UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN Y DISEÑO DIGITAL

CARRERA DE TELEMÁTICA

Trabajo de Integración Curricular previa la obtención del Grado Académico de Ingeniera en Telemática.

PROYECTO DE INVESTIGACIÓN:

“MODELO DE PREDICCIÓN PARA LA ESTIMACIÓN DE LA PRODUCCIÓN DEL BANANO (MUSA PARADISIACA) A TRAVÉS DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO”

AUTOR:

SILVANA MARISOL ESPINOZA ZAMORA

DIRECTOR DE PROYECTO DE INVESTIGACIÓN:

ING. ANGEL IVÁN TORRES QUIJIJE, M.Sc.

QUEVEDO – LOS RÍOS – ECUADOR

2025

UTEQ

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo, **ESPINOZA ZAMORA SILVANA MARISOL,** declaro que la investigación aquí descrita es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo, puede hacer uso de los derechos correspondientes a este documento, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

ESPINOZA ZAMORA SILVANA MARISOL

C.I: 1206811356

UTEQ

CERTIFICACIÓN DE CULMINACIÓN DEL PROYECTO DE INVESTIGACIÓN

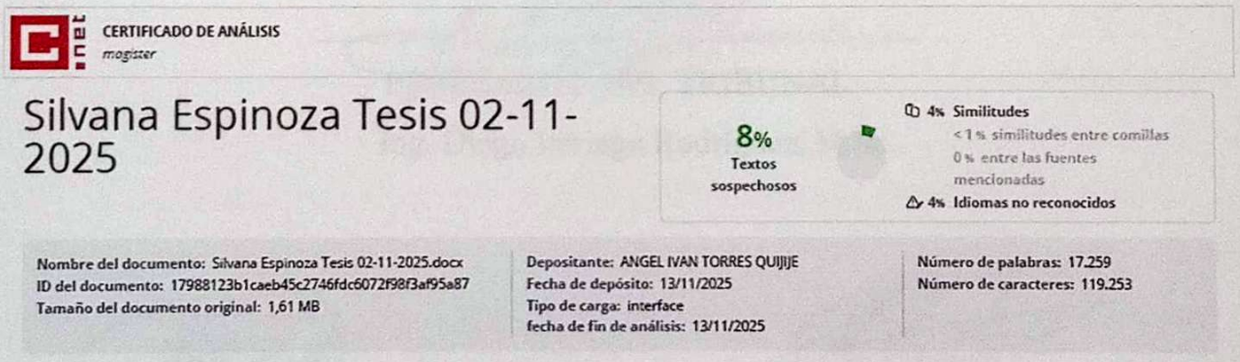
El suscrito, **Ing. Ángel Iván Torres Quijije. M.Sc.** Docente de la Universidad Técnica Estatal de Quevedo, certifica que el estudiante **Silvana Marisol Espinoza Zamora**, realizó el Proyecto de Investigación de grado titulado **"MODELO DE PREDICCIÓN PARA LA ESTIMACIÓN DE LA PRODUCCIÓN DEL BANANO (MUSA PARADISIACA) A TRAVÉS DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO"**, previo a la obtención del título de **Ing. Telemática**, bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

Ing. Ángel Iván Torres Quijije, M.Sc.
DIRECTOR DEL PROYECTO DE INVESTIGACIÓN

UTEQ

CERTIFICADO DEL REPORTE DE LA HERRAMIENTA DE PREVENCIÓN DE COINCIDENCIA Y/O PLAGIO ACADÉMICO

El suscrito, **Ing. Ángel Iván Torres Quijije. M.Sc.**, mediante el presente cumplo en presentar a usted, el informe de proyecto de Investigación titulado **"MODELO DE PREDICCIÓN PARA LA ESTIMACIÓN DE LA PRODUCCIÓN DEL BANANO (MUSA PARADISIACA) A TRAVÉS DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO"**, Presentado por el estudiante **Silvana Marisol Espinoza Zamora,** egresado de la Carrera de Telemática, que fue revisado bajo mi dirección según resolución del Consejo Directivo de la Facultad de Ciencias de la Computación y Diseño Digital, que se ha desarrollado de acuerdo al Reglamento de la Unidad de Integración Curricular de la Universidad Técnica Estatal de Quevedo y cumple con el requerimiento de análisis de Compilatio el cual avala los niveles de originalidad en un 92 % y similitud 8 %, del trabajo investigativo. Valido este documento para que el estudiante siga con los trámites pertinentes, de acuerdo como lo estable el Reglamento.


CERTIFICADO DE ANÁLISIS
magister

Silvana Espinoza Tesis 02-11-2025

8% Textos sospechosos

4% Similitudes
< 1 % similitudes entre comillas
0 % entre las fuentes mencionadas
4% Idiomas no reconocidos

Nombre del documento: Silvana Espinoza Tesis 02-11-2025.docx
ID del documento: 17988123b1caeb45c2746fdc6072f98f3af95a87
Tamaño del documento original: 1,61 MB

Depositante: ANGEL IVAN TORRES QUIJIJE
Fecha de depósito: 13/11/2025
Tipo de carga: interface
fecha de fin de análisis: 13/11/2025

Número de palabras: 17.259
Número de caracteres: 119.253

Ing. Ángel Iván Torres Quijije. M.Sc.

DIRECTOR DEL PROYECTO DE INVESTIGACIÓN

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN Y DISEÑO DIGITAL

CARRERA DE TELEMÁTICA

PROYECTO DE INVESTIGACION

TÍTULO:

"MODELO DE PREDICCIÓN PARA LA ESTIMACIÓN DE LA PRODUCCIÓN DEL BANANO (MUSA PARADISIACA) A TRAVÉS DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO"

Presentado al Consejo Directivo de la Facultad de Ciencias de la Computación y Diseño Digital, como requisito previo a la obtención del título de Ingeniero en Telemática.

Aprobado por:

PRESIDENTE DEL TRIBUNAL

Ing. Diego Intriago Rodríguez, M.Sc.

MIEMBRO DEL TRIBUNAL

Ing. Paola Benítez Navarrete. M.Sc.

MIEMBRO DEL TRIBUNAL

Ing. Santiago Meneses Narváez. M.Sc.

QUEVEDO – LOS RIOS – ECUADOR

2025

AGRADECIMIENTO

Agradezco profundamente a Dios, fuente de vida y sabiduría, por acompañarme en cada etapa de mi formación universitaria. Él me concedió la salud necesaria para mantenerme firme, así como la fortaleza y la claridad para continuar cuando el camino se tornaba difícil. En cada desafío encontré su guía, en cada decisión su consejo, y en cada logro su bendición. Su presencia constante me sostuvo en los momentos de incertidumbre y me dio la motivación para no rendirme, recordándome siempre que con fe y perseverancia todo es posible.

Agradezco con todo mi corazón a mi familia por su apoyo incondicional a lo largo de este camino. A mi madre, por su compañía constante y por ser mi guía en los momentos en que no sabía cómo avanzar; a mi padre, por su esfuerzo y generosidad al brindarme la ayuda necesaria para cubrir mis estudios. Y a mis hermanos, que con su cariño y solidaridad estuvieron presentes en cada etapa, acompañándome en largas noches de estudio y con gestos sencillos que me recordaban que no estaba sola. Gracias a cada uno de ellos pude mantenerme firme y motivada, pues su apoyo incondicional fue el pilar que me sostuvo para alcanzar esta meta.

Agradezco a mi pareja por su apoyo constante y por estar a mi lado en cada etapa de este camino. Su compañía, comprensión y cariño me dieron la fuerza necesaria para afrontar los momentos difíciles y la motivación para continuar hasta alcanzar esta meta.

Agradezco de manera especial a la ingeniera Sara Franco, por su apoyo y orientación en los inicios de mi carrera universitaria, así como por su valiosa amistad. Su confianza y compañía fueron un apoyo significativo que me impulsó a continuar en mi formación académica.

Agradezco de manera especial a Gina y Magela, por la amistad compartida durante esta etapa, una experiencia llena de alegría y momentos agradables que hicieron más ameno mi camino universitario.

Agradezco a mi amiga Nathaly por ser un pilar importante en mi etapa universitaria, brindándome su apoyo tanto en lo académico como en mi maternidad. Su compañía y cariño fueron fundamentales en este proceso, y siempre guardaré un especial aprecio por ella.

Agradezco con especial cariño a mis amigos Johanna, William y Luis, quienes fueron de mis primeros compañeros en el preuniversitario. Su amistad y compañía marcaron el inicio de esta etapa académica, convirtiéndose en un recuerdo valioso que me acompañó a lo largo de mi formación universitaria.

Agradezco profundamente a César Torres por su apoyo incondicional a lo largo de este camino, tanto en el ámbito emocional como en los momentos en que más necesité de su ayuda. Su generosidad y respaldo constante fueron fundamentales para superar las dificultades y continuar con firmeza hasta alcanzar esta meta.

Finalmente, agradezco a la Universidad Técnica Estatal de Quevedo por haberme abierto sus puertas y permitirme formarme como profesional. Extiendo mi gratitud a todas las personas que me brindaron su apoyo en el desarrollo de mi emprendimiento dentro de la institución, así como a los docentes que compartieron con generosidad sus conocimientos y experiencias. De manera especial, al ingeniero Ángel Torres, por su paciencia, sus consejos y su constante apoyo, que fueron guía y motivación en este camino académico.

DEDICATORIA

Dedico este trabajo a mi hija Angelianis, quien llegó al mundo al culminar mi etapa universitaria, llenando mi vida de alegría y convirtiéndose en la mayor bendición que pudo acompañar este logro. Su nacimiento representa un nuevo comienzo y me inspira a seguir luchando cada día para darle un mejor futuro.

De manera especial, dedico este trabajo a mi madre, por su amor incondicional y su apoyo constante; a mi padre, por su esfuerzo y respaldo a lo largo de este camino; a mis hermanos, por su compañía y por estar siempre presentes en los momentos de mayor necesidad; y a mi pareja, por su paciencia, comprensión y aliento en los días más difíciles.

Este triunfo no es solo mío, sino también de quienes caminaron a mi lado con amor, confianza y fortaleza.

RESUMEN Y PALABRAS CLAVES

La presente investigación tuvo como propósito desarrollar un modelo de predicción orientado a estimar la producción del banano (Musa paradisiaca) mediante la aplicación de técnicas de aprendizaje automático. Para ello, se identificaron inicialmente las variables morfológicas, fisiológicas, edáficas y climáticas con mayor incidencia en el rendimiento del cultivo, a partir de un análisis comparativo de estudios previos y un análisis de correlación múltiple. Posteriormente, se evaluaron diversas técnicas de aprendizaje supervisado, entre ellas Regresión Lineal, Árbol de Decisión, Máquinas de Vectores de Soporte, K-Vecinos Más Cercanos, Random Forest, Gradient Boosting, AdaBoost, XGBoost y LightGBM, aplicadas al conjunto de datos provenientes de la Hacienda El Rosario (Valencia, Los Ríos) correspondientes al periodo 2023–2024. La validación del desempeño predictivo se realizó mediante validación cruzada de cinco folds y métricas estadísticas como RMSE, MAE, MAPE, MASE y R^2, lo que permitió comparar objetivamente la exactitud y estabilidad de cada algoritmo. Los resultados evidenciaron que el modelo Random Forest presentó el mejor comportamiento predictivo, destacándose por su alta precisión, menor margen de error y mayor capacidad de generalización frente a los demás métodos evaluados. La investigación demuestra que la integración de datos agronómicos y herramientas de aprendizaje automático fortalece la agricultura de precisión, optimiza la planificación productiva y constituye una alternativa eficiente para mejorar la estimación del rendimiento bananero.

Palabras clave: Modelo predictivo, Musa paradisiaca, producción de banano, aprendizaje automático, correlación múltiple, agricultura de precisión, Random Forest.

ABSTRACT AND KEYWORD

This research aimed to develop a predictive model to estimate banana (Musa paradisiaca) production using machine learning techniques. First, the key morphological, physiological, edaphic, and climatic variables influencing crop yield were identified through a comparative analysis of previous studies and multiple correlation analysis. Subsequently, several supervised learning algorithms were evaluated—including Linear Regression, Decision Tree, Support Vector Regression, K-Nearest Neighbors, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and LightGBM—using the dataset collected from Hacienda El Rosario, located in Valencia, Los Ríos, during the 2023–2024 production period. Model performance was assessed through five-fold cross-validation and statistical metrics such as RMSE, MAE, MAPE, MASE, and R^2, enabling an objective comparison of accuracy and stability among the algorithms. The results demonstrated that the Random Forest model achieved the best predictive performance, showing higher accuracy, lower error rates, and improved generalization capacity compared to the other evaluated techniques. This study highlights that integrating agronomic data with machine learning tools strengthens precision agriculture, enhances production planning, and provides an efficient alternative for improving banana yield estimation.

