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

Citación: 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

Recibido: Abril 7, 2020; **Aceptado:** Mayo 13, 2020;
Publicado: Mayo 15, 2020

Conflicto de intereses: Los autores declaran no tener conflicto de intereses.

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

Revista Facultad de Ingeniería (Rev. Fac. Ing.) Vol. 29 (54), e10477. 2020. Tunja-Boyacá, Colombia.
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 ($kg \cdot ha^{-1}$) 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. INTRODUCCIÓN

El cacao, es un producto agrícola tropical demandado a nivel mundial por diferentes industrias, representando una importante fuente de sustento económico para pequeños agricultores. En Colombia, se registró un incremento de 3.750 toneladas en la producción para 2017 en comparación con el año inmediatamente anterior, echo que representa un hito para el país y es consecuente con los esfuerzos de agricultores, agremiaciones y gobierno nacional. Sumado a esto, ha recibido la nominación de cultivo para la paz al ser una opción para la sustitución de cultivos ilícitos y generación de oportunidades laborales. No obstante, las causas del aumento en la producción se deben a la ampliación de la frontera agrícola y no por la mejora de la productividad y prácticas agrícolas, renovación de cultivos o el uso de nuevas tecnologías.

El aprendizaje automático se ha constituido como una alternativa para estudiar los rendimientos agrícolas e identificar los factores que explican su variabilidad, incluyendo condiciones de suelo y clima. Esta alternativa, considera cada cultivo como un experimento diferente y sus datos asociados proporcionan información adecuada para definir relaciones agrícolas que permitan realizar predicciones [1-3]. Drummon et al. [4] propusieron el uso de redes neuronales, *stepwise linear regression* y *project pursuit regression* para pronosticar el rendimiento del maíz y soja en Missouri, Estados Unidos, considerando las características físicas y químicas del suelo además de las condiciones climáticas. De la misma forma, De Paepe et al. [5] plantean analizar los efectos de las características del suelo y las condiciones climáticas en el rendimiento del trigo en las Pampas Argentinas utilizando redes neuronales. Por otro lado, diversos autores han optado por modelar el rendimiento en función de las características físicas de las plantas (fenotipo) [6-8]. Romero et al. [8] emplean algoritmos de clasificación OneR, IBK, C4.5 y Apriori para proporcionar reglas de asociación que permitan a los agricultores saber si su producción de trigo va a ser alta, baja o media de acuerdo al número espigas, altura de la planta, longitud del pedúnculo y fertilidad de la espiga. Otros autores han evaluado variables tales como la cantidad de fertilizante, fuente de fertilización, manejo de pestes y enfermedades y la variedad de la semilla [7, 9-10].

En cuanto a los cultivos de cacao, la predicción de los rendimientos ha sido abordada desde diferentes perspectivas. Corrales et al. [11] pronostican el nivel del rendimiento de cacao para Santander. Los autores evalúan la temperatura promedio diaria, la humedad relativa diaria y la tasa de precipitaciones diarias, utilizando diez algoritmos diferentes que son implementados en el software WEKA. Para ellos, Random Forest es la mejor alternativa de pronóstico para clasificar los niveles de rendimiento. Otros estudios relacionados con el cacao [12-13] evalúan el rendimiento usando modelos de regresión lineal, análisis de varianza (ANOVA) y modelos mecánicos como SUCROS, encontrando que las condiciones climáticas: temperatura, radiación y lluvias son las más críticas en la productividad del cacao.

De acuerdo con la literatura, los algoritmos de aprendizaje automático han resultado satisfactorios en diferentes cultivos tradicionales, tales como, el trigo, maíz, soja y arroz. Sin embargo, pocos estudios han evaluado los factores que afectan el rendimiento del cacao utilizando este enfoque y particularmente, evaluando la influencia de la sombra en sistemas agroforestales. Por lo tanto, la presente investigación evalúa algunos de los algoritmos más potentes y populares: Support Vector Machines, Random Forest, Gradient Boosting, y regresión LASSO, para pronosticar los rendimientos de cacao e identificar los factores que influyen en él. A su vez, a partir de un análisis de la influencia marginal se establecen los factores que afectan el rendimiento de cacao dependiendo del tipo de exposición solar (exposición a sol o exposición a sombra), lo cual es clave para diferenciar el manejo agronómico y así maximizar la productividad.

