Ciencias Agrarias/ Agricultural Sciences
Revista Ciencia y Tecnología (2026) 19(2) p 1 - 10 ISSN 1390-4051; e-ISSN 1390-4043 https://doi.org/10.18779/cyt.v19i2.1149

Predicción del crecimiento de plantas de *Coffea* mediante machine learning basado en variables climáticas y agronómicas

Prediction of *coffee* plant growth using machine learning based on climatic and agronomic variables

Ricardo Augusto Luna Murillo[1], Bryan Alexander Alvarez Real[2], Joshua Emanuel Vinces Manrique[2], Johnny Xavier Bajaña Zajia[2]

[1]*Universidad Técnica de Cotopaxi, Ecuador*
[2]*Universidad Técnica de Cotopaxi - Extension La Maná, Ecuador*
Autor de correspondencia: patoricardo@yahoo.es

Recibido: 10/07/2025. Aceptado: 05/05/2026.
Publicado el 3 de julio de 2026.

Resumen

El cultivo de café en zonas subtropicales enfrenta desafíos crecientes debido a la variabilidad climática y a la necesidad de toma de decisiones agronómicas más precisas, considerando este punto, las técnicas de aprendizaje automático se posicionan como una herramienta crucial para apoyar la gestión agrícola, este estudio expone el desarrollo e integración de un modelo de refuerzo de gradiente para predecir el crecimiento en altura y diámetro del tallo de plantas de café de las variedades Manabí 01 y Sarchimor, utilizando variables climáticas y agronómicas recolectadas mediante sensores instalados en el sector de Sacha Wiwa, La Maná. El modelo con coeficiente de determinación de R^2=0,554 para altura y R^2=0,535 para diámetro, mostró un desempeño adecuado para predecir el crecimiento de las plantas de café utilizando variables climáticas y agronómicas. Estos resultados, si bien pueden mejorar mediante la optimización del algoritmo y la inclusión de mayores cantidades de datos, evidencian el potencial del machine learning como herramienta para la toma de decisiones agronómicas. Como aplicación práctica, el modelo fue posteriormente integrado en una aplicación móvil para permitir la realización de predicciones en tiempo real y facilitar el monitoreo del desarrollo de las plantas en campo. Aunque se observaron márgenes de error propios de complejidad del entorno agrícola, los resultados mostraron un alto potencial de mejora mediante el uso de conjuntos de datos, más robustos y la optimización del algoritmo, y con ello, esta investigación busca acercar el machine learning al campo, fortaleciendo la toma de decisiones agronómicas basadas en datos concretos.

Palabras clave: café, algoritmo, refuerzo de gradiente, agricultura, aprendizaje supervisado, factores agroclimáticos.

Abstract

Coffee cultivation in subtropical regions faces increasing challenges due to climate variability and the need for more precise agronomic decision-making. Considering this, machine learning techniques are emerging as a crucial tool to support agricultural management.This study presents the development and integration of a gradient boosting model to predict plant growth in terms of height and stem diameter for coffee varieties Manabí 01 and Sarchimor, using climatic and agronomic variables collected through sensors installed in the Sacha Wiwa sector, La Maná. The model, with a coefficient of determination of R^2 = 0.554 for height and R^2 = 0.535 for diameter, showed adequate performance in predicting coffee plant growth based on climatic and agronomic variables. Although these results can be improved through algorithm optimization and the inclusion of larger datasets, they demonstrate the potential of machine learning as a tool for agronomic decision-making.As a practical application, the model was later integrated into a mobile application to enable real-time predictions and facilitate field monitoring of plant development. Although some margins of error were observed due to the complexity of the agricultural environment, the results showed strong potential for improvement through the use of more robust datasets and algorithm optimization. Ultimately, this research aims to bring machine learning closer to the field, strengthening agronomic decision-making based on concrete data.

Keywords: Coffee cultivation, algorithm, gradient boosting, agriculture, growth prediction, agroclimatic factors.

Introducción

El cultivo de café es un pilar fundamental de la economía agrícola en numerosas regiones subtropicales, en Sacha Wiwa, La Maná, la variación climática y la necesidad de una gestión eficiente de los recursos limitan constantemente la productividad y la calidad de la cosecha de café (Ahmed *et al.*, 2021). En los últimos años, el sector agrícola tiene cada vez más a soluciones innovadoras para abordar estos desafíos, la inteligencia artificial (IA) y machine learning (ML), emergen como herramientas poderosas, al aprovechar datos sobre el clima, las propiedades del suelo y las condiciones de las plantas, estas tecnologías permiten a los productos prever posibles dificultades y optimizar la toma de decisiones.

La ventaja de predecir una temporada seca con semanas de antelación, permite una planificación del riego estratégica, comprender la humedad del aire influye en el desarrollo de las plantas ayudando a identificar los momentos ideales para la poda y la fertilización como lo menciona Astri Muliasari *et al*. (2021), esta mejora no solo aumenta la eficiencia de la producción, sino que también fomentan la sostenibilidad, un factor clave para garantizar que el café de lugares como La Maná, Ecuador, se mantenga en el mercado nacional e internacional.

En regiones cafeteras semejantes al cantón La Maná, la aplicación de machine learning (aprendizaje automático) ha permitido anticipar el ciclo de forestación y estimar rendimientos con varios meses de antelación reduciendo pérdidas asociadas a eventos climáticos extremos, en zonas con limitada conectividad o acceso a servicios técnicos, estos modelos pueden integrar en aplicaciones móviles o sistemas locales respaldando a pequeños y medianos productores en su labor cotidiana, el machine learning funciona como herramienta tecnológica y mecanismo de democratización del conocimiento y resiliencia productiva (Mena Iza *et al.*, 2024).

Este proyecto surge de la colaboración entre dos actores clave: La Universidad Técnica de Cotopaxi- Extensión La Maná, en representación del sector académico, y FIASA, líder en el sector productivo, juntos comparten la visión de elevar la capacidad productiva y la competitividad internacional del café ecuatoriano, más allá de los resultados inmediatos, esta colaboración refleja un compromiso más amplio con la integración del conocimiento científico con la aplicación práctica, ofreciendo beneficios tangibles a los cafeteros y promoviendo el avance de la agricultura de precisión.

