Early prediction of frost events in high altitude crops, using machine learning methods[1]

Evelin Calderón Caro[2*], *Darío Antonio Castañeda Sánchez*[3], *John Willian Branch Bedoya*[2]

10.1590/0034-737X2024710040

ABSTRACT

In the tropic, many crops are distributed in the highlands of provinces of the Andean regions at heights of 2,500 m asl and constitute the areas with the highest susceptibility to the frost events occurrence. The study objective was to propose an early frost prediction model based on the relationships between frost events and climatic variables, modeled with machine learning methods. The climatic variables were obtained from thirteen meteorological stations located inside flower crops and distributed in nine municipalities of the Cundinamarca Department. The variables registered were temperature, relative humidity, dew point, photosynthetically active radiation, and precipitation, entered as explanatory variables of frost events. The metrics used for predictive performance evaluation of the five machine learning methods examined were precision, recall, true negative rate, accuracy, and F1 score. The variables' climatic behavior of previous hours to a frost event are low humidity, wind speed and cloudiness, and high thermal radiation. The fourth of the five trained models performed well due to their classification evaluation metrics, greater than 91%. The cross-validation and statistical analysis demonstrated the higher accuracy of the GBDT model on frost events detection.

Keywords: forecast; artificial neural networks; gradient boosting; climatic variables.

INTRODUCTION

Temperature drives the latitudinal and elevational limits of plant species distribution across the earth. In some plants, freezing conditions define the lower temperature limit, suggested as the main driver of the species reduction, which increases with the latitude. Likewise, in tropical mountains, the upper elevational limit of many vascular plants is determined by freezing condition that expands in open areas such as tropical alpine ecosystems (e.g., Paramos) or transformed landscapes for agriculture. The tissues of the tropical plants, exposed to frost events (temperature < 0°C), can be damaged due to cellular dehydration and membrane disintegration. At this temperature, the water in the extracellular spaces freezes to form ice crystals, causing cell death (Kochhar & Gujral, 2020; Rout, 2020). Effects associated with this phenomenon include reduction in leaf size, wilting, chlorosis, necrosis, and poor reproductive development (Kochhar & Gujral, 2020). In addition, photosynthetic rate, respiration, and protein synthesis rate can decrease due to temperature-dependent processes (Li *et al*., 2018; Kochhar & Gujral, 2020). Then, the magnitude of these effects varies according to plant species traits, plant age, and exposition time to these low temperatures.

Frost events have the highest probability of occurrence in paramos or transformed highland landscapes of the tropical alpine ecosystems. These events can affect the performance of many plants' species, including multiple

Submitted on May 26th, 2022 and accepted on July 29th, 2024.
[1] This work is part of the first author's Master Dissertation and it was funded by Soluciones Wiga and Growers Hub Trading Group Companies.
[2] Universidad Nacional de Colombia, departamento de Ciencias de la Computación y de la Decisión, Grupo de Investigación y Desarrollo en Inteligencia Artificial, Medellín, Antioquia, Colombia, evcalderonca@unal.edu.co; jwbranch@unal.edu.co
[3] Universidad Nacional de Colombia, departamento de Ciencias Agronómicas, Grupo de Investigación Fitotecnia Tropical, Medellín, Antioquia, Colombia. dacasta4@unal.edu.co
*Corresponding author: evcalderonca@unal.edu.co

crops, that cohabit in these ecosystems. In Colombia, many crops, such as flowers, potatoes, cane, coffee, and corn, are distributed in the highlands of provinces of the Colombian Andean, such as Cundinamarca, Boyacá, Norte de Santander, and Nariño, and are the main economic activity of these. In the Andes, the distribution of the referred crops can happen at heights of 2,500 m asl and constitute the areas with the highest susceptibility to the occurrence of frost events, favored for being highly transformed areas and unprotected from exposure to wind and cold (González & Torres, 2012). Historically, frost events have occurred more frequently between December and February (González & Torres, 2012). This period coincides with the flower production cycle for Valentine's season, corresponding to one of the highest sales peaks of the year for flowers, about 15% of annual sales. For example, in Colombia, the sale of flowers in Valentine's season in 2021 was 700 million stems worth US$654 billion (Becerra, 2021). Therefore, understanding the drivers of frost events and forecasting these across Colombian highlands is essential to safeguard flower crops and maintain their production during this critical period.

Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) is a public body linked to Colombian Ministerio de Ambiente y Desarrollo Sostenible's focus on generating knowledge and guaranteeing access to information on hydrometeorological conditions throughout the country. Moreover, the Agroclimatic Technical Table, a government initiative to avoid losses in the agricultural sector caused by climate variability, has provided recommendations for risk prevention due to frost events. Among the main recommendations are planting plans that consider climatic phenomena, establishment reduction of susceptible crops, live fence implementation, and irrigation system adaptation (Ministerio de Agricultura y Medio Ambiente, 2020). Other mitigation strategies include soil adaptations, energy supply through heat sources or air heaters, micro-sprinklers, or fan use (Simnitt *et al*., 2017; González & Torres, 2012). In addition, other strategies have been proposed based on genetics and cell engineering (Marmolejo & Ruiz, 2018), chemical treatment and thermal conditioning by intermediate heating and heat stress (Joshi *et al*., 2020; Kochhar & Gujral, 2020). The development of some of these strategies can take time and are expensive. Traditional strategies require accurate forecasting methods that allow producers to anticipate, prepare for and mitigate a possible frost event on flower production.

The lack of preparation for frost events can cause economic losses, especially in the quantity and quality of crops, and can even cause exposed crops ipso facto destruction (Juurakko *et al*., 2021). Specifically, in 2020, frost events generated losses in 2.0% of the productive area in Cundinamarca (5400 hectares), especially in the northwestern region of the department (Vargas, 2021). In this way, improving our ability to forecast frost events in Colombian highlands, particularly in Cundinamarca and Boyacá, is essential, as support, for the mitigation strategies implementation of previous frost events.

Among the most recent methodologies used to climate prediction and specifically for frost events prediction, supervised machine learning techniques such as artificial neural network (Diedrichs *et al*., 2018; Latif *et al*., 2020), decision tree (Lee *et al*., 2016), random forest (Diedrichs *et al*., 2018) and support vector machine (Ding *et al*., 2019) have been used. These techniques have been used to identify when a reduction in temperature below 0°C is going to occur, and to predict the temperature behavior and the minimum value that temperature will reach.

