Artificial Intelligence and Machine Learning Model for Spatial and Temporal Prediction of Drought Events in the Department of Magdalena, Colombia

Article *in* Inge CUC · December 2022

DOI: 10.17981/ingecuc.18.2.2022.20

Edier Aristizábal
Universität Potsdam

Artificial Intelligence and Machine Learning Model for Spatial and Temporal Prediction of Drought Events in the Department of Magdalena, Colombia

DOI: http://doi.org/10.17981/ingecuc.18.2.2022.20

Artículo de Investigación Científica. Fecha de Recepción: 27/08/2021, Fecha de Aceptación: 11/11/2021

Daissy Milenys Herrera Posada
Universidad Nacional de Colombia, Medellín. (Colombia)
dmherrerap@unal.edu.co

Edier Aristizábal
Universidad Nacional de Colombia, Medellín. (Colombia)
evaristizabalg@unal.edu.co

To cite this paper

Hererra posada D. Aristizábal E “Artificial Intelligence and Machine Learning Model for Spatial and Temporal Prediction of Drought Events in the Department of Magdalena, Colombia”. DOI: http://doi.org/10.17981/ingecuc.18.2.2022.20

Abstract

Drought is one of the most critical hydrometeorological phenomenon in terms of its impacts on society. Although Colombia is a tropical country, there are areas of the territory which have periods of drought, and this causes significant economic damage. Due to recent advances in terms of the spatial and temporal resolutions of remote sensing, and artificial intelligence techniques, it is possible to develop automatic learning models supported by historical information. In this study, a Random Forest (RF) and Bagged Decision Tree Classifier (DTC) model was built to perform spatial and temporal drought prediction in the department of Magdalena using the following features: Normalized Difference Vegetation Index (NDVI), land surface temperature (LST), precipitation, Normalized Difference Water Index (NDWI), Normalized Multiband Drought Index (NMDI), evapotranspiration (ET), surface soil moisture (SSM), subsurface soil moisture (SUSM), Multivariate ENSO Index (MEI), Southern Oscillation Index (SOI), and Oceanic Niño Index (ONI). For labelling, which allows one to train and evaluate the model, the Standardized Precipitation Index (SPI) was used to identify drought events. The implementation of the developed model can allow governmental entities to take actions to mitigate impacts generated by recurring droughts in their territories.

Keywords

Drought forecasting, Standardized Precipitation Index, satellite imagery, Google Earth Engine, machine learning, random forest, decision tree classifier, spatial interpolation

I. INTRODUCTION

Drought is a common and frequently occurring characteristic of the climate [1], and it is defined as a lack of moisture caused by the absence of precipitation during a given period of time [2]. When the length of time without precipitation increases significantly, the amount of available water cannot meet the water demand of the environment and population [3]. According to the Food and Agriculture Organization of the United Nations [4], drought is one of the most critical natural phenomena in terms of impacts on society, since it halts food production, prevents the development of pastures, affects markets, and causes the deaths of living beings and population migration.

An increase in the number of droughts has been pointed out as one of the main consequences of climate change. There has not just been an increase in their frequency: an increase in the intensity of droughts is also expected [5] [6]. This is due to changes in the hydro-climatological variables that condition or determine the occurrence of drought events, as mentioned in [7], which estimates a reduction in precipitation of between 2 and 8% for each degree kelvin that the temperature increases due to global warming. In addition, [8] predicts an increase in evapotranspiration and an increase in the surface temperature.

In Colombia, drought has had a significant impact on the population. During the 2014 drought event, there were 642 fires in the departments of La Guajira, Magdalena, Córdoba, Atlántico, Sucre, Bolívar, and Cesar, as well as the deaths of 3,200 head of cattle, the loss of about 47% of rice crops, and water shortages in 48 municipalities in Colombia [9]. In the case of the department of Magdalena, the drought event that occurred in 2015 reduced the water supply of the department by 60%, decreasing the water supply available for the population and for agricultural, fish farming, and livestock farming activities in the department, which led to food shortages [10].

Considering the negative effects of drought and the possible increase in its frequency as a consequence of climate change, it is necessary to implement mechanisms to monitor the phenomenon and search for strategies for implementing adaptation measures and reducing its effects. As part of these mechanisms, there are early warning systems supported by automatic learning techniques that make it possible to evaluate or predict the occurrence of droughts [11] [12] [13] [14] [15] [16].

Machine learning (ML) is a group of computational techniques within artificial intelligence (AI), which takes inputs from statistics to learn from past events, recognize patterns and predict new observations [17]. Within ML techniques, there are several supervised and unsupervised methods, such as Neural Networks, Decision Trees, Logistic Regression, Principal Component Analysis, Clustering, among others. ML applications range from cancer prediction and forecasting [17], automatic speech recognition [18], daily flow forecasting [19], remote sensing [20], among others.

Regarding the application of artificial intelligence for drought prediction, papers such as [11] stand out: for the prediction of agricultural drought in Australia, they implement classification and regression trees, random forests, flexible discriminant analysis, and support vector machines. In [12], the authors run random forest models and regression trees in different climatic regions of the United States for drought prediction, which is determined by the Standardized Precipitation Index (SPI). In [13], in order to evaluate and reduce the potential impact of drought on palm crops, the authors implement different types of support vector machines for the prediction of the Standardized Precipitation Evapotranspiration Index (SPEI). Similarly, [14] predicts the SPEI to assess agricultural drought over South-Eastern Australia, making use of models such as random forest, support vector machines, and neural networks. In [15], the authors implement bootstrapping and boosting in artificial neural networks and support vector machines for SPI prediction at different time scales over the Awash River basin in Ethiopia. Finally, [16] proposes a new index to evaluate drought, called the Integrated Agricultural Drought Index (IDI), and uses neural networks with back-propagation for the recognition of non-stationary patterns in the occurrence of droughts. In Colombia, the National Observatory, for the follow-up and monitoring of drought in Colombia and through unsupervised methods such as principal component analysis, estimated the threat due to meteorological drought at the national scale [21].

Several studies [14][16][12] incorporate satellite information as an input variable for the execution of different models. This is because remote sensors provide data with a high temporal resolution that covers large extensions of land [3], thus allowing access to areas where field data is scarce or where there is a low density of specific data. The Colombian territory is no exception to the problems mentioned above; for this reason, the implementation of strategies or models for the evaluation of droughts using open, free, and quality information, such as the use of satellite images, becomes a necessary strategy for carrying out hydrometeorological studies distributed in any part of the territory. Additionally, droughts are natural phenomena that have an impact on large portions of land [1]; therefore, it is relevant to obtain continuous maps where the spatiality of the phenomenon is detailed. The use of remote sensing data facilitates obtaining these results and reduces the uncertainty involved in making distributed maps by implementing interpolation techniques where the density of point data is low.