Keywords: Predictive model, Musa paradisiaca, banana production, machine learning, multiple correlation, precision agricultura, Random Forest.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

CÓDIGO DUBLIN

Título:	Modelo de predicción para la estimación de la producción del banano (musa paradisiaca) a través de técnicas de aprendizaje automático.		
Autor:	Silvana Marisol Espinoza Zamora		
Palabras claves:	Modelo Banano	Predicción Aprendizaje automático	Estimación
Fecha de publicación:	Noviembre 2025		
Editorial:	Quevedo- UTEQ “La María”, 2025		
Resumen:	La presente investigación tuvo como propósito desarrollar un modelo de predicción orientado a estimar la producción del banano (Musa paradisiaca) mediante la aplicación de técnicas de aprendizaje automático. Para ello, se identificaron inicialmente las variables morfológicas, fisiológicas, edáficas y climáticas con mayor incidencia en el rendimiento del cultivo, a partir de un análisis comparativo de estudios previos y un análisis de correlación múltiple. Posteriormente, se evaluaron diversas técnicas de aprendizaje supervisado, entre ellas Regresión Lineal, Árbol de Decisión, Máquinas de Vectores de Soporte, K-Vecinos Más Cercanos, Random Forest, Gradient Boosting, AdaBoost, XGBoost y LightGBM, aplicadas al conjunto de datos provenientes de la Hacienda El Rosario (Valencia, Los Ríos) correspondientes al periodo 2023–2024. La validación del desempeño predictivo se realizó mediante validación cruzada de cinco folds y métricas estadísticas como RMSE, MAE, MAPE, MASE y R^2, lo que permitió comparar objetivamente la exactitud y estabilidad de cada algoritmo. Los resultados evidenciaron que el modelo Random Forest presentó el mejor comportamiento predictivo, destacándose por su alta precisión, menor margen de error y mayor capacidad de generalización frente a los demás métodos evaluados. La investigación demuestra que la integración de datos agronómicos y herramientas de aprendizaje automático fortalece la agricultura de precisión, optimiza la planificación productiva y constituye una alternativa eficiente para mejorar la estimación del rendimiento bananero.		
Descripción:	93 hojas; tamaño A4 (29 × 21 cm); anexos y material digital complementario.		
URI:			

INTRODUCCIÓN

La producción de banano (Musa paradisiaca) es una de las actividades agrícolas más relevantes a nivel mundial, ocupando el cuarto lugar entre los alimentos más consumidos en el planeta. Este cultivo no solo es fundamental en la dieta de millones de personas, sino que también desempeña un papel estratégico en la economía de numerosos países. En la actualidad, Ecuador, Filipinas y Costa Rica se destacan como los principales países exportadores, mientras que Estados Unidos, Alemania y Bélgica figuran entre los mayores importadores [1].

La industria bananera ecuatoriana es una de las más importantes del país, tanto en términos de generación de empleo como de aporte a la economía nacional. Se estima que el banano ecuatoriano representa alrededor del 20% de las exportaciones totales de Ecuador [2]. Esta participación se mantiene por la infraestructura agrícola, con más de 200.000 hectáreas cultivadas, principalmente en la región costera Los Ríos, Oro y Guayas, donde se aplican prácticas intensivas de riego y fumigación para asegurar cosechas semanales [3].

Por otra parte, los factores climáticos, edafológicos y de manejo agrícola influyen en la variabilidad del rendimiento, lo que dificulta la planificación y optimización de la producción. Por tal motivo, las técnicas de aprendizaje automático han emergido como herramientas clave para desarrollar modelos predictivos capaces de estimar con precisión la producción del banano y tomar de decisiones [4]. El aprendizaje automático es un pilar de la inteligencia artificial (IA) en la agricultura, ofreciendo flexibilidad e impulsando prácticas agrícolas con mayor productividad y mejor calidad con una mínima intervención [5].

Danilo Yánez y Ramiro Gaibor expertos en la agricultura indica que un modelo predictivo para estimar la producción del banano es una herramienta valiosa en la agricultura moderna, que contribuye a mejorar la calidad del producto, optimizar recursos, y aumentar la eficiencia y rentabilidad del proceso agrícola.

En este trabajo de integración curricular se desarrolla un modelo de predicción orientado a estimar la producción de banano. Para ello, se identificaron las variables clave que influyen en el rendimiento del cultivo y se evaluaron diez técnicas de aprendizaje automático, entre ellas Bosque Aleatorio y Regresión Lineal. Los resultados evidenciaron que Bosque

Aleatorio se destacó por su alta precisión y estabilidad en las estimaciones. Finalmente, se validó el desempeño de los modelos mediante pruebas de rendimiento, presentando una descripción clara y metódica del proceso desarrollado.

CAPÍTULO I
CONTEXTUALIZACIÓN DE LA INVESTIGACIÓN

1.1 Problema de la investigación.

1.1.1 Planteamiento del problema.

La producción del banano enfrenta diversos desafíos debido a factores ambientales, climáticos y de manejo agrícola que afectan su rendimiento. Según la Organización de las Naciones Unidas para la Alimentación y la Agricultura [6], el banano es uno de los cultivos más comercializados a nivel mundial, con una demanda creciente en los mercados internacionales. Sin embargo, su producción se ve afectada por variables como la temperatura, la humedad, las precipitaciones, la calidad del suelo y la presencia de plagas o enfermedades, lo que dificulta la predicción precisa de los volúmenes de cosecha y puede generar pérdidas económicas significativas [7].

En Ecuador los agricultores y productores de banano han dependido históricamente de la experiencia empírica y de modelos simples para estimar su producción, esto resultan insuficiente para representar la compleja interacción de variables que determinan el rendimiento del cultivo, ocasionando desperdicio de insumos y disminución de la competitividad en el mercado.

Además, la dificultad para determinar con precisión la producción debido a la carencia de herramientas técnicas y conocimientos especializados, lo que conlleva a subestimaciones que afectan el precio de venta, la organización de la cosecha y el transporte. Por lo tanto, surge la necesidad de desarrollar y evaluar un modelo de predicción para la estimación de la producción del banano (Musa paradisiaca), fundamentado en la identificación de las variables más relevantes, la selección de la técnica de aprendizaje automático más adecuada y la validación de su desempeño a través de métricas estadísticas de precisión.

Diagnóstico.

La estimación de la producción de banano suele apoyarse en la experiencia de los agricultores y en registros básicos, lo que reduce su precisión y complica la planificación. A esto se suma el impacto del cambio climático, que incrementa la incertidumbre debido a variaciones en la temperatura, las precipitaciones y la aparición de plagas. Asimismo, la ausencia de herramientas para la estimación del cultivo limita la eficiencia de la gestión agrícola.

Pronóstico.

La implementación del Modelo Predictivo para la estimación de la producción de banano constituye un avance estratégico en la gestión agrícola, al permitir una asignación más eficiente de los recursos, una planificación de cosechas con mayor precisión y una contribución directa a la sostenibilidad, rentabilidad del cultivo e innovación tecnológica en el sector bananero, fortaleciendo su competitividad en el ámbito internacional.

1.1.2 Formulación del problema.

¿Cómo desarrollar un modelo de predicción para estimar la producción del banano (Musa paradisiaca) mediante técnicas de aprendizaje automático y validarlo con métricas estadísticas?

1.1.3 Sistematización del problema.

- ¿Cuáles son las variables que presentan mayor relación con la producción del banano (Musa paradisiaca)?
- ¿Qué técnica de aprendizaje automático se utiliza para la estimación de la producción del banano (Musa paradisiaca)?
- ¿Cómo se evalúa el desempeño del modelo de predicción a través de pruebas de rendimiento?

1.2 Objetivos

1.2.1 Objetivo general.

Desarrollar un modelo de predicción para la estimación de la producción del banano (Musa paradisiaca) mediante técnicas de aprendizaje automático, validando su precisión con métricas estadísticas.

1.2.2 Objetivos específicos

- Identificar las principales variables con mayor relación en la producción del banano (Musa paradisiaca) mediante revisión bibliográfica y análisis de correlación múltiple.
- Determinar la técnica de aprendizaje automático para la estimación de la producción del banano (Musa paradisiaca), a partir de la comparación de su desempeño predictivo
- Evaluar el desempeño del modelo de predicción a través de métricas estadísticas, con el propósito de validar la exactitud de sus estimaciones.

1.3 Justificación.

La producción de banano constituye un eje fundamental de la economía mundial, pero su rendimiento se ve afectado por factores climáticos, edafológicos y fitosanitarios que dificultan su estimación precisa. Por este motivo, el uso de modelos predictivos basados en aprendizaje automático son una alternativa innovadora para mejorar la exactitud en la proyección de la producción.

Desde el ámbito tecnológico, estas herramientas fortalecen la agricultura de precisión al optimizar procesos productivos; en lo económico, favorecen una mejor planificación de cosechas, reducen pérdidas y aumentan la rentabilidad; y en lo ambiental, promueven un uso eficiente de insumos, minimizan el impacto ecológico y refuerzan la sostenibilidad del cultivo.

Por ello, esta investigación es relevante, ya que busca integrar técnicas de inteligencia artificial en la predicción de la producción bananera, impulsando la competitividad del sector y aportando soluciones sostenibles frente a los desafíos actuales del cambio climático.

CAPÍTULO II
FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN.

2.1 Marco Conceptual

2.1.1 El banano

El banano (*Musa paradisiaca*) es una planta herbácea perenne de la familia Musaceae, caracterizada por un pseudotallos formado por vainas foliares y un fruto partenocárpico agrupado en racimos. Su origen se encuentra en el sudeste asiático, aunque actualmente es cultivado en todas las regiones tropicales del mundo [8]. La Musa Paradisiaca es una variedad de plátano muy consumido en la región Costa del medio ecuatoriano [9].

2.1.2 Variables

La estimación de la producción de banano requiere la identificación de las variables clave que determinan su rendimiento; a continuación, se detallan dichas variables.

Variables morfológicas

Las variables morfológicas corresponden a las características estructurales y físicas de la planta, las cuales reflejan su desarrollo vegetativo y su capacidad para sostener el peso del racimo. Estas variables se consideran indicadores directos del potencial productivo del cultivo, ya que están asociadas con la biomasa acumulada y el vigor general de la planta. Entre las variables morfológicas más utilizadas en modelos predictivos agrícolas se encuentran la altura de la planta, el diámetro o circunferencia del pseudotallo, el número de manos por racimo, el peso de la planta y el peso del racimo. De acuerdo con Ortiz-Ulloa et al. [10], estos parámetros permiten estimar la robustez y la eficiencia fisiológica de la planta, influyendo de manera significativa en la productividad del banano.

Variables fisiológicas

Las variables fisiológicas describen los procesos funcionales de la planta y permiten evaluar su estado metabólico, crecimiento activo y capacidad fotosintética. Estos indicadores permiten identificar condiciones de estrés, eficiencia en la utilización de recursos y nivel de vigor del cultivo. Entre las variables fisiológicas más relevantes se encuentran el índice de vegetación de diferencia normalizada (NDVI), el estado fenológico y la tasa de desarrollo. Según Jayasinghe et al. [11], el uso de variables fisiológicas en modelos predictivos mejora significativamente la capacidad de estimación, ya que capturan dinámicas internas del crecimiento vegetal que no son visibles únicamente mediante parámetros morfológicos.

Variables edáficas

Las variables edáficas representan las propiedades físicas, químicas y biológicas del suelo, las cuales determinan la disponibilidad de nutrientes, el desarrollo de las raíces y la retención de agua. Estas variables influyen directamente en el rendimiento del banano, ya que condicionan la absorción de nutrientes esenciales y la estabilidad de la planta. Entre las variables edáficas más utilizadas se encuentran la humedad del suelo, la porosidad, la densidad aparente, el contenido de nitrógeno y otros elementos minerales. De acuerdo con Valenzuela [12], las características químicas y físicas del suelo tienen una correlación directa con la variabilidad del rendimiento en sistemas bananeros, por lo que su inclusión en modelos predictivos mejora la robustez de las estimaciones.

Variables climáticas

Las variables climáticas incluyen los factores ambientales que regulan el crecimiento, desarrollo y productividad del banano, como la temperatura máxima y mínima, la precipitación y la humedad relativa. Estos elementos influyen sobre procesos fisiológicos clave como la fotosíntesis, la transpiración y la formación del racimo. En cultivos tropicales como el banano, las variaciones climáticas pueden generar incrementos o disminuciones significativas en la producción, especialmente cuando ocurren condiciones extremas. Según Patrick et al. [13], la integración de variables climáticas en modelos predictivos permite mejorar la precisión de las estimaciones al considerar la influencia del entorno sobre el comportamiento productivo del cultivo.

2.1.2.1 NDVI (Índice de Vegetación de Diferencia Normalizada)

El NDVI es un índice espectral que mide la vigorosidad y biomasa de la planta mediante la diferencia en la reflexión de la luz roja e infrarroja cercana. En el cultivo de banano, un NDVI alto indica un estado vegetativo saludable y mayor capacidad fotosintética, mientras que valores bajos reflejan estrés hídrico o deficiencia nutricional. Su incorporación en modelos predictivos permite monitorear de forma no destructiva el estado de los cultivos y anticipar la producción [14].

2.1.2.2 Área (ha)

El área cultivada, medida en hectáreas, constituye un factor básico para la estimación de la productividad. Su análisis permite establecer la relación entre la extensión sembrada y la cantidad total de fruta producida. En estudios de rendimiento agrícola, esta variable se

emplea para normalizar los datos y expresar la producción en términos comparables (kg/ha) [15].

2.1.2.3. Densidad de siembra (plantas/ha)

La densidad de siembra define la cantidad de plantas establecidas por unidad de superficie. Un exceso de densidad puede generar competencia por agua, luz y nutrientes, reduciendo la productividad, mientras que una densidad baja desaprovecha el potencial del terreno. En banano, se ha demostrado que la densidad óptima varía según las condiciones del suelo y la variedad cultivada [16].

2.1.2.4. Porcentaje de pendiente (%)

El porcentaje de pendiente mide la inclinación del terreno, la cual influye en la erosión, drenaje y distribución del agua. En banano, terrenos con fuerte inclinación presentan mayor riesgo de pérdida de nutrientes y menor retención hídrica, lo que impacta directamente en la productividad [17].

2.1.2.5. Categoría del terreno

Es una clasificación cualitativa que agrupa los lotes según su nivel de inclinación: plano, inclinado y fuertemente inclinado. Esta variable permite integrar factores topográficos en los modelos predictivos, pues la pendiente afecta la mecanización agrícola y la eficiencia de las labores de campo [15].

