Uptc®
Universidad Pedagógica y Tecnológica de Colombia

Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia

Henry Lamos-Díaz; David-Esteban Puentes-Garzón; Diego-Alejandro Zarate-Caicedo

Citation: H. Lamos-Díaz, D.-E. Puentes-Garzón, and D.-A. Zarate-Caicedo, “Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia,” *Revista Facultad de Ingeniería*, vol. 29 (54), e10853, 2020. https://doi.org/10.19053/01211129.v29.n54.2020.10853

Received: April 7, 2019; **Accepted:** May 13, 2020; **Published:** May 15, 2020

Copyright: This is an open access article distributed under license CC BY

Conflict of interest: The authors state there is no conflict of interest.

Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia

Henry Lamos-Díaz[1]
David-Esteban Puentes-Garzón[2]
Diego-Alejandro Zarate-Caicedo[3]

Abstract

The identification of influencing factors in crop yield (kg·ha^{-1}) provides essential information for decision-making processes related to the prediction and improvement of productivity, which gives farmers the opportunity to increase their income. The current study investigates the application of multiple machine learning algorithms for cocoa yield prediction and influencing factors identification. The Support Vector Machines (SVM) and Ensemble Learning Models (Random Forests, Gradient Boosting) are compared with Least Absolute Shrinkage and Selection Operator (LASSO) regression models. The considered predictors were climate conditions, cocoa variety, fertilization level and sun exposition in an experimental crop located in Rionegro, Santander. Results showed that Gradient Boosting is the best prediction alternative with Coefficient of determination (R^2) = 68%, Mean Absolute Error (MAE) = 13.32, and Root Mean Square Error (RMSE) = 20.41. The crop yield variability is explained mainly by the radiation one month before harvest, the accumulated rainfall on the harvest month, and the temperature one month before harvest. Likewise, the crop yields are evaluated based on the kind of sun exposure, and it was found that radiation one month before harvest is the most

[1] Ph. D. Universidad Industrial de Santander (Bucaramanga-Santander, Colombia). hlamos@uis.edu.co. ORCID: 0000-0003-1778-9768
[2] M. Sc. Universidad Industrial de Santander (Bucaramanga-Santander, Colombia). david.puentes1@correo.uis.edu.co. ORCID: 0000-0001-8178-2339
[3] Ph. D. Corporación Colombiana de Investigación Agropecuaria (Rionegro-Santander, Colombia). dzarate@corpoica.org.co. ORCID: 0000-0001-9630-3927

L-ISSN: 0121-1129, e-ISSN: 2357-5328, DOI: https://doi.org/10.19053/01211129.v29.n54.2020.10853

influential factor in shade-grown plants. On the other hand, rainfall and soil moisture are determining variables in sun-grown plants, which is associated with the water requirements. These results suggest a differentiated management for crops depending on the kind of sun exposure to avoid compromising productivity, since there is no significant difference in the yield of both agricultural managements.

Keywords: agricultural yield; agroforestry system; cocoa; machine learning; prediction; productivity.

Comparación de modelos de aprendizaje automático para la predicción de rendimientos agrícolas en cultivos de cacao en Santander, Colombia

Resumen

La identificación de los factores que influyen en el rendimiento (kg·ha^{-1}) de un cultivo provee información esencial para la toma de decisiones orientadas al mejoramiento y predicción de la productividad, proporcionando posibilidades a los agricultores para mejorar sus ingresos económicos. En este estudio, se presenta la aplicación y comparación de diversos algoritmos de aprendizaje automático para la predicción del rendimiento agrícola en cultivos de cacao y la identificación de los factores que influyen sobre éste. Se comparan los algoritmos de máquinas de soporte vectorial (SVM), modelos ensamblados (Random Forest, Gradient Boosting) y el modelo de regresión *Least Absolute Shrinkage and Selection Operator* (LASSO). Los predictores considerados fueron: condiciones climáticas de la región, variedad de cacao, nivel de fertilización y exposición al sol para un cultivo experimental ubicado en Rionegro, Santander. Los resultados identifican a Gradient Boosting como la mejor alternativa de pronóstico con un coeficiente de determinación (R^2) = 68 %, Error Absoluto Medio (MAE) = 13.32 y Raíz Cuadrada del Error Medio (RMSE) = 20.41. La variabilidad del rendimiento del cultivo es explicada principalmente por la radiación y la temperatura un mes previo a la cosecha, además de las lluvias acumuladas el mes de la cosecha. De igual manera, los rendimientos de los cultivos son evaluados con base en el tipo de exposición al sol, encontrando que la radiación un mes previo a la cosecha es el factor más influyente para los cultivos bajo sombra. Por otro lado, la lluvia y la humedad son las variables determinantes en las plantas con exposición plena a sol, lo que está asociado a los requerimientos hídricos. Estos resultados sugieren un manejo diferenciado de los cultivos dependiendo del tipo de exposición, sin tener que comprometer la productividad, dado que no se evidencia diferencia significativa en los rendimientos de ambos manejos agrícolas.