Therefore, our objective was to propose an early frost prediction model based on climatic variables and machine learning methods, as support, for the anticipation of the implementation of mitigation and adaptation strategies, against the damage caused by this factor in Colombian floriculture.

MATERIAL AND METHODS

Area of study

Altiplano Cundiboyacense presents an altitude range from 460 to 4,240 m a.s.l., therefore, there is a wide range of biogeographical units (Rivera *et al*., 2004 cited by Gómez *et al*., 2021). The region bears a typical humid tropical mountain climate with seasonal rainy periods. The precipitation varies between 580 and 1,000 mm year^{-1}, with a bimodal pattern, with higher values from April to May and between October to November, influencing the ecosystems strongly (Aguilar & Torres, 2010; Guhl, 2013). The region has an annual average temperature of 13.5 °C and keeps a wide range of daily oscillations between 3 °C and 28 °C. Temperatures below 2 °C generally occurs in the early mornings of the dry months (Aguilar & Torres, 2010; Guhl, 2013).

Thirteen meteorological stations located inside flower crops, distributed in nine municipalities of the Cundinamarca department (Chía, Facatativá, Guachancipá,

Madrid, Nemocón, Sesquilé, Suesca, Tocancipá, Ubaté, and Zipaquirá) at heights between 2,562 and 2,584 m a.s.l, were used to obtain climatic information (Figure 1). Wiga SAS and Growers Hub Trading (GHT) are the companies that manage the stations.

The records acquired between 2017 and 2021 had a resolution of one hour (24 records day^{-1}). The total number of records was 99,336. Given the periodic nature of the frost phenomenon, this study focused on the dry season of the year, which corresponds to November, December, January and February, months. Frost events as response variable and five climatic variables as explanatory variables were collected from all stations. Climatic variables were temperature (T, °C), photosynthetically active radiation or PAR (µmol m^{-2} s^{-1}), relative humidity (RH, %), dew point (DP, °C) and precipitation (PP, mm m^{-2} h^{-1}) (Figure 2). The response variable, frost events, was managed as a binary variable, taking the value of one (1) from the previous 24 hours until the day in which the temperature was less than 0°C. In other words, by the time a frost event occurred, the magnitude of the explanatory climatic variables 24 hours of the previous day was associated with a value of one in the response variable. In any other case, the response variable took the value of zero, associated with the corresponding climatic variables' magnitudes. Each explanatory variable refers to a combination of climatic variables and hour of the day; therefore, the resulting dataset for training the models was composed of 123 columns (5 climatic variables x 24 hours of the day), including the day, month, and year of each record.

Descriptive analysis

We used the most common statistics (minimum, 1st, 2nd, and 3rt quartile, maximum, kurtosis, skewness) to describe the annual behavior of each climatic variable and compare with frost season (from November to February) behavior. It was made a detailed description of the climatic variables to differentiate the average hourly behavior of each variable, during the 24 hours of the day prior to a frost event occurrence (FE) and the average behavior of each variable for the 24 hours of the day prior to a frost event non-occurrence (FEN). The hours were labeled using the 24-hour time system where 0 corresponds to 12 A.M. and 23 indicates 11 P.M. We also evaluated the type (direct or indirect) and relationship degree between frost events and climatic variables per hour through standardized beta coefficient of logistic regression.

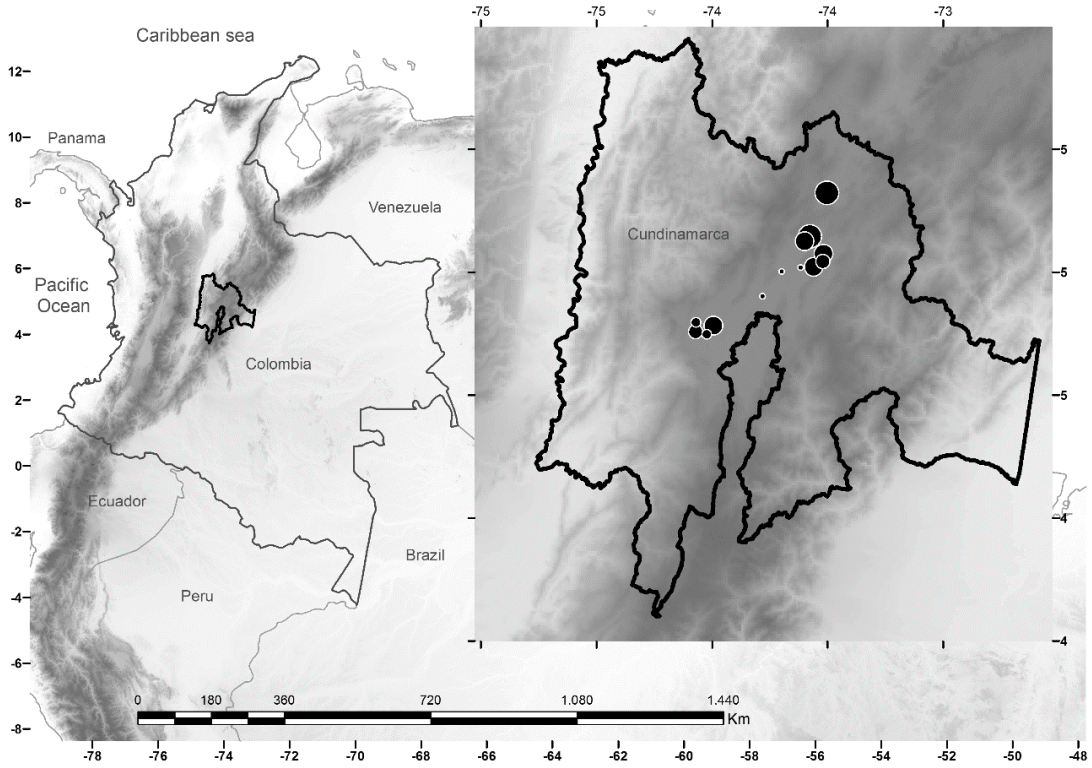

Figure 1: Area of study in Colombia and meteorological station's location in Cundinamarca department in the right map (circle size corresponds to amount of data).