Este proyecto tiene como objetivo dotar a los caficultores de herramientas para planificar y gestionar las etapas de crecimiento de las plantas de café de forma más eficaz, mediante la integración de inteligencia artificial, se busca optimizar la precisión de sus decisiones, impulsando tanto las prácticas agrícolas como la sostenibilidad regional.

Metodología

Sobre esta base, nuestra investigación se centra en la implementación de un sistema predictivo basado en técnicas de ML, diseñado para estimar el crecimiento en altura y el diámetro del tallo de las plantas de café, las variedades de café presentes en el conjunto de datos incluyen Manabí 01 (896 registros) y Sarchimor (768 registros), cultivadas en Sacha Wiwa, La Maná y para lograrlo, nuestro equipo empleó diversos algoritmos, incluyendo stacking regressor, gradient boosting y XGBoost regressor, todo esto se entrenará con una combinación de datos históricos y en tiempo real, capturando factores climáticos como temperatura, precipitación y humedad relativa, atributos del suelo como pH, textura y niveles de nutrientes, prácticas agronómicas como la fertilización y el riego.

El procedimiento de datos se realiza mediante el lenguaje de programación Python, con el apoyo de bibliotecas especializadas: Pandas para la organización y preprocesamiento de datos, y Scikit-learn para la implementación de modelos de aprendizaje automático como lo fue gradient boosting (refuerzo por gradiente), la metodología CRISP-DM guía este esfuerzo y estructura nuestro flujo de trabajo desde la recopilación inicial de datos hasta la validación de modelos predictivos, garantizando rigor y fiabilidad (León Chilito *et al*., 2025).

La metodología del proyecto se estructuró siguiendo el estándar CRISP-DM, que organiza el ciclo de desarrollo en fases iterativas: entendimiento del problema, comprensión y preparación de los datos, modelado, evaluación y despliegue (Schröer, 2021). Durante el proceso de recolección de datos, se instalaron sensores climáticos y agronómicos en el sector de Sacha Wiwa (La Maná) y se monitoreó el cultivo de café en campo durante el periodo de octubre de 2023 a julio de 2024. Las variables registradas incluyeron temperatura ambiental, precipitación, humedad relativa, pH, textura del suelo, fertilización y frecuencia de riego, así como datos de crecimiento (altura y diámetro del tallo) de las plantas de las variedades estudiadas.

En la fase de entendimiento del problema se definieron claramente los objetivos: predecir el crecimiento de las plantas de café (altura de la planta y diámetro del tallo) a partir de variables climáticas (temperatura, precipitación, humedad relativa) y agronómicas (pH, textura del suelo, fertilización, riego). Para ello, se instaló un conjunto de sensores meteorológicos y sensores del suelo en el sector de Sacha Wiwa (0°47’45.7”S 79°09’10.6”W con una altitud aproximada es de 500 m sobre el nivel del mar) en La Maná, los cuales permitieron recopilar datos precisos y en tiempo real, sobre las condiciones ambientales y edáficas de la zona de cultivo en base a Waqas *et al. (*2025), de esta manera se logró obtener una base de datos con toda la información necesaria para poder alimentar el modelo de aprendizaje automático. Asimismo, se establecieron los indicadores de desempeño del

modelo: RMSE, MAE y R^2, que guiarán su evaluación final. Esta fase inicial aseguró que los resultados tuvieran relevancia práctica para los caficultores de Sacha Wiwa (Botero-Valencia *et al*., 2025).

La investigación se enmarca dentro del enfoque cuantitativo, apoyándose en el análisis numérico de datos registrados mediante dispositivos especializados en condiciones reales de cultivo. El conjunto de datos utilizado para el entrenamiento y evaluación del modelo está compuesto por 1.664 registros y 28 variables. De estas, 21 son numéricas como temperatura, humedad del aire y del suelo, presión atmosférica, nutrientes y las variables objetivo: altura de planta (AP) y diámetro del tallo (DT), y 7 categóricas —como la variedad de café (Manabí 01, Sarchimor), la edad en días, temperatura ambiental, humedad ambiental, humedad del suelo, presión atmosférica, temperatura suelo, índice de lluvia, pH (potencial de hidrogeno) CE (conductividad eléctrica), MO (materia orgánica), NH4 (amonio), P (fósforo), S (azufre), K (potasio), Ca (calcio), Mg (magnesio), Cu (cobre), B (boro), Fe (hierro), Zn (zinc), Mn (manganeso), N total (nitrógeno total), arena, limo, arcilla, altura de planta (AP), diámetro de tallo (DT).

La diversidad y volumen de los datos permitió capturar una representación completa de los factores agronómicos y climáticos que influyen en el desarrollo del cafeto, proporcionando así una base sólida para el aprendizaje y la predicción por parte del modelo.

Recolección y preparación de datos

Los datos se obtuvieron de tres fuentes complementarias extraídas de Sacha Wiwa en el cantón La Maná: (i) una estación meteorológica local que registró variables climáticas horarias (temperatura, precipitación, humedad relativa); (ii) sensores de suelo instalados en parcelas de cafeto, que aportaron mediciones de humedad, porosidad y otros parámetros físicos; y (iii) registros agronómicos de campo que incluyeron características del suelo (pH, textura) y prácticas de manejo (esquema de fertilización y riego). Cada conjunto de datos se exportó en formato estructurado y se documentó su esquema.

En la fase de comprensión de datos se realizó un análisis exploratorio preliminar para verificar rangos plausibles, detectar valores faltantes y estimar correlaciones iniciales entre variables. A continuación, se ejecutó la preparación de datos, aplicando técnicas estándar de limpieza y preprocesamiento, esto incluyó las siguientes tareas principales:

Tratamiento de valores faltantes y atípicos

Se identificaron registros con datos incompletos (como, datos climáticos ausentes) y se imputaron valores cuando fue apropiado (usando la mediana o promedio) o bien se eliminaron registros muy incompletos. También se detectaron valores atípicos fuera de los rangos esperados, los cuales se revisaron con base en el conocimiento agronómico y se corrigieron o eliminaron si correspondían (Omoseebi *et al*., 2025) (Tabla 1).