Palabras clave: aprendizaje automático; cacao; predicción; productividad; rendimientos agrícolas; sistemas agroforestales.

Comparação de modelos de aprendizado de máquina para a previsão de produção agrícola em cacau em Santander, Colômbia

Resumo

A identificação de fatores que influenciam o rendimento (kg·ha^{-1}) de uma safra fornece informações essenciais para a tomada de decisões com o objetivo de melhorar e prever a produtividade, oferecendo possibilidades aos agricultores de melhorar sua renda econômica. Neste estudo, são apresentadas a aplicação e comparação de vários algoritmos de aprendizado de máquina para a previsão do desempenho agrícola em cultivos de cacau e a identificação dos fatores que o influenciam. Os algoritmos de máquinas de suporte de vetores (SVM), modelos montados (floresta aleatória, reforço de gradiente) e o modelo de regressão Operador de seleção e contração mínimos absolutos (LASSO) são comparados. Os preditores considerados foram: condições climáticas da região, variedade de cacau, nível de fertilização e exposição ao sol para uma cultura experimental localizada em Rionegro, Santander. Os resultados identificam o Gradient Boosting como a melhor alternativa de prognóstico com um coeficiente de determinação (R^2) = 68%, Erro Absoluto Médio (MAE) = 13.32 e Erro Médio de Raiz Quadrada (RMSE) = 20.41. A variabilidade do rendimento das culturas é explicada principalmente pela radiação e temperatura um mês antes da colheita, além das chuvas acumuladas no mês da colheita. Da mesma forma, os rendimentos das culturas são avaliados com base no tipo de exposição ao sol, constatando que a radiação um mês antes da colheita é o fator mais influente para as culturas sombreadas. Por outro lado, chuva e umidade são as variáveis determinantes em plantas com exposição solar total, as quais estão associadas às necessidades de água. Esses resultados sugerem um manejo diferenciado das culturas, dependendo do tipo de exposição, sem comprometer a produtividade, uma vez que não há diferença significativa nos rendimentos de ambos os manejos agrícolas.

Palavras chave: aprendizado de máquina; cacau; predição; produtividade; rendimentos agrícolas; sistemas agroflorestais.

I. INTRODUCTION

Cocoa, which is a tropical agricultural product in worldwide demand by different industries, represents an important source of economic sustenance for small farmers. In 2017, Colombia registered an increase of 3.750 tons in production compared to the previous year, which marks a milestone for the country consistent with the efforts of farmers, guilds, and the national government. In addition, cocoa has was nominated for "crop for peace", which allows the substitution of illicit crops and the generation of job opportunities. However, the causes of the production increase lie in the expansion of the harvested area and not in the improvement of productivity and agricultural practices, crop renewal or use of new technologies.

Machine learning has become an alternative for studying agricultural yields and identifying the factors that explain their variability, including climatic and soil conditions. This alternative considers each crop as a different experiment and their associated data is adjusted to a certain function to make predictions [1-3]. Drummon et al. [4] proposed the use of neural networks, stepwise linear regression, and projection pursuit regression to predict the yield of corn and soybean in Missouri, United States, by considering physical and chemical characteristics of the soil, as well as climatic conditions. Similarly, De Paepe et al. [5] analyzed the effects of soil characteristics and climatic conditions on wheat yields in the Argentine pampas using neural networks. On the other hand, several authors have modelled crop yields according to the phenotype of the plants [6-8]. Romero et al. [8] suggested OneR, IBK, C4.5, and Apriori classification algorithms to provide association rules in order to predict the level of wheat production, according to spikelet number, plant height, peduncle length, and spike fertility. Other authors have evaluated variables such as quantity of fertilizer, fertilizer source, pest and disease management, and seed variety [7, 9-10].