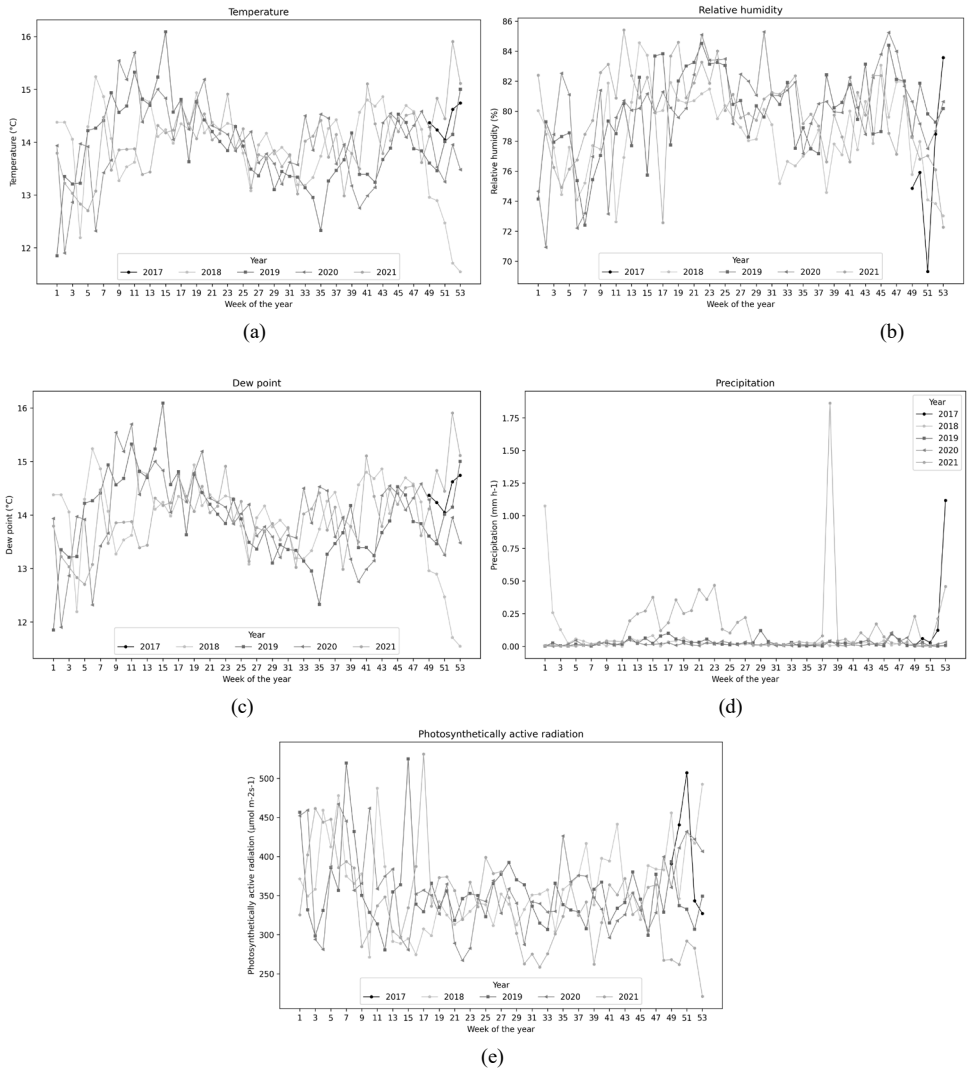

Figure 2: Temporal distribution of climatic variables grouped by the mean value per week for each year: (a) temperature, (b) relative humidity, (c) dew point, (d) precipitation and (e) photosynthetically active radiation.

Preprocessing data

The missing data were imputed and reinstated with the nearest neighbor approximation of the KNNImputer, Sklearn Impute library of python, and standardized with Sklearn StandarScale function to avoid the effect of their magnitude. Machine learning classifiers are sensitive to the imbalance presented by the frost events response variable between positive and negative cases. The imbalance was improved by resampling the positive events in 2% of the database, a procedure executed using the oversampling method SMOTE (Synthetic Minority Oversampling Technique). The climatic variables selection most related to frost events and the degree of relationship according to the time of day were based on standardized beta coefficient of logistic regression, principal component analysis (PCA), and features selection with the highest scores (Sklearn SelectKBest function).

Predictive models

We split dataset into three subsets; the first with 70% for the model's training performing five-fold cross-validation, the second (20%) for testing, and the third (10%) for validation. Performance metrics (precision, recall, positive negative rate, accuracy, and F1 score) were calculated based on test and validation data. The predictive models used were the Logistic regression (Lee *et al*., 2016), Decision Tree (Charbuty & Abdulazeez, 2021), Gradient Boosting Decision Trees (Danandeh, 2021), Support Vector Machine (Pedregosa *et al*., 2011; Dinh *et al*., 2021), and Neural Network (Pedregosa *et al*., 2011; Fuentes *et al*., 2018). The optimization of the learning and fit of the algorithms was done with the pipeline and gridsearcCV functions of the Sklearn library and the grid search method as the hyperparameter tuning.

Model performance evaluation metrics and comparison

The quantification and comparison performance metrics of prediction models were precision (proportion of correctly classified cases out of all the actual positives), recall (proportion of positive instances correctly classified), true negative rate (proportion of negative instances correctly classified), accuracy (proportion of correctly classified cases out of all the data points), and F1 score (weighted average between precision and sensitivity) (Fuentes *et al*., 2018). Model performance evaluation metrics was computed for all models with test and validation datasets.

TP is the number of predicted values; in the other words, it is true positives or correctly predicted frost events; FP is the false positive values or false forecasted frosts events; FN is false negatives forecasted frosts events. TN corresponds to true negatives or days without frosts occurrence correctly predicted. These categories were specified using the contingency table and confusion matrix for a binary classifier (Diedrichs *et al*., 2018). Two additional performance parameters used were the proportion of true positives versus false positives proportion (Cho *et al*., 2021), denominated receiver operating characteristic curve (ROC), and the area under the curve precision-recall curve (AUC-PR) (DeVries *et al*., 2021).

