.8. A pesar de ello es importante la disponibilidad de datos sobre la ocurrencia de incendios activos medidos en tierra, para aumentar la veracidad de los resultados. Del mismo modo, estas herramientas tienen potencial para apoyar los sistemas de gestión de riesgo.

Palabras clave: Incendios Forestales, Machine learning, Riesgo, Predicción, .

ABSTRACT

The incidence of forest fires has increased with the effects of climate change worldwide, because the natural interactions that lead to fire are complex, predicting the occurrence of this phenomenon is complex. Therefore, being able to determine the place and time of these events is an important task for the planning of the territory and delimitation of risk areas. These analyses have been carried out under various models such as physical, dynamic (interaction between variables) and based on machine learning as well. As the occurrence of these events is not so recurrent, the dataset used for this analysis is unbalanced, which infers in the results. For this case, 9 variables were taken, obtained from satellite data with a daily temporal resolution, during 2015-2019, with a pixel size of 10km, evaluated in 10 supervised machine learning models. The results show that assembled models such as the Random Forest, the Xgboost and the Stocastic Gradient are the ones with the best metrics, presenting a recall and an accuracy of more than 0.8. Despite this, the availability of data on the occurrence of active fires measured

on land is important to increase the veracity of the results. Similarly, these tools have the potential to support risk management systems.

Key words Wild Fires, Machine learning, Risk, Prediction, .

1. INTRODUCCIÓN

Los incendios forestales tienen gran relevancia a nivel mundial, ya que representan una fuente importante de gases de efecto invernadero, afectan los ecosistemas y emiten grandes cantidades de contaminantes que pueden afectar la calidad del aire (Chacón, 2015). La magnitud de los incendios forestales es de tal alcance que alrededor de 4 millones de Km2 fueron quemados entre el 2000 y el 2006, y se estima que entre el 5% y el 9% del total de áreas incendiadas a nivel mundial tienen lugar en Sudamérica (Armenteras et al., 2011). Para países sudamericanos, Colombia es uno de los de mayor relevancia cuando se tratan incendios forestales, donde en los años 2004 y 2007 se tuvieron más quemas extensivas, lo que produjo que se consumieran 19449 Km2 del territorio, correspondiendo al 12.7% del total de quemas para ese mismo año en Sudamérica (Chuvieco et al., 2008): Dentro de las regiones con mayores áreas quemadas en Colombia, la que se lleva el liderazgo es la región Orinoquia (Figura 1), donde se presentan los incendios forestales de mayor magnitud en Colombia, gracias a sus coberturas de vegetación herbáceas (propensas a la quema), a su topografía principalmente plana, temperaturas elevadas en épocas de sequía y a las prácticas culturales donde utilizan la quema para expandir la frontera agrícola (Armenteras et al., 2011).

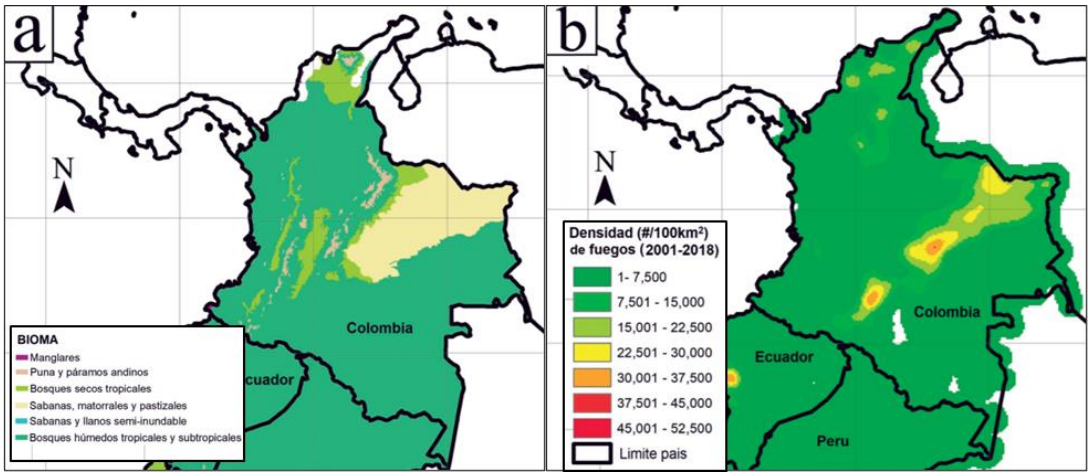

Figura 1. a. Distribución de biomas en Colombia, Ecuador y Perú; b. densidad de fuegos activos detectados con MODIS para el periodo 2001-2018 (# fuegos/100 km2).(Ocampo & Beltrán, 2018).