Regarding cocoa crops, the yield prediction has been approached from different perspectives. Corrales et al. [11] predicted the cocoa yield level in Santander. The authors evaluated the daily average temperature, daily relative humidity, and total daily precipitations rate, using ten different algorithms implemented in WEKA software. For them, Random Forest was the algorithm that generates the best model in order to classify cocoa yield levels. Other studies [12-13] evaluate the yield using linear regression models, ANOVA and mechanistic models like SUCROS, finding that climatic conditions (such as temperature, radiation and rainfall) are the most critical in the cocoa productivity.

According to the literature, machine learning algorithms have had satisfactory results in different traditional crops, such as wheat, corn, soybean, and rice. However, few studies have assessed the factors that affect the cocoa yield using this approach and, particularly, evaluating the influence of shadow on agroforestry systems. Therefore, the present research study evaluates some of the most powerful and popular algorithms: Support Vector Machines, Random Forest, Gradient Boosting, and LASSO regression, to predict cocoa yields and identify the factors that influence them. Similarly, from a marginal influence analysis, these algorithms are used to determine the factors that affect cocoa yield depending on the kind of sun exposure (shade-grown or sun-grown), which is key to differentiated agricultural management and productivity maximization.

II. Materials and Methods

Data is the most important input for predictive model construction based on machine learning. This section describes the experimental design used to obtain the data, and the secondary sources consulted. Furthermore, it shows the algorithms implemented and the metrics used to compare their different performances.

A. Data Acquisition

The data analyzed in this research study was obtained from an experimental crop during the period 2015-2017. This crop was stablished in 2008 in the research center "La Suiza" of the Colombian Corporation for Agricultural Research – Agrosavia - in the municipality of Rionegro (Santander, Colombia), at an altitude of 550 meters above sea level. The experimental design was completely randomized in a block design, with three replicates, ten cocoa varieties (5 universal and 5 regional) [14], three levels of fertilization, and two kinds of sun exposure (Table 1).

Table 1. Factors and experimental design levels.

Clones	Fertilization	Exposition
Regionals clones: SCC-19, SCC-52, SCC-61, SCC-64, SCC-83 *Universal clones:* CNN-51, EET-8, ICS-1, ICS-95, TSH-565	50%, 100%, 50%	Sun / Shade

Fertilization is related to the percentage of basic criteria which includes urea, Diammonium phosphate (DAP), KCl, organic matter, sulfur (S), magnesium sulfate ($MgSO_4$) and borate.

The treatments were applied to five plants per replication, for a total of 900 plants per hectare (450 sun-grown and 450 shade-grown). The shade is supplied by *Cariniana pyriformis Miers* and *Tectona grandis L.f*, with an average height of 12 [m] and a density of 340 [trees/ha].

Also, the models consider the physical characteristics of the soil and the climatic conditions of the area (Table 2), which were measured daily by meteorological stations located in the region (Watchdog 2000, Spectrum Technologies Inc, Aurora, IL, USA) and data from the "Instituto de Hidrología, Meteorología y Estudios Ambientales" (Ideam).

Table 2. Inputs for development of cocoa yield model

Variable name	Meaning	Type	Variable name	Meaning	Type
Cocoa_v	Cocoa variety	Cat[a]	P_accu_prev	Accumulated rainfall one month before harvest	Con[b]
Exp	Exposition	Cat[a]	T_avg	Temperature average on harvest month	Con[b]
F_level	Fertilization level	Cat[a]	T_avg _prev	Temperature average one month before harvest	Con[b]
EC_avg	Electrical conductivity on harvest month	Con[b]	Rad_accu	Accumulated photosynthetic active radiation (PAR)	Con[b]
Hum_avg	Soil humidity average on harvest month	Con[b]	Rad_accu_prev1	Accumulated photosynthetic active radiation (PAR) one month before harvest	Con[b]
P_accu	Accumulated rainfall on harvest month	Con[b]	Rad_accu_prev2	Accumulated photosynthetic active radiation (PAR) two months before harvest	Con[b]

a Categorical variable, b Continuous variable

B. Linear Regression Models

The linear regression LASSO (Least Absolute Shrinkage and Selection Operator) is a statistical model that relates a set of independent variables (predictors) to one dependent variable (response variable). Unlike the classical linear regression model, LASSO includes a regularization factor (α) in the regression coefficients, using the L1 norm (absolute value) equation (1).

$$\widehat{\beta_l} = argmin_{\beta}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j)^2 + \alpha \sum_{j=1}^{p}\left|\beta_j\right|\} \qquad (1)$$

Where y is a vector of observations (yield), x_{ij} are vectors of independent variables (independent variables), β are the regression coefficients, and *α* is the penalty. A high value of α implies low or almost zero coefficients, while a low value, a classical linear regression. Therefore, the value of α is determined by cross-validation.