This work seeks to implement another type of machine learning (ML) tool for drought prediction over the Colombian territory. In this case, the department of Magdalena is used as a pilot study area, and, by implementing satellite information, the intent is to predict the droughts determined by the SPI, using decision trees and random forests. This allows one to model the temporal occurrence and spatial distribution of drought events over the department. The objective is to establish an early warning system that allows authorities to take measures to reduce a region's vulnerability to drought events and to establish a methodology for drought prediction that can be implemented in other territories of Colombia.

II. STUDY AREA

Each year, Unidad Nacional para la Gestión del Riesgo de Desastres (UNGRD) prepares a consolidated report on emergencies reported in each municipality of Colombia [22], in which the damage and human and economic losses are recorded. Droughts are among the events reported. Between 2010 and 2019, 144 drought events were recorded (see Table 1), and the departments of Cauca and Magdalena are the departments with the highest numbers of events. Therefore, considering the high recurrence of events and territorial extension, the department of Magdalena was selected as a pilot study area.

Table 1. Droughts reported by departments to the UNGRD from 2010 to 2019

Department	Droughts reported	Department	Droughts reported
Atlántico	8	Guajira	13
Bolívar	12	Magdalena	17
Boyacá	4	Nariño	5
Caldas	1	Norte de Santander	2
Cauca	17	Quindío	1
Cesar	7	Risaralda	9
Córdoba	14	Santander	13
Cundinamarca	2	Sucre	6
Guaviare	1	Tolima	4
Huila	1	Valle de cauca	5

* Note: The consolidation of emergencies for the year 2010 presented problems for the visualization of the information; therefore, the reports for the year 2010 could not be considered.

The department of Magdalena is in northern Colombia, with a territorial extension of 23,188 km^2 [23] and a population of 1,263,788 inhabitants [24]. The average temperature varies by sector within the department; in the south it can exceed 28°C, and in the center and north, it is between 26–28°C; meanwhile, in the Sierra Nevada, it decreases according to the elevation with respect to sea level, reaching -8°C. With a bimodal regime, the average annual precipitation varies by sector within Magdalena: in the flat area of the department, the average annual precipitation is between 1,000 and 1,500 mm, and in the south and in the vicinity of the Sierra Nevada, the average precipitation exceeds 2,000 mm [25].

III. DATA

A. *Remote sensing data*

In [12], the Normalized Difference Vegetation Index (NDVI), land surface temperature (LST), precipitation, Normalized Difference Water Index (NDWI), Normalized Multiband Drought Index (NMDI), and evapotranspiration (ET) were used as predictor variables. Additionally, the surface soil moisture (SSM) and subsurface soil moisture (SUSM) were considered. All these variables are directly related to the water content present in the soil and hydrological systems [26]. In addition, vegetation responds to moisture changes; therefore, the vegetation indices NDVI, NDWI, and NMDI were considered to understand the state of the vegetation and how it is affected by a moisture deficit.

The previously mentioned variables are obtained through the Google Earth Engine (GEE) platform [27], from which satellite information is extracted for the department of Magdalena for the period from 2010 to 2019 on a monthly time scale. For precipitation, the Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS) database was accessed, specifically UCSB-CHG/-CHIRPS/DAILY, with a spatial resolution of 0.05°, which is equivalent to 5.5 km (see Fig. 1A). Soil moisture data were provided by the National Aeronautics and Space Administration (NASA) in conjunction with the U.S. Department of Agriculture (NASA_USDA/HSL/soil_moisture) at a resolution of 0.25°, which corresponds to 28 km (see Fig. 1G and H). The other variables were obtained through information provided by the MODIS satellite at spatial resolutions of 1 km for LST (MODIS/006/MOD11A2) (see Fig. 1B), 0.5 km for ET (MODIS/006/MOD16A2) (see Fig. 1C), and 0.5 km for reflectance (MODIS/006/MOD09A1).

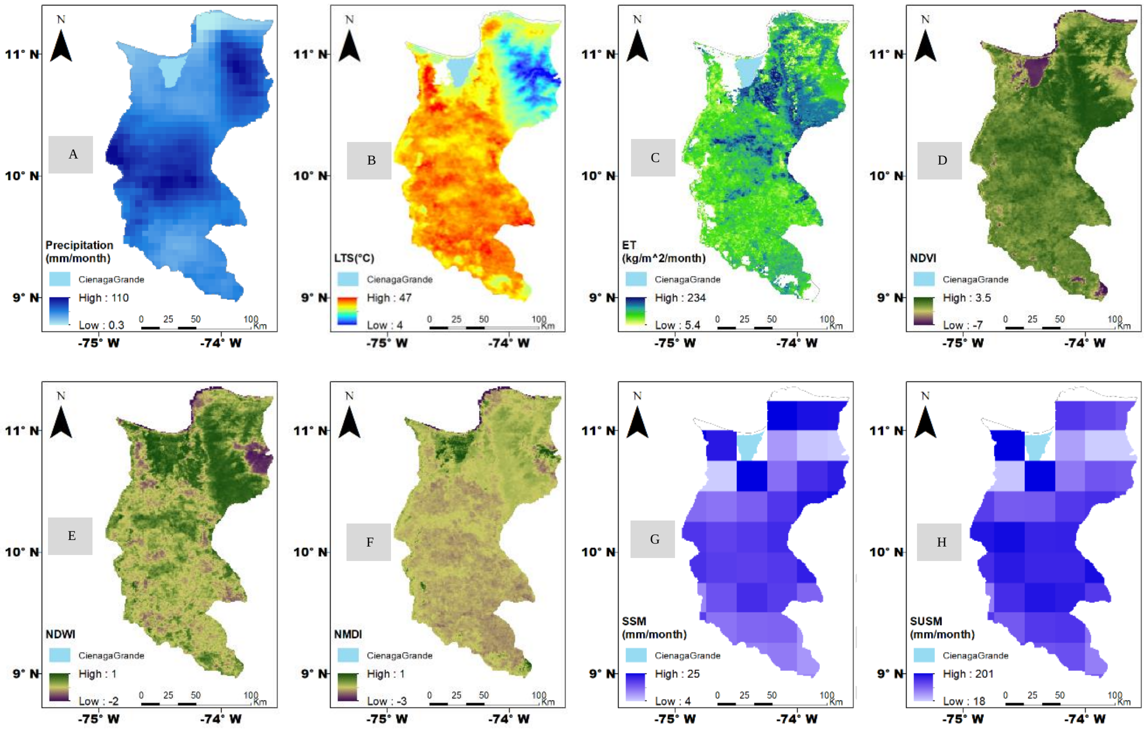

Fig. 1. Satellite predictor variables: A) Precipitation. B) Land surface temperature (LST). C) Evapotranspiration (ET). D) Normalized Difference Vegetation Index (NDVI). E) Normalized Difference Water Index (NDWI). F) Normalized Multiband Drought Index (NMDI). G) Surface soil moisture (SSM). H) Subsurface soil moisture (SUSM)