Tabla 1. Tratamiento de valores faltantes y atípicos

Columnas con valores nulos	Imputación aplicada	Registros eliminados	Variables con valores atípicos	Método para detectar outliers
Sin valores nulos	Media/Mediana según el tipo de variable y criterio agronómico	Registros con múltiples nulos o inconsistencia severa	Temperatura_ambiental=256 Temperatura_suelo: 384 Altura_planta: 44 Diametro_tallo: 56	Rango intercuartílico (IQR)

Normalización de características numéricas: Las variables continuas se escalaron mediante normalización (min-max) y estandarización (media 0, varianza 1) para facilitar el entrenamiento de los modelos y evitar que variables de mayor magnitud dominen el ajuste (Prity, 2024).

Codificación de variables categóricas: Las variables cualitativas nominales (como el tipo de café) se transformaron mediante codificación one-hot, creando variables binarias (0/1) para cada categoría mientras se eliminaba una categoría base para evitar multicolinealidad. Para evitar multicolinealidad tras aplicar codificación one-hot, se elimina una categoría base. No obstante, en modelos con regularización Ridge (usado como meta-modelo en el stacking), la multicolinealidad residual es gestionada por la propia penalización del modelo, lo que aporta robustez y estabilidad a las estimaciones de los coeficientes. Para variables ordinales con jerarquía inherente (como textura del suelo), se aplicó codificación ordinal asignando valores numéricos que preservan el orden de las categorías. Esta transformación permite la incorporación de estas variables en modelos de aprendizaje supervisado manteniendo su significado semántico (Bolikulov *et al*., 2024). En este estudio, las variables objetivo o etiquetas utilizadas para el entrenamiento y evaluación de los modelos de aprendizaje supervisado fueron “altura de planta (AP)” y “diámetro del tallo (DT)”, ambas registradas como variables continuas que reflejan el crecimiento y desarrollo morfológico del cafeto bajo diferentes condiciones ambientales y agronómicas (Figura 1).

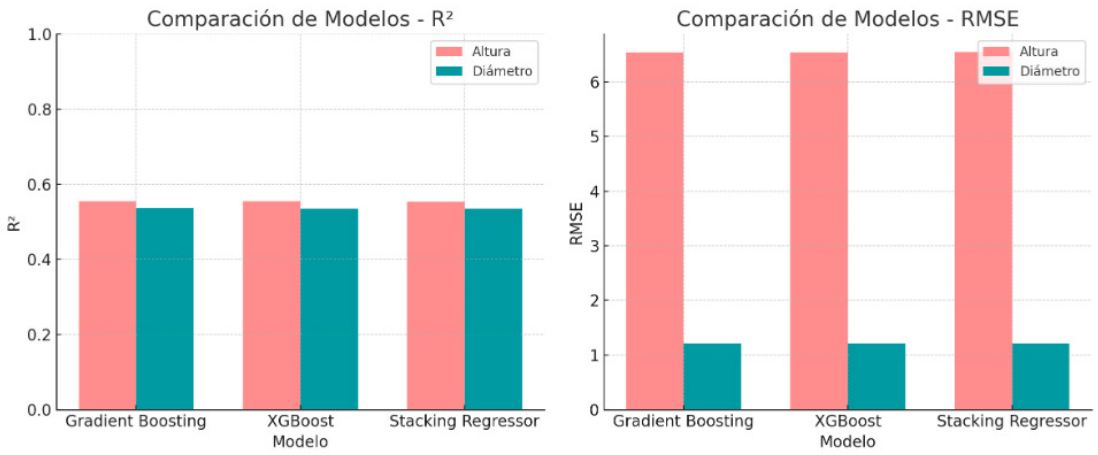

Figura 1. Registro de modelos y su coeficiente de promedio

Tabla 2. Codificación de variable categórica en el conjunto de datos de café

Información	Valor
Variables categóricas	TIPO_DE_CAFE
Distribución por categoría	Manabi 01: 896 Sarchimor: 768
Estrategia de codificación	One-Hot Encoding
Nuevas variables generadas	TIPO_DE_CAFE_Sarchimor

División en conjuntos de entrenamiento y prueba: Finalmente, el conjunto de datos limpio se dividió aleatoriamente en subconjuntos de entrenamiento del 80 % y un subconjunto de prueba del 20 %. Esta partición inicial permitió reservar datos independientes para validar el desempeño final. Además, se configuró la validación cruzada de k- pliegues (k=5) sobre el subconjunto de entrenamiento para ajustar hiper parámetros y evaluar la generalización de los modelos (Prity *et al*., 2024) (Tabla 2).

Evaluación comparativa de modelos de aprendizaje automático

Para predecir las variables objetivo como la altura y el diámetro del tallo de las plantas de café, se implementaron tres enfoques avanzados de aprendizaje automático utilizando el lenguaje de programación Python y scikit-learn, los algoritmos de machine learning, gradient boosting, XGBoost y stacking regressor, fueron elegidos por su capacidad para modelar relaciones complejas y su robustez estadística (Sharma *et al*., 2023). Cada modelo fue optimizado mediante validación cruzada, utilizando métricas estandarizadas (R^2, MSE, MAE) para garantizar comparabilidad, este enfoque metodológico permitió evaluar sistemáticamente el rendimiento predictivo de cada técnica antes de mostrar la selección final.

La selección de los hiperparámetros del modelo se realizó considerando el tamaño y la variabilidad del conjunto de datos, así como la necesidad de evitar tanto el sobreajuste como el subajuste. Un número de 200 árboles (n_estimators) proporciona suficiente capacidad de aprendizaje sin incurrir en excesivo tiempo de cómputo, mientras que una tasa de aprendizaje de 0,1 permite un ajuste gradual de los árboles que favorece la generalización del modelo. La profundidad máxima de 3 niveles (max_depth=3) fue elegida para limitar la complejidad de cada árbol, controlando el riesgo de sobreajuste en base de datos de tamaño moderado. Para XGBoost, se exploraron valores de max_depth entre 4 y 8 y tasas de aprendizaje entre 0,01 y 0,1 usando grid search, buscando el mejor equilibrio entre bajo error y estabilidad del modelo. Otros hiperparámetros como gamma y colsample_bytree se optimizaron para mejorar la robustez y la generalización, ajustando la selección de características y la regularización durante el entrenamiento.