2.1.2.6. Altura de la planta (m)

La altura es un indicador morfológico del crecimiento vegetativo. Plantas más altas suelen estar asociadas con un mayor vigor, aunque no siempre con mayor productividad, ya que una altura excesiva puede incrementar el riesgo de volcamiento por viento. En predicción de rendimiento, se relaciona directamente con la capacidad de sostener racimos de mayor peso [18].

2.1.2.7. Circunferencia de la planta (cm)

La circunferencia del pseudotallo es un indicador del grosor y robustez de la planta. Una mayor circunferencia se asocia con una mayor acumulación de reservas y, por ende, con racimos más pesados. Es una de las variables más utilizadas en estudios de predicción del rendimiento en musáceas [19].

2.1.2.8. Humedad del suelo (%)

La humedad refleja la disponibilidad de agua para el cultivo. En banano, un nivel adecuado de humedad es esencial para la fotosíntesis, el transporte de nutrientes y el llenado de los frutos. Déficits hídricos prolongados reducen drásticamente el tamaño de los racimos y el rendimiento por hectárea [20].

2.1.2.9. Porosidad (%)

La porosidad del suelo indica el volumen de espacios vacíos que permiten la circulación de aire y agua. Una porosidad equilibrada asegura un adecuado intercambio gaseoso para las raíces y evita problemas de compactación. En banano, la porosidad está directamente relacionada con la disponibilidad hídrica y la absorción de nutrientes [6].

2.1.2.10. Densidad del suelo (g/cm^3)

Es el grado de compactación del suelo. Valores altos de densidad dificultan el desarrollo radicular y la infiltración de agua, mientras que valores bajos pueden limitar la estabilidad de la planta. La densidad del suelo es una variable crítica para relacionar la fertilidad y productividad en banano [21].

2.1.2.11. Nitrógeno del suelo (mg/kg)

El nitrógeno es uno de los nutrientes más importantes para el desarrollo de las musáceas, ya que interviene en la formación de hojas y en la fotosíntesis. Una deficiencia reduce el crecimiento y el rendimiento, mientras que un exceso puede generar desequilibrios nutricionales [22].

2.1.2.12. Peso de la planta (lb)

El peso fresco de la planta refleja la biomasa acumulada. Una planta con mayor peso generalmente dispone de más reservas para sostener un racimo grande, aunque la relación no siempre es proporcional. Esta variable se utiliza como indicador indirecto de la productividad [18].

2.1.2.13. Peso del racimo (lb)

Es una de las variables más representativas de la producción, pues constituye el producto comercializable. Está influenciado por la nutrición, el manejo del cultivo y las condiciones

ambientales. En los modelos predictivos se emplea como variable clave o como parte de la variable de salida [19].

2.1.2.14. Número de manos

Corresponde a la cantidad de manos (subconjuntos de frutos) en un racimo. Es un indicador directo de la productividad, ya que más manos suelen significar mayor volumen de fruta. Sin embargo, también depende de la genética del cultivar y de la densidad de siembra [17].

2.1.2.15. Ratio planta-racimo

Esta relación compara el peso total de la planta con el peso del racimo. Una ratio equilibrada indica que la planta destina de manera eficiente sus recursos a la producción de fruta. Se emplea para evaluar la eficiencia fisiológica del cultivo [15].

2.1.2.16. Producción (kg fruta/ha)

Es la variable de salida principal en la mayoría de modelos agrícolas, ya que refleja el rendimiento económico y productivo del cultivo. Se calcula como el total de fruta obtenida por unidad de superficie y constituye la meta de predicción en este trabajo [14].

2.1.2.17. Temperatura mínima (°C)

La temperatura mínima corresponde al valor más bajo registrado durante un periodo determinado y constituye un factor climático esencial para el desarrollo fisiológico del banano. Valores muy bajos pueden afectar procesos como la respiración, el crecimiento foliar y la emisión de frutos, generando estrés térmico que reduce el rendimiento del cultivo. De acuerdo con Patrick et al. [20], la temperatura mínima influye directamente en la tasa de desarrollo del banano y presenta efectos acumulativos cuando se mantiene por debajo de los rangos óptimos establecidos para musáceas.

2.1.2.18. Temperatura máxima (°C)

La temperatura máxima representa el valor más alto alcanzado durante el periodo de medición. En el cultivo de banano, temperaturas elevadas pueden ocasionar estrés hídrico, reducción de la eficiencia fotosintética y afectación en la formación y llenado del racimo. Además, incrementan la evapotranspiración y la demanda de agua del cultivo. Según Jayasinghe et al. [11], la temperatura máxima es una variable crítica en modelos predictivos

debido a su influencia directa en la fisiología de la planta y su impacto en la variabilidad del rendimiento agrícola.

2.1.3. Python

Python es un lenguaje de programación que se ha convertido en una herramienta clave para aplicación del aprendizaje automático en la agricultura gracias a su sencillez y sus librerías especializadas como scikit-learn, TensorFlow, Keras y XGBoost. En estudios recientes sobre banano, como los de Muñoz Torres (2024) y Zamora Cáliz (2025), utilizaron Python para entrenar y evaluar modelos predictivos, alcanzando altos niveles de precisión en la estimación del rendimiento. Su uso permite procesar grandes volúmenes de datos y combinar información edáfica, fisiológica y ambiental, fortaleciendo la agricultura de precisión orientada a mejorar la productividad del banano (Musa paradisiaca) [18], [19].

2.1.4. Análisis de correlación múltiple

El análisis de correlación múltiple es una técnica estadística que permite estudiar la relación entre una variable dependiente y variables independientes de manera simultánea. A diferencia de la correlación simple, que evalúa el vínculo entre dos factores, esta herramienta ofrece una visión más completa al identificar qué conjunto de variables explica mejor los cambios observados en un fenómeno. En investigaciones agrícolas, como en el caso del banano (Musa paradisiaca), se utiliza para determinar qué factores edáficos, fisiológicos y ambientales influyen con mayor peso en la productividad del cultivo [23].

2.1.5. Técnicas de Aprendizaje Automático

Las técnicas de aprendizaje automático representan un conjunto de enfoques computacionales que permiten analizar datos complejos y generar modelos predictivos. En el ámbito agrícola, estas metodologías resultan clave para estimar la producción del banano a partir de variables edáficas, fisiológicas y ambientales. [24]

2.1.5.1. Aprendizaje supervisado

Es una técnica en la que el modelo se entrena con datos de entrada y sus resultados esperados (etiquetas). Su objetivo es predecir o clasificar nuevas observaciones a partir de lo aprendido. En el caso del aprendizaje supervisado, se dispone de diversos algoritmos ampliamente utilizados, entre los que destacan los siguientes:

Regresión lineal: Es un algoritmo que establece una relación matemática entre variables independientes y una dependiente, permitiendo predecir valores continuos como el rendimiento de un cultivo.

Árboles de decisión: Organizan los datos en una estructura jerárquica de reglas, facilitando la interpretación de cómo ciertas variables afectan al resultado.

Bosque Aleatorio (Random Forest): Combina múltiples árboles de decisión para generar predicciones más precisas y robustas, reduciendo el riesgo de sobreajuste.

Máquinas de vectores de soporte (SVM): Buscan la mejor frontera que separe los datos en clases, optimizando la clasificación incluso en espacios de alta dimensión.

Redes neuronales artificiales (ANN): Imitan el funcionamiento del cerebro humano, aprendiendo relaciones complejas entre variables para mejorar la predicción.

K-Vecinos Más Cercanos (KNN): Clasifica o predice en función de la similitud con sus vecinos más cercanos en el conjunto de datos.

Regresor AdaBoost: Método de potenciación que combina modelos débiles para mejorar progresivamente la exactitud.

Regresor de Potenciamiento de Gradiente (Gradient Boosting Regressor): Técnica secuencial que ajusta errores residuales para aumentar la precisión.

Regresor XGBoost: Versión optimizada del Gradient Boosting, reconocida por su eficiencia computacional y capacidad de generalización.

Regresor LightGBM: Algoritmo basado en árboles de decisión optimizado para grandes volúmenes de datos y alta velocidad de entrenamiento. [25]

2.1.5.2. Aprendizaje no supervisado

Se utiliza cuando los datos no tienen etiquetas. El modelo busca patrones, agrupaciones o estructuras ocultas dentro de los datos, como clasificar muestras similares entre sí. En lo que respecta al aprendizaje no supervisado, los algoritmos más empleados incluyen los siguientes:

K-Medias (K-Means): Agrupa datos en un número definido de clústeres según su similitud, facilitando la identificación de patrones.

Agrupamiento jerárquico (Clustering jerárquico): Crea una estructura en forma de árbol que muestra cómo los datos se agrupan de manera progresiva.

DBSCAN: Detecta grupos de datos con densidades similares y es capaz de identificar valores atípicos.

PCA (Análisis de Componentes Principales): Reduce la cantidad de variables sin perder información esencial, facilitando la visualización y análisis.

SOM (Mapas autoorganizados): Son redes neuronales que transforman datos complejos en representaciones visuales fáciles de interpretar. [25]

2.1.5.3. Aprendizaje semisupervisado

Combina un pequeño conjunto de datos etiquetados con una gran cantidad de datos sin etiquetas. Permite aprovechar la información incompleta y reducir costos de etiquetado.

En el ámbito del aprendizaje semisupervisado, destacan técnicas como el autoentrenamiento:

Autoentrenamiento (Self-training): El modelo usa los pocos datos etiquetados para etiquetar los no conocidos.

Entrenamiento conjunto (Co-training): Entrena dos modelos en paralelo que se retroalimentan con nuevas etiquetas.

Métodos basados en grafos: Relacionan datos como nodos y usan conexiones para propagar etiquetas.

Máquinas de Vectores de Soporte Semisupervisadas (S3VM (Semi-Supervised SVM): Versión mejorada de SVM para datos con pocas etiquete [25]

2.1.5.4. Aprendizaje por refuerzo

Se basa en la interacción de un agente con un entorno. El modelo aprende a tomar decisiones mediante recompensas o penalizaciones, mejorando su estrategia con la experiencia. En cuanto al aprendizaje por refuerzo, los algoritmos más representativos son:

Aprendizaje Q (Q-Learning): El agente aprende qué acción tomar maximizando recompensas acumuladas.

SARSA: Similar a Q-Learning, pero aprende considerando la acción realmente tomada.

Red Q profunda DQN (Deep Q-Network): Combina Q-Learning con redes neuronales profundas.

Métodos de gradiente de política Policy Gradient Methods: Ajustan directamente la política de decisiones del agente, útil en problemas complejos. [25]

2.1.5.5. Aprendizaje profundo (Deep Learning)

Es una rama avanzada que emplea redes neuronales artificiales con múltiples capas para procesar grandes volúmenes de datos y detectar patrones complejos, especialmente en

imágenes, voz o series temporales. Dentro del aprendizaje profundo, sobresalen arquitecturas como:

Redes neuronales convolucionales CNN (Convolutional Neural Networks): Analizan imágenes, muy usadas en agricultura de precisión con drones.

Redes neuronales recurrentes RNN (Recurrent Neural Networks): Procesan secuencias de datos, como crecimiento o clima.

Memoria a largo corto plazo LSTM (Long Short-Term Memory): Versión avanzada de RNN, maneja dependencias de largo plazo.

Autoencoders: Reducen datos y detectan anomalías en información compleja.

Redes generativas antagónicas GANs (Generative Adversarial Networks): Crean datos sintéticos realistas (ej. imágenes de cultivos). [25]

2.1.6 Métricas para comparación de modelos

2.1.6.1 RMSE (Root Mean Squared Error/Raíz del Error Cuadrático Medio)

El RMSE es una de las métricas más utilizadas en la validación de modelos predictivos porque mide la magnitud promedio de los errores, penalizando con mayor fuerza los errores grandes. Se calcula como la raíz cuadrada de la media de las diferencias al cuadrado entre los valores reales y los estimados. Según [26], el RMSE "se utiliza como una de las métricas más confiables para medir la precisión de los modelos de predicción agrícola", porque refleja con claridad qué tan alejadas están las predicciones de la realidad en términos absolutos.

Su principal ventaja es la sensibilidad ante desviaciones grandes, lo que lo hace útil para aplicaciones donde es crítico minimizar los errores más altos. Por lo tanto, el RMSE mide el promedio de la magnitud de los errores de predicción, penalizando más los errores grandes. Valores más bajos = mayor exactitud. Es útil para comparar el desempeño de diferentes algoritmos en la estimación del rendimiento.

2.1.6.2 R^2 (Coeficiente de determinación)

El R^2 es un indicador estadístico que mide la proporción de la varianza de la variable dependiente explicada por el modelo. Su valor varía entre 0 y 1, donde valores más cercanos a 1 indican un mejor ajuste del modelo a los datos real.

El coeficiente de determinación (R^2) sirve como indicador del ajuste del modelo a los datos, mostrando qué tan bien las variables explicativas logran capturar la variabilidad del fenómeno estudiado [26]. En agricultura, esta métrica se usa para evaluar la capacidad de

los modelos de predecir rendimientos, siendo fundamental en la comparación de algoritmos de aprendizaje automático.

2.1.6.3 MAE (Mean Absolute Error/ Error absoluto medio)

El MAE mide el error medio en términos absolutos, calculando la diferencia promedio entre los valores observados y los predichos sin importar la dirección del error. A diferencia del RMSE, no eleva los errores al cuadrado, por lo que cada desviación contribuye de manera lineal al resultado final. Esto lo convierte en una métrica más robusta frente a valores atípicos.

El artículo "*Predicción del rendimiento de los cultivos en la agricultura: una revisión exhaustiva de los enfoques de aprendizaje automático y aprendizaje profundo, con perspectivas para la investigación futura y la sostenibilidad"* [26]. Afirman que el MAE "proporciona una medida simple y fácilmente interpretable del error de predicción", siendo especialmente útil cuando se busca expresar los errores de manera clara para la toma de decisiones en agricultura de precisión.