RESULTS AND DISCUSSION

Climatic characterization of the region

The study region presented an annual average temperature of 13.6 °C with values displayed between -3.1 and 28.7 °C, similar to those reported by Aguilar & Torres (2010), Guhl (2013) and Mayorga *et al*. (2020); however, 75% of temperature values were below 17.1 °C in the period evaluated. Concerning RH, the annual average was 79.8%, with a median of 86.5% and values between 8.3 and 100%; this annual mean value was within the average range specified by Luengas *et al*. (2021) for Cundinamarca department (60 to 80%). In addition, the annual mean PAR, DP, and PP were 370.9 (0 to 3,300) µmol m^{-2} s^{-1}, 9.7 (-10.6 to 19.9) °C, and 0.02 (0 to 60.4) mm m^{-2} h^{-1} respectively, as shown in table 1. The annual mean PAR was among the values reported by Mayorga *et al*. (2020) for the municipality of Pasca (Cundinamarca), located at an altitude of 2,498 m a.s.l. These authors recorded average monthly PAR values between 300 and 500 µmol m^{-2} s^{-1}.

According to the kurtosis coefficient for annual evaluation (Table 1), T and RH presented a low concentration of values around their mean and therefore associated with platykurtic curve distribution (Kurt < 0). However, this coefficient can vary for the same climatic variable monitored in a specific location by different sensors, as evidenced by Trilles *et al*. (2021). These authors determined kurtosis values between -0.7 and 11.98 for temperature and between -1.2 and -0.8 for humidity for Castellón, Spain location.

On the other hand, PAR and DP presented a leptokurtic distribution due to the high concentration of values around the mean (Kurt > 0). The high kurtosis coefficient value for PP was because 75% of values correspond to zero; in other words, PP events correspond to less than 25% for the period evaluated. The kurtosis behavior in the year, frost season, and non-frost season periods was similar. However, some authors have reported differences in data distribution when comparing various seasons (Brito *et al*., 2019). The kurtosis coefficient for each climatic variable was different from zero, and thus, their distribution is not normal.

Skewness coefficient showed a temperature distribution close to symmetry (Skew = 0.1), while RH and DP presented a negative bias; with lengthening to values lower than the mean in the distribution (Skew < 0); on the contrary Brito *et al*. (2019) found negative skewness coefficients for temperature and positive values for relative humidity. Otherwise, PAR and PP showed a positive asymmetry (Skew > 0). Additionally, skewness behavior did not vary for the three periods evaluated.

The temperature showed in the frost season an average temperature equal to that average annual (13.6 °C); however, the minimum T for months not associated with frost events was 3.1 °C higher than the minimum recorded in the frost season. The RH, DP, and PP average values were higher at 3.7, 8.7 and 66.7%, respectively, for the months, between March and October compared to frost season (table 1). PAR was the only climatic variable whose average value was higher in the frost season (367.7 *µmol* m^{-2} s^{-1}) compared to the annual average value (364.9 µmol m^{-2} s^{-1}) and non-frost season average value (362.6 µmol m^{-2} s^{-1}). On the other hand, stations that recorded the higher number of frost events were PL (25.4%), UB (12.3%), FV (10.9%) and ER (10.9%), located in the northwestern of the Cundinamarca department.

Additionally, the average climatic variables during the 24 hours before a FE day and FEN day evidenced differences, as shown in Figure 3. The hour of the day with the

Table 1: Descriptive analysis of climatic variables throughout the year (annual) compared to climatic behavior in frost season (Nov – Feb) and non-frost season (Mar – Oct)

Climatic variable*	Statistics									
	Mean	SD[1]	Min[2]	Median	Q1[3]	Q2[4]	Q3[5]	Max[6]	Kurt[7]	Sw[8]
Annual										
T (°C)[9]	13.6	4.7	-3.1	12.9	10.5	12.9	17.1	28.7	-0.3	0.1
RH (%)[10]	79.8	17.8	8.3	86.5	66.3	86.5	95	100	-0.3	-0.8
PAR (***μmol*** m^{-2} s^{-1})[11]	364.9	587.1	0	9.7	0	9.7	564.5	3300	2.9	1.8
DP (°C)[12]	9.7	2.5	-10.6	10.2	8.8	10.2	11.3	19.9	8.3	-1.9
PP (mm m^{-2} h^{-1})[13]	0.02	0.5	0	0	0	0	0	60.4	8626.3	78.47
Frost season										
T (°C)	13.6	5.3	-3.1	12.8	9.9	12.8	18.0	28.7	-0.7	0.1
RH (%)	78.7	18.6	10.5	86	63.6	86.0	94.6	100	-0.5	-0.8
PAR (***μmol*** m^{-2} s^{-1})	367.7	580	0	3.4	0	3.4	568.3	3300	2.2	1.7
DP (°C)	9.4	2.7	-14.4	10	8.1	10	11.3	19.9	1.4	-1
PP (mm m^{-2} h^{-1})	0.01	0.7	0	0	0	0.2	0.4	0.9	7237.3	118.6
Non-frost season										
T (°C)	13.7	4	0	13	11	13	16.7	28.1	-0.1	0.3
RH (%)	81.1	15.7	10	86.6	69.3	86.6	94.6	100	-0.4	-0.8
PAR (***μmol*** m^{-2} s^{-1})	362.6	572.8	0	12.9	0	12.9	551.6	3300	3.5	1.9
DP (°C)	10.3	1.8	0.4	10.5	9.4	10.5	11.4	15.3	3.3	-1.3
PP (mm m^{-2} h^{-1})	0.03	0.6	0	0	0	0	0	60.4	6571	72

[1]Standard deviation. [2]Minimum value. [3]1st quartile. [4]2nd quartile. [5]3rd Quartile. [6]Maximum value. [7]Kurtosis. [8]Skewness. [9]Temperature. [10]Relative humidity. [11]Photosynthetic active radiation. [12]Dew point. [13]Precipitation.

highest frequency of frost events was at 5:00. A FE day presented an average temperature per hour increasing between 6:00 and 14:00 from 1.7 to y 23.0 °C, while after 14:00 until 6:00 decreased. The temperature curve behavior of the FEN day compared to one FE day was below, between 10:00 and 16:00, and on top from 16:00 to 10:00 (Figure 3a). The RH curves presented a convex decrease from 7:00 with a minimum inflection point (33.8%, 53.1% for FE and FEN) at 13:00; from this time, a convex growth begins with a maximum inflection point at 7:00 (96.8%, 94.2% for FE and FEN).