En este trabajo se propone abordar los incendios forestales el departamento del Casanare, desde machine learning, debido a que en el proceso que origina y desarrolla el incendio forestal intervienen muchas variables que se relacionan entre sí, como: Las especies de flora en la zona, la radiación incidente, la precipitación, las actividades antrópicas entre otras (Ocampo & Beltrán, 2018), haciendo de este

problema un sistema complejo que relaciona variables como son las condiciones climáticas o el tipo y estado de la vegetación (Corredor, 2017).

Se han hecho esfuerzos por aplicar Machine Learning para el estudio de estos fenómenos con grandes resultados, a partir de datos de sensores remotos (MODIS), datos meteorológicos y de topografía, como es el caso del estudio realizado por (Sayad, Mousannif &Moatassime, 2019), obteniendo alta precisión en la prediccion 98,32%. En Colombia, se han empleado estudios en la región Orinoquia con el uso del producto MCD64A1 del espectrorradiómetro de imágenes de resolución moderada (MODIS) para el período 2015-2019, obteniendo una acuracy de 94%, lo que indica que el desempeño del modelo fue excelente y a partir de esos resultados se generaron productos para el manejo del fuego en la región, para identificar áreas de priorización de esfuerzos y atención. Los resultados de la zonificación de probabilidad de ocurrencia indican que la categoría muy baja cubre la mayor área (28,2%), seguida de baja (23,2%), muy alta (17,6%), moderada (17,2%) y alta (13,8%) (Barreto& Armenteras, 2020).

La aproximación que se busca con este estudio es determinar esas zonas de incendios potenciales, en el departamento del Vichada, el cual cuenta con 107,808 habitantes, en un área de 105,947 km² , la cual se encuentra dentro de la región Orinoquia, la cual se caracteriza por sus fuertes incendios forestales, haciendo que este departamento presente altos riesgos por incendios forestales debido a las fuertes temporadas secas.