El modelo stacking regressor (apilamiento de regresores) combinó los modelos GBR y XGB junto con lightGBM como estimadores base, utilizando ridge regression (regresión de crestas) como meta-modelo, esta estrategia de ensamblado jerárquico demostró superioridad al integrar las fortalezas predictivas de cada algoritmo individual, cada componente del modelo stacking fue previamente ajustado mediante validación cruzada, asegurando coherencia en la pipeline, es importante destacar que modelos lineales ya árboles de decisión individuales fueron descartados en fases preliminares por su menor capacidad predictiva en este contexto específico (Sharma *et al*., 2023).

Validación y selección del modelo de aprendizaje

El proceso de validación del modelo gradient boosting seleccionado se desarrolló mediante una metodología estructurada, que garantizó la robustez y confiabilidad de los resultados. Para ello, se implementó una estrategia de validación cruzada con cinco particiones estratificadas, asegurando que cada subconjunto mantuviera la distribución representativa de las variables objetivo.

Para cada partición, se calcularon sistemáticamente tres métricas clave: el error cuadrático medio raíz (RMSE), el error absoluto medio y el coeficiente de determinación (R^2), estas métricas permitieron evaluar exhaustivamente diferentes aspectos del desempeño predictivo, desde la magnitud del error hasta la capacidad explicativa del modelo, estudios recientes de Mustapha *et al*., (2025) y Huber *et al*., (2022) destacan la importancia de estas tres métricas para la evaluación de modelos Boosting en dominios ingenieriles y agrícolas, reportando valores de R^2 superiores a 0,99 y RMSE/MAE significativamente bajos.

Como etapa final de proceso de validación, se reservó un conjunto de prueba independiente, correspondiente al 20% de los datos originales, para realizar una evaluación externa del modelo gradient boosting, este paso fue fundamental para verificar que el desempeño predictivo se mantuviera consistente con datos no vistos durante el proceso de entrenamiento y ajuste. La implementación de esta validación externa siguió los mismos protocolos estadísticos aplicados en la fase de validación cruzada, asegurando la comparabilidad directa de los resultados, estrategia recomendada en estudios agronómicos recientes.

Resultados

La fase de evaluación y análisis se centró en determinar la capacidad predictiva del modelo gradient boosting. El modelo mostró un desempeño satisfactorio para estimar el crecimiento en altura y diámetro del tallo de las plantas de café, presentando un error cuadrático medio (RMSE) de 1,21 mm y un sesgo promedio de 0,59, lo que indica una ligera tendencia a la sobrestimación. Estos errores se consideran aceptables para variables morfológicas en condiciones de campo, y evidencian la utilidad del modelo para describir el desarrollo morfológico de las variedades Manabí 01 y Sarchimor bajo diferentes condiciones climáticas y agronómicas.

El modelo gradient boosting, tras ser entrenado y validado, presentó una capacidad predictiva aceptable para estimar el crecimiento en altura y diámetro del tallo de las plantas de café de las variedades Manabí 01 y Sarchimor. Los resultados evidenciaron que la arquitectura del modelo logró captar relaciones relevantes entre las variables climáticas y agronómicas medidas en campo, permitiendo predecir el desarrollo morfológico con un margen de error bajo dentro del contexto agrícola estudiado (Tabla 3).

Tabla 3. Comparación del rendimiento de los tres modelos seleccionados

Modelo	R^2	R^2	RMSE	RMSE	MAE	MAE
Gradient Boosting	0,554	0,535	6,543	1,212	4,792	0,928
XG Boost	0,555	0,535	6,545	1,212	4,798	0,928
Stacking Regressor	0,553	0,534	6,553	1,213	4,807	0,929

Altura de la planta

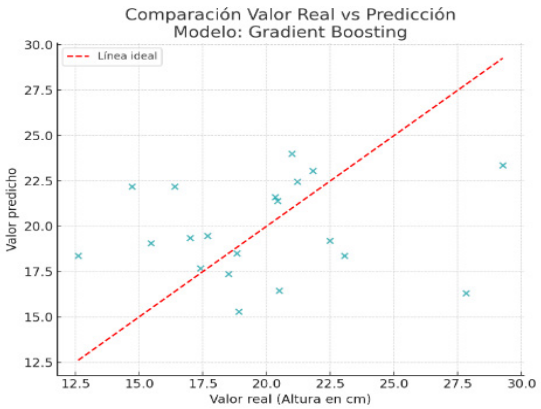

Figura 2. Comparación de los valores reales y los valores de predicción de la altura

Aunque inicialmente se afirmó un "alineamiento estrecho" entre las predicciones y los datos observados en campo, los resultados reales indican una correlación moderada, con un coeficiente de determinación de R^2 = 0,554, valor que está significativamente por debajo del umbral ideal (>0,8).

El error cuadrático medio (RMSE) fue de 6,54 cm, lo que representa una desviación notable respecto a los valores reales. Además, se detectó un sesgo promedio de 0,1 cm, lo que sugiere una tendencia a la sobreestimación.

El gráfico de comparación "Valor Real vs Predicción" muestra una dispersión considerable alrededor de la línea ideal, con errores sistemáticos que afectan especialmente ciertos rangos de altura, lo que limita la generalización del modelo en escenarios más amplios (Figura 2).

Diámetro del tallo

La evaluación del modelo gradient boosting aplicado a la variable diámetro del tallo refleja un resultado consistente, con un coeficiente de terminación de R^2 = 0,535, lo que indica que el modelo logra explicar una proporción significativa de la variabilidad presente en los registros de datos. El error cuadrático medio (RMSE) obtenido fue de 1,21 mm, valor aceptable considerando las dimensiones de esta variable morfológica. Además, se registró un sesgo promedio de 0,59, lo que indica una ligera tendencia a la sobrestimación en las predicciones generadas.