C. Support Vector Machines (SVM)

SVM is a non-parametric algorithm based on statistical learning theory that seeks to identify a decision hyperplane where the margin of separation between positive and negative observations is maximum. Initially, Vapnik [15] proposed this algorithm for classification problems, however, it has been extended to regression problems [16]. The objective is to minimize the error between observed data (dependent variable – cocoa yield) and a family of functions *F(x,w)* parameterized by *w,* and x, which is the input space (independent variables).

D. Ensemble Learning Models

Ensemble methods are based on the premise that multiple algorithms are better than one, since they improve predictive performance by aggregating multiple and independent learning algorithms [17]. Base and aggregation algorithms are used to build an ensemble. The first ones are used to generate multiple predictions that are adding. The algorithm is usually a regression tree. On the other hand, the latter manipulates the inputs of the base algorithms to generate independent models. In the development of the present research study, the following aggregation algorithms are considered:

1) Boosting: iterative procedure to adaptively change the distribution of training samples, so that the basic algorithm focuses on samples that are difficult to predict. In each iteration, weights are assigned to each training-observation, which are updated according to the error with respect to the observed values. Two of the most popular boosting algorithms are AdaBoost and Gradient Boosting, the latter does the training of the base algorithms with the errors of the previous iteration, and maximizes the predictive accuracy by means of gradient descent [18].

2) Random Forest: it was proposed by professor Leo Breiman [19]. This algorithm is a combination of predictions of multiple regression trees, where each one depends on a set of independent random vectors and has the same probability distribution.

E. Evaluation Metrics

These metrics evaluate the model performance and compare it with other proposals. Some of these metrics include the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the coefficient of determination (R^2), as shown in equations (2), (3) and (4).

$$R^2 = \frac{\sum_{i=1}^{n}(O_i-\bar{O})^2 \cdot (P_i-\bar{P})^2}{\sum_{i=1}^{n}(O_i-\bar{O})^2 \cdot \sum_{i=1}^{n}(P_i-\bar{P})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad (4)$$

$$RI_{RMSE} = \frac{RMSE_R - RMSE_C}{RMSE_R} \quad (5)$$

Where: Oi and Pi are the observed and predicted values for the ith observation, respectively; $\bar{O}$ and $\bar{P}$ are the average values of the observed and predicted yield; and n is the number of observations. The relative improvement of RMSE (RI_{RMSE}) shown in equation (5) was used to measure and compare the algorithms and establish the alternative that best fits the yields observed in the crop. $RMSE_R$ and $RMSE_c$ are the performance values of the reference and comparative algorithms, respectively.

III. Results and Discussion

Initially, the predictive models are built for the complete dataset (cocoa yield and inputs described in table 1), including the kind of sun exposure as an independent variable. In a second scenario, the dataset is divided according to the kind of sun exposure: sun-grown (284 observations) and shade-grown (274 observations).

In the training phase, 80% of the data is used as the training set for each model, and the remaining 20% as the test set (hold-out validation). In the same way, a cross-validation with k=10 was performed, together with grid search, to establish the best hyper parameters for each algorithm. The module of model_selection in the sklearn package is used [20] for this process.

A. Model Evaluation

Table 3 shows the average results for performance metrics in hold-out validation. On average, the performance of Gradient Boosting is higher than the other algorithms, with the lowest values for MAE and RMSE, and the highest value for R^2. Moreover, the relative improvement in RMSE is 20.99%, 8.54%, and 5.93% compared to LASSO, SVM, and Random Forest, respectively.

Table 3. Average performance of the algorithms for the complete dataset.