The RH curve of a FEN day showed similar behavior to a FE day, between 00:00 to 9:00, and it was below between 9:00 and 12:00 (Figure 3b). The DP curve of the FEN day was always under the FE day curve, and both have distinctive stages along the day. At 6:00, the DP curves showed the lowest moments (1.3 °C FE day, 7.5 °C FEN day), and three hours later, at 9:00 a.m, their highest moments (8.6 °C, FE Day, 10.9 °C FEN day), exhibiting a high rate of change. From maximum points, both curves decreased until 6:00 a.m. on the following day; however, the FE curve with the highest decay rate (Figure 3c). Concerning the precipitation, the FE days were dry, while FEN days were wet, with events between 11:00 to 19:00 principally (Figure 3d). The radiation curves of the FE and FEN day were similar form; nonetheless, between 7:30 and 17:00, the radiation of the FE curve day was above the FEN curve, maximum at 11:00 (1548.7 μmol m^{-2} s^{-1} FE days, and 1283.5 μmol m^{-2} s^{-1} FEN days) (Figure 3e).

In summary, the climatic characteristics of previous hours to a FE are dry, with low RH, marked by a reduction in cloudiness between 11:00 and 14:00 that generates an increase in thermal radiation emitted from the ground into space (Arribillaga *et al*., 2020). In the before conditions, under low wind speed (<7.2 km h-1), the loss of energy by radiation is not possible to compensate, leading to a gradual drop in the temperature from the first night hours to the first hours of dawn, reaching values below 0°C (González & Torres, 2012). Water vapor contained in atmosphere also influences the amount of energy emitted as radiation heat from the earth's surface; since the water can absorb this energy and output it back to surface. Therefore, a higher

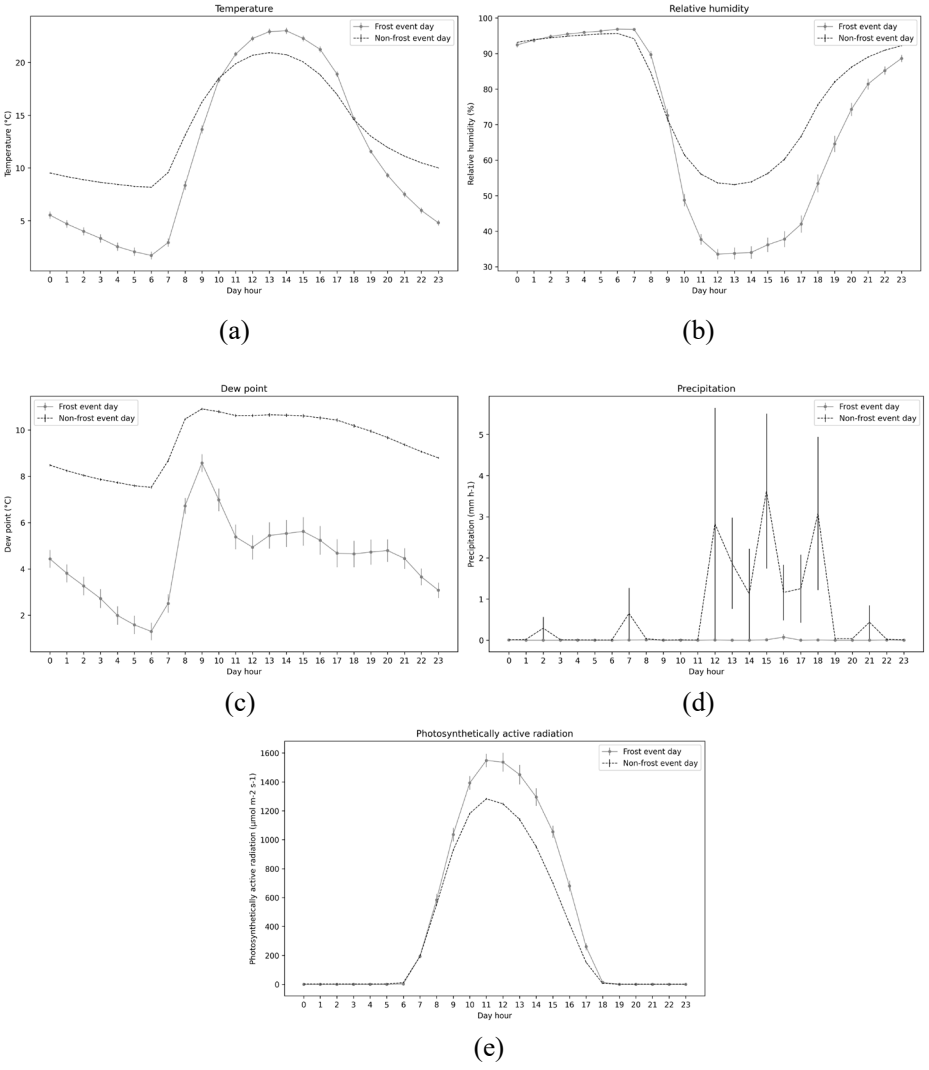

Figure 3: Climatic variables behavior 24 hours of the day prior to a FE and 24 hours prior to a FEN: (a) temperature, (b) relative humidity, (c) dew point, (d) precipitation and (e) photosynthetically active radiation. The length of the error bars corresponds to the standard error of the mean (SEM).

amount of water vapor generates less heat loss. In this stage, the temperature reduction on a night with low cloudiness and calm conditions will be more gradual and even avoid reaching values below 0°C (González & Torres, 2012).

On the other hand, the temperature's standardized beta coefficient of logistic regression shows a direct growth relationship with FE from 10:00 to 14:00 and an inverse rising from 18:00 to 23:00 and 0:00 to 7:00. From this, the inverse relationship begins a weaken of the inverse relationship until 9:00, when it becomes positive (Figure 4a). Similarly, DP presented an inverse rising relationship from 17:00 to 23:00 and between 0:00 and 7:00; the relationship was lower from 8:00 until 14:00 (-1.0 a 1.0) (Figure 4c). In addition, RH and FE exhibited a positive growth relationship from 5:00 to 9:00. Subsequently, a weakening trend began, turning negative between 10:00 to 14:00. It increases again and becomes positive between 15:00 and 16:00. From 17:00 until 5:00 of the next day, the relationship remained within the range of -1.0 to 1.0 (Figure 4b). Notably, RH from 8:00 to 9:00 and from 15:00 and 14:00 presented the highest standardized beta coefficients (from 2.51 to 3.31).