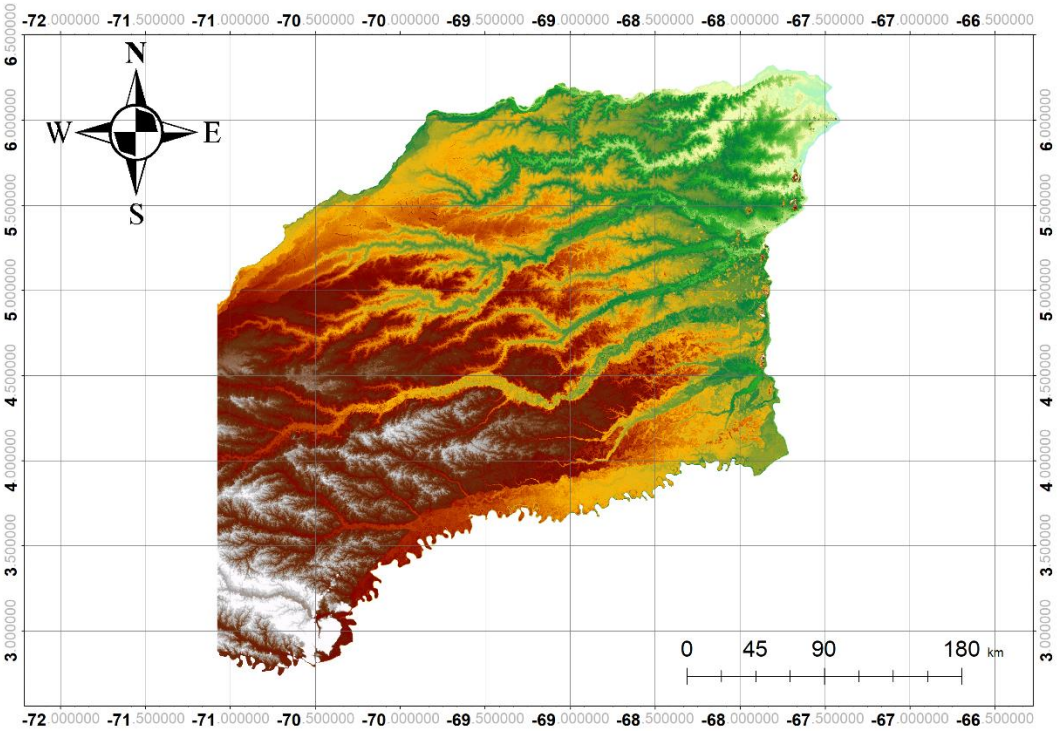

Figura 2: Zona de estudio

2. MATERIALES Y MÉTODOS

Para llevar a cabo el análisis emplearon 9 variables independientes (tabla 1), teniendo en cuenta la combustibilidad (coberturas), las variables meteorológicas y topográficas. Asimismo se usaron 2 variables más para construir la variable independiente, emisiones de CO2 y áreas quemadas realizando una composición de las 2 para determinar si en los pixeles hay o no incendio, entonces donde existan valores mayores a 0 de emisión de CO2 sobre un área quemada, se tomará ese pixel como zona de incendio.

Tabla 1. Variables analizadas en la predicción de la ocurrencia de un incendio.

Variable	Tipo	Resolución Espacial	Resolución Temporal	Fuente
Precipitación	Continua	0.1° x 0.1° ≈10 km	Diaria	(Muñoz Sabater, 2019)
Temp 2metros	Continua		Diaria	
Temp del suelo	Continua		Diaria	
Temp superficial	Continua		Diaria	
Evaporación	Continua		Diaria	
Humedad	Continua		Diaria	
Radiación	Continua		Diaria	
Pendientes	Continua	30 m	-	(ASF DAAC, 2015)
Coberturas	Categórica	0.001° x 0.001° ≈ 0.1 km	Anual	(Marcel Buchhorn et al., 2020)
Emisiones de CO2	Numérica	0.1° x 0.1° ≈10 km	Diaria	(Inness et al., 2019)
Burned área	Numérica	500 m	Diaria	(Giglio et al., 2017)

Los pasos a seguir para implementar los algoritmos de machine-learning, al problema primero se descargan, los datos de las distintas variables, posteriormente se realiza un pre procesamiento de los datos, para llevarlos a la misma resolución espacial y temporal, asimismo un análisis de las unidades de cada variable. Luego con las variables preprocesadas, se hace una validación de los datos para determinar si el conjunto de datos es suficiente para el análisis y se seleccionan las variables que representen mejor el problema.

Una de características del problema, es que dentro de la zona de estudio serán más los pixeles donde no hay incendios, por lo tanto es un problema desbalanceado. El cual será tratado bajo el método RandomUnderSampler, que minimiza la cantidad de datos de la clase mayoritaria seleccionando muestras al azar e igualando las 2 categorías. Con ello, se tienen los datos listos para implementar los modelos de clasificación binaria que para este caso solo usarán los supervisados, ya que, los no supervisados dan resultados con métricas bajas.