Model	MAE	RMSE	R^2 (%)	RI_{rmse} (%)
LASSO	20.65	31.73	20.65	20.99
SVM	15.69	27.41	41.17	8.54
Random Forest	14.7	26.65	44.19	5.93
Gradient boosting	12.94	25.07	49.29	-

For each of the 100 repetitions in holdout validation, the algorithm is trained and tested with a random sample, where the best alternative is Gradient Boosting with 480 trees. This algorithm explained the 68% of the variability, and presented a MAE of 13.32, and a RMSE of 20.41 (Figure 1).

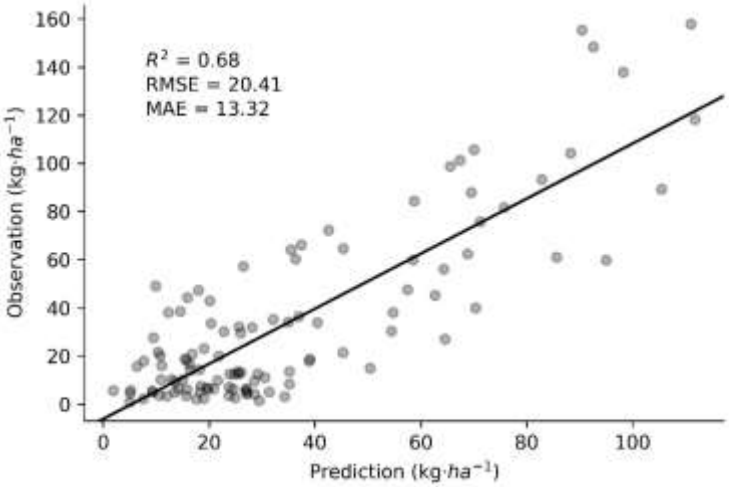

Fig. 1. Cocoa yield prediction with Gradient Boosting.

Once the best model alternative is identified, the next step is to evaluate the importance of the variables. For Gradient Boosting, the MSE Friedman metric allows to choose the variables which will improve the quality in a split. However, to make a comparison between the impact of each variable in the different algorithms and quantify this impact, it is necessary to use another strategy. The partial dependence plots illustrate the marginal influence when a variable change and the other variables remain constant. Logan et al. [21] proposed an alternative to measure the marginal influence with equation (6).

$$Oscillation_n = \frac{\max(V_n) - \min(V_n)}{\sum_n Oscillation_n} \tag{6}$$

Table 4 shows the variables with the highest oscillation value considering the best model identified in the validation phase.

Table 4. Variables for the complete dataset considering Gradient Boosting.

Variable	Swing	Variable	Swing
P_accu	0.17	Clon_CCN51	0.07
Rad_accu_prev1	0.16	Clon_TSH565	0.07
Rad_accu_prev2	0.10	T_avg _prev	0.06

These results indicate that the average temperature one month before harvest, the accumulated radiation one and two months before harvest, and the accumulated rainfall on harvest month are the factors with greatest impact on crop yields. According to [22], temperature is one of the factors that limit cocoa production, since it causes stress on the plants, increases seasonal variability, and is responsible for the reduction in photosynthetic rates. Radiation and rainfall are related to the final stage of cocoa crop growth, where rainfall is more important than radiation [13]. Concerning the sun exposure variable, the oscillation is close to 0. Thus, it can be assumed that the type of sun exposure is not representative for the predictive model. The variable influence evaluation is performed using Gradient Boosting as well as partial dependence plots for the interaction between precipitation, temperature, and radiation one month before harvest (Figure 2).

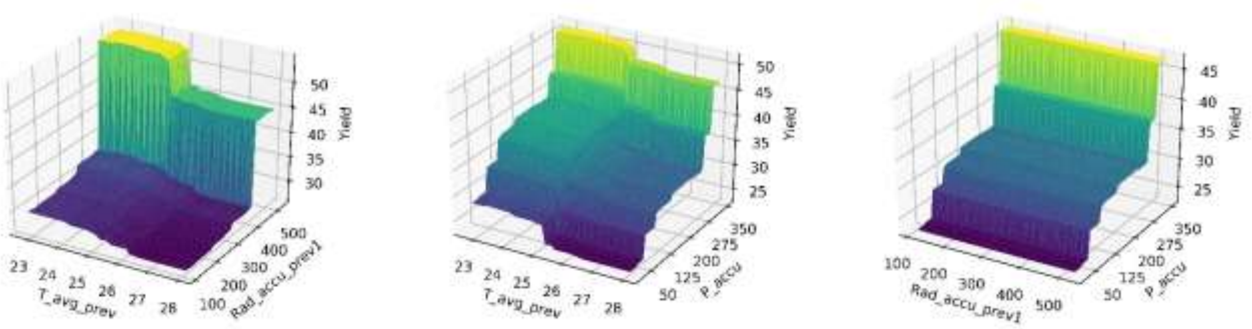

Fig. 2. Partial dependence plots for the interaction of identified variables.