II. Materiales y métodos

Los datos representan el insumo más importante para la construcción de los modelos predictivos basados en algoritmos de aprendizaje automático. A continuación, se describe el diseño experimental empleado para la obtención de los datos y las fuentes secundarias consultadas. Igualmente, se presentan los algoritmos utilizados y las métricas con las que se comparan los diferentes desempeños.

A. Adquisición de datos

En el desarrollo de la investigación se utilizó una plantación experimental localizada en el municipio de Rionegro y establecida en el 2008 a una altitud de 550 msnm en el centro de investigación “La Suiza” propiedad de la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA). Los datos fueron tomados para el periodo 2015-2017. El diseño experimental consistió en un diseño factorial aleatorizado 10 x 3 x 2 por bloques con 3 repeticiones. Los tratamientos consistieron en la combinación de 10 variedades de cacao (5 universales y 5 regionales) [14], 3 niveles de fertilización y 2 tipos de exposición solar (Tabla 1).

Tabla 1. Factores y niveles del diseño experimental.

Clones	Fertilización	Exposición
Clones regionales: SCC-19, SCC-52, SCC-61, SCC-64, SCC-83 *Clones universales:* CNN-51, EET-8, ICS-1, ICS-95, TSH-565	50%, 100%, 50%	Sol/sombra

La fertilización está asociada con el porcentaje del criterio base, el cual incluye urea, diamonio de fosfato (DAP), cloruro de potasio (KCl), materia orgánica, azúfre (S), sulfato de magnesio ($MgSO_4$) y boro (B).

Los tratamientos fueron aplicados a cinco plantas por réplica, para un total de 900 plantas por hectárea (450 exposición a sol y 450 exposición a sombra). La sombra es suministrada por *Cariniana pyriformis Miers* and *Tectona grandis L.f*, con un promedio de altura de 12 [m] por árbol y una densidad de 340 [árboles/ha].

También, los modelos consideran las características físicas y las condiciones climáticas (Tabla 2), las cuales fueron medidas diariamente por una estación meteorológica ubicada en la región (Watchdog serie 2000, Spectrum Technologies Inc, Aurora, IL, USA) y datos de fuentes secundarias como el Instituto de Hidrología, Meteorología y Estudios ambientales (IDEAM).

Tabla 2. Entradas para el desarrollo de los modelos predictivos del rendimiento de cacao.

Nombre de la variable	Significado	Tipo	Nombre de la variable	Significado	Tipo
Cocoa_v	Variedad de Cacao	Cat[a]	P_accu_prev	Lluvias acumuladas un mes previo a la cosecha	Con[b]
Exp	Exposición	Cat[a]	T_avg	Temperatura promedio el mes de la cosecha	Con[b]

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

Nombre de la variable	Significado	Tipo	Nombre de la variable	Significado	Tipo
F_level	Nivel de fertilización	Cat[a]	T_avg _prev	Temperatura promedio un mes previo a la cosecha	Con[b]
EC_avg	Electro conductividad el mes de la cosecha	Con[b]	Rad_accu	Radiación fotosintéticamente activa (PAR) acumulada	Con[b]
Hum_avg	Humedad promedio del suelo en el mes de la cosecha	Con[b]	Rad_accu_prev1	Radiación fotosintéticamente activa (PAR) acumulada un mes previo a la cosecha	Con[b]
P_accu	Lluvia acumulada en el mes de la cosecha	Con[b]	Rad_accu_prev2	Radiación fotosintéticamente activa (PAR) acumulada dos meses previos a la cosecha	Con[b]

a Variable categórica, b Variable continua

B. Modelo de regresión lineal

La regresión lineal LASSO (Least Absolute Shrinkage and Selection Operator) es un modelo estadístico que permite relacionar un conjunto de variables independientes (predictores) con uno de variables dependientes (respuesta). A diferencia del modelo de regresión lineal clásico, LASSO incluye un factor de regularización en los coeficientes de regresión utilizando la norma L1 (valor absoluto), como se muestra en la ecuación (1).