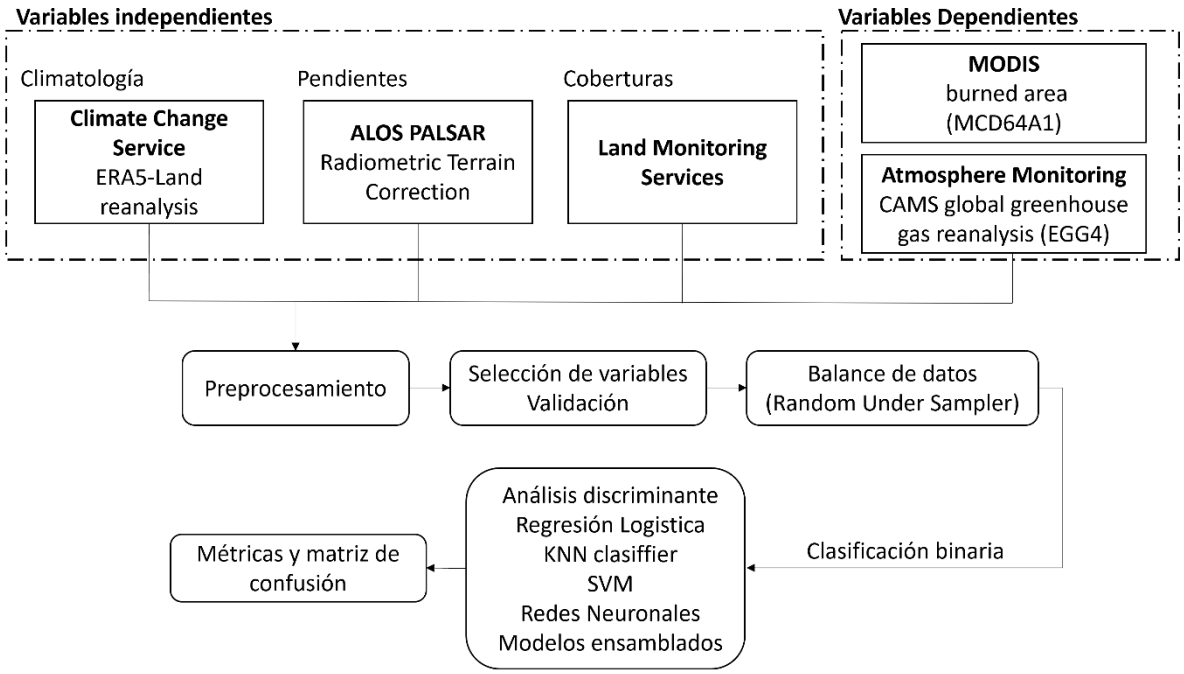

Figura 3 : Diagrama metodológico.

Finalmente, se evalúa el modelo mediante las métricas como el Recall (1) , la precisión (2) para la variable positiva, el número de falsos negativos (FN) y verdaderos positivos (TP). Esto debido a que el problema se enfoca en la reducción del riesgo, por ende, se pretende que el modelo pronostique la mayoría de los pixeles con potencial de incendio, y se seleccionará el que cometa menos errores en el proceso.

- Recall: mide qué tan bien el clasificador puede detectar observaciones positivas.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

- Precisión: la relación entre las observaciones positivas predichas correctamente y el total de observaciones positivas previstas.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. RESULTADOS Y DISCUSIÓN

En el análisis inicial de variables mediante, posterior al preprocesamiento, se empleó la correlación para determinar las variables que realmente representan el fenómeno, como resultado se obtuvo que las temperaturas estaban altamente correlacionadas por ello se procedió a eliminar 2 de estas y dejar solo 1 que represente esa dimensión del problema (figura 4).