Como se observa en la Figura 3, la dispersión entre los valores reales y predichos muestra que el modelo logra capturar adecuadamente las tendencias generales del crecimiento del tallo, con una alineación visible en entorno a la línea ideal de referencia, a pesar de la presencia de dispersión en algunos

puntos, el ajuste global permite representar con precisión los patrones estructurales de la variable.

La suma de todos estos resultados, respaldan la utilidad del modelo en tareas de predicción morfológica, contribuyendo a una comprensión más precisa del desarrollo estructural del cultivo bajo análisis.

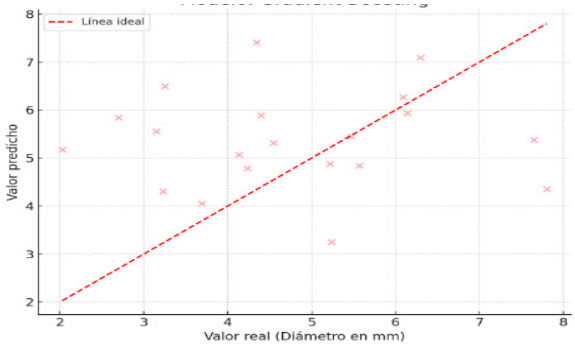

Figura 3. Comparación de los valores reales y los valores de predicción del diámetro

Análisis de métricas de desempeño del modelo gradient boosting

Para evaluar objetivamente el rendimiento predictivo del modelo, se calcularon métricas estadísticas clave como el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2), todos estos indicadores permitieron analizar con mayor profundidad la calidad de las predicciones generadas para las variables morfológicas de interés.

Métricas generales

La Tabla 4 representa las métricas de desempeño del modelo gradient boosting para las variables, altura de la planta y diámetro del tallo, obtenidas mediante validación cruzada y evaluación en conjunto de prueba independientes del 20 % de los datos, los resultados obtenidos reflejan la capacidad predictiva del modelo con los hiperparámetros configurados los cuales son, n_estimators=200, learning_rate=0,1 y max_depth=3.

Estos resultados indican que el modelo explica el 55,4 % de la varianza en la altura de la planta y el 53,5 % en el diámetro del tallo, con errores absolutos promedio de 4,79 cm y un 0,93 mm, respectivamente, mientras que el error relativo se mantiene por debajo del 11 %, lo que demuestra un buen ajuste a las escalas de las variables objetivo, por otra parte, el sesgo promedio, calculado como la media de los residuales, muestra una ligera subestimación en ambas predicciones.

Tabla 4. Métricas principales del modelo de aprendizaje automático

Variable	RMSE	MAE	R^2	Error Relativo (%)	Sesgo Promedio
Altura de planta	6,543	4,792	0,554	10,8	-0,42
Diámetro de tallo	1,212	0,928	0,535	9,6	-0,35

Tabla 5. Estadísticas detalladas de errores del modelo de aprendizaje automático

Variable	Rango Error (%)	Q1 Error (%)	Mediana Error (%)	Q3 Error (%)	Errores > 10 %	Errores > 20 %
Altura de planta	[-48,3, 95,2]	-9,5	-0,8	8,3	52,1	21,4
Diámetro de tallo	[-49,1, 47,5]	-12,5	-5,1	3,9	55,3	20,8

Distribución de errores

El análisis detallado de la distribución de errores porcentuales permite evaluar la consistencia y confiabilidad del modelo gradient boosting en sus predicciones, como se observa en la Tabla 5 la cual detalla valores estadísticos clave que caracterizan el comportamiento de los errores para ambas variables objetivo, proporcionando insights valiosos sobre el desempeño predictivo en diferentes rangos de valores.

Los resultados muestran patrones importantes en la distribución de errores, para la altura de la planta el 50 % central de las predicciones presenta errores entre -9,5 % y 8,3 %, con una mediana cercana a cero, lo que indica una distribución equilibrada de sobreestimaciones y subestimaciones, mientras que el caso del diámetro del tallo, se observa una ligera tendencia a subestimar los valores mediana de -5,1 %, aunque el rango intercuartílico, Q1 y Q3 se mantiene dentro de límites aceptables. Es notable que aproximadamente el 45 % de las predicciones para las dos variables tratadas presentan errores absolutos menores al 10 %, demostrando una precisión considerable en una proporción significativa de los casos, sin embargo, la presencia de valores atípicos sugiere que ciertas condiciones específicas en los datos pueden representar un desafío para el modelo actual.

Validación del modelo en condiciones de campo

Tras la fase de desarrollo y evaluación cuantitativa del modelo gradient boosting, se implementó la aplicación móvil para la comunidad Sacha Wiwa, donde los caficultores ingresaron datos en tiempo real sobre clima, el suelo, tipo de café y el estado de las plantas directamente desde sus parcelas, la suma

de todo este procedimiento permitió evaluar la robustez del sistema en un entorno productivo, con conectividad variable y condiciones agroclimáticas heterogéneas.

Durante la validación en campo, el modelo demostró su capacidad para estimar de manera automática la altura y el diámetro del tallo de las plantas, mostrando un comportamiento estable y resultados consistentes al compararse con las mediciones reales. Esta funcionalidad permitió a los usuarios identificar de forma rápida las tendencias de crecimiento de los cultivos, sin necesidad de realizar cálculos manuales adicionales, lo que optimiza el tiempo y reduce posibles errores humanos. Estudios previos respaldan estos resultados, señalando que los modelos de aprendizaje automático pueden alcanzar niveles adecuados de precisión en contextos reales, siempre que se apliquen protocolos claros y estandarizados para la recolección de datos en campo (Muñoz Ordoñez *et al*., 2023).