The vertical axis (crop yield) shows that the interaction between radiation and accumulated rainfall has the lowest effect, while interactions with temperature generate higher yields. These results suggest that the control of temperature, radiation and accumulated rainfall are determinant for increasing crop productivity. Likewise, the effect of radiation decreases when it interacts with rainfall, which ratifies accumulated rainfall as the most influential variable on crop yields.

B. Sunshade Exposure Models

In this second scenario, for each kind of exposition (sunshade) the best identified algorithm is ran again. Table 5 shows that variability is best explained in the shade-grown model with an average R^2 of 54.27%, and lower values of MAE and RMSE, compared to the sun-grown model.

Table 5. Average performance for two different kinds of sun exposure.

Metric	R^2	RMSE	MAE
Sun	42.28%	25.98	14.67
Shade	54.27%	24.51	11.25

To evaluate the importance of the variables in the models associated with the kind of sun exposure, the procedure described in the previous section is applied once again. Table 6 suggests that the accumulated rainfall on harvest month and the average soil moisture are the most influential variables in the sun-grown predictive model. This result evidences the higher water requirements of sun grown plants. In fact, sun-grown crops have a higher leaf transpiration and soil water evaporation, which lead to lower photosynthetic activity and higher stomatal closure. The last affirmation implies shorter production cycles, higher nutrient requirements, better management of irrigation systems, and, therefore, a higher investment [23].

For the shade-grown case, radiation one month before harvest has the highest oscillation value, which indicates a strong relationship between this variable and crop yield. As stated by Zuidema et al. [13], shade must be properly managed in this kind of crops, to avoid yield reduction due to lack of radiation.

Table 6. Importance of variables for sun-grown and shade-grown models.

SUN		SHADE	
Variable	Swing	Variable	Swing
P_accu	0.20	Rad_accu_prev1	0.39
Hum_avg	0.13	P_accu	0.11
Clon_CCN51	0.10	Rad_accu	0.08
Clon_TSH565	0.07	Clon_TSH565	0.06

Furthermore, in Table 6 some cocoa clones are identified (CNN-51 and TSH-565). As a result, the cocoa yield is evaluated and compared in the two kinds of sun exposure. A Kruskal-Wallis test is used with a significance level of 5%, stating as a null hypothesis that the medians of the samples are equal. With values of $p < 0.05$, the results (Table 7) suggest that CNN51 and SCC64 clones perform dissimilarly in both kinds of sun exposure. Likewise, the average yield indicates that both clones perform better under sun-grown conditions.

Table 7. Kruskal-Wallis test.

Cocoa clone	Yield mean Sun	Yield mean Shade	Yield median Sun	Yield median Shade	P_value
CNN51	62.39	33.71	44.651	19.224	0.017
EET8	32.42	36.18	27.641	13.532	0.145
ICS1	30.95	25.18	17.859	18.760	0.715
ICS95	31.75	37.49	16.900	21.699	0.481
SCC19	34.77	36.92	20.649	14.953	0.984
SCC52	33.23	29.67	24.966	18.205	0.742
SCC61	57.86	44.90	37.670	23.521	0.751
SCC64	43.53	30.43	27.615	16.211	0.012
SCC83	28.22	45.16	21.718	20.673	0.575
TSH565	45.58	41.35	37.930	24.078	0.269

In general, there is no difference between the crop yield under sun and shade-grown conditions, which is positive for the promotion of agroforestry crops. These findings are consistent with [12-13, 24], who state that shade does not affect cocoa yield, as long as it is adequately provided. In addition, [25-26] state that moderate shade-grown crops have positive implications for soil management, moisture and temperature control, and for the creation of environments to improve cocoa physiology and reduce the impact of pests and diseases.