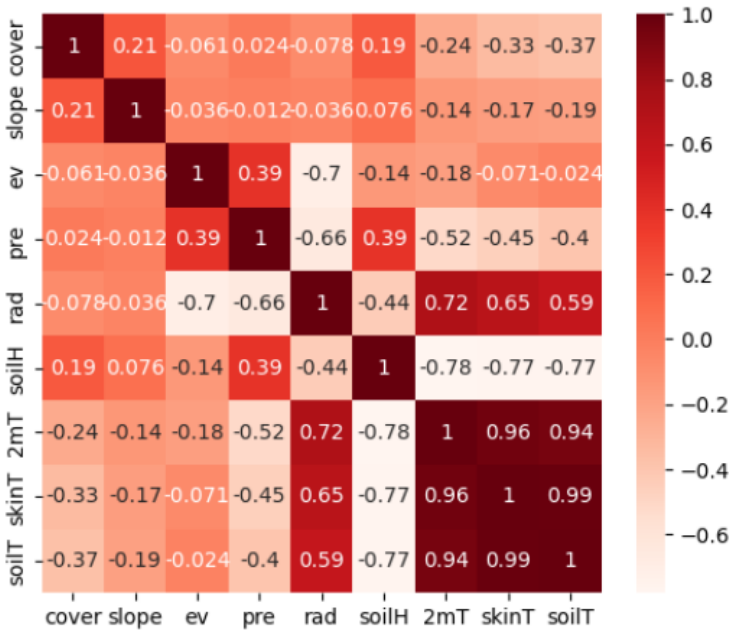

Figura 4 : Curva de aprendizaje

Ahora bien, se traza la curva de aprendizaje para determinar si el conjunto de datos es suficiente, se obtuvo que el modelo aprenderá con tan solo 680 mil datos y que a partir de ahí se estabiliza, lo que representa un buen ajuste (figura 5). Igualmente, de ahí en adelante se puede presentar un sobreajuste del modelo.

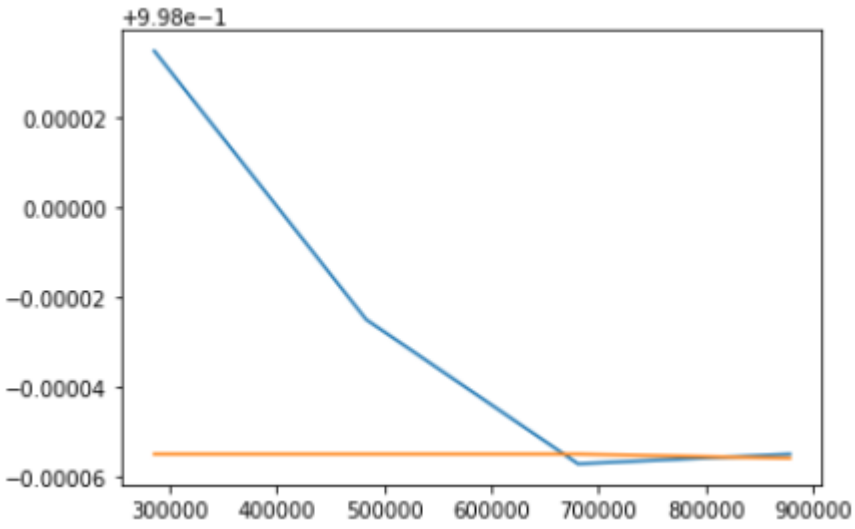

Figura 5 : Curva de aprendizaje

Así pues, con las variables seleccionadas y el tamaño de datos óptimo para la estimación, se aplicaron cada uno de los algoritmos a la base de datos tomando los hiperparámetros que dieran los mejores resultados y un test size del 20%. En consecuencia, se encontraron los resultados expuestos en la siguiente tabla con los valores de las métricas de evaluación para cada método, donde se puede apreciar que los modelos ensamblados dan mejores resultados que el resto, y dentro de ellos se resaltan las redes neuronales, los Árboles de Decisión, el Árbol con Baggin, el Random Forest, Stocastic Gradient y el XGBoost.

Tabla 2. Evaluación de métodos

Modelo	**Recall**	**Precision**	**FN**	**FP**	**TP**	**TF**
Análisis Discriminante Lineal	0.79	0.7	30	48	110	277
Regresión Logística	0.73	0.71	38	42	105	293
KNN	0.43	0.48	91	76	70	213
SVC	0.15	0.17	126	116	23	175
Redes Neuronales	0.8	0.79	29	32	117	284
Arboles Decisión	0.83	0.69	24	52	117	231
Baggin Arbol	0.82	0.69	25	51	116	232
Random Forest	0.84	0.77	22	36	119	247
AdaBoost	0.78	0.72	31	42	110	241
Stocastic Gradient	0.84	0.75	23	39	118	244
XGBoost	0.85	0.71	21	48	120	235

Como la finalidad del trabajo es determinar el mejor modelo bajo estas métricas de evaluación, se tiene que, el que presenta in mejor rendimiento es el XGBoost , gracias a su estructura de varios arboles de decisión en paralelo, que van iterando el algoritmo y en cada iteración usa los residuos de error del árbol anterior para optimizar los resultados con cada paso.