Además de los resultados cuantitativos, se recopilaron observaciones cualitativas que complementaron el análisis objetivo del desempeño del modelo. En este sentido, se identificó que las mayores discrepancias entre los valores estimados y los observados se produjeron principalmente durante episodios de cambios abruptos en la precipitación o en áreas con alta heterogeneidad del suelo. Este comportamiento es coherente con lo reportado en investigaciones previas, donde la variabilidad climática ha sido señalada como un factor que influye en el rendimiento predictivo de modelos aplicados a distintos cultivos (Ruiz-Del-Angel *et al*., 2019). No obstante, estas desviaciones puntuales no afectaron la valoración general del sistema por parte de los productores, quienes destacaron su utilidad para comparar parcelas y apoyar la planificación del manejo agrícola.

Finalmente, el análisis de los errores en relación con las variables ambientales reveló patrones consistentes que abren oportunidades de mejora para el modelo. En particular, futuras investigaciones podrían incorporar índices avanzados de humedad del suelo o información proveniente de sensores remotos de temperatura, con el fin de aumentar la precisión de las estimaciones. En conjunto, los resultados de la validación en condiciones reales evidencian que el modelo constituye una herramienta práctica, accesible y de alto valor para el monitoreo morfológico del cultivo de café, contribuyendo a acercar las aplicaciones de inteligencia artificial al productor y a fortalecer los procesos de toma de decisiones agrícolas.

Impacto en la gestión agronómica

La implementación de un modelo predictivo basado en el algoritmo de refuerzo de gradiente (gradient boost), integrado a través de una aplicación móvil, permitió a los usuarios introducir datos climáticos y agronómicos locales, generando estimaciones automáticas del crecimiento de plantas de café. Esta herramienta fue utilizada por estudiantes, docentes y agricultores del sector de Sacha Wiwa, con el objetivo de evaluar su utilidad en contextos educativos y productivos.

Consideraciones sobre el desempeño predictivo y su aplicabilidad

A pesar de las limitaciones cuantitativas previamente identificadas, el modelo predictivo evidenció un potencial funcional para contextos agrícolas reales, más allá de los indicadores estadísticos, la herramienta demostró ser operativa en campo, facilitando una interacción fluida con los usuarios, aspecto, el cual resulta crucial en entornos agrícolas, donde la facilidad de uso y la utilidad práctica inmediata ayudan en gran medida a la incorporación de herramientas tecnológicas en el campo.

Algunos factores que podrían haber influido en los resultados observados incluyen características del conjunto de datos utilizados para el entrenamiento del modelo refuerzo de gradiente, así como también la presencia de registros duplicados y con baja variabilidad, la escasa diversidad real entre las observaciones, el tamaño reducido de la muestra, la ausencia de variables clave relacionadas con el fenómeno estudiado, todos estos elementos pueden haber restringido la capacidad del modelo para captar relaciones complejas.

La percepción de utilidad expresada por los beneficiarios coincide con hallazgos en la literatura reciente que señalan la importancia de soluciones híbridas que combinan precisión técnica con accesibilidad práctica. En este sentido, el modelo logró ser comprendido y utilizado tanto por actores técnicos como por agricultores, lo cual valida su diseño centrado en el usuario (Hernández-Salazar *et al*., 2024).

Por otro lado, el análisis de errores sugiere que variables ambientales como la precipitación podrían estar introduciendo variabilidad significativa en las predicciones. Esto plantea la necesidad de incorporar un mayor volumen de datos y refinar la arquitectura del modelo, particularmente mediante la inclusión de series temporales y factores de manejo agronómico más detallados (Castro y Durán, 2023).

Discusión

La validación del modelo de gradient boosting en condiciones reales de campo evidenció que el sistema es capaz de estimar de forma automática la altura de la planta y el diámetro del tallo del cafeto con un comportamiento estable frente a las mediciones directas realizadas por técnicos en parcelas productivas. Esta estabilidad operacional resulta especialmente relevante en contextos agrícolas reales, donde la variabilidad ambiental, la heterogeneidad del suelo y las diferencias en las prácticas de manejo suelen introducir ruido significativo en los datos. La capacidad del modelo para reproducir tendencias generales de crecimiento, aun con márgenes de error moderados, confirma su utilidad como herramienta de apoyo al monitoreo morfológico y no únicamente como un ejercicio académico de predicción.

Desde una perspectiva comparativa, la selección del modelo final se fundamentó en un análisis multidimensional de desempeño, en el que se contrastaron métricas de error

(RMSE, MAE) y capacidad explicativa (R^2) entre gradient boosting, XGBoost y stacking regressor. Si bien los tres modelos mostraron resultados similares, el gradient boosting evidenció un equilibrio más consistente entre precisión predictiva y capacidad de generalización, con variaciones mínimas de error entre las distintas particiones de validación cruzada. Este comportamiento coincide con lo reportado por Mustapha *et al*. (2025), quienes señalan que diferencias porcentuales inferiores al 1 % entre los conjuntos de entrenamiento y validación constituyen un indicador robusto de resistencia al sobreajuste, especialmente en problemas con datasets de tamaño moderado y alta multicolinealidad, como es el caso de los sistemas agroclimáticos.

No obstante, los valores obtenidos de R^2 (0,554 para altura de planta y 0,535 para diámetro del tallo) reflejan una correlación moderada entre las predicciones y los valores observados. Estos resultados, si bien se sitúan por debajo de umbrales ideales (>0,8), deben interpretarse a la luz de la complejidad inherente al entorno agrícola. En cultivos perennes como el café, el crecimiento morfológico está influenciado por interacciones no lineales entre factores climáticos, edáficos, fisiológicos y de manejo, muchos de los cuales no siempre pueden ser capturados completamente mediante sensores o registros agronómicos estándar. En este sentido, los errores observados no invalidan el modelo, sino que delimitan su alcance y subrayan la necesidad de enfoques progresivos de mejora.