The NDVI, NDWI, and NMDI (Fig. 1D, E, and F) were estimated using reflectance and the equations in Table 2 at a spatial resolution of 0.5 km. The NDVI indicates the state of the vegetation, based on the radiation reflected by the vegetation in near-infrared wavelengths with respect to the red band of the visible spectrum [28]; the NDWI estimates the moisture content of the vegetation based on the radiation reflected by the surface in the infrared wavelength [29], and the NMDI is used to track the moisture content of the soil and vegetation from the infrared [30].

Table 2: Calculation and bands needed to find NDVI, NDWI and NMDI.

Índices de vegetación	Fórmula
NDVI	$\frac{\rho_{Banda\ 2} - \rho_{Banda\ 1}}{\rho_{Banda\ 2} + \rho_{Banda\ 1}}$
NDWI	$\frac{\rho_{Banda\ 2} - \rho_{Banda\ 5}}{\rho_{Banda\ 2} + \rho_{Banda\ 5}}$
NMDI	$\frac{\rho_{Banda\ 6} - \rho_{Banda\ 7}}{\rho_{Banda\ 2} + (\rho_{Banda\ 6} - \rho_{Banda\ 7})}$

As can be seen, each variable has a different spatial resolution, so it is necessary to homogenize the resolutions; for this purpose, the GEE Scale function was used and all the satellite information was resampled at a spatial resolution of 1 km.

As for the missing values within the data series, these were filled in with the average value of each image. Anomalous data, caused by information processing errors or errors in data collection, were replaced by the 99th percentile, as the maximum value, and the 1st percentile, as the minimum value, within the data series of each predictor variable.

B. Macro-climatic variables

The El Niño Southern Oscillation (ENSO) phenomena La Niña and El Niño are closely related to the hydrological anomalies that have developed in the South American tropics [31]. High precipitation and maximum flows are associated with the occurrence of La Niña, while El Niño is characterized by long-lasting dry periods, modifying the intensity and prolongation of droughts within the territory [31].

To evaluate the influence of ENSO, variables such as the Multivariate ENSO Index (MEI), Southern Oscillation Index (SOI), and Oceanic Niño Index (ONI), which are elaborated using a monthly temporal resolution by the U.S. Oceanic and Atmospheric Administration (NOAA) and consider different parameters that allow the status of each month to be classified as El Niño, La Niña, or Neutral, were selected for the study.

IV. MODEL FOR DROUGHT FORECASTING

A. Reference data: SPI

To evaluate drought conditions within the department of Magdalena, the Standardized Precipitation Index (SPI) is selected as the model response variable. The SPI evaluates precipitation anomalies at various time scales, allowing the study of various types of droughts [3].

For the calculation of the SPI, precipitation information on a monthly scale and a continuous information record of at least 30 years are required [2]. Therefore, a search was made for data provided by the Instituto de Estudios Ambientales (IDEAM) from the existing precipitation stations within the department of Magdalena that meet the requirements for the calculation of the index. Thus, 43 precipitation stations that present rainfall information on a monthly scale since before 1989 were identified, as shown in Fig. 2.

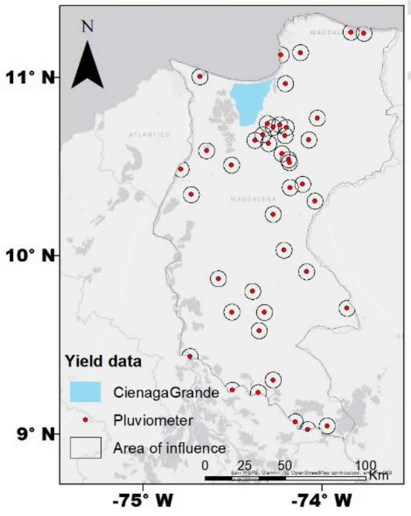

Fig. 2. Yield data precipitation stations within the department of Magdalena used in the study, and their respective areas of influence.

The rainfall stations represent point values, so they do not allow one to establish the spatial distribution of rainfall in the department, which is necessary to establish the spatial distribution of SPI, preventing the analysis of this variable with the previously mentioned distributed variables. For each IDEAM rainfall station within the department of Magdalena, an area of influence with a radius of 5 km was established (see Fig. 2), which makes it possible to establish an SPI value for each area of influence, as well as to determine the satellite information corresponding to each evaluated area. This allows the development of the analysis.

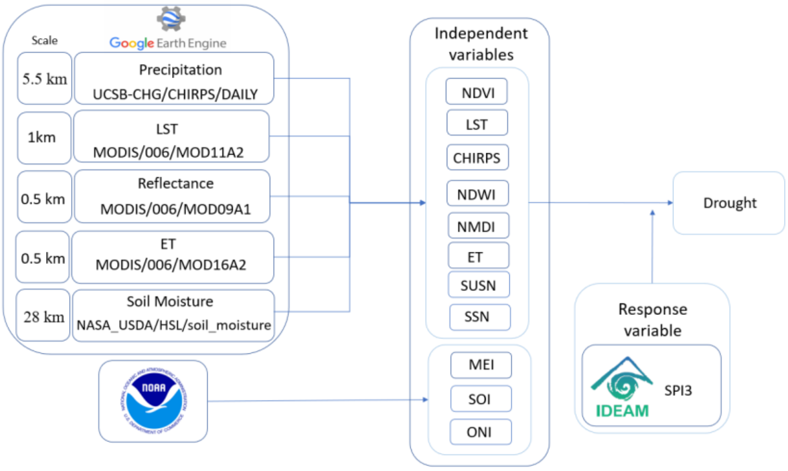

Fig. 3. Conceptual scheme of the sources of information for the predictor variables and the response variable