2.1.6.4 MAPE (Mean Absolute Percentage Error/Error porcentual absoluto medio)

El MAPE es una métrica expresada en porcentaje que representa el error promedio absoluto relativo al valor real. Se calcula dividiendo el error absoluto entre el valor real y multiplicando por 100, lo que permite obtener una interpretación intuitiva del error en términos porcentuales.

En el artículo *"Arquitectura de red neuronal recurrente para la previsión de precios del banano en Gujarat, India"* [27], indican que el MAPE "es ampliamente utilizado en modelos de predicción porque facilita la interpretación del error relativo". No obstante, presenta limitaciones cuando los valores reales son cercanos a cero, ya que puede inflar los porcentajes de error. A pesar de esto, su facilidad de interpretación lo convierte en una herramienta recurrente en estudios de predicción agrícola y de series temporales.

2.1.6.5 MASE (Mean Absolute Scaled Error/ Error absoluto medio escalado)

El MASE es una métrica más reciente que evalúa la precisión de un modelo escalando el MAE con respecto a un modelo de referencia simple (naive model). De esta forma, se puede comparar la calidad de diferentes modelos, incluso entre conjuntos de datos con distintas escalas o unidades de medida.

Se calcula el MASE dividiendo el error absoluto medio del modelo por el error absoluto medio de un modelo de referencia ingenuo, permitiendo comparar modelos de predicción en diferentes conjuntos de datos y escalas [27]. Su principal ventaja es que evita los problemas de interpretación que presentan métricas como el MAPE y permite una evaluación más justa entre distintos enfoques.

2.1.6.6. Precisión

La precisión en la predicción es un aspecto clave en la validación de modelos de estimación aplicados al banano, ya que determina la cercanía entre los valores reales de producción y los obtenidos por el modelo.

En el artículo *"Estimación del crecimiento y rendimiento del banano mediante modelado matemático: una revisión sistemática"* [11]. destacan que el uso de métricas como el coeficiente de determinación (R^2), el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE) permite evaluar con rigor el desempeño de distintos enfoques matemáticos y de aprendizaje automático en este cultivo, aportando mayor confiabilidad a las proyecciones.

Según la investigación de Yánez-Cajo [14] evidencia que la aplicación de estas métricas en Ecuador posibilita validar modelos predictivos basados en datos multifuente —suelo, fenología e imágenes UAV— alcanzando altos niveles de exactitud en la estimación del rendimiento bananero.

2.2. Marco Referencial

En el trabajo titulado "*Predicción del rendimiento del banano mediante bosque aleatorio, integrando datos fenológicos, propiedades del suelo, tecnología espectral e imágenes de drones en el litoral ecuatoriano*" realizado por Yánez Cajo, Vásconez Montúfar, Villamar Torres, Godoy Montiel, Jazayeri, Pérez Porras y Mesas Carrascosa desarrollaron un modelo predictivo de rendimiento de banano basado en Random Forest. El estudio integró datos fenológicos, propiedades del suelo, imágenes espectrales y tecnología UAV (Vehículo Aéreo No Tripulado) en la región litoral del Ecuador. Los resultados mostraron que Random Forest superó a los métodos estadísticos tradicionales en precisión. El trabajo evidencia que la combinación de variables diversificado con aprendizaje automático fortalece la agricultura de precisión en contextos tropicales [14].

Según Muñoz Torres, en su artículo *"Modelo predictivo basado en algoritmos de machine learning para la estimación del peso de racimos de banano en una hacienda"*, aplicó distintos algoritmos de aprendizaje automático supervisado, entre ellos Random Forest, Árboles de Decisión y Regresión Lineal, con el fin de estimar el peso de racimos de banano a partir de variables agronómicas. Los resultados indicaron que Random Forest superó al resto de los modelos, al ofrecer mayor exactitud en la predicción. El estudio demuestra que el uso de técnicas de aprendizaje automático mejora considerablemente la precisión respecto a los métodos tradicionales, consolidándose como una alternativa eficaz para la gestión de la productividad en sistemas agrícolas [19].

Zamora Cáliz, en su tesis titulada *"Diseño de un modelo predictivo basado en algoritmos de Machine Learning para la estimación del rendimiento en banano orito"*, desarrolló un modelo de predicción empleando diferentes algoritmos de aprendizaje automático, entre ellos Random Forest, XGBoost y Redes Neuronales Artificiales. El modelo con mejor desempeño fue Random Forest, alcanzando un coeficiente de determinación $R^2 = 0.894$. La investigación resalta la importancia de la integración de variables como diámetro del pseudotallo, número de manos, peso del racimo y NDVI, para obtener predicciones más fiables. Este trabajo constituye un aporte significativo en el contexto ecuatoriano, al evidenciar cómo los métodos de inteligencia artificial potencian la agricultura de precisión en cultivos tropicales [18].

La investigación de Valenzuela, titulada *"Caracterización de la productividad del banano: un enfoque basado en variables químicas del suelo"*, utilizó modelos de regresión múltiple y análisis correlacional para relacionar el rendimiento con nutrientes y propiedades edáficas como nitrógeno, potasio y densidad del suelo. Sus resultados confirmaron que las condiciones químicas del suelo tienen una influencia significativa en la variabilidad de la producción, aportando evidencia clave para la selección de variables en modelos predictivos de banano [12].

En el artículo Agricultura de precisión en la producción de banano. Revisión sistemática, Romero García, Saraguro Reyes, Mazón Olivo, Morocho Román examinaron el uso de tecnologías emergentes en la cadena productiva del banano, incluyendo sensores, UAV (Vehículo Aéreo No Tripulado), IoT e inteligencia artificial. El estudio propone una arquitectura de tres capas (percepción, red y aplicación) para integrar datos agronómicos, edáficos y ambientales, subrayando que la digitalización del sector bananero es fundamental para lograr sostenibilidad y competitividad internacional [16].

Según Quiloango-Chimarro, Gioia y Oliveira Costa, en su artículo llamado *"Tipología de Unidades Productivas para el Mejoramiento de la Gestión Agronómica del Banano en Ecuador"*, analizaron los distintos tipos de unidades productivas de banano en dos provincias del Ecuador, basándose en las prácticas agronómicas adoptadas por los productores. Con una muestra de 319 unidades productivas tomadas del censo agrícola del INEC 2021, aplicaron análisis de componentes principales y análisis de conglomerados para clasificar las unidades productivas en cuatro grupos: convención de alta tecnología, convención equilibrada, convención intensiva y agroecológica. Además, realizaron regresión múltiple para cada grupo para identificar las variables que más influyen en el rendimiento (Mg/ha). Este estudio aporta al entender cuáles prácticas (por ejemplo, uso de fungicidas, mejoras genéticas, eficiencia del riego) tienen mayor impacto según el tipo de unidad productiva, lo que permite proponer políticas diferenciadas para mejorar la productividad bananera [23].

Según Ortiz Ulloa, Abril González, Pelaez Samaniego y Zalamea Piedra en su artículo con el título *"Rendimiento de biomasa y potencial de reducción de carbono del cultivo de banano (Musa spp.) en Ecuador"* cuantificaron la biomasa residual del banano y desarrollaron modelos para estimar dicha biomasa, tomando medidas físicas de 36 plantas

de la variedad Cavendish en las provincias de Los Ríos, Guayas y El Oro. Midieron altura, circunferencia, número de hojas, entre otros, y ajustaron modelos exponenciales usando Python para estimar la biomasa, obteniendo coeficientes de determinación (R^2) de hasta 0.85. También calcularon el potencial de mitigación de carbono de esta biomasa residual, lo que da soporte a prácticas agrícolas más sostenibles. Este artículo demuestra cómo variables morfológicas de planta y suelo pueden usarse en modelos predictivos aplicados al cultivo de banano, no solo para producción sino también para sostenibilidad ambiental [10].

Los estudios recientes evidencian que los modelos estadísticos convencionales, como la regresión lineal o múltiple, han servido como base para los primeros intentos de predicción del rendimiento, pero presentan restricciones en su capacidad de generalización y en la interpretación de fenómenos complejos. Investigaciones como la de Soares et al. [28] han demostrado que los modelos no lineales ofrecen un mejor ajuste en cultivos de banano, lo que respalda la necesidad de explorar metodologías más avanzadas. Esta evolución metodológica abre paso a los algoritmos de aprendizaje supervisado, que permiten mayor flexibilidad y precisión en escenarios agrícolas dinámicos.

Entre las técnicas de aprendizaje automático, destacan algoritmos como Random Forest, redes neuronales artificiales, Support Vector Machines (SVM) y XGBoost, que han mostrado resultados superiores en la predicción del rendimiento frente a los métodos estadísticos clásicos. Estos algoritmos permiten manejar simultáneamente múltiples variables como la altura de planta, el diámetro del pseudotallo, el número de manos por racimo, el índice de vegetación (NDVI) y condiciones del suelo, logrando estimaciones más confiables Santhosh et al. [29]. La capacidad de estos modelos para evaluar la importancia relativa de las variables resulta especialmente útil para identificar los factores más determinantes en la productividad del banano.

Por otra parte, la integración de tecnologías de teledetección y sensores remotos ha revolucionado la agricultura de precisión. El uso de imágenes satelitales y de drones permite obtener datos en tiempo real sobre la cobertura vegetal, el estado de salud de las plantas y la humedad del suelo. Estudios como el de Zhang et al. [30] han demostrado que la combinación de imágenes Sentinel-2 con modelos de regresión y machine learning incrementa significativamente la precisión de las predicciones, mostrando que la fusión de fuentes de datos heterogéneas mejora el monitoreo y la gestión de los cultivos

De manera complementaria, diversas revisiones sistemáticas han destacado que tanto las técnicas de machine learning como las de deep learning —particularmente redes neuronales convolucionales (CNN), recurrentes (RNN) y de memoria a largo plazo (LSTM)— poseen un alto potencial en la predicción agrícola, ya que permiten procesar imágenes multiespectrales, series temporales y grandes volúmenes de datos de sensores Waqas et al. [31]; Romero-García et al., [16] Estas metodologías no solo mejoran la exactitud en la estimación de rendimientos, sino que también facilitan la automatización de procesos y la optimización de recursos en sistemas agrícolas inteligentes.

En la investigación de Jayasinghe et al. [11], la estimación de la producción agrícola requiere integrar múltiples variables que influyen en el rendimiento, ya que los métodos estadísticos tradicionales, como la regresión lineal simple, presentan limitaciones al no capturar la complejidad de las relaciones entre factores fisiológicos, edáficos y ambientales. En este estudio, se demuestra que la aplicación de técnicas avanzadas como la regresión múltiple y los algoritmos de aprendizaje automático, entre ellos Random Forest y redes neuronales, permite mejorar la precisión de las predicciones en cultivos como el banano. Los autores resaltan la importancia de combinar información proveniente de características de la planta, condiciones del suelo y parámetros climáticos, lo cual contribuye a generar modelos más robustos y confiables. Además, el trabajo resalta el uso de métricas como R^2, RMSE y MAE para validar el desempeño de los modelos, concluyendo que la incorporación de estas herramientas tecnológicas constituye un avance clave para optimizar la planificación agrícola y la sostenibilidad productiva en contextos tropicales.

2.3. Marco legal

El desarrollo de un modelo predictivo para la estimación de la producción de banano requiere alinearse con las normativas nacionales que regulan el uso de datos, la protección ambiental y la actividad agrícola. A continuación, se presentan los principales cuerpos legales que fundamentan el alcance y aplicación de la presente investigación.

2.3.1. Ley Orgánica de Protección de Datos Personales

Esta ley regula el manejo responsable de la información personal y garantiza los derechos de los titulares frente al tratamiento de datos utilizados en entornos digitales y tecnológicos.

"***Art. 7***. *Establece las condiciones para el tratamiento legítimo de datos personales, definiendo las circunstancias en las cuales dicho tratamiento es considerado legal y ético.*"

"***Art. 8***. *El tratamiento y comunicación de datos personales requiere el consentimiento del titular, el cual debe ser libre, específico, informado e inequívoco. El titular tiene el derecho de revocar el consentimiento en cualquier momento sin justificación, y el responsable debe establecer procedimientos eficientes para ello. El tratamiento previo a la revocación es válido, ya que el consentimiento no tiene efectos retroactivos.*" [32]

2.3.2. Ley Orgánica de Agrobiodiversidad, Semillas y Fomento de la Agricultura

Esta ley establece principios para la conservación de la biodiversidad, la gestión sostenible de los recursos agrícolas y el derecho de las comunidades al acceso a alimentos de calidad.