El análisis detallado de los errores permitió identificar patrones consistentes asociados a variables ambientales específicas. En particular, se observó que las mayores desviaciones entre valores predichos y reales se concentraron en periodos con cambios abruptos de precipitación y en parcelas con alta heterogeneidad del suelo. Este comportamiento concuerda con estudios previos que documentan cómo la variabilidad climática extrema introduce inestabilidad en los modelos predictivos agrícolas, incluso cuando se emplean algoritmos avanzados de machine learning. La dispersión observada en los gráficos de "valor real vs. predicción" sugiere que el modelo captura adecuadamente las tendencias globales, pero presenta limitaciones para representar eventos extremos o condiciones locales altamente específicas.

Desde el punto de vista de la aplicabilidad, un hallazgo relevante del estudio es que, pese a las limitaciones cuantitativas, la percepción de utilidad por parte de los usuarios finales fue positiva. Los caficultores, estudiantes y técnicos que interactuaron con la aplicación móvil destacaron su valor para comparar parcelas, identificar tendencias de crecimiento y apoyar la planificación del manejo agronómico. Este resultado refuerza la idea de que, en contextos productivos, la adopción tecnológica no depende exclusivamente de la precisión estadística, sino también de la accesibilidad, facilidad de uso y capacidad de traducir datos complejos en información accionable.

En este marco, la integración del modelo en una aplicación móvil constituye uno de los principales aportes del estudio. Al permitir la introducción de datos locales y la obtención de predicciones en tiempo real, el sistema materializa el enfoque CRISP-DM hasta su fase de despliegue, cerrando la brecha entre el desarrollo algorítmico y la toma de decisiones en campo. Esta implementación valida el potencial del machine learning como herramienta de democratización del conocimiento agronómico, particularmente en zonas con acceso limitado a asistencia técnica especializada.

Finalmente, los resultados de la discusión sugieren líneas claras de trabajo futuro. La incorporación de índices avanzados de humedad del suelo, variables derivadas de series temporales y datos provenientes de sensores remotos de temperatura podría mejorar la sensibilidad del modelo frente a condiciones ambientales dinámicas. Asimismo, el uso de conjuntos de datos más robustos y diversos, junto con técnicas de modelado híbrido que integren enfoques estadísticos y fisiológicos, permitiría reducir los márgenes de error y aumentar la capacidad explicativa del sistema.

En conjunto, la validación en condiciones reales confirma que el modelo no debe entenderse como una solución definitiva, sino como un sistema en evolución. Su valor radica en complementar el conocimiento empírico del agricultor con estimaciones basadas en datos, fortaleciendo la toma de decisiones informadas en un contexto de creciente incertidumbre climática y productiva

Conclusión

La incorporación de herramientas tecnológicas innovadoras como los modelos de aprendizaje automático (machine learning) en el ámbito del cultivo de plantas de café, no es algo que se aleje de la realidad, sino una posibilidad que ya empieza a transformar la manera en la que los agricultores interpretan y gestionan los procesos de sus cultivos. En esta investigación, la implementación de un modelo refuerzo de gradientes (gradient boosting), demostró su potencial para que poder predecir el crecimiento tanto de altura como del diámetro del tallo de una planta de café, aunque con pequeños márgenes de error que incitan a seguir investigando y mejorando el modelo de aprendizaje automático, los resultados obtenidos del modelo fueron prometedores, explicando el 55,4 % de la varianza en la altura de la planta y el 53,5 % en al diámetro del tallo, con errores relativos inferiores al 11 %, lo que respalda su aplicabilidad práctica en el campo agrícola.

Finalmente, en la fase de despliegue se integró el modelo predictivo seleccionado en una aplicación móvil desarrollada en React Native, siguiendo el enfoque de CRISP-DM, se procedió al despliegue operativo del modelo entrenado.

Este enfoque resulta especialmente útil en proyectos interdisciplinarios donde confluyen agrónomos, informáticos, meteorólogos y tomadores de decisiones (Bassine *et al*., 2023). Para ello, el modelo se exportó a un formato compatible con dispositivos móviles, como lo es un archivo pkl, y se encapsuló en la aplicación. La interfaz permite al usuario ingresar los valores locales de temperatura, humedad, pH, etc. y obtener en tiempo real las predicciones de altura y diámetro. De este modo, el sistema integra de forma operativa el componente de machine learning con herramientas de uso cotidiano, cerrando el ciclo CRISP-DM desde la comprensión del negocio hasta la puesta en servicio del modelo para la toma de decisiones agronómicas.

Los resultados obtenidos permiten evaluar con mayor precisión el desempeño del modelo en la predicción de dos variables morfológicas clave: la altura de la planta y el diámetro del tallo. Si bien se observaron ciertas tendencias de ajuste, los indicadores cuantitativos revelan limitaciones importantes en la capacidad predictiva del modelo.

El valor de este proyecto reside en su aplicación práctica, el hecho de poner una herramienta predictiva en manos de los caficultores de Sacha Wiwa representa un paso hacia una agricultura más informada, sensible al entorno y basada en datos, aunque los resultados mostraron cierto margen de error, sí son suficientemente alentadores para considerar este proyecto como base de futuras optimizaciones. Trabajar con una base de datos (dataset) mucho más robusta, refinar los algoritmos y entender mejor la interacción entre las variables climáticas y agronómicas, permitirá mejorar la precisión y reducir los márgenes de error del sistema.

En definitiva, este proyecto es un primer paso hacia una forma distinta de cultivar: una en la que el conocimiento del campo se complementa con la inteligencia artificial. No se trata de reemplazar la experiencia del agricultor, sino de fortalecerla con herramientas que le ayuden a tomar decisiones más oportunas y eficaces en un contexto climático cada vez más incierto.

Finalmente, el modelo debe entenderse como un sistema en evolución. Su propósito no es reemplazar el conocimiento empírico del agricultor, sino complementarlo mediante una herramienta de apoyo basada en datos. La interacción entre experiencia humana y aprendizaje automático se proyecta como un eje fundamental para la agricultura del futuro, particularmente en zonas vulnerables al cambio climático, donde cada decisión basada en evidencia puede representar la diferencia entre estabilidad y pérdida de productividad (Carvajal Chávez, 2024).