En la Figura 6 se puede apreciar el comportamiento de las métricas mas a fondo del modelo XGBoost, donde se muestra que para la predicción se equivoca en 21datos, clasificándolos como puntos de no incendio cuando si hay presencia del fenómeno, este es el resultado más bajo en esta métrica. Además, es el modelo que mas valores acertados tiene con 120 TP, por otro lado la curva ROC tiene un comportamiento bueno debido a que las tasas de valores verdaderos positivos son mucho mayores a las de los falsos positivos, lo que muestra q el modelo a medida que va acertando en la predicción comete pocos errores.

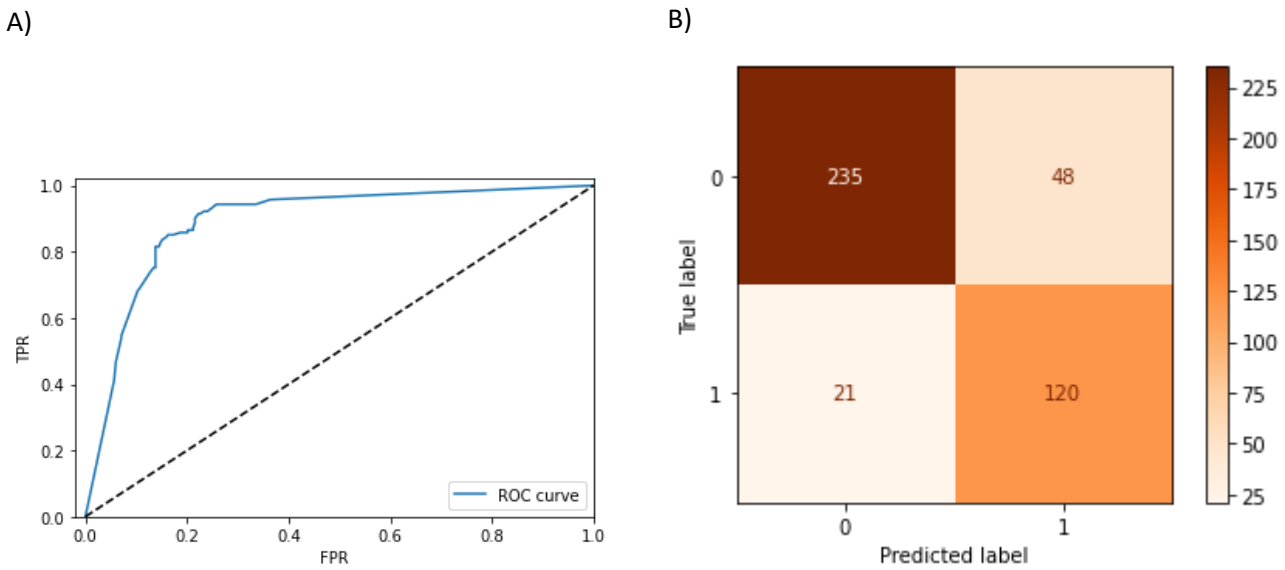

Figura 6 A) Curva ROC B) Matriz de confusión para el modelo XGBoost.

4. CONCLUSIONES

Para concluir, se considera que el volumen de datos no siempre es garantía de un buen desempeño del modelo, como se aprecia en la figura 5, una gran cantidad de datos puede propiciar sobre ajuste del modelo, de igual forma la reducción de datos permite disminuir el consumo de máquina. También se encontró que el RandomUnderSampler, para la predicción de este fenómeno, es el mejor método para el tratamiento de datos desbalanceados.

Por otra parte, los modelos ensamblados, tienen un gran potencial para ajustarse a estos fenómenos dinámicos y con tantas variables. Debido a su robustes y que se comparten los errores paso a paso, arrojan resultados interesantes en la evaluación.

Para mejorar los resultados y poder mayor veracidad de los resultados obtenidos, es necesario contar con mediciones en tierra con fecha y lugar de incendios activos o zonas quemadas. Si bien, los sensores remotos son una buena fuente de información, no son mediciones directas de ocurrencia sino son datos estimados a partir de algoritmos que interpretan las anomalías en la radiación y estos pueden tener errores en sus cálculos.