"***Art. 13***. *Establece el derecho de las personas y colectividades al acceso seguro y permanente de alimentos sanos, suficientes y nutritivos; preferentemente producidos a nivel local y en correspondencia con sus diversas identidades y tradiciones culturales.*"

"***Art. 14***. *Establece el derecho a vivir en un medio ambiente sano y ecológicamente equilibrado, declarando de interés público la preservación del ambiente, la conservación de los ecosistemas, la biodiversidad y la integridad del patrimonio genético del país, así como la prevención del daño ambiental y la recuperación de los espacios naturales degradados.*" [33]

2.3.3. Relación del marco legal con la investigación

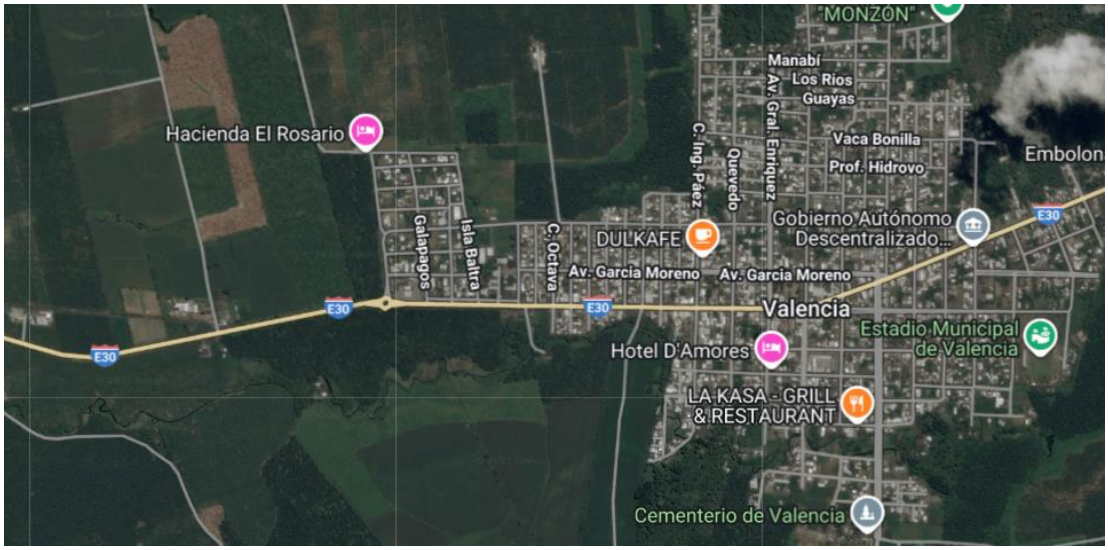
La normativa descrita garantiza el manejo adecuado de la información recopilada, así como el cumplimiento de los principios de sostenibilidad y protección ambiental. En este sentido, la construcción de un modelo predictivo para la estimación de la producción del banano se desarrolla bajo lineamientos éticos, respetando la confidencialidad de los datos y promoviendo prácticas agrícolas responsables conforme a lo establecido por la legislación ecuatoriana.

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN.

3.1 Localización

El proyecto se realizó en la Hacienda El Rosario, ubicada en el cantón Valencia, provincia de Los Ríos, Ecuador, en las coordenadas geográficas latitud -.9481741342559311 y longitud -79.36618514418014. En esta zona se llevó a cabo la recolección del conjunto de datos relacionados con la producción de banano (Musa paradisiaca L.).

Figura 1. *Ubicación hacienda El Rosario*



Fuente: Google maps

3.2 Tipo de investigación

3.2.1 Investigación Exploratoria

La investigación exploratoria se centró en la identificación inicial de los factores que influyen en la producción del banano. Se llevó a cabo revisión de la literatura agrícola, así como consulta con agrónomos y expertos en el cultivo de banano. Esta fase proporcionó una base sólida para la selección de variables a considerar en el modelo.

3.2.2 Investigación Descriptiva

La investigación es descriptiva porque busca la caracterización de las variables en la producción del banano. Se analizaron parámetros como altura de planta, número de manos, peso del racimo, contenido de nutrientes, índice de vegetación (NDVI), entre otros, describiendo su comportamiento mediante análisis estadísticos.

3.2.3 Investigación Correlacional

La investigación correlacional permitió establecer la relación entre las variables independientes y la producción de banano, identificando los factores con mayor incidencia

en el rendimiento mediante análisis estadísticos, lo que sirvió de base para la construcción y validación del modelo predictivo.

3.3 Métodos de investigación

3.3.1 Método Deductivo

El método deductivo se utilizó partiendo de principios teóricos generales relacionados con el aprendizaje automático y su aplicabilidad en la estimación de rendimientos agrícolas. Permitió vincular los fundamentos conceptuales con el análisis particular de la producción de banano (Musa paradisiaca) en la Hacienda El Rosario, contrastando las teorías con la evidencia empírica obtenida en campo. Este procedimiento facilitó la identificación de las variables con mayor influencia en la producción, proporcionando una base sólida para el llevar a cabo los análisis.

3.3.2 Método Inductivo

El método inductivo se aplicó mediante la observación de los resultados y el análisis de los datos de la producción de banano recopilados durante los años 2023–2024. Se identificaron patrones, tendencias y relaciones entre las variables independientes y la producción, que se generalizaron en conclusiones válidas para la estimación del rendimiento y la construcción de modelos de predicción.

3.3.3 Método Analítico

El método analítico se empleó para descomponer el conjunto de variables en sus componentes específicos y estudiar de manera individual su comportamiento frente a la producción. Variables como altura de planta, número de manos, peso del racimo, precipitación y concentración de nutrientes fueron analizadas estadísticamente para determinar su nivel de correlación con el rendimiento del cultivo. Este proceso permitió identificar los factores más influyentes y minimizar la presencia de variables poco significativas en el modelo predictivo.

3.3.4 Método Comparativo

El método comparativo se utilizó para evaluar y contrastar diferentes algoritmos de aprendizaje automático en su capacidad para estimar la producción de banano. Técnicas como Regresión Lineal, Árbol de Decisión, Bosque Aleatorio, Máquinas de Vectores de Soporte (SVR), K-Vecinos Más Cercanos (KNN), Regresor AdaBoost, Regresor de

Potenciamiento de Gradiente, Regresor XGBoost (Potenciación de Gradiente Extrema), Regresor LightGBM, Redes Neuronales Artificiales (ANN) – modo supervisado fueron entrenadas y validadas bajo un mismo esquema metodológico, utilizando métricas estadísticas estandarizadas (R^2, RMSE, MAE y MAPE). De esta manera, fue posible seleccionar el modelo más adecuado en función de su precisión y aplicabilidad en la productividad bananera de la Hacienda El Rosario.

3.4 Fuentes de recopilación de información

3.4.1 Fuentes Primarias

Las fuentes primarias se basaron en los datos experimentales de la producción de banano de la Hacienda El Rosario durante el periodo 2023-2024, que proporcionaron información cuantitativa, y entrevista con experto en el cultivo de banano, que ofrecieron una visión detallada sobre los factores que afectan la producción, complementando así los datos numéricos obtenidos.

3.4.2 Fuentes Secundarias

La investigación se complementó con información secundaria obtenida de artículos científicos, tesis, normativas y reportes técnicos relacionados con la producción bananera y el uso de técnicas de aprendizaje automático en la agricultura. Estas fuentes, consultadas en bases de datos como Scopus, ScienceDirect, SpringerLink y Google Scholar, permitieron contextualizar el estudio, fortalecer el marco teórico y respaldar la selección de variables, metodologías aplicadas en la investigación y técnicas de aprendizaje automáticos.

3.5 Diseño de la Investigación

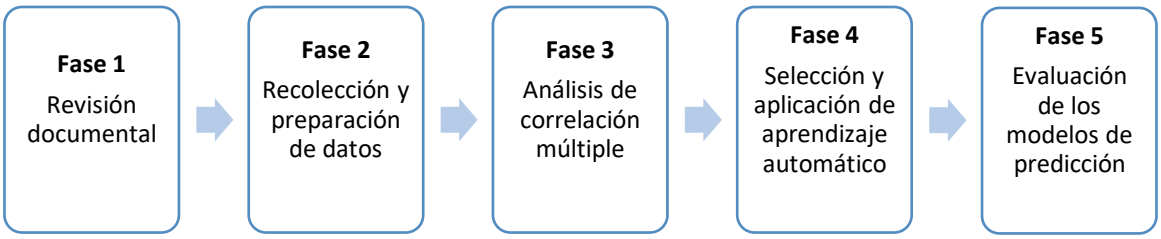
El diseño utilizado es no experimental, transversal y de tipo predictivo. Se considera no experimental porque no existe manipulación de las variables, sino que se analizan registros ya existentes con el fin de establecer relaciones entre las variables independientes y dependiente, correspondiente a la producción de banano.

Su carácter transversal radica en la recolección y el análisis de la información es de un periodo temporal, lo que permitió obtener una visión integrada de las condiciones del cultivo. En el caso de predictivo porque el propósito central de la investigación es desarrollar y validar un modelo de estimación de la producción a partir de técnicas de aprendizaje automático aplicadas sobre los datos recolectados.

3.5.1 Fases de la investigación

La investigación se estructuró en cinco fases interrelacionadas que permitieron alcanzar los objetivos planteados de manera ordenada. En la siguiente figura se muestra cada fase de investigación.

Figura 2. Fases de investigación



Fuente: Elaborado por autora

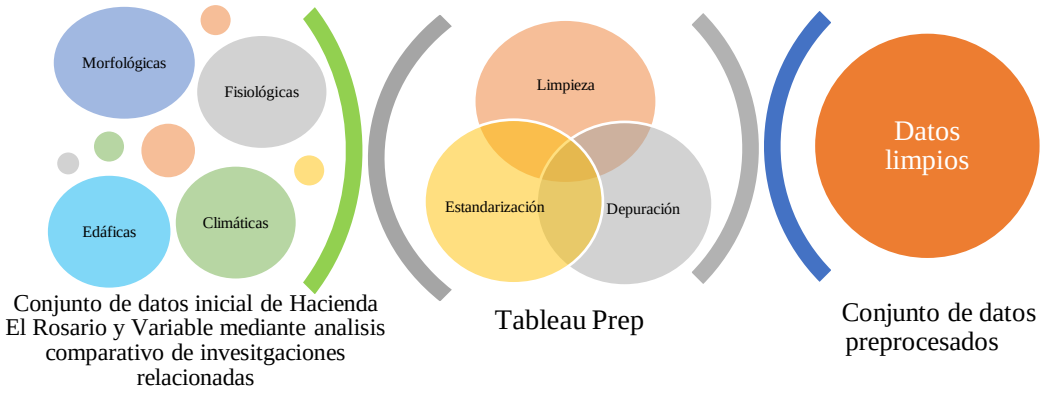
Fase 1: Revisión documental

Se efectuó una búsqueda y análisis exhaustivo de literatura científica indexada en bases de datos reconocidas, tales como Web of Science, Scopus, ScienceDirect y SciELO. Permitiendo construir el marco conceptual, metodológico del estudio, análisis comparativos de investigaciones, así como identificar antecedentes relevantes relacionados con la estimación de rendimientos agrícolas mediante técnicas de aprendizaje automático y su aplicación en el cultivo de banano (Musa paradisiaca).

Fase 2: Recolección y preparación de datos

En esta fase comprendió la obtención de información correspondiente a una hacienda del cantón Valencia, provincia de Los Ríos. Los registros recolectados incluyeron variables morfológicas, fisiológicas, edáficas y climáticas, los cuales fueron sometidos a procesos de limpieza, depuración y estandarización, con el propósito de asegurar la consistencia, fiabilidad y pertinencia del conjunto de datos para los análisis posteriores. En la figura se muestra una representación de la fase.

Figura 3. Recopilación y preparación del conjunto de datos



Fuente: Elaborado por autora

Fase 3: Análisis de correlación múltiple

Se aplicó técnica estadística orientadas a establecer el grado de asociación entre las variables independientes y la producción de banano como variable dependiente para la predicción. Este procedimiento permitió la identificación de los factores con mayor significancia estadística, optimizando así la selección de características y reduciendo la dimensionalidad de los datos, lo que constituye un paso crítico en la construcción de modelos predictivos eficientes.

Fase 4: Selección y aplicación de técnicas de aprendizaje automático

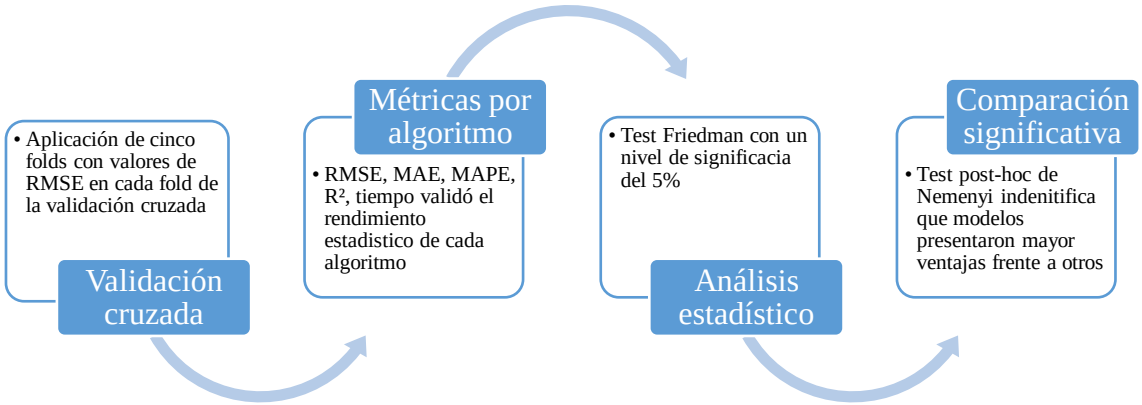
Se consideró en esta fase las referencias bibliográficas relacionadas con el tema para la selección y aplicación de los algoritmos automáticos: Regresión Lineal, Árbol de Decisión, Bosque Aleatorio, Máquinas de Vectores de Soporte (SVR), K-Vecinos Más Cercanos (KNN), Regresor AdaBoost, Regresor de Potenciamiento de Gradiente, Regresor XGBoost (Potenciación de Gradiente Extrema), Regresor LightGBM, Redes Neuronales Artificiales (ANN).

Fase 5: Evaluación comparativa del desempeño de los modelos

En esta fase se realizó la evaluación comparativa del desempeño de los modelos de predicción para la estimación de la producción del banano utilizando las métricas estadísticas de Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE), el Porcentaje de Error Absoluto Medio (MAPE) y el Coeficiente de Determinación (R^2). Además, se aplicó una

validación cruzada de cinco particiones (5-fold cross-validation), con el análisis estadístico del test de Friedman, y el test post-hoc de Nemenyi, se determinó qué algoritmos presentaron ventajas estadísticamente relevantes frente a los demás. Este proceso comparativo permitió validar de manera objetiva la técnica que mostró mayor precisión, menor error y una mayor consistencia entre los valores observados y estimados. Tal como se visualiza en la figura siguiente.