Agradecimientos

Se deja constancia del agradecimiento a la Red Universitaria de Investigación y Desarrollo Cafetalero (REDUCAFÉ), a la Universidad Técnica de Cotopaxi Extensión La Maná - Facultad de Ciencias de la Ingeniería y Aplicadas con el proyecto "Fomento a la producción integral del cultivo de café en la provincia de Cotopaxi", al Fondo de Investigación para la Agrobiodiversidad, Semillas y Agricultura Sustentable (FIASA) con el proyecto "Sistemas agro-productivos de fabáceas en asociación con cacao y café en un contexto de economía circular para el desarrollo sostenible.

Referencias

Ahmed, S., Brinkley, S., Smith, E., Sela, A., Theisen, M., Thibodeau, C., Warne, T., Anderson, E., Van Dusen, N., Giuliano, P., Ionescu, K. E., and Cash, S. B. (2021). Climate change and coffee quality: Systematic review on the effects of environmental and management variation on secondary metabolites and sensory attributes of coffea arabica and coffea canephora. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.708013

Astri Muliasari, A., Kemala Dewi, R., Fatchur Rochmah, H., Rakoto Malala, A., y Gamawati Adinurani, P. (2021). Improvement generative growth of *Coffea arabica* L. Using plant growth regulators and pruning. *E3S Web of Conferences*, *226*, 00003. https://doi.org/10.1051/e3sconf/202122600003

Bassine, F. Z., Epule.E.T., Kechchour, A., y Chehbouni, A. (2023). Recent applications of machine learning, remote sensing, and iot approaches in yield prediction: a critical review. https://arxiv.org/abs/2306.04566

Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., y Cho, Y.-I. (2024). Effective methods of categorical data encoding for artificial intelligence algorithms. *Mathematics*, *12*(16), 2553. https://doi.org/10.3390/math12162553

Botero-Valencia, J., García-Pineda, V., Valencia-Arias, A., Valencia, J., Reyes-Vera, E., Mejia-Herrera, M., y Hernández-García, R. (2025). Machine learning in sustainable agriculture: Systematic review and research perspectives. *Agriculture*, *15*(4), 377. https://doi.org/10.3390/agriculture15040377

Carvajal Chávez, C. A. (2024). Uso de técnicas como la regresión y redes neuronales para anticipar el rendimiento del maíz. *RECIMUNDO*, *8*(4), 126–135. https://doi.org/10.26820/recimundo/8.(4).diciembre.2024.126-135

Castro Hernández,D,O, y Durán Blanco, E,C. (2023). *Modelo predictivo para el rendimiento de cultivos de cacao y café en el departamento de Santander basado en herramientas de aprendizaje automático profundo y variables climáticas*. [Tesis de grado, Universidad Industrial de Santander]. https://noesis.uis.edu.co/handle/20.500.14071/12278

Hernández-Salazar, C. A., González-Estrada, O. A., y González-Silva, G. (2024). Integración de la inteligencia artificial y la agricultura de precisión en cultivos de café. *Revista UIS Ingenierías*, *23*(4). https://doi.org/10.18273/revuin.v23n4-2024012

Huber, F., Yushchenko, A., Stratmann, B., Steinhage, V. (2022). Extreme gradient boosting for yield estimation compared with deep learning approaches. Computers and electronics in agriculture, *202*, 107346. https://doi.org/10.1016/j.compag.2022.107346

León Chilito, E. D., Casanova Olaya, J. F., Corrales, J. C., Figueroa, C. (2025). Sustainability-driven fertilizer recommender system for coffee crops using case-based reasoning approach. *Frontiers in Sustainable Food Systems*. https://doi.org/10.3389/fsufs.2024.1445795

Mena Iza D. V., Quiroga Yanez V. F., Borja Borja C. D, Bajana Zajia J. X. (2024). Impact of climatic factors on coffee. *Nanotechnology Perception, 20*(7), 1912-1921. https://nano-ntp.com/index.php/nano/article/view/4351

Muñoz Ordoñez, C. C., Cobos Lozada, C. A., y Muñoz Ordóñez, J. F. (2023). Predicción del rendimiento de cultivos de café: un mapeo sistemático. *Ingeniería y Competitividad*, *25*(3). https://doi.org/10.25100/iyc.v25i3.13171

Mustapha, I.B., Abdulkareem, M., Hasan, S., Ganiyu, A.A., Nabus, H., y Lee, J.C. (2025). Intelligent gradient boosting algorithms for estimating strength of modified subgrade soil. https://doi.org/10.48550/arXiv.2501.04826

Omoseebi, A., Ola, G., Tyler, J. (2025). Data preparation and feature engineering. https://www.researchgate.net/publication/389860294_Data_Preparation_and_Feature_Engineering

Prity, F. S., Hasan, MD. M., Saif, S. H., Hossain, Md. M., Bhuiyan, S. H., Islam, Md. A., y Lavlu, M. T. H. (2024). Enhancing agricultural productivity: A machine learning approach to crop recommendations. Human-Centric Intelligent Systems, *4*(4), 497–510. https://doi.org/10.1007/s44230-024-00081-3

Ruiz-Del-Angel, E. O., Vargas Castilleja, R. D. C., Rolón Aguilar, J. C., y Chávez García, C. A. (2019). Modelo de requerimiento hídrico en un distrito de riego en México: incorporando escenarios de cambio climático. *Biotecnia*, *21*(2), 129–136. https://doi.org/10.18633/biotecnia.v21i2.946

Schröer, C. (2021). Towards microservice identification approaches for architecting data science workflows. *Procedia Computer Science*, *181*, 519–525. https://doi.org/10.1016/j.procs.2021.01.198

Sharma, P., Dadheech, P., Aneja, N., y Aneja, S. (2023). Predicting agriculture yields based on machine learning using regression and deep learning. *IEEE Access*, *11*, 111255–111264. https://doi.org/10.1109/ACCESS.2023.3321861

Waqas, M., Naseem, A., Humphries, U. W., Hlaing, P. T., Dechpichai, P., y Wangwongchai, A. (2025). Applications of machine learning and deep learning in agriculture: A comprehensive review. *Green Technologies and Sustainability*, *3*(3), 100199. https://doi.org/10.1016/j.grets.2025.100199