REFERENCIAS

Armenteras-Pascual, D., Retana-Alumbreros, J., Molowny-Horas, R., Roman-Cuesta, R. M., Gonzalez-Alonso, F., & Morales-Rivas, M. (2011). Characterising fire spatial pattern interactions with climate and vegetation in Colombia. *Agricultural and Forest Meteorology*, *151*(3), 279–289. https://doi.org/10.1016/j.agrformet.2010.11.002

Barreto, J. S., & Armenteras, D. (2020). Open Data and Machine Learning to Model the Occurrence of Fire in the Ecoregion of "Llanos Colombo–Venezolanos." *Remote Sensing*, *12*(23), 3921. MDPI AG. Retrieved from http://dx.doi.org/10.3390/rs12233921

Chacón, L. M. (2015).Efecto de los Incendios forestales sobre la calidad del aire en dos ciudades colombianas.

Chuvieco, E., Giglio, L., & Justice, C. (2008). Global characterization of fire activity: Toward defining fire regimes from Earth observation data. *Global Change Biology*, *14*(7), 1488–1502. https://doi.org/10.1111/j.1365-2486.2008.01585.x

Corredor Llano, X. (2017). *Desarrollo de un modelo para la dispersión del fuego en la Orinoquía Colombiana usando autómatas celulares*.

Gholamnia, K., Gudiyangada Nachappa, T., Ghorbanzadeh, O., & Blaschke, T. (2020). Comparisons of Diverse Machine Learning Approaches for Wildfire Susceptibility Mapping. *Symmetry*, *12*(4), 604. MDPI AG. Retrieved from http://dx.doi.org/10.3390/sym12040604

Dataset: ASF DAAC 2015, ALOS PALSAR_Radiometric_Terrain_Corrected_low_res; Includes Material © JAXA/METI 2007. Accessed through ASF DAAC 11 November 2015. DOI: https://doi.org/10.5067/JBYK3J6HFSVF

Giglio, L., Justice, C., Boschetti, L., Roy, D,(2017)."MODIS/Terra+Aqua Direct Broadcast Burned Area Daily L3 Global 500m SIN Grid V006".Revisor NASA EOSDIS Land Processes DAAC. December 2017.DOI : https://doi.org/10.5067/MODIS/MCD64A1.006

Inness, A, Ades, M, Agustí-Panareda, A, Barré, J, Benedictow, A, Blechschmidt, A, Dominguez, J, Engelen, R, Eskes, H, Flemming, J, Huijnen, V, Jones, L, Kipling, Z, Massart, S, Parrington, M, Peuch, V-H, Razinger M, Remy, S, Schulz, M and Suttie, M (2019): CAMS global reanalysis (EAC4). Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overview

Louis, G.; Wilfrid, S.; Joanne, V.H.; Christopher, O. Justice (2018), "MODIS Collection 6 Active Fire Product User's Guide", Revision B, NASA. Available online: http://modis-fire.umd.edu/files/MODIS_C6_Fire_User_Guide_B.pdf (accessed on 7 September 2020).

Marcel Buchhorn, Bruno Smets, Luc Bertels, Bert De Roo, Myroslava Lesiv, Nandin-Erdene

Tsendbazar, … Steffen Fritz. (2020). Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2015-2019: Globe (Version V3.0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3518036

Muñoz Sabater, J., (2019): ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on < DD-MMM-YYYY >), 10.24381/cds.e2161bac

Pérez-Porras, F.-J., Triviño-Tarradas, P., Cima-Rodríguez, C., Meroño-de-Larriva, J.-E., García-Ferrer, A., & Mesas-Carrascosa, F.-J. (2021). Machine Learning Methods and Synthetic Data Generation to Predict Large Wildfires. *Sensors*, *21*(11), 3694. MDPI AG. Retrieved from http://dx.doi.org/10.3390/s21113694

Sayad, Y. O., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire safety journal*, *104*, 130-146.

Urbanski, S. (2014). Wildland fire emissions, carbon, and climate: Emission factors. *Forest Ecology and Management*, *317*, 51-60.

Wang, S. S. C., Qian, Y., Leung, L. R., & Zhang, Y. (2021). Identifying key drivers of wildfires in the contiguous US using machine learning and game theory interpretation. *Earth's future*, *9*(6), e2020EF001910.