IV. Conclusions

This research study proved the ability of machine learning algorithms to represent agricultural crop relations and predict their yields. Therefore, they are an adequate alternative to support farmers and stakeholders in the cocoa production chain. Comparative results indicated that the Gradient Boosting algorithm performs best with the highest value of R^2 and the lowest of MAE and RMSE. Also, relationships between variables are identified to improve the specific management of crops and, therefore, their productivity.

Variables such as radiation one month before harvest, rainfall on the harvest month, temperature one month before harvest, and soil moisture are the most important to explain the variability of crop yields. Sun-grown crops should have adequate management in their irrigation and fertilization systems, while shade-grown crops should have careful management of their forest plants. These results provide valuable information to make decisions targeted at crop requirements, which allows the implementation of a specific agriculture management that may not only improve the productivity, but also reduce costs. For instance, if there is no significant difference between the sun and shade yield, farmers should choose agroforestry systems with positive implications over soil management and moisture and temperature control. By doing so, the crop productivity won't be compromised. It is important to mention that results must be carefully interpreted, since the models are based on data taken from a specific site, and the performance of clones may vary according to geographic and environmental conditions. However, the methodological approach can be replicated in other study sites.

Future researches can consider multiple study sites to determine changes in crop yield influential variables according to crop location. Also, it is recommendable to incorporate other predictor variables in the models, such as the age of the cocoa plants, agricultural practices, or geographical location.

AUTHOR'S CONTRIBUTIONS

Lamos-Díaz provided the study methodology, statistical interpretation of results and review of the final manuscript. Puentes-Garzón performed the algorithm programming, data recollection and writing of the manuscript. Zarate-Caicedo carried out the experiment, agronomic interpretation of the results and review of the final manuscript.

REFERENCES

[1] D. Jiménez, J. Cock, A. Jarvis, J. Garcia, H. F. Satizábal, P. Van-Damme, A. Peréz-Uribe, and M. Barreto-Sanz, “Interpretation of commercial production information: A case study of lulo (Solanum quitoense), an under-researched Andean fruit,” *Agricultural Systems*, vol. 104 (3), pp. 258-270, Mar. 2011. https://doi.org/10.1016/j.agsy.2010.10.004

[2] J. W. Jones, J. M. Antle, B. Basso, K. J. Boote, R. T. Conant, I. Foster, H. C. J. Godfay, M. Herrero, R. E. Howitt, S. Janssen, B. A. Keating, R. Munoz-Carpena, C. H. Porter, C. Rosenzweig, and T. R. Wheeler, “Brief history of agricultural systems modeling,” *Agricultural Systems*, vol. 155, pp. 240-254, Jul. 2017. https://doi.org/10.1016/j.agsy.2016.05.014

[3] I. Diaz, S. M. Mazza, E. F. Combarro, L. I. Gimenez, and J. E. Gaiad, “Machine learning applied to the prediction of citrus production,” *Spanish Journal of Agricultural Research*, vol. 15 (2), e0205, Jun. 2017. https://doi.org/10.5424/sjar/2017152-9090

[4] S. T. Drummond, K. A. Sudduth, A. Joshi, S. J. Birrell, and N. R. Kitchen, “Statistical and neural methods

for site-specific yield prediction," *Transactions of the ASAE*, vol. 46 (1), pp. 5-14, 2003. https://doi.org/10.13031/2013.12541

[5] J. L. De Paepe, and R. Alvarez, "Wheat Yield Gap in the Pampas: Modeling the Impact of Environmental Factors," *Agronomy, Soils & Environmental Quality*, vol. 108 (4), pp. 1367-1378, 2016. https://doi.org/10.2134/agronj2015.0482

[6] J. D. R. Soares, M. Pasqual, W. S. Lacerda, S. O. Silva, and S. L. R. Donato, "Comparison of techniques used in the prediction of yield in banana plants," *Scientia Horticulturae*, vol. 167, pp. 84-90, Mar. 2014. https://doi.org/10.1016/j.scienta.2013.12.012

[7] A. Shekoofa, Y. Emam, N. Shekoufa, M. Ebrahimi, and E. Ebrahimie, "Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture," *PLoS One*, vol. 9 (5), e97288, May 2014. https://doi.org/10.1371/journal.pone.0097288

[8] J. R. Romero, P. F. Roncallo, P. C. Akkiraju, I. Ponzoni, V. C. Echenique, and J. A. Carballido, "Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires," *Computers and Electronics in Agriculture*, vol. 96, pp. 173-179, Aug. 2013. https://doi.org/10.1016/j.compag.2013.05.006