Precipitation and PAR had a weak impact on frost events' occurrence. The relationship evaluated through standardized beta coefficient between PP and frost events was positive and presented values between 0.29 and 0.46 (Figure 4d). The relationship direction between PAR and frost events occurrence was positive between 6 A.M. and 6 P.M. and, exhibiting maximum values at 7:00 and 18:00 (0.67 and 1.23, respectively); and negative from 19:00 to

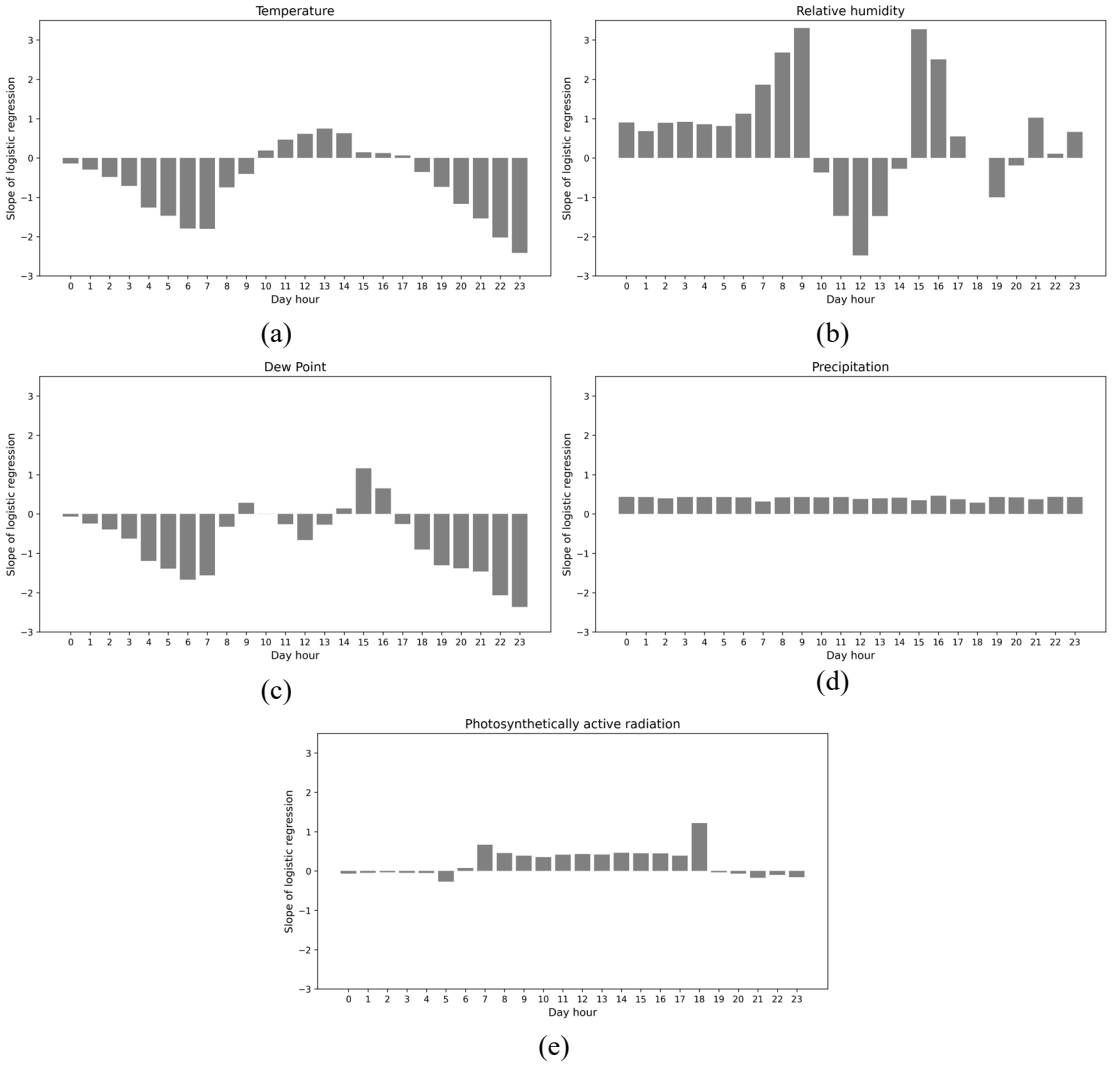

Figure 4: Relationship between frost events and climatic variables per hour through standardized beta coefficient of logistic regression: (a) temperature, (b) relative humidity, (c) dew point, (d) precipitation and (e) photosynthetically active radiation.

5:00 (-0.03 to -0.27) as shown in Figure 4e. The hours of the day for which the SelectKBest function detected the highest score were for temperature between 02:00 and 10:00, RH between 10:00 and 16:00, DP between 01:00 and 04:00 and PAR between 9:00 and 16:00.

Otherwise, Lee *et al*. (2016) established a positive relationship between DP and RH with frost occurrence probability. In addition, Ding *et al*. (2019) found that air temperature, net radiation and wind speed had a negative correlation with frost (-0.37, -0.2, and -0.28, respectively) and positive correlation with relative humidity (0.35), with a data frequency of 60 data per hour (one record per minute). Other authors have considered wind speed and minimum temperature for frost prediction because when speed exceeds the very low threshold value, it prevents the formation of a thermal inversion layer near the ground (Ding *et al*., 2019; Fuentes *et al*., 2018). However, Ding *et al*. (2019) have found lower correlation values between wind speed and frost events.

Predictive models

The most closely related variables with FE were included, in the training dataset until 16:00 of the day before to have a window of time available to deploy prevention strategies. The best score for LR model (0.948) was found with C = 5.0, penalty = l1, and liblinear solver. Similarly, for DT (0.917), GBDT (0.995), SVM (0.993) and NN (0.927) were found with the optimal parameters. In addition, in the five models training, it was determined that number of components to maintain, resulting from PCA, was 30.