The SPI is calculated by fitting the monthly precipitation data (x) to a gamma distribution function (1), where the alpha (α) and beta (β) parameters of the function are estimated for each precipitation station and time scale. With the precipitation series fitted to a distribution, we proceed to calculate the cumulative probability of each event (2). Since the gamma distribution function is not defined for events in which x = 0, the factor q (3) is added, which represents the probability in the case that precipitation events have values of 0 [32]. The cumulative probability is transformed into a standard normal random variable (Z), with a mean of zero and a standard deviation of 1; the Z values found represent the values corresponding to the SPI [32]. In other words, the SPI represents how many standard deviations, above or below, an event is from the average rainfall fitted to the gamma distribution function [32].

$$g(x) = \frac{1}{\beta^{\alpha} * \Gamma(\alpha)} * x^{\alpha-1} * e^{\frac{-x}{\beta}} \quad (1)$$

$$G(x) = \int_0^x g(x)dx \quad (2)$$

$$H(x) = q + (1 - q) * G(x) \quad (3)$$

By considering the normal distribution of SPI values, drought or wetness events can be defined [2]. The World Meteorological Organization (WMO) [3], based on [2], categorizes the SPI from extremely wet, for SPI values greater than 2, to extremely dry, for values less than -2 (see Table 3).

Tabla 3. SPI values. Fuente: [3].

Value	Category
2.0 y más	Extremely wet
1.5 a 1.99	Very wet
1.0 a 1.49	Moderately wet
-0.99 a 0.99	Near normal
-1.0 a -1.49	Moderately dry
-1.5 a -1.99	Severely dry
-2.0 y menos	Extremely dry

The SPI can be evaluated for different time scales depending on the type of drought to be studied, which can be meteorological, hydrological, or agricultural [33]. Meteorological drought occurs when, during a period of time, precipitation is lower than expected; hydrological drought refers to a decrease in river flow and the levels of reservoirs and lakes due to a deficit of precipitation; and agricultural drought refers to the fact that, due to the deficit of precipitation, there is not enough moisture in the soil for the normal functioning of crops [33]. According to [3], to study agricultural drought, the SPI should be calculated with accumulated precipitation between 1 and 6 months; for meteorological drought, the accumulated precipitation should be between 1 and 2 months, and for hydrological drought, between 6 and 24 months.

In this sense, the present work evaluates agricultural drought, due to the social repercussions of this event. For this, the SPI was calculated with a precipitation accumulation period of 3 months (SPI3), since this allows one to understand the changes of agricultural drought [34] and provides a seasonal approximation of precipitation [3]. Additionally, the time span evaluated is relevant for annual crops [35] and the intra-seasonal study of precipitation is relevant for herbaceous and low-cut crops [36].

In this way, and with the IDEAM precipitation information, SPI3 was calculated for each precipitation station through the SPI Generator program developed by the National Drought Monitoring Center of the University of Nebraska–Lincoln [37], and the SPI3 values from 2010 to 2019 in the areas of influence of each station were selected.

The SPI3 value calculated is a continuous value that can be classified according to Table 3; however, the classification of the drought magnitude is a function of local conditions [3]. Therefore, in order to establish which SPI3 values represent drought in the department of Magdalena, SPI3 was calculated for each month and municipality that reported drought emergencies to the UNGRD. The frequency histogram of the SPI3 values obtained is shown in Fig. 4. As can be seen, most of the reported months have an SPI3 of -1; therefore, the continuous variable is transformed into a categorical variable. SPI values below -1 indicate a drought and values above -1 indicate normal or wet conditions, as shown in Fig. 5.

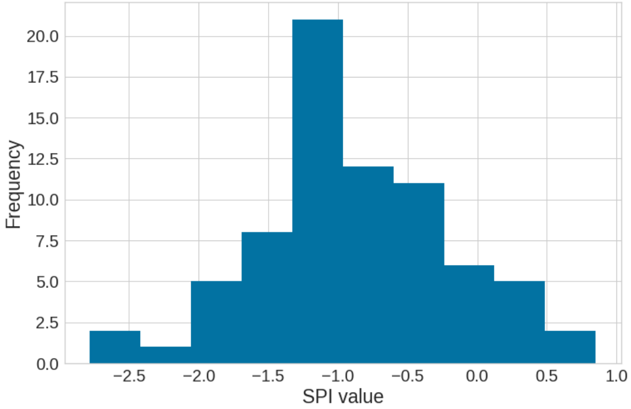

Fig. 4. SPI3 frequency histogram of months with reported droughts in the department of Magdalena from 2010 to 2019

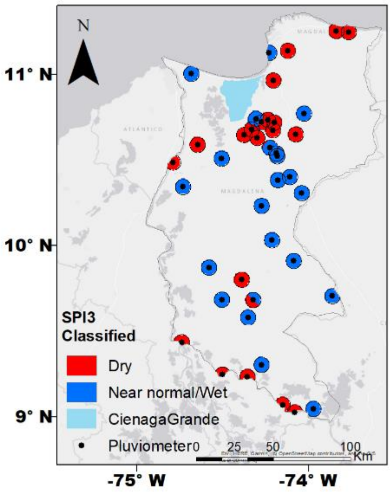

Fig. 5. Map of SPI3 classified into drought and normal/wet conditions

B. Predictor variables

Initially, 11 predictor variables were considered: the ONI, MEI, SOI, LST, precipitation, ET, SSM, SUSM, NDVI, NDWI, and NMDI. For the selection of the variables with the highest predictive capacity, the free Feature Selector from scikit-learn in Python was used, where the importance of each variable is calculated according to a gradient boosting machine (GBM) [38]. The results are presented in Fig. 6. The variables that contribute the least to the drought prediction are the NDWI, ONI, and MEI, so they were eliminated from the model. To establish the collinearity between the remaining variables, the Pearson correlation coefficient was calculated, as shown in Fig. 7. The SUSM and SSM variables show a high positive correlation (0.94), so the SUSM variable was not considered in the model due to its high collinearity and lower importance compared to SSM (see Fig. 6).