Figura 4. Principales variables con mayor correlación



Fuente: Elaborado por autora

3.6 Recursos y Materiales

Para el desarrollo de la investigación, se consideraron los recursos y materiales esenciales que se detallan a continuación.

Tabla 1. Materiales y recursos utilizados en el proyecto de investigación

Material / Recurso	Detalle	Descripción
Base de datos productiva	Registros de la Hacienda El Rosario (2023–2024).	Constituyó la fuente primaria de información cuantitativa para identificar variables relevantes y entrenar los modelos predictivos.
Equipo de cómputo	PC con procesador Intel i7/Ryzen 5, 16 GB RAM, 512 GB SSD.	Permitió ejecutar software especializado y procesar grandes volúmenes de datos sin comprometer el rendimiento.
Software estadístico	RStudio y librerías estadísticas (ggplot2, caret, tidyverse).	Utilizados para el análisis exploratorio de datos, validación de correlaciones y pruebas estadísticas.
Software de aprendizaje automático	Python con librerías scikit-learn, pandas, numpy, matplotlib, xgboost.	Implementación de algoritmos predictivos, entrenamiento de

		modelos y generación de métricas de desempeño.
Bibliografía científica	Artículos, libros y tesis disponibles en Scopus, ScienceDirect, SciELO y Google Scholar.	Fundamentación teórica y revisión de antecedentes metodológicos sobre agricultura y aprendizaje automático.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1 Principales variables con mayor relación en la producción del banano (Musa paradisiaca) mediante revisión bibliográfica y análisis de correlación múltiple.

4.1.1 Variables identificadas a partir del análisis comparativo de estudios previos sobre la producción del banano.

Se revisaron inicialmente 30 artículos científicos relacionados con la estimación de la producción del banano y la aplicación de técnicas de aprendizaje automático. Para asegurar la pertinencia de la información, se aplicaron criterios de selección basados en la indexación de los artículos en bases de datos académicas de alto impacto (Web of Science, Scopus, ScienceDirect y SpringerLink).

Además, el año de publicación, la relevancia temática, el aporte metodológico, la coherencia con las variables del estudio y la aplicabilidad en contextos agrícolas. Como resultado, se eligieron 14 artículos pertinentes para el análisis comparativo, el cual se describe en la siguiente tabla.

Tabla 2. Comparación de investigaciones relacionadas para la identificación de variables

Investigaciones relacionadas	Autor	Año	Variables analizadas	Cita
Estimación del crecimiento y rendimiento del banano mediante modelos matemáticos: una revisión sistemática	Jayasinghe et al.	2022	Altura, diámetro, clima, fisiología	[11]
Predicción del rendimiento del banano utilizando Random Forest integrando suelo, fenología e imágenes UAV	Yánez-Cajo et al.	2025	Suelo, fenología, imágenes UAV, espectros	[14]
Modelado predictivo del peso del racimo de banano mediante técnicas de aprendizaje automático	Muñoz Torres	2024	Diámetro del pseudotallo, altura, número de manos	[19]
Estimación de la producción del banano orito mediante técnicas de aprendizaje automático	Zamora Cáliz	2025	NDVI, altura, diámetro, densidad	[18]
Agricultura de precisión en la producción de banano: una revisión sistemática	Romero-García et al.	2025	Sensores, UAV, IoT	[16]
Tipología de unidades de producción para mejorar el manejo agronómico del banano en Ecuador	Quiloango-Chimarro et al.	2024	Factores productivos y socioeconómicos	[23]
Rendimiento de biomasa y potencial de captura de carbono de los cultivos de banano (Musa spp.) en Ecuador	Ortiz-Ulloa et al.	2021	Biomasa, carbono, residuos	[10]
Análisis de la brecha de rendimiento en sistemas bananeros de pequeños productores en Ecuador	Bernal-Monterrosa et al.	2022	Fertilización, riego, clima	[21]

Modelos de series temporales y ensamble para pronosticar el rendimiento del banano en Tanzania	Patrick et al.	2023	Temperatura, precipitación, series de tiempo	[13]
Comparación de técnicas utilizadas en la predicción del rendimiento en plantas de banano	Soares et al.	2014	Parámetros fisiológicos y de campo	[28]
Predicción del rendimiento del banano combinando datos temporales Sentinel-2 y modelos de regresión	Zhang et al.	2023	Imágenes Sentinel-2, NDVI	[30]
Modelado predictivo del rendimiento del banano mediante aprendizaje automático: un análisis comparativo	Santhosh et al.	2025	Variables fisiológicas, climáticas y edáficas	[29]
Aplicaciones de aprendizaje automático y profundo en agricultura (revisión)	Waqas et al.	2025	Datos agrícolas generales	[31]
Aprendizaje automático para agricultura de precisión y optimización del rendimiento de cultivos: técnicas y aplicaciones	Pankaj Roy et al.	2025	Cultivos tropicales (banano incluido)	[32]

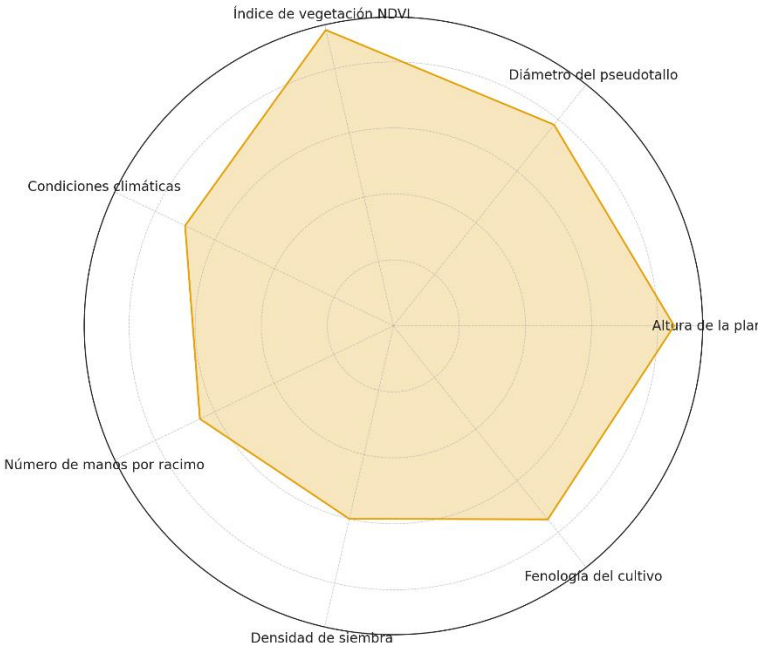
Fuente: Elaborado por autora

En la tabla muestra una tendencia al uso de variables agronómicas estructurales (altura, diámetro, número de manos) y variables derivadas de tecnologías emergentes (NDVI, sensores IoT, imágenes UAV). Estas variables se emplean para estimar el rendimiento del banano, ya sea a nivel de planta o racimo. Por otra parte, las investigaciones más recientes integran datos multifuente, como sensores remotos y plataformas IoT, aumentando la precisión de los modelos predictivos. En el anexo 1 se describen cada variable.

Los estudios de Jayasinghe et al. (2022) y Muñoz Torres (2024) se centran principalmente en parámetros morfológicos (altura, diámetro, número de manos), mientras que trabajos más recientes como los de Yánez-Cajo et al. (2025) incorporan datos espectrales y fenológicos mediante UAV. Por su parte, Zamora Cáliz (2025) utiliza variables vegetativas como el NDVI, combinadas con medidas estructurales, evidenciando la utilidad de índices de vegetación para evaluar vigor y productividad.

En el caso de Romero-García et al. (2025) incorporan sensores e infraestructura IoT, demostrando la transición hacia sistemas de monitoreo inteligente para agricultura de precisión. El gráfico siguiente visualiza de manera integral el comportamiento de las variables.

Figura 5. Radar Comportamiento de las variables en la producción del banano



Fuente: Elaborado por Autora

El gráfico visualiza el comportamiento de las variables más empleadas en las investigaciones sobre estimación de la producción del banano. Cada eje del radar corresponde a una variable y la figura central representa la relevancia que cada una tiene dentro de los estudios comparados. Además, se observa que el índice de vegetación NDVI presenta el valor más alto, lo que evidencia su papel como uno de los indicadores más utilizados y efectivos para evaluar el vigor vegetativo del cultivo. Le siguen la altura de la planta y el diámetro del pseudotallo, variables estructurales que reflejan la biomasa y la capacidad de sostén del racimo, situándose también en niveles elevados dentro del radar.

Las condiciones climáticas y la fenología del cultivo muestran valores intermedios, indicando que son factores relevantes en varios estudios, especialmente en aquellos que analizan la influencia del ambiente en los ciclos de crecimiento y producción. En contraste, las variables como el número de manos por racimo y la densidad de siembra presentan valores ligeramente menores, aunque continúan siendo elementos importantes para afinar la precisión de los modelos predictivos.

4.1.2 Estructura del conjunto de datos de la producción de banano

La recolección de la información se basó en datos que fueron obtenidos a través de un archivo Excel proporcionado por el encargado de la hacienda. Donde se registra la información de manera manual que abarca un área de 35 hectáreas correspondiente a la producción del banano, el archivo contiene información correspondiente a los años 2023 y 2024, con un total de 10318 filas y 19 columnas. Anexo 2 se visualiza base de datos inicial, en la tabla 3 se muestran la estructura del conjunto de datos con sus respectivas variables.

Tabla 3. Estructura del conjunto de datos inicial con las variables

Variable	Tipo de datos	Definición	Medida
Fecha	Fecha	Año de registro	Año
area2	Numérico	Área cultivada	Superficie destinada al cultivo de banano (ha)
Ndvi	Numérico	Índice de Vegetación de Diferencia Normalizada	Nivel de vigor y salud de la planta (-1 a 1)
Densidad	Numérico	Densidad de siembra	Número de plantas por hectárea
porcent_p	Numérico	Porcentaje de pendiente	Grado de inclinación del terreno (%)
categoria	Texto	Categoría de terreno	Clasificación según condiciones topográficas
Alt_pl_	Numérico	Altura de la planta	Altura del pseudotallo en metros
Circunferencia de planta	Numérico	Circunferencia del pseudotallo	Grosor de la planta en centímetros
Humedad actual	Numérico	Porcentaje de humedad del suelo	Disponibilidad de agua en el suelo (%)
Porosidad (%)	Numérico	Proporción de espacios vacíos en el suelo	Capacidad del suelo para almacenar agua y aire (%)
Densidad (g/cm3)	Numérico	Densidad aparente del suelo	Compactación del suelo (g/cm3)
Nitrógeno (mg/kg)	Numérico	Concentración de nitrógeno en el suelo	Disponibilidad de nutrientes esenciales (mg/kg)
Peso de planta (libras)	Numérico	Peso total de la planta	Biomasa aérea de la planta en libras
Peso de racimo (libras)	Numérico	Peso total del racimo	Peso de la unidad de cosecha en libras
Número de manos	Numérico	Número de manos por racimo	Número de grupos de frutos en el racimo
Ratio	Numérico	Relación entre peso de racimo y peso de planta	Indicador de eficiencia productiva
Producción (kg fruta/ha)	Numérico	Producción total por hectárea	Rendimiento expresado en kilogramos por hectárea
TempMin	Numérico	Temperatura mínima registrada	Condición climática mínima en °C
TempMax	Numérico	Temperatura máxima registrada	Condición climática máxima en °C

Fuente: Elaborado por autora

En la tabla se detalla la estructura del conjunto de datos inicial para la investigación, donde cada variable se encuentra definida con su tipo de dato, descripción técnica y unidad de medida. La variable fecha corresponde al año de registro de la producción, definida como un dato de tipo temporal en formato (aaaa), la variable área cultivada (area2) se presenta como un valor numérico expresado en hectáreas, representando la superficie destinada al cultivo de banano durante el periodo de análisis.

El índice de vegetación de diferencia normalizada (NDVI), constituye un indicador numérico del vigor y la salud de las plantas, con valores que oscilan entre –1 y 1. A este se suma la densidad de siembra, expresada como el número de plantas por hectárea, la cual permite evaluar la distribución y aprovechamiento del terreno agrícola.

Entre las variables de carácter edafoclimático se encuentra el porcentaje de pendiente (porcent_p), que mide el grado de inclinación del terreno en porcentaje, siendo un factor determinante en la productividad y en la gestión de labores agrícolas. De igual forma, otras variables complementarias incluidas en el conjunto de datos abarcan parámetros de clima, nutrición del suelo y rendimiento de la planta, como peso del racimo, número de manos y número de dedos por planta, todos expresados en sus respectivas unidades de medida.

En conjunto, esta estructura de datos garantiza la organización sistemática de la información y la consistencia en su análisis, constituyéndose en un insumo fundamental para el desarrollo del modelo predictivo de estimación de la producción de banano mediante técnicas de aprendizaje automático.

4.1.3 Preprocesamiento del conjunto de datos

El conjunto de datos correspondiente al periodo 2023–2024 presentó variaciones en la disponibilidad, precisión y homogeneidad de los registros. La revisión inicial permitió detectar valores nulos, inconsistencias en formatos, duplicados, datos fisiológicamente incorrectos y diferencias en unidades de medida.

Análisis de valores nulos del conjunto de datos

El análisis de valores nulos permitió evaluar la calidad interna del conjunto de datos. Este procedimiento es fundamental, dado que la presencia de registros incompletos puede afectar

la consistencia de los análisis estadísticos, la robustez de los modelos de aprendizaje automático y la validez de las conclusiones derivadas.