In some experiments, the predictive performance comparisons between traditional prediction systems and

hyperparameters tuning show that search for optimal parameters through hyper-parameters tuning provides better results and higher quality predictions (Shahhosseini *et al*., 2022). However, the optimal parameters depend on dataset and arbitrary defaults of the respective algorithm, and its search, through random quest, may incur an increase in computational costs (Shahhosseini *et al*., 2022). For this study, the higher run time among the five trained models was for NN (24,300 seconds), followed by GBDT (4,206.5) and SVM (1,816.6). The run times training in the case of LR and DT models were lower than 600 seconds.

Model performance evaluation metrics and comparison

The five trained models performed very well due to their classification evaluation metrics (accuracy, precision, recall, TNR, and F_1 score), greater than 91% except for the SVM model, whose metrics fluctuated between 64 and 100%. All the tested models were effective and efficient in detecting FE. The cross validation and statistical analysis demonstrated the higher accuracy of the GBDT model on FE detection with Ac = 0.95 in the validation set, the LR and NN ranks second with Ac = 0.92, and DT ranks third with Ac = 0.91. The GBDT was also superior in other metrics such as p, TPR, and F1-score in both test as validation dataset (Table 2). More specifically, GBDT has p = 0.94, TPR = 0.96, and F1-score = 0.95, while SVM model has achieved lower performance evaluation metrics with Ac = 0.82, p = 1.00, TPR = 0.64, TNR = 1.00, and F1-score = 0.78.

However, LR model also presented high performance metrics in validation dataset (Ac = 0.92, p = 0.91, TPR = 0.94, TNR = 0.91, and F1-score = 0.92) and a shorter run time compare with GBDT and NN; indicating that GBDT, NN and LR models can be used to predict FE with a very low misclassification error. Also, the SVM model exhibits lower performance compared to the models developed by Ding *et al*. (2019). These authors found classification evaluation metrics of 0.92 (Ac), 0.9 (p), and 0.99 (TPR) for frost forecast with a support vector machine approach and temperature, humidity, and radiation as input features.

The AUC value reported for GBDT (0.96) was higher than those of the NN (0.94), LR (0.94), DT (0.94), and SVM (0.91). Similarly, the ROC value for GBDT model was higher (0.95) followed by NN (0.93), LR (0.92), DT (0.91) and SVM (0.82). Therefore, all the trained binary classification models performed very well. However, TPR values constitute the most appropriate metric for the models' performance evaluation since the model aims to predict the frost events occurrence and thus generate a timely warning to deploy prevention mechanisms of frost damage. In this sense, it's more serious to fail to predict a frost than to generate a false alarm. Therefore, could be used the GBDT model as one of the main tools to generate alarms of damage prevention of the frost events on the high-altitude crops in the country.

Although the models presented high performance metrics when evaluated with training and test dataset, their monitored in real-time is advisable because overfitting is one of the threats to machine learning algorithms. Overfitting occurs when there is a performance high in training and low in testing with a poor generalization to independent datasets. The most common causes are a small sample

Table 2: Models' performance evaluation metrics in test and validation datasets.

Dataset	Metric	LR	DT	SVM	GBDT	NN
Test dataset	p	0.95	0.93	1.00	0.95	0.95
	TPR	0.92	0.94	1.00	0.92	0.92
	TNR	0.95	0.93	0.71	0.94	0.95
	Ac	0.93	0.93	0.80	0.93	0.93
	F_1 score	0.93	0.93	0.75	0.93	0.94
Validation dataset	p	0.91	0.91	1.00	0.94	0.92
	TPR	0.94	0.91	0.64	0.96	0.93
	TNR	0.91	0.91	1.00	0.93	0.92
	Ac	0.92	0.91	0.82	0.95	0.92
	F_1 score	0.92	0.91	0.78	0.95	0.93

size, high-dimensional features, and models with many parameters; however, one strategy recommended to avoid overfitting is the model training using cross validation (Gao *et al*., 2018).

While vigilance against overfitting remains crucial in the application of machine learning algorithms, their advantages become apparent when compared to traditional regression approaches. Machine learning algorithms offer advantages compared with longitudinal data modeling approaches by accommodating various data complexities, such as non-independence, non-normal distributions, multicollinearity, and missing values (Sheetal *et al*., 2023).

From the perspectives of this work, we plan to apply our findings to prevent frost occurrence in high altitude crops located in other tropical regions and obtain predictions of the minimum temperature and frost events duration for optimal implementation of frost mitigation strategies. It´s also expected to include other climatic variables and obtain predictions in larger time windows.

Exploring simulation models would represent another future investigation to determine whether the frequency of frost events is increasing over time and to evaluate the potential impacts of different climate change scenarios on their frequency and intensity.

CONCLUSIONS

Both machine learning models and logistic regression methods implemented for early frost predictions based on climatic variables have high performance, except for SVM. The GBDT showed the highest performance, with a true positives rate and accuracy (>92%).

Dew point was the climatic variable with the highest relationship with frost events. The results show the possibility of generating an early frost prediction, as support to producers, in the anticipation and implementation of mitigation and adaptation strategies, to frost damage in high altitude crops of the tropic.

ACKNOWLEDGEMENTS, FINANCIAL SUPPORT AND FULL DISCLOSURE

The authors express their gratitude to Soluciones Wiga and Growers Hub Trading Group Companies for providing the research grant for the project realization.

REFERENCES

Aguilar M & Torres SB (2010) Protocolo de uso y aprovechamiento de la uva de anís, *Cavendishia bracteata* (Ruiz y Pavón ex Jaume Saint. Hillaire) Horeold, en matorrales andinos del Altiplano Cundiboyacense. Bogotá, Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Ministerio de Agricultura y Desarrollo Rural y Cámara de Comercio de Bogotá. 32p.

Arribillaga D, Bravo R, Campos C, Fuentes M, Gatica J, Luchabeche P, Quintana J, Reyes M, Salacar C, Salvo del Pedregal J & Vidal M (2020) Heladas. Factores, tendencias y efectos en frutales y vides. Chile, Instituto de Investigaciones Agropecuarias INIA. 102p. (Technical Bulletin, 417).

Becerra LL (2021) San Valentín, el desquite de los floricultores en pandemia. Available at: <https://www.portafolio.co/economia/san-valentin-el-desquite-de-los-floricultores-en-pandemia-con-las-exportaciones-de-flores-548989>. Accessed on: October 18th, 2021.