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

Donde y es un vector (n x 1) de observaciones, x son vectores de las variables independientes, β corresponde a los coeficientes de regresión e α indica que tan grande es la penalización, por lo que un valor elevado lleva a coeficientes bajos o casi nulos, mientras que un valor reducido llevaría a tener una regresión lineal clásico, motivo el que se su valor se determina mediante validación cruzada.

C. Support Vector Machines (SVM)

SVM es un algoritmo no paramétrico basado en la teoría de aprendizaje estadístico, que busca identificar un hiperplano de separación donde el margen de separación entre las observaciones positivas y negativas es máximo. Inicialmente, Vapnik [15] propuso este algoritmo para problemas de clasificación, sin embargo, esto ha sido

extendido a problemas de regresión [16]. El objetivo es minimizar la medida de error entre los datos observados (variable dependiente) y una familia de funciones F(x,w) parametrizada por w, y x como datos de entrada (variables independientes).

D. Modelos de aprendizaje ensamblados

Los métodos ensamblados parten de la premisa que varios son mejores que uno, por lo que mejoran la precisión en la predicción mediante la agregación de múltiples algoritmos de aprendizaje, siempre y cuando estos sean independientes [17]. En la construcción de un ensamble se distingue una jerarquía de algoritmos: algoritmos base y algoritmos de agregación, los primeros serán utilizados múltiples veces para generar múltiples predicciones que se irán agregando. En general, suele utilizarse los árboles de regresión. Por otro lado, los algoritmos de agregación manipulan las entradas para los algoritmos base de tal manera que los modelos generados sean independientes. Para este trabajo se consideran los algoritmos de agregación:

1) Boosting: Es un procedimiento iterativo para cambiar de manera adaptativa la distribución de las muestras de entrenamiento, de tal manera que el algoritmo base se enfoque en las muestras difíciles de pronosticar. En cada iteración, se asignan pesos a cada observación de entrenamiento y se actualizan de acuerdo con el error respecto a los valores observados. Dos de los algoritmos más populares son AdaBoost y Gradient Boosting, donde este último hace el entrenamiento de los algoritmos base con los errores de la iteración anterior y mediante el gradiente de descenso maximiza la precisión de la predicción [18].

2) Random Forest: Fue propuesto por el profesor Leo Breiman [19], este algoritmo es una combinación de predicciones de múltiples árboles de regresión, donde cada uno depende de un conjunto de vectores de variables independientes aleatorias y con la misma distribución de probabilidad.

E. Métricas de evaluación

Permiten evaluar el desempeño de un modelo y comparar diferentes propuestas. Algunas de estas métricas son: raíz cuadrada del error medio (RMSE), error medio

absoluto (MAE) y el coeficiente de determinación (R^2), como se muestra en las ecuaciones (2) a (4).

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

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

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

Donde Oi y Pi son los valores medidos y pronosticados para la i-ésima observación, respectivamente. $\bar{O}$ y $\bar{P}$ son los valores medios del rendimiento medido y pronosticado, y n es el número total de observaciones. La mejora relativa de RMSE (RI_{RMSE}), dada en la ecuación (5), fue usada para medir y comparar los algoritmos y establecer la alternativa que mejor se ajusta a los rendimientos observados en el cultivo.

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

$RMSE_R\ y\ RMSE_c$ son los valores de desempeño del algoritmo de referencia y del que se quiere comparar.

III. Resultados y discusión

Inicialmente, se ajustan los modelos para el conjunto de datos completo (rendimiento de cacao y las variables descritas en la tabla 1), es decir, se incluye el tipo de exposición como variable independiente. Posteriormente, se evalúa individualmente la exposición bajo sol (284 observaciones) y bajo sombra (274 observaciones) para determinar si son influenciados por las mismas variables. El entrenamiento de cada modelo se hace usando el 80% de los datos y el 20% restante para prueba, de igual manera, se realiza validación cruzada con k=10 para establecer los parámetros de cada algoritmo. Para esto, se utilizó el módulo "model_selection" del paquete sklearn [20].