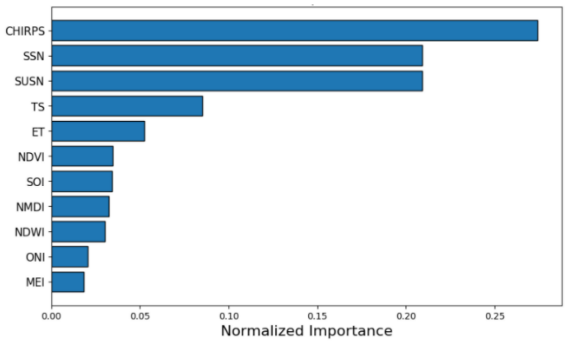

Fig. 6. Importance of each variable according to the Feature Selector

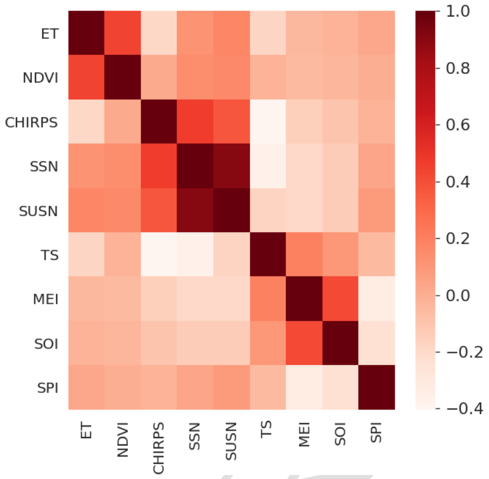

Fig. 7. Correlation of selected variables

C. *Assembled models*

For the construction of the drought prediction model, ensemble machine learning models were used. These models retain the properties of the base estimator but reduce the variance or fit problems that can affect model performance. Bagging-type ensemble models use estimators with good performance and build multiple models simultaneously, randomly selecting the observations and variables, in some cases. This is why they are used for problems where variance reduction is desired [39]. Boosting methods, on the other hand, use weak estimators, i.e., with poor performance, and build a new consecutive model, in which they assign weights to the observations erroneously predicted by the base estimator. In this way, in the end, a robust model is obtained that reduces the fitting problems of the initial models [39] [40].

In this work, we chose to select bagging models with decision tree-type estimators because of the good performance of these estimators when there is a large volume of observations (they yield excellent fits). The following is a brief description of the two bagging models used.

Bagging decision trees

The decision tree classifier (DTC) is based on fragmenting a complex decision into multiple simple decisions, with the objective that the final result gives a reason for the solution of the initial complex decision [41]. It is called a decision tree because simpler decisions are derived from the complex decision, and these in turn become even simpler decisions, thus forming a tree-shaped scheme in which the leaves represent the final answer to each question and the roots represent the complex decision to be addressed [41].

To reduce the variance associated with the decision tree model, subsets of data will be created by randomly extracting observations from the training data, thus creating different predictive models with each data set; the final result is the most repeated prediction within each subset [39].

Random forest

The random forest (RF) method uses the same concept of bagging decision trees, but the difference is that in RF, in addition to randomly selecting the observations of each subset, it also performs a random selection of variables to be used in each subset of data [42].

To run the supervised models, the sklearn package was implemented through Python [43]. According to the scheme shown in Fig. 8, initially the observations are randomly divided into training data (75%), which are used for validation curves, hyper-parameter fitting, learning curves, and re-training the model, and evaluation data (25%).

A validation curve refers to the result obtained by varying a hyper-parameter over a wide range of values, in order to delimit where the model performs best [44] by modifying only one hyper-parameter. On the other hand, with a learning curve, it is possible to visualize the behaviour or performance of the model as the number of observations increases, which makes it possible to establish whether the model has problems with fit or variance [44]. Within the two procedures described and in the search for the best set of hyper-parameters, cross-validation is used, as shown in Fig. 8. This consists of dividing the training data into subsets, in this case, 5, and in each iteration 4 subsets are used for training and the remaining subset is used for validation; thus, all the observations are used to both train and validate the model. The metric used to determine the performance in all the procedures described was recall, since it focuses on evaluating the accuracy or predictive ability of the class of interest, which in this case is the drought class. Finally, the final model, already calibrated, uses the evaluation data, i.e., 25% of the data, to predict the response variable; then, it is possible to discover the predictive capacity of the model by comparing the simulated SPI3 with the measured values.

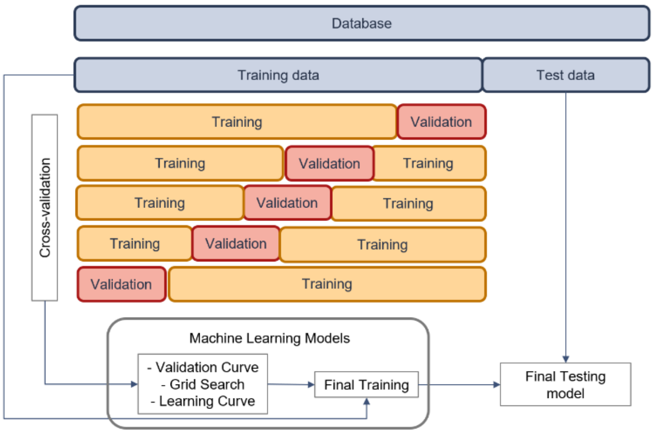

Fig. 8. Framework of the application of RF and DTC models

V. DROUGHT FORECASTING MODEL

A. Bagging decision tree

Table 4 presents the hyper-parameters that optimize the model results in terms of recall. Fig. 9 shows the learning curve of the model: it is possible to observe that, as more observations are added, both the validation curve and the training curve increase the recall value; likewise, both curves tend to approach each other, which indicates a reduction in the variance of the problem.

Table 4. Variation of hyper-parameters for bagging decision trees

Hyper-parameters		
Changing	Range of values	Best value
min samples leaf	20, 30, 40, y 50	40
splitter	Best o Random	Best
max features	Sqrt, Log2, o None	Sqrt
Constant	Value	
Class weight	Balanced	
Criterion	Entropy	
Random state	0	

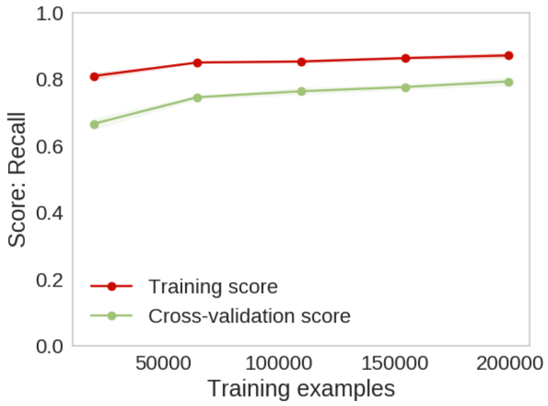

Fig. 9. Learning curve for bagging decision trees

Table 5 presents the classification report and confusion matrix using the evaluation data (25%): it can be observed that false negatives (FNs) – drought events not identified by the model – represent 2.5% of the total evaluation data. False positives (FPs) – drought events erroneously identified by the model – represent 21% of the total evaluation data. In fact, the number of FPs is higher than the number of true positives (TPs) – drought events identified by the model –. This percentage is significant and indicates that the model tends to overestimate the number of drought events. The accuracy for predicting drought is 0.33, while the recall is 0.8, which is in accordance with the percentages of FPs and FNs, respectively.