[9] X. Huang, G. Huang, C. Yu, S. Ni, and L. Yu, "A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging," *Field Crops Research*, vol. 211, pp. 114-124, Sep. 2017. https://doi.org/10.1016/j.fcr.2017.06.011

[10] A. A. V. da Silva, I. A. F. Silva, M. C. M. Teixeira Filho, S. Buzetti, and M. C. M. Teixeira, "Estimate of wheat grain yield as function of nitrogen fertilization using neuro fuzzy modeling," *Revista Brasileira de Engenharia Agrícola e Ambiental*, vol. 18 (2), pp. 180-187, Feb. 2014. https://doi.org/10.1590/S1415-43662014000200008

[11] I. Lopez, J. Plazas, and J. C. Corrales, "A tool for classification of cacao production in Colombia based on multiple classifier systems," in *17th International Conference Computational Science and Its Applications – ICCSA 2017*, Trieste, Italy, Jul. 2017. https://doi.org/10.1007/978-3-319-62395-5_5

[12] E. Somarriba, and J. Beer, "Productivity of Theobroma cacao agroforestry systems with timber or legume service shade trees," *Agroforestry Systems*, vol. 81, pp. 109-121, 2011. https://doi.org/10.1007/s10457-010-9364-1

[13] P. A. Zuidema, P. A. Leffelaar, W. Gerritsma, L. Mommer, and N. P. R. R. Anten, "A physiological production model for cocoa (Theobroma cacao): model presentation, validation and application," *Agricultural Systems*, vol. 84 (2), pp. 195-225, May 2005. https://doi.org/10.1016/j.agsy.2004.06.015

[14] L. F. García Carrión, *Catalogo de cultivares de cacao del Perú*, Lima: Ministerio de Agricultura y Riego, 2010.

[15] V. Vapnik, *The nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.

[16] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support Vector Regression Machines," *Neural Information Processing Systems*, vol. 9, pp. 1-11, 1997.

[17] T. Dietterich, *Ensemble Methods in Machine Learning. In: Multiple Classifier Systems*, Heidelberg: Springer Berlin, 2000.

[18] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29 (5), pp. 1189-1232, 2001.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45 (1), pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[21] T. M. Logan, S. McLeod, and S. Guikema, “Predictive models in horticulture: A case study with Royal Gala apples,” *Scientia Horticulturae*, vol. 209, pp. 201-213, Sep. 2016. https://doi.org/10.1016/j.scienta.2016.06.033

[22] A. Daymond, and P. Hadley, “The effects of temperature and light integral on early vegetative growth and chloroplyll fluorescence of four contrasting genotypes of cacao,” *Annals of Applied Biology*, vol. 145 (3), pp. 257-262, 2004. https://doi.org/10.1111/j.1744-7348.2004.tb00381.x

[23] Y. Ahenkorah, B. Halm, M. Appiah, and G. Akrofi, “Twenty Years’ Results from a Shade and Fertilizer Trial on Amazon Cocoa (Theobroma cacao) in Ghana,” *Experimental Agriculture*, vol. 23 (1), pp. 31-39, Jan. 1987. https://doi.org/10.1017/s0014479700003380

[24] O. Deheuvels, J. Avelino, E. Somarriba, and E. Malezieux, “Vegetation structure and productivity in cocoa-based agroforestry systems in Talamanca, Costa Rica,” *Agriculture, Ecosystems & Environment*, vol. 149, pp. 181-188, Mar. 2012. https://doi.org/doi: 10.1016/j.agee.2011.03.003

[25] W. Vanhove, N. Vanhoudt, and P. Van Damme, “Effect of shade tree planting and soil management on rehabilitation success of a 22-year-old degraded cocoa (Theobroma cacao L.) plantation,” *Agriculture, Ecosystems & Environment*, vol. 219, pp. 14-25, Mar. 2016. https://doi.org/doi: 10.1016/j.agee.2015.12.005

[26] B. Utomo, A. A. Prawoto, S. Bonnet, A. Bangviwat, and S. H. Gheewala, “Environmental performance of cocoa production from monoculture and agroforestry systems in Indonesia,” *Journal of Cleaner Production*, vol. 134 (Part B), pp. 583-591, Oct. 2016. https://doi.org/10.1016/j.jclepro.2015.08.102