A. Evaluación de modelos

En la Tabla 3, se presentan los resultados medios para las métricas desempeño en la validación Hold-out. En promedio, el desempeño de Gradient Boosting es superior al de los demás algoritmos con los menores valores de MAE, RMSE y el valor más

alto de R^2. Por otro lado, la mejora relativa en el RMSE es de 20.99, 8.54 y 5.93 % sobre LASSO, SVM y Random Forest, respectivamente.

Tabla 3. Desempeño promedio de los algoritmos para el conjunto de datos completo.

Modelo	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	14.70	26.65	44.19	5.93
Gradient boosting	12.94	25.07	49.29	-

Para cada una de las 100 repeticiones en la validación Hold-out, el algoritmo es entrenado y probado con una partición aleatoria, donde la mejor alternativa resultó ser Gradient Boosting, con 480 árboles. Esto explica el 68% de la variabilidad, y presenta un MAE de 13.32, y RMSE de 20.41 (Figura 1)

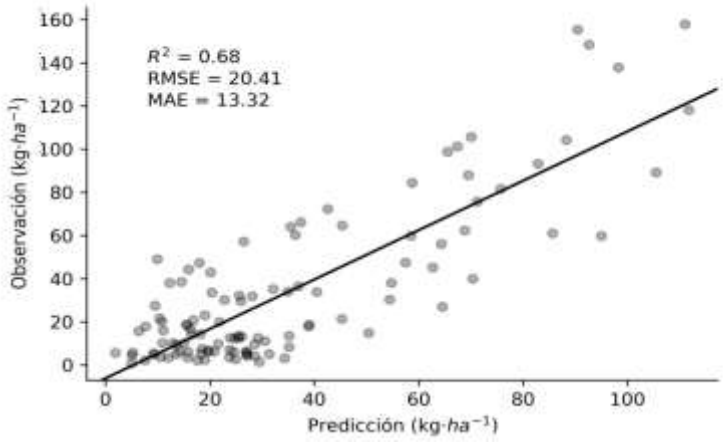

Fig. 1. Predicción del rendimiento de cacao con Gradient Boosting.

Una vez el mejor modelo es identificado, el siguiente paso es evaluar la importancia de las variables. Para Gradient Boosting, la métrica MSE Friedman es utilizada para seleccionar las variables que mejoran la calidad de la partición. No obstante, para hacer comparaciones en el impacto de cada variable en los diferentes algoritmos y cuantificar su impacto, es necesario acudir a otras estrategias. Los gráficos de dependencia parcial ilustran la influencia marginal cuando una variable cambia y

mantiene las demás variables constantes. Logan et al. [21] proponen una alternativa para medir la influencia marginal dada por la ecuación 6.

$$Oscilacaión_n = \frac{\max(V_n) - \min(V_n)}{\sum_n Oscilación_n} \quad (6)$$

La Tabla 4 muestra las variables con las oscilaciones más grandes considerando el mejor modelo identificado en la fase de validación.

Tabla 4. Variables para el conjunto de datos complete considerando Gradient Boosting

Variable	Oscilación	Variable	Oscilación
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

Los resultados de la Tabla 4 indican que la temperatura promedio un mes previo a la cosecha, la radiación acumulada un mes previo a la cosecha y las lluvias acumuladas el mes de la cosecha son los factores que mayor incidencia tienen sobre el rendimiento del cultivo. De acuerdo con Daymon et al. [22], la temperatura es uno de los factores que limita la producción de cacao al causar un estrés sobre la planta que afecta la variabilidad estacional. En cuanto a la radiación y la lluvia, se demuestra que están relacionadas con la etapa final del crecimiento de la mazorca de cacao donde la lluvia es más importante que la radiación [13]. En cuanto a las variables de exposición al sol y sombra la oscilación es cercana a 0, por lo que no son representativas para el modelo predictivo.

La evaluación de la influencia de las variables es desarrollada utilizando gráficos de dependencia parcial para las interacciones entre las lluvias, temperatura y radiación un mes antes de la cosecha y Gradient Boosting (Figura 2).