Table 5. Classification report and confusion matrix results for bagging decision trees

	Precision	Recall	F1 Score	Support
Near normal / wet	0.96	0.76	0.85	72374
Dry	0.33	0.8	0.47	10636
Average	0.88	0.76	0.8	83010
Evaluation data: 83010	TN: 54919	FN: 2089	TP: 8547	FP: 17455

B. Random forest

The set of hyper-parameters for random forest that yielded the best drought prediction are presented in Table 6.

Table 6. Variation of hyper-parameters for random forest

Hyper-parameters		
Changing	Range of values	Best value
min_samples_leaf	20, 30, 40, y 50	20
n estimators	90, 100, y 150	100
class weight	Balanced o Balanced subsample	Balanced
max features	Sqrt, Log2, o None	Sqrt
Constant	Value	
min samples leaf	20	
Criterion	Entropy	
Random State	0	

Fig. 10 presents the learning curve obtained with RF, showing that the model improves performance, in terms of recall, as the amount of data increases, in both the training and validation curves.

Likewise, it is possible to observe that both curves tend to converge, which indicates that the variance of the problem is reduced as the model improves the learning process by increasing the amount of data.

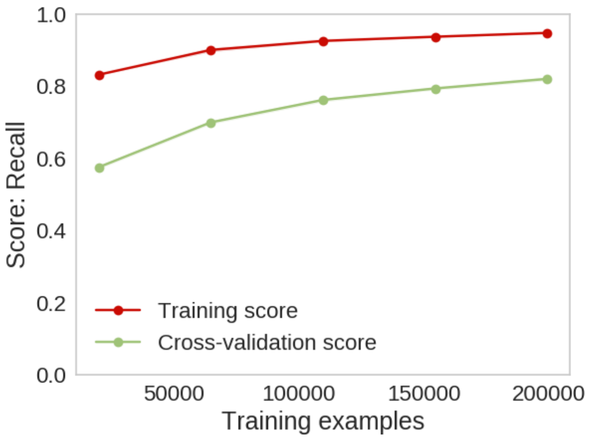

Fig. 10. Learning curve for random forest

The results of the model with the evaluation data (25%) are shown in Table 7. The percentages of FNs and FPs are 2% and 7.5%, respectively, indicating that the model with RF tends to reduce FNs over FPs. This is reflected in the precision and recall obtained (0.59 and 0.84, respectively).

Tabla 7. Reporte de la Clasificación y Resultados de la Matriz de Confusión para Bosque Aleatorio.

	Precision	Recall	F1 Score	Support
Near normal / wet	0.97	0.91	0.94	72374
Dry	0.59	0.84	0.69	10636
Average	0.93	0.9	0.91	83010
Evaluation data: 83010	TN: 66100	FN: 1706	TP: 8930	FP: 6274

C. Spatial prediction of drought

To discover the spatial distribution of droughts in the entire department of Magdalena, we proceeded to take the distributed values for each of the selected predictor variables, and, using the RF model constructed, we estimated the SPI3 value for the entire department.

The month of July 2014 was selected for the spatial validation of the SPI3 results. The press reports of the month in question stated that the municipalities of Santa Marta, Plato, Zapayán, Concordia, and Tenerife declared a public calamity due to drought; and the municipalities of San Sebastián, San Zenón, San Ángel, Nueva Granada, and Pivijay were close to declaring it. Reports indicated that 70% of the department’s crops were affected, 4,300 head of cattle died, and there were large forest fires [45]. In response to the emergency, 65,000 liters of water were provided to 100,000 families in the municipalities of Santa Marta, Zapallán, and Concordia [46]. Due to the state of emergency, on 1 August 2014, the UNGRD declared a public calamity for the entire department of Magdalena.

The results of the prediction model, using RF, are shown in Fig. 11; Fig. 11A shows the prediction of the response variable and Fig. 11B presents the probability of the occurrence of drought.

To evaluate the spatial predictive capacity of the model, the modelled SPI3 pixels within the area of influence of each station were taken and compared with the SPI3 calculated from IDEAM rainfall stations. Table 8 presents the results, in which the percentages of FNs (0.7%) and FPs (29.6%) with respect to the total amount of data evaluated indicate an overestimation of the pixels with drought. Additionally, the accuracy and recall for predicting drought are 0.59 and 0.98.

Table 8. Report of the classification and confusion matrix results for random forest in July 2014

	Precision	Recall	F1 Score	Support

Near normal / wet	0.97	0.48	0.64	1559
Dry	0.59	0.98	0.74	1208
Average	0.78	0.73	0.69&	2767
Evaluation data: 2767	TN:741	FN:19	TP:1189	FP:818

According to Fig. 11A, 60.4% of the territory of the department of Magdalena experienced drought conditions. Fig. 11 highlights the municipalities that reported drought conditions for that month. It can be seen that, according to the model, drought is not present in all the municipalities reported, i.e., there are sectors with wet-to-normal conditions, which indicates that there may be areas that are more affected than others by the occurrence of the climatic phenomenon. Additionally, there are sectors that, according to the model used, present a high probability of drought, and these were not reported within the municipalities that declared a public calamity in the month in question.

A more detailed description of the events would help to strengthen the validation of the results. This would facilitate both the evaluation of SPI3 as an index to measure agricultural drought and a more detailed validation of any method developed to evaluate drought within the department.

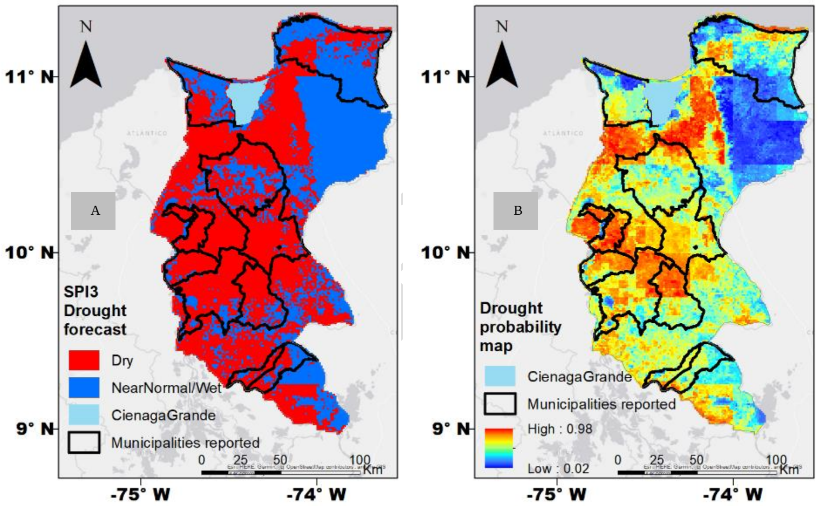

Fig. 11. Application of the random forest model for July 2014 and for municipalities experiencing a public calamity due to drought for the month in question. A) Drought forecast, according to SPI3. B) Drought probability map

VI. CONCLUSIONS AND DISCUSSION

The results obtained in this work indicate that ML methods are a tool with an important potential for predicting the temporal and spatial occurrence of droughts in the Colombian territory.