Brito A de A, de Araújo HA & Zebende GF (2019) Detrended Multiple Cross-Correlation Coefficient applied to solar radiation, air temperature and relative humidity. Scientific Reports, 9:19764.

Charbuty B & Abdulazeez A (2021) Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2:20-28.

Cho S, Kim YJ, Lee M, Woo JH & Lee HJ (2021) Cut-off points between pain intensities of the postoperative pain using receiver operating characteristic (ROC) curves. BMC Anesthesiol, 21:29.

Danandeh A (2021) Drought classification using gradient boosting decision tree. Acta Geophysica, 69:909-918.

DeVries Z, Locke E, Hoda M, Moravek D, Phan K, Stratton A, Kingwell S, Wai EK & Phan P (2021) Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability. Spine Journal, 21:1135-1142.

Diedrichs AL, Bromberg F, Dujovne D, Brun-Laguna K & Watteyne T (2018) Prediction of Frost Events Using Machine Learning and IoT Sensing Devices. IEEE Internet of Things Journal, 5:4589-4597.

Ding L, Noborio K & Shibuya K (2019) Frost forecast using machine learning - From association to causality. Procedia Computer Science, 159:1001-1010.

Dinh TV, Nguyenm H, Tran XL & Hoang ND (2021) Predicting rainfall-induced soil erosion based on a hybridization of adaptive differential evolution and support vector machine classification. Mathematical Problems in Engineering, 2021:01-20.

Fuentes M, Campos C & García S (2018) Application of artificial neural networks to frost detection in central Chile using the next day minimum air temperature forecast. Chilean Journal of Agricultural Research, 78:327-338.

Gao S, Calhoun VD & Sui J (2018) Machine learning in major depression: From classification to treatment outcome prediction. CNS Neuroscience and Therapeutics, 24: 1037-1052.

Gómez D, Araujo G, Martínez FE, Rodríguez AO, Estupiñan JM & Deantonio LY (2021) Análisis de eventos climáticos extremos asociados a excesos de lluvia y heladas meteorológicas en el Altiplano Cundiboyacense de Colombia. Revista de Climatología, 21:112-126.

González OC & Torres CF (2012) Actualización nota técnica heladas. Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Available at: <http://www.ideam.gov.co/documents/21021/21147/Documento+FINAL+actualizacion+nota+tecnica+heladas.pdf/e10a0183-62e6-410a-8e96-7e0739f6f06b>. Accessed on: October 21th, 2021.

Guhl E (2013) La región hídrica de Bogotá. Revista de La Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 37:327-341.

Joshi NC, Yadav D, Ratner K, Kamara I, Aviv E, Irihimovitch V & Charuvi D (2020) Sodium hydrosulfide priming improves the response of photosynthesis to overnight frost and day high light in avocado (*Persea americana* Mill, cv. 'Hass'). Physiologia Plantarum, 168:394-405.

Juurakko CL, diCenzo GC & Walker VK (2021) Cold acclimation and prospects for cold-resilient crops. Plant Stress, 2:100028.

Kochhar SL & Gujral SK (2020) Plant Physiology: Theory and Applications. 2nd ed. Cambridge, Cambridge University Press. 866p.

Latif RMA, Belhaouari SB, Saeed S, Imran LB, Sadiq M & Farha M

(2020) Integration of Google Play Content and Frost Prediction Using CNN: Scalable IoT Framework for Big Data. IEEE Access, 8:6890-6900.

Lee H, Chun JA, Han HH & Kim S (2016) Prediction of Frost Occurrences Using Statistical Modeling Approaches. Advances in Meteorology, 2016:01-09.

Li X, Ahammed JG, Li Z, Zhang L, Wei J, Yan P, Zhang LP & Han WY (2018) Freezing stress deteriorates tea quality of new flush by inducing photosynthetic inhibition and oxidative stress in mature leaves. Scientia Horticulturae, 230:155-160.

Luengas E, Guhl A, Castro JC, González LN & Restrepo S (2021) Modeling the correlation between potato disease spread and climate variables to guide fungicide applications in Cundinamarca, Colombia. Naturaleza y Sociedad. Desafios Medioambientales, 1:07-42.

Marmolejo D & Ruiz JE (2018) Tolerance of native potatoes (*Solanum* spp.) to ice creams in the context of climate change. Scientia Agropecuaria, 9:393-400.

Mayorga M, Fischer G, Melgarejo LM & Parra A (2020) Growth, development and quality of Passiflora tripartita var. Mollissima fruits under two environmental tropical conditions. Journal of Applied Botany and Food Quality, 93:66-75.

Ministerio de Agricultura y Medio Ambiente (2020) "Debemos mantener la guardia con medidas preventivas frente a bajas temperaturas y fenómeno de La Niña": ministro Rodolfo Zea. Available at: <https://www.minagricultura.gov.co/noticias/Paginas/%E2%80%9CDebemos-mantener-la-guardia-con-medidas-preventivas-frente-a-bajas-temperaturas-y-fen%C3%B3meno-de-La-Ni%C3%B1a%E2%80%9D-ministro-Rodolfo-Ze.aspx>. Accessed on: September 29th, 2021.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M & Duchesnay É (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825-2830.

Rout BM (2020) Advances in Freezing Stress Resistance in Vegetable Crops. Biotica Research Today, 2:261-263.

Shahhosseini M, Hu G & Pham H (2022) Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. Machine Learning with Applications, 7:100251.

Sheetal A, Jiang Z & Di Milia L (2023) Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. Applied Psychology, 72:1339-1364.

Simnitt S, Borisova T, Chavez D & Olmstead M (2017) Frost protection for Georgia Peach varieties: Current practices and information needs. HortTechnology, 27:344-353.

Trilles S, Juan P, Chaudhuri S & Fortea ABV (2021) Data on CO2, temperature and air humidity records in Spanish classrooms during the reopening of schools in the COVID-19 pandemic. Data in Brief, 39:107489.

Vargas C (2021) 10 mil hectáreas de cultivos se han perdido en Cundinamarca por las heladas y la sequía. Available at: <https://www.rcnradio.com/colombia/region-central/10-mil-hectareas-de-cultivos-se-han-perdido-en-cundinamarca-por-las-heladas>. Accessed on: November 15th, 2021.