Presenta un nivel general de completitud adecuado, con porcentajes de ausencia inferiores al 1 % en todas las variables. Sin embargo, se identificó una mayor incidencia de valores nulos en Fecha, NDVI, densidad, TempMin y TempMax, lo cual sugiere deficiencias puntuales en la captura temporal, climática o espectral. La Tabla resume la cantidad y el porcentaje de valores ausentes en cada variable.

Tabla 4. Cantidad y porcentaje de valores nulos por variable

Variable	Valores nulos	Porcentaje (%)	Método Aplicado
Fecha	**38**	**0.359**	**No se utilizó porque tiene datos que no permite identificar día, mes, año**
area2	**10**	**0.095**	Imputación por media
Ndvi	**16**	**0.151**	Imputación por media
Densidad	**12**	**0.113**	Imputación por media
porcent_p	4	0.038	Imputación por media
Categoría	7	0.066	Imputación por moda
Alt_pl_(m)	6	0.057	Imputación por media
Circunfencia de planta (cm)	**9**	**0.085**	Imputación por media
Humedad actual (%)	6	0.057	Imputación por media
Porosidad (%)	6	0.057	Imputación por media
Densidad (g/cm3)	6	0.057	Imputación por media
Nitrógeno (mg/kg)	3	0.028	Imputación por media
Peso de planta (libras)	7	0.066	Imputación por media
Peso de racimo (libras)	4	0.038	Imputación por media
Número de manos	6	0.057	Imputación por media
Ratio	6	0.057	Imputación por media
Producción (kg fruta/ha)	**9**	**0.085**	Imputación por media
TempMin	**9**	**0.085**	Imputación por mediana
TempMax	**10**	**0.095**	Imputación por mediana
Total	**174**	**1,646**	

Fuente: Elaborada por autora

La tabla muestra las variables con mayor incidencia de valores nulos, están asociadas a procesos sensibles a fallas operativas, como sensibilidad del NDVI a condiciones atmosféricas, en el caso de la variable Fecha existen registros incompletos, sin formato y vacíos a pesar de tener una gran cantidad de registro no se utilizó por tener valores que no permiten identificar el año del registro.

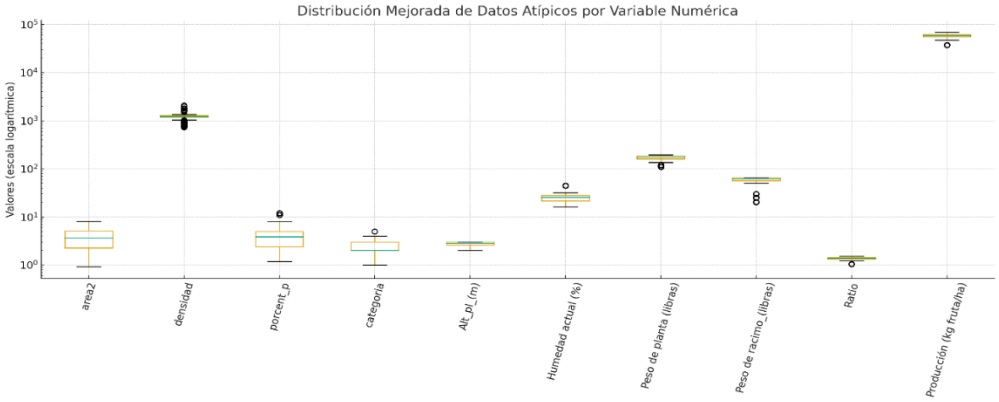
En el caso de las otras variables se aplicó imputación estadística simple (media o mediana), interpolación temporal para variables climáticas y verificación cruzada en registros

experimentales cuando sea necesario. No se detectaron patrones sistemáticos de ausencia, lo cual indica que los valores nulos no comprometen la representatividad del conjunto de datos ni generan sesgos significativos.

Detección de datos atípicos (outliers)

El análisis de datos atípicos es fundamental para evitar que valores extremos afecten la precisión de los algoritmos predictivos. En el conjunto de datos se identificaron outliers especialmente en Producción (kg/ha), Densidad, Peso de planta y Peso de racimo.

Figura 6. Datos atípicos en variables numéricas.



Fuente: Elaborado por autora

La gráfica muestra la presencia de valores atípicos en varias variables numéricas del conjunto de datos. La producción (kg/ha) muestra la mayor dispersión y concentra los outliers más extremos, las variables como densidad del suelo, peso de planta y peso de racimo también presentan valores inusuales. Fueron depurados mediante imputación estadística, interpolación temporal y verificación cruzada con registros experimentales. En la tabla siguiente se muestra las técnicas aplicadas.

Tabla 5. Tratamiento de datos aplicados a valores nulos por variable

Variable	Tipo de dato atípico	Valor atípico	Técnica aplicada	Aplicación de técnica	Casos depurados
Producción (kg/ha)	Valor extremadamente alto	70 000 kg/ha	Recorte IQR + Winsorización	58 500 kg/ha (valor límite ajustado)	3 casos
Densidad del suelo (g/cm³)	Valor fuera del rango físico	1.45 g/cm³	Eliminación + Estandarización	Reemplazado por 1.25 g/cm³ (media)	2 casos

Peso de planta (lb)	Valor muy superior al promedio	240 lb	Recorte IQR + Estandarizar	210 lb (valor suavizado)	4 casos
Peso de racimo (lb)	Valor atípico alto	90 lb	Winsorización + Normalización	75 lb (valor ajustado)	3 casos
NDVI	Valor fuera del rango 0–1	1.12	Eliminación + Reescalado	0.87 (valor válido)	1 caso
Porosidad (%)	Variación abrupta	60 %	Suavizado por IQR	47 % (valor suavizado)	2 casos
Humedad actual (%)	Variación moderada	45 %	Estandarización	Z-score = 1.12	5 casos
Circunferencia de planta (cm)	Valor aislado	35 cm	Recorte IQR	28 cm (límite superior)	2 casos
Número de manos	Valor fuera del rango	3 manos	Corrección con moda	8 manos (valor corregido)	1 caso

Fuente: Elaborado por autora

El tratamiento de los datos atípicos permitió corregir valores extremos que podían distorsionar la calidad del conjunto de datos y afectar la precisión del modelo predictivo. El recorte por IQR y la winsorización fueron las técnicas más utilizadas para ajustar valores excesivamente altos en variables como Producción, Peso de planta y Peso de racimo, logrando suavizar su impacto sin perder información relevante. En casos con valores físicamente imposibles, como NDVI > 1 o densidades superiores a 1.4 g/cm³, se aplicó eliminación y sustitución por valores válidos.

La estandarización y normalización garantizaron que variables con escalas muy diferentes mantuvieran una estructura uniforme para el modelado, mientras que la corrección mediante moda se utilizó en variables categóricas como el Número de manos. En conjunto, estas técnicas contribuyeron a depurar el conjunto de datos, mejorar su coherencia estadística y optimizar las condiciones para el entrenamiento del modelo predictivo.

Aplicación de limpieza, depuración y estandarización a variables del conjunto de datos

Se identificó que la mayoría de las variables del conjunto de datos requieren, en mayor o menor medida, procesos de limpieza, estandarización y depuración. Tal como se muestra en la siguiente tabla.

Tabla 6. Lista de variables con aplicación de limpieza, depuración y estandarización

Variable	Limpieza	Estandarización	Depuración	Descripción
Fecha	Sí	No	Sí	Fechas codificadas en 1900; corrección del formato y verificación de año.
area2	Sí	No	Sí	Presentó valores nulos; variable categórica codificada numéricamente.
Ndvi	Sí	Sí	Sí	Almacenada como texto; contiene nulos y posibles valores atípicos; debe convertirse a numérico.
densidad	Sí	Sí	Sí	Nulos y valores extremos; requiere estandarización por rango amplio.
porcent_p	Sí	No	Sí	Nulos aislados; verificar consistencia de proporciones.
categoria	Sí	No	Sí	Categórica numérica; requiere documentación de códigos.
Alt_pl_(m)	Sí	Sí	Sí	Variable clave; nulos y posibles outliers; requiere escalamiento.
Circunfencia de planta (cm)	Sí	Sí	Sí	Está en texto; requiere conversión a numérico y estandarización.
Humedad actual (%)	Sí	No	Sí	Nulos y algunos valores extremos; requiere revisión agronómica.
Porosidad (%)	Sí	Sí	Sí	Registrada como texto; requiere conversión y depuración de valores.
Densidad (g/cm3)	Sí	Sí	Sí	En formato textual; requiere conversión y control de rangos.
Nitrógeno (mg/kg)	Sí	Sí	Sí	Texto con nulos; variable clave para fertilidad.
Peso de planta (libras)	Sí	Sí	Sí	Presenta outliers y nulos; variable clave.
Peso de racimo_(libras)	Sí	Sí	Sí	Valores extremos requieren verificación; indispensable para estimación de rendimiento.
Número de manos	Sí	No	Sí	Nulos puntuales y posibles valores atípicos.
Ratio	Sí	Sí	Sí	Variable derivada; requiere estandarización y revisión de outliers.
Producción (kg fruta/ha)	Sí	Sí	Sí	Rango amplio; se recomienda estandarización y análisis de valores extremos.
TempMin	Sí	No	Sí	Nulos; revisar coherencia con registros climáticos oficiales.

TempMax	Sí	No	Sí	Nulos y variabilidad marcada por lote; requiere validación.

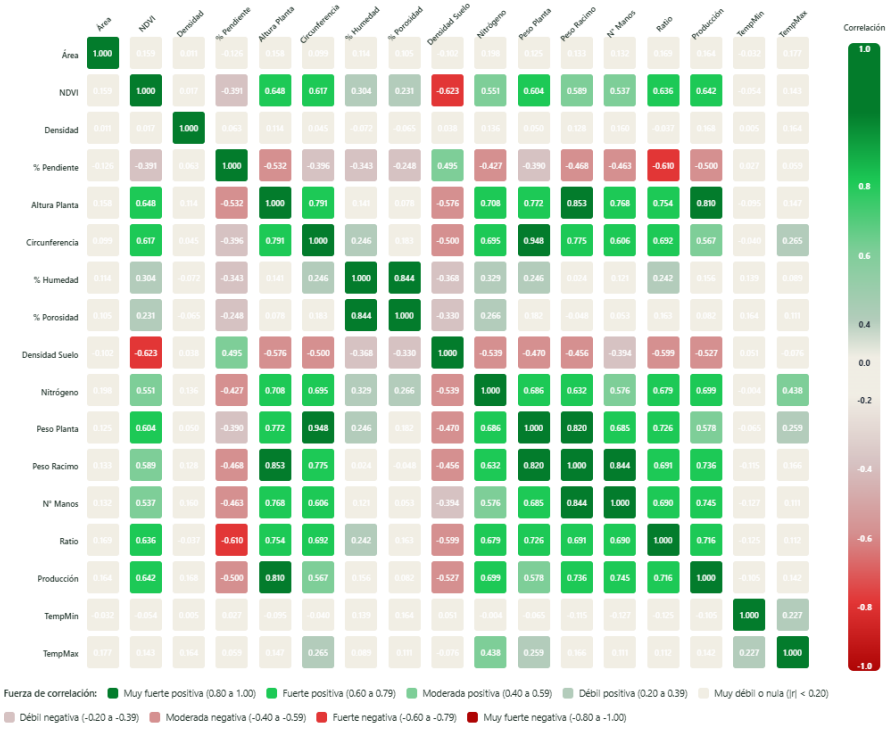
Fuente: Elaborado por autora

Las necesidades más frecuentes se relacionan con: Conversión de variables almacenadas como texto (ndvi, circunferencia, porosidad, nitrógeno); Presencia de valores atípicos, especialmente en producción, peso de racimo y peso de planta; Nulos mínimos pero significativos, que deben ser imputados para evitar pérdida de datos durante el entrenamiento del modelo; Diferencias en escalas, que justifican la estandarización de variables cuantitativas clave.

4.1.3 Análisis de correlación múltiple para identificar las principales variables para el modelo de predicción

Con la finalidad de identificar las principales variables que influyen en la producción del banano (Musa paradisiaca), se aplicó un análisis de correlación múltiple a las 18 variables: área, ndvi, Densidad de la planta, porcent_p, categoría, Alt_pl_, circunferencia de planta, humedad actual, porosidad, densidad del suelo, nitrógeno, peso de planta, peso de racimo, número de manos, ratio, producción, tempMin y tempMax. Se identificaron a partir del análisis comparativo de investigaciones relacionadas y el conjunto de datos. Los resultados obtenidos se presentan en la figura 7, donde se observa la magnitud y dirección de las correlaciones, clasificadas según su fuerza y significancia.

Figura 7. Matriz de correlación



Fuente: Elaborado por autora

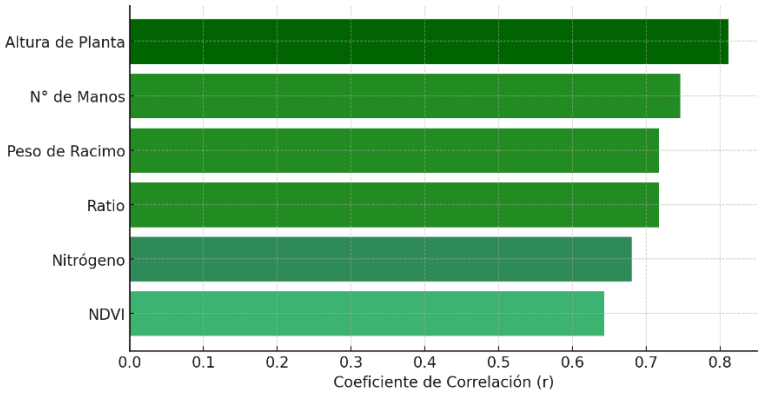
La figura muestra las variables con los valores más altos de correlación positiva con la variable Producción se registraron en altura de la planta (r = 0.810), Número de manos (r = 0.745) y Peso del racimo (r = 0.736). Estas variables muestran una relación muy fuerte y directa, lo que indica que el incremento en sus valores se asocia con un aumento en la producción de banano. Asimismo, la variable Circunferencia del pseudotallo (r = 0.567), y el Peso de la planta (r = 0.578), evidenciaron una correlación moderada positiva consolidándose también como un factor determinante.