There is a wide range of ML methods. The present study indicates that the assembled methods perform adequately, with appropriate values for both model performance and spatial and temporal predictive capability.

Both RF and DTC models predict drought within the department in a timely manner; however, the accuracy of DTC is much lower than that obtained by RF, indicating that DTC greatly overestimates the occurrence of drought events compared to RF.

The model developed, in addition to providing a spatialized map of the occurrence of drought within the department, provides a map of the probabilities of the occurrence of the event, which could help local authorities to make decisions about how the emergency is distributed in the territory and to discover the sites with the highest probability of occurrence. This would allow them to determine the sectors most affected by the event and thus to deliver resources to these priority locations in a drought emergency within the department.

The information necessary to develop the proposed methodology is free and accessible to the public. This fact is relevant because it opens up the possibility of replicating the described workflow in other departments of Colombia and making a continuous and progressive follow-up of the behaviour of the phenomenon, facilitating studies based on what has been observed and the implementation of mitigation strategies. The above can be implemented within the national strategy for the integral management of drought in Colombia, as a mechanism within the objective of strengthening the monitoring and follow-up of drought.

As mentioned previously, the use of satellite images brings with it a great advantage in terms of the spatial distribution of information, but it is important to highlight the limitations that this type of data entails. For example, the low spatial resolution of some implemented variables such as the SSM, SUSM, and precipitation, which have resolutions of 28 km, 28 km, and 5.5 km, respectively, prevents the application of the proposed methodology in areas with little geographical extension, because there would not be enough information that was representative of the conditions of the territory. In this same sense, to the extent that satellites or variables that provide a better spatial resolution are integrated, a more detailed study and applicability in areas with low territorial extension may be possible.

On the other hand, the possibility of extending the applicability of the SPI is planned, i.e., taking advantage of the fact that the index has a normal distribution, it is possible to simultaneously evaluate both dry and wet conditions within the territory. This in turn makes it possible to study the impacts produced not only by drought, but also by an excess of humidity or precipitation within the territory.

The type of drought evaluated within the study is agricultural drought, so the precipitation deficit is studied in an accumulated period of 3 months. This is an initial approach according to the literature, but it is possible to adjust the accumulated months to evaluate the behaviour of certain vegetation or relevant crops within the economy of each department; within the Colombian territory, bananas and potatoes, among others, are especially relevant. Thus, it is feasible to implement within the models specific information related to the behaviour and needs of the crops, such as water requirements and in which seasons the harvest, cultivation, and growth occur. This allows a comprehensive assessment of drought from the perspective of the food security of the population and economic impacts.

The variable to be implemented to evaluate drought within the study is the SPI, due to the ease with which this variable is obtained. It would be important in later studies to implement or evaluate different indexes that allow the evaluation of drought, such as the Standardized Precipitation and Evapotranspiration Index (SPEI), Effective Drought Index (EI), and Palmer Drought Index (PDSI), among others, which integrate, in addition to rainfall, other types of meteorological variables that would allow researchers to cover different aspects that are key in the occurrence or determination of drought.

REFERENCES

[1] D. A. Wilhite, M. Sivakumar, D. A. Woodet, et al., “Early warning systems for drought preparedness and drought management,” in Proceedings of an expert group meeting held in Lisbon, Portugal, vol. 57, 2000.

[2] T. B. McKee, N. J. Doesken, J. Kleistet, et al., “The relationship of drought frequency and duration to time scales,” in Proceedings of the 8th Conference on Applied Climatology, Boston, vol. 17, no. 22, pp. 179–183, 1993.

[3] M. Svoboda, M. Hayes, and D. Wood, “Índice normalizado de precipitación,” Guía de usuario, Organización Meteorológica Mundial, pp. 1–23, 2012.

[4] FAO, Sequía, 2020. [Online]. Available: http://www.fao.org/emergencies/tipos-de-peligros-y-de-emergencias/sequia/es/

[5] W. Cramer, G. Yohe, C. B. Field, et al., Detection and Attribution of Observed Impacts. Cambridge University Press, 2014.

[6] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tig-nor, and H. Miller, “Climate change 2007: the physical science basis,” Inter-governmental Panel on Climate Change (IPCC), Cambridge University Press, Cambridge, 2007.

[7] A. Dai, T. Zhao, and J. Chen, “Climate change and drought: A precipitation and evaporation perspective,” Current Climate Change Reports, vol. 4, no. 3, pp. 301–312, 2018.

[8] S. Mukherjee, A. Mishra, and K. E. Trenberth, “Climate change and drought: a perspective on drought indices,” Current Climate Change Reports, vol. 4, no. 2, pp. 145–163, 2018.

[9] Revista SEMANA, Las graves secuelas económicas de la sequía, 2014. [Online]. Available: https://www.semana.com/nacion/articulo/las-graves-secuelas-economicas-de-la-sequia/396750-3

[10] E. Heraldo, Magdalena, el más azotado por la temporada de sequía, 2015. [Online]. Available: https://www.elheraldo.co/magdalena/magdalena-el-mas-azotado-por-la-temporada-de-sequia-219590

[11] O. Rahmati, F. Falah, K. S. Dayal, R. C. Deo, F. Mohammadi, T. Biggs, D. D.Moghaddam, S. A. Naghibi, and D. T. Bui, "Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia," Science of the Total Environment, vol. 699, p. 134230, 2020

[12] S. Park, J. Im, E. Jang, and J. Rhee, "Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions," Agricultural and Forest Meteorology, vol. 216, pp. 157–169, 2016.

[13] K. F. Fung, Y. F. Huang, C. H. Koo, and M. Mirzaei, "Improved SVR machine learning models for agricultural drought prediction at downstream of Langat River Basin, Malaysia," Journal of Water and Climate Change, 2019.

[14] P. Feng, B. Wang, D. Li Liu, and Q. Yu, "Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in south-eastern Australia," Agricultural Systems, vol. 173, pp. 303–316, 2019.