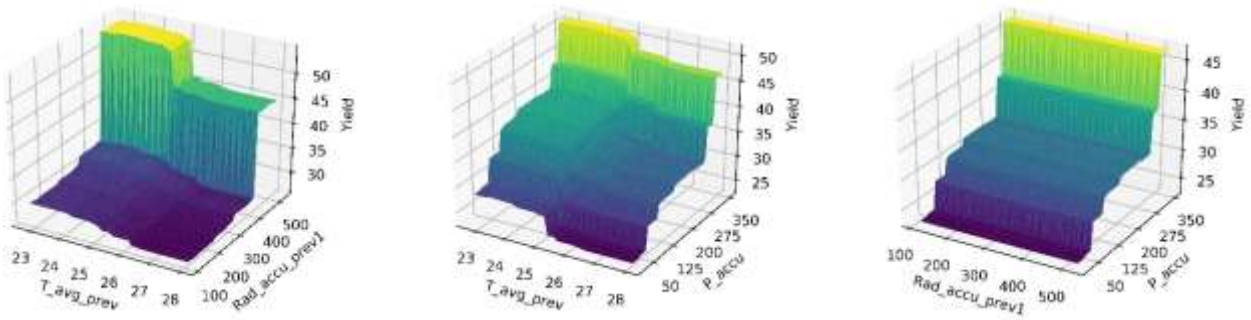

Fig. 2. Gráficos de dependencia parcial para la interacción de las variables identificadas.

Al observar el eje vertical o rendimiento se identifica que la interacción entre la radiación un mes previo a la cosecha y las lluvias acumuladas tienen el menor efecto, mientras que las interacciones con la temperatura llevan a rendimientos más altos. De acuerdo con este resultado, en el mejoramiento del cultivo se debería incluir el manejo de la temperatura junto a la radiación o lluvias acumuladas como elementos a controlar para aumentar la productividad. A su vez, se observa que el efecto de la radiación se ve mermado cuando interactúa con las lluvias, lo que ratifica a las lluvias acumuladas como la variable más influyente sobre los rendimientos del cultivo.

B. Modelos para exposición a Sol/Sombra

En el segundo escenario, para cada tipo de exposición (sol-sombra) el mejor algoritmo identificado se prueba una vez más. La Tabla 5 muestra que la variabilidad es mejor explicada en el modelo de cultivo bajo sombra con un valor medio de R^2 de 54.27% y menores valores de MAE y RMSE comparado con el cultivo expuesto a sol.

Tabla 5. Desempeño promedio para los dos tipos de exposición.

Métrica	R^2	RMSE	MAE
Sol	42.28%	25.98	14.67
Sombra	54.27%	24.51	11.25

El procedimiento descrito en la sección anterior para evaluar la importancia de las variables es utilizado esta vez para identificar si el rendimiento de los cultivos bajo

sol y bajo sombra está relacionado con las mismas variables. La Tabla 6 sugiere que la lluvia acumulada el mes de cosecha y la humedad promedio son quienes mayor influencia ejercen sobre el modelo predictivo cuando el cultivo está expuesto a sol. Este resultado se asocia a los mayores requerimientos de agua dado que hay una mayor exposición a sol, lo que aumenta la transpiración de las hojas y la evaporación en el suelo, teniendo como consecuencia menores tasas fotosintéticas y mayor cierre estomático. La última afirmación implica periodos cortos de producción, mayores requerimientos de nutrientes, manejo de sistemas de irrigación y por lo tanto, mayor inversión [23].

Para el caso bajo sombra, la radiación un mes previo a la cosecha tiene el valor más alto de oscilación, lo cual indica una fuerte relación entre esta variable y el rendimiento del cultivo. Como afirmó Zuidema et al. [13], la sombra debe ser administrada de forma adecuada para garantizar que no haya una reducción en el rendimiento por la falta de radiación.

Tabla 6. Importancia de las variables para los modelos con exposición bajo sol y bajo sombra.