[15] A. Belayneh, J. Adamowski, B. Khalil, and J. Quilty, "Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction," Atmospheric Research, vol. 172, pp. 37–47, 2016.

[16] X. Liu, X. Zhu, Q. Zhang, T. Yang, Y. Pan, and P. Sun, "A remote sensing and artificial neural network-based integrated agricultural drought index: Index development and applications," Catena, vol. 186, p. 104394, 2020.

[17] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Informatics, vol. 2, p. 117693510600200030, 2006.

[18] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, pp. 1060–1089, 2013.

[19] K. Rasouli, W. W. Hsieh, and A. J. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," Journal of Hydrology, vol. 414, pp. 284–293, 2012.

[20] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," Geoscience Frontiers, vol. 30, p. 1e9, 2015.

[21] Min Ambiente, UNGRD, IDEAM, "Estrategia Nacional para la gestión integral de la sequía en Colombia," 2018. [Online]. Available: https://knowledge.unccd.int/sites/default/files/countryprofiledocuments/ENGIS%2520para%2520publicaci%25C3%25B3nColombia.pdf

[22] UNGRD, Consolidado anual de emergencias, 2020. [Online]. Available: http://portal.gestiondelriesgo.gov.co/Paginas/Consolidado-Atencion-de-Emergencias.aspx

[23] Gobernación del Magdalena, Nuestro departamento, 2020. [Online]. Available: http://www.magdalena.gov.co/departamento/nuestro-departamento

[24] DANE, Resultados censo nacional de población y vivienda, 2018. [Online]. Available: https://www.dane.gov.co/files/censo2018/informacion-tecnica/presentaciones-territorio/191004-CNPV-presentacion-Magdalena.pdf

[25] M. y. E. A. I. Instituto de Hidrología, Magdalena, 2020. [Online]. Available: http://atlas.ideam.gov.co/basefiles/magdalenatexto.pdf

[26] K. E. Trenberth, A. Dai, G. Van Der Schrier, P. D. Jones, J. Barichivich, K. R. Briffa, and J. Sheffield, "Global warming and changes in drought," Nature Climate Change, vol. 4, no. 1, pp. 17–22, 2014.

[27] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sensing of Environment, vol. 202, pp. 18–27, 2017.

[28] NASA, Normalized difference vegetation index (NDVI), 2020. [Online]. Available: https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuringvegetation2.php

[29] T. L. T. Du, D. D. Bui, M. D. Nguyen, and H. Lee, "Satellite-based, multi-indices for evaluation of agricultural droughts in a highly dynamic tropical catchment, central Vietnam," Water, vol. 10, no. 5, p. 659, 2018.

[30] L. Wang and J. J. Qu, "NMDI: A normalized multi-band drought index for monitoring soil and vegetation moisture with satellite remote sensing," Geophysical Research Letters, vol. 34, no. 20, 2007.

[31] G. Poveda and ´O. J. Mesa, "Las fases extremas del fenómeno ENSO (El Niño y La Niña) y su influencia sobre la hidrología de Colombia," Tecnología y ciencias del agua, vol. 11, no. 1, pp. 21–37, 2015.

[32] D. C. Edwards, "Characteristics of 20th century drought in the United States at multiple time scales," Air Force Institute of Tech Wright-Patterson AFB OH, Tech. Rep., 1997.

[33] O. M. Valiente, "Sequía: definiciones, tipologías y métodos de cuantificación," Investigaciones geográficas, no. 26, pp. 59–80, 2001.

[34] Mayorga, R., y Hurtado, G. IDEAM, La sequía en Colombia, 2006. [Online]. Available: http://www.ideam.gov.co/web/tiempo-y-clima/notas-tecnicas-sequia?p_p_id=110_INSTANCE_emIh6vh8JqRp&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_count=1&_110_INSTANCE_emIh6vh8JqRp_struts_action=%2Fdocument_library_display%2Fview_file_entry&_110_INSTANCE_emIh6vh8JqRp_redirect=http%3A%2F%2Fwww.ideam.gov.co%2Fweb%2Ftiempo-y-clima%2Fnotas-tecnicas-

sequia%3Fp_p_id%3D110_INSTANCE_emIh6vh8JqRp%26p_p_lifecycle%3D0%26p_p_state%3Dnormal%26p_p_mode%3Dview%26p_p_col_id%3Dcolumn-1%26p_p_col_count%3D1&_110_INSTANCE_emIh6vh8JqRp_fileEntryId=69501430

[35] Hurtado, G, Sequía meteorológica y sequía agrícola en Colombia: Incidencia y tendencias, 2012. [Online]. Available: http://www.ideam.gov.co/documents/21021/21138/Sequias+Incidencias+y+Tendencias.pdf/3e72c86c-cf4a-42f9-95f1-07e7cf88861a

[36] J. Gómez and M. Cadena, Actualización de las estadísticas de la sequía en Colombia, 2017. [Online]. Available: http://www.ideam.gov.co/web/tiempo-y-clima/notas-tecnicas-sequia?p_p_id=110_INSTANCE_emIh6vh8JqRp&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_count=1&_110_INSTANCE_emIh6vh8JqRp_struts_action=%2Fdocument_library_display%2Fview_file_entry&_110_INSTANCE_emIh6vh8JqRp_fileEntryId=76813238

[37] University of Nebraska, SPI program, 2020. [Online]. Available: https://drought.unl.edu/droughtmonitoring/SPI/SPIProgram.aspx

[38] W. Koehrsen, A feature selection tool for machine learning in Python, 2020. [Online]. Available: https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0

[39] C. D. Sutton, “Classification and regression trees, bagging, and boosting,” Handbook of Statistics, vol. 24, pp. 303–329, 2005.

[40] E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,” Machine Learning, vol. 36, no. 1-2, pp. 105–139, 1999.

[41] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660–674, 1991.

[42] M. Pal, “Random forest classifier for remote sensing classification,” International Journal of Remote Sensing, vol. 26, no. 1, pp. 217–222, 2005.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourget, et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[44] Scikit-Learn, scikit-learn machine learning in Python, 2020. [Online]. Available: https://scikit-learn.org/stable/index.html

[45] WRadio, Declaran calamidad pública por sequía en cinco municipios del magdalena, 2014. [Online]. Available: https://www.wradio.com.co/noticias/actualidad/declaran-calamidad-publica-por-sequia-en-cinco-municipios-del-magdalena/20140727/nota/2341212.aspx

[46] El Colombiano, La sequía impacta a 7 departamentos, 2014. [Online]. Available: https://www.elcolombiano.com/historico/lasequiaimpactaa7departamentos-IGec30364923