SOL		**SOMBRA**	
Variable	**Oscilación**	**Variable**	**Oscilación**
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

En la Tabla 6, además de variables asociadas al clima también se incluyen variedades de cacao, por lo que se evalúa y compara el rendimiento de los clones en los dos tipos de exposición. Se utiliza una prueba Kruskal-Wallis con un nivel de significancia del 5%, planteando como hipótesis nula que las medianas de las muestras son iguales. Con valores $p < 0.05$ los resultados de la Tabla 7 sugieren que los clones CNN51 y SCC64 tienen un comportamiento diferente para los dos tipos de exposición, y de acuerdo con el valor promedio del rendimiento ambos clones tienen mejor desempeño con exposición plena a sol.

Tabla 7. Test de Kruskal-Wallis.

Clon de cacao	Rendimiento promedio sol	Rendimiento promedio sombra	Mediana del rendimiento Sol	Mediana del rendimiento Sombra	Valor p
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

En general, se observa que no hay diferencia en el rendimiento de los cultivos cuando están bajo sol o sombra, lo cual es positivo para la promoción de los cultivos agroforestales y concuerda con los hallazgos de [12-13, 24], quienes manifiestan que la sombra en los cultivos agroforestales no afectan el rendimiento del cacao siempre y cuando sea proporciona de manera adecuada, mientras que, [25-26] afirman que en la exposición moderada a sombra el rendimiento no se ve afectado y tiene implicaciones positivas en el manejo de suelos, control de la humedad y temperatura del ambiente y en la creación de entornos que mejoren la fisiología del cacao y reduzcan el impacto de las pestes y enfermedades.

IV. Conclusiones

Esta investigación demuestra la habilidad de los algoritmos de aprendizaje automático para representar relaciones en cultivos agrícolas y predecir sus rendimientos. Esto es porque estos son, una alternativa adecuada para apoyar a los agricultores e interesados en la cadena de suministro del cacao. Los resultados comparativos indican que el algoritmo Gradient Boosting se desempeña mejor con el valor más alto de R^2 y el más bajo de MAE y RMSE. A su vez, se identifican relaciones entre las variables buscando mejorar el manejo específico del cultivo, y, por lo tanto, su productividad.

Variables como la radiación un mes previo a la cosecha, lluvias un mes previo a la cosecha, temperatura un mes previo a la cosecha y la humedad del suelo son las más importantes para explicar la variabilidad del rendimiento del cultivo. En

particular, los cultivos expuestos al sol deberían tener un manejo adecuado de sus sistemas de irrigación y fertilización, mientras que los cultivos en sombra deberían ser cuidadosos en el manejo de las plantas forestales. Esos resultados suministran información valiosa orientada para la toma de decisiones asociadas a los requerimientos del cultivo, lo cual permite no solamente mejorar la productividad, sino también, reducir los costos. Por lo tanto, si no hay diferencia significativa entre los cultivos a sol o sombra, los agricultores podrían optar por los sistemas agroforestales con implicaciones positivas en el manejo del suelo, humedad y control de la temperatura. Al hacer esto, la productividad del cultivo no se verá comprometida.

Es importante resaltar que los resultados deben ser interpretados con cuidado, dado que al ser modelos basados en datos los hallazgos aplican para el sitio en el que fueron tomados, ya que el comportamiento de los clones podría variar de acuerdo a las condiciones geográficas y ambientales. Ahora bien, la metodología empleada puede ser replicada para estudiar otras locaciones.

Como futuros trabajos, se recomienda la consideración de múltiples ubicaciones para determinar si las variables identificadas en este trabajo tienen la misma influencia sobre el rendimiento. También sería interesante la inclusión de otras variables predictoras como la edad de las plantas de cacao, las prácticas agrícolas o la posición geográfica.

CONTRIBUCIÓN DE LOS AUTORES

Lamos-Díaz propuso la metodología de estudio, interpretación estadística de los resultados y la revisión del manuscrito final. Puentes-Garzón desempeñó la programación de los algoritmos, recolección de los datos y escritura del manuscrito. Zarate-Caicedo llevó a cabo el experimento, interpretación agronómica de los resultados y revisión del documento final.

REFERENCIAS

[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