En cuanto a las variables de tipo edáfico y ambientales, el índice NDVI (r = 0.642) y el contenido de nitrógeno (r = 0.699) mostraron correlaciones positivas relevantes, reflejando la importancia de la salud foliar y la disponibilidad de nutrientes en la productividad del cultivo. Por otro lado, ratio (r = 0.716) se destacó como un indicador clave de eficiencia productiva, al mostrar también una fuerte correlación con el rendimiento.

En contraste, la variable densidad del suelo presentó una correlación negativa moderada (r = -0.527) con la producción, lo que sugiere que una mayor compactación del suelo podría limitar el desarrollo radicular y, por ende, afectar negativamente el rendimiento. De igual manera, el porcentaje de pendiente (r = -0.500) también se asoció de manera inversa con la

productividad, lo que evidencia la influencia de factores topográficos en el cultivo. En la figura se muestra las principales variables para el modelo.

Figura 8. Principales variables con mayor correlación



Fuente: Elaborado por autora

La figura identifica las variables con mayor incidencia en la producción del banano son de naturaleza morfológica (altura de planta, peso de racimo, número de manos), complementadas por indicadores fisiológicos (NDVI, nitrógeno) y de eficiencia (ratio). Estas variables conforman un conjunto robusto para el desarrollo del modelo predictivo, al integrar dimensiones relacionadas con el crecimiento de la planta, la nutrición y las condiciones del suelo. En la figura se muestra las principales variables para el modelo.

4.1.4 Discusión

El análisis de correlación múltiple evidenció que las variables con mayor relación en la producción del banano son principalmente de tipo morfológico, destacando la altura de la planta (r = 0.810), el número de manos (r = 0.745) y el peso del racimo (r = 0.736). Estas variables muestran una influencia directa sobre el rendimiento, en concordancia con lo señalado por Jayasinghe et al. (2022), quienes afirman que los indicadores estructurales mejoran la precisión de los modelos predictivos.

De igual forma, la circunferencia del pseudotallo (r = 0.567) y el peso de la planta (r = 0.578) presentaron una relación moderada pero significativa, resultados que coinciden con lo expuesto por Zamora Cáliz (2025) sobre la importancia de la biomasa y el grosor del tallo en la productividad.

En cuanto a las variables fisiológicas, el NDVI (r = 0.642) y el contenido de nitrógeno (r = 0.699) se consolidaron como factores relevantes al reflejar el estado de salud de la planta y la disponibilidad de nutrientes, tal como lo destacan Yánez-Cajo et al. (2025). Asimismo, el ratio planta-racimo (r = 0.716) demostró ser un indicador eficiente de la asignación de recursos productivos.

Por otro lado, variables edáficas como la densidad del suelo (r = -0.527) y la pendiente del terreno (r = -0.500) mostraron correlaciones negativas, evidenciando que la compactación y la topografía afectan de manera adversa la productividad, lo cual concuerda con Valenzuela (2021).

4.2 Técnicas de aprendizaje automático para la estimación de la producción del banano (Musa x paradisiaca L.), a partir de la comparación de su desempeño predictivo.

La determinación de las técnicas de aprendizaje automático se efectuó con un análisis de comparación de investigaciones relacionadas con el fin de seleccionar las técnicas más utilizadas para el modelo de predicción del banano. Además, se utilizó las variables con mayor correlación para aplicar las técnicas aprendizaje automático.

4.2.1 Comparación de investigaciones relacionados para selección de las técnicas de aprendizaje automático

Se realizó un análisis comparativo con 14 estudios científicos que fueron revisados en la Tabla 2 Comparación de investigaciones relacionadas para la identificación de variables en este apartado de la investigación se identificó patrones comunes en cuanto a los métodos aplicados y los resultados obtenidos. Tal como se muestra en la siguiente tabla.

Tabla 7. Comparación de investigaciones relacionadas para la identificación de técnicas de aprendizaje automático

Investigaciones relacionadas	Autor	Año	Técnicas aplicadas	Resultados principales	Cita
Estimación del crecimiento y rendimiento del banano mediante modelos matemáticos: una revisión sistemática	Jayasinghe et al.	2022	Modelos matemáticos, regresión múltiple	Integrar variables multivariadas mejora la predicción de rendimiento.	[11]

Predicción del rendimiento del banano utilizando Random Forest integrando suelo, fenología e imágenes UAV	Yánez-Cajo et al.	2025	Random Forest	Alta precisión en predicción de rendimiento combinando datos multifuente.	[14]
Modelado predictivo del peso del racimo de banano mediante técnicas de aprendizaje automático	Muñoz Torres	2024	Random Forest, XGBoost	RF fue el más exacto en predicción de peso de racimos.	[19]
Estimación de la producción del banano orito mediante técnicas de aprendizaje automático	Zamora Cáliz	2025	Random Forest, Redes Neuronales	NDVI y diámetro fueron clave; R^2=0.894.	[18]
Agricultura de precisión en la producción de banano: una revisión sistemática	Romero-García et al.	2025	Revisión sistemática	IA e IoT mejoran sostenibilidad y competitividad del banano.	[16]
Tipología de unidades de producción para mejorar el manejo agronómico del banano en Ecuador	Quiloango-Chimarro et al.	2024	Clustering, PCA	Identificaron 4 tipologías de productores para orientar políticas.	[23]
Rendimiento de biomasa y potencial de captura de carbono de los cultivos de banano (Musa spp.) en Ecuador	Ortiz-Ulloa et al.	2021	Modelos estadísticos	Cuantificaron el potencial de captura de carbono en banano.	[10]
Análisis de la brecha de rendimiento en sistemas bananeros de pequeños productores en Ecuador	Bernal-Monterrosa et al.	2022	Regresión múltiple, yield gap	Un buen manejo de nutrientes reduce la brecha de rendimiento.	[21]
Modelos de series temporales y ensamble para pronosticar el rendimiento del banano en Tanzania	Patrick et al.	2023	SARIMAX, LSTM, ensambles	Modelos híbridos reducen incertidumbre bajo variabilidad climática.	[13]
Comparación de técnicas utilizadas en la predicción del rendimiento en plantas de banano	Soares et al.	2014	Regresión lineal y no lineal	Modelos no lineales obtuvieron mejor ajuste que los lineales.	[28]
Predicción del rendimiento del banano combinando datos temporales Sentinel-2 y modelos de regresión	Zhang et al.	2023	Regresión múltiple	Teledetección + regresión mejoran predicciones de campo.	[30]
Modelado predictivo del rendimiento del banano mediante aprendizaje automático: un análisis comparativo	Santhosh et al.	2025	Random Forest, XGBoost, ANN	RF/XGBoost > regresión tradicional en precisión.	[29]
Aplicaciones de aprendizaje automático y profundo en agricultura (revisión)	Waqas et al.	2025	Revisión ML/DL	DL (CNN, RNN, LSTM) destacan en agricultura de precisión.	[31]

Aprendizaje automático para agricultura de precisión y optimización del rendimiento de cultivos: técnicas y aplicaciones	Pankaj Roy et al.	2025	Revisión de ML en agricultura	ML mejora la toma de decisiones y predicción de rendimientos.	[32]

Fuente: Elaborado por autora

El análisis comparativo de los 14 estudios revisados permitió identificar patrones comunes en cuanto a los métodos aplicados para la estimación y predicción de la producción del banano.

Se evidenció que los modelos de tipo ensamble, particularmente Random Forest, fueron los más recurrentes, apareciendo en cuatro investigaciones y consolidándose como la técnica preferida debido a su capacidad para manejar datos multivariados y relaciones no lineales con elevada precisión. Asimismo, se observó que XGBoost y las redes neuronales artificiales (ANN) se emplearon en combinación con Random Forest en más de un estudio, lo que refleja una tendencia hacia la exploración de modelos híbridos que potencian la robustez y la capacidad predictiva.

Por otro lado, la regresión múltiple se posicionó como el segundo método más utilizado, presente en tres investigaciones, lo que evidencia que, pese al auge de técnicas más complejas, los modelos estadísticos continúan siendo relevantes como línea base interpretativa. También se encontraron aplicaciones de regresiones lineales y no lineales en conjunto, reafirmando su utilidad en escenarios comparativos. En estudios específicos orientados a dinámicas temporales, destacaron los modelos SARIMAX y LSTM, generalmente integrados en esquemas de ensamble, lo que sugiere la pertinencia de enfoques híbridos para capturar estacionalidades y variabilidad climática.

De forma complementaria, se identificó el uso de PCA y clustering como herramientas de reducción de dimensionalidad y clasificación de unidades productivas, aportando valor al preprocesamiento de datos. Finalmente, tres investigaciones correspondieron a revisiones sistemáticas y de aprendizaje automático, donde se concluye que los modelos no lineales (RF, XGBoost, ANN) ofrecen consistentemente un mejor desempeño frente a regresiones tradicionales.

Se identificó 10 técnicas de aprendizaje automático para la aplicación del modelo de predicción de la producción del banano por las variables cuantitativas y la aplicación de investigaciones relacionadas. En la siguiente tabla se muestra las técnicas seleccionadas.

Tabla 8. Técnicas de aprendizaje automático seleccionados

No	Técnica de aprendizaje automático	Descripción
1	Random Forest	Utiliza árboles de decisión para mejorar la precisión y la robustez de las predicciones
2	Regresión lineal	Relación lineal entre predictores y variable respuesta.
3	Support Vector Regression	Maximiza márgenes para regresión, útil en problemas complejos.
4	Árbol de decisión	Modelo basado en reglas jerárquicas, interpretable, puede sobreajustar.
5	Gradient Boosting	Árboles secuenciales que corrigen errores de predicción anteriores.
6	XGBoost	Boosting optimizado con regularización y alto rendimiento.
7	LightGBM	Se destaca por su alta velocidad de entrenamiento, menor uso de memoria, soporte para aprendizaje paralelo
8	K-Nearest Neighbors	Predice en función de la similitud con vecinos más cercanos.
9	AdaBoost	Ensamble que ajusta pesos para reducir errores sucesivos.
10	Redes Neuronales	Aprende relaciones no lineales mediante capas ocultas.

Fuente: Elaborado por autora

La tabla muestra una diversidad de técnicas de aprendizaje automático que van desde modelos simples hasta enfoques avanzados. Los árboles de decisión y sus variantes, como Random Forest, Gradient Boosting, XGBoost y LightGBM, destacan por su robustez y precisión en contextos agrícolas. Métodos clásicos como la regresión lineal siguen siendo útiles como referencia, mientras que algoritmos como SVR y KNN aportan alternativas para relaciones complejas o basadas en similitud. Por otro lado, ensambles como AdaBoost y modelos de redes neuronales ofrecen un mayor potencial para capturar patrones no lineales. En conjunto, esta clasificación refleja que la elección del modelo depende del balance entre interpretabilidad, complejidad de los datos y nivel de precisión requerido.

4.2.2 Aplicación de las técnicas de aprendizaje automático

Se aplicaron las 10 técnicas de aprendizaje automático con las 6 variables con mayor correlación teniendo como resultado la siguiente tabla.

Tabla 9. Aplicación de técnicas de aprendizaje

Modelo	R^2	RMSE (kg/ha)	Precisión (%)	Observación
Random Forest	0.991	570.35	99.8%	Mejor ajuste, menor error. Excelente estabilidad.
Regresión Lineal	0.735	3166.26	95.7%	Modelo simple, pero bajo desempeño.
Support Vector Regression	0.151	5662.75	92.9%	Muy deficiente para este caso.
Decision Tree	0.862	2277.54	97.3%	Aceptable, pero inestable frente a RF.
Gradient Boosting	0.902	1918.80	97.6%	Buen desempeño, menor que RF y XGBoost.
XGBoost	0.980	863.01	99.1%	Muy bueno, con bajo error y alta precisión.
LightGBM	0.947	1408.75	98.4%	Bueno, pero menor que XGBoost.
KNN	0.985	748.37	99.7%	Muy cercano a RF, buen ajuste, pero menos robusto.
AdaBoost	0.805	2712.93	96.2%	Desempeño medio.
Neural Network	0.894	2003.88	97.5%	Promedio, requiere más datos para mejorar.

Fuente: Elaborado por autora

La tabla muestra tres indicadores de aplicación en la comparación de los modelos de aprendizaje automático evidencia diferencias significativas en términos de ajuste, error y precisión. El Random Forest se destacó como la técnica con el mejor desempeño global (R^2=0.991, RMSE=570.35 kg/ha, precisión=99.8%), confirmando su robustez y estabilidad frente a la variabilidad de los datos. De manera similar, el K-Nearest Neighbors (KNN) mostró un rendimiento muy cercano (R^2=0.985, RMSE=748.37 kg/ha, precisión=99.7%), aunque con menor consistencia que Random Forest.

Entre los modelos de ensamble, XGBoost alcanzó un alto nivel de exactitud (R^2=0.980, RMSE=863.01 kg/ha, precisión=99.1%), superando a LightGBM (R^2=0.947, RMSE=1408.75) y a Gradient Boosting (R^2=0.902, RMSE=1918.80), lo que confirma su eficiencia en escenarios de predicción agrícola. El Decision Tree obtuvo un rendimiento aceptable (R^2=0.862), pero evidenció inestabilidad frente a modelos más avanzados,

mientras que AdaBoost presentó un desempeño intermedio con errores relativamente altos (RMSE=2712.93).

En contraste, la Regresión Lineal (R^2=0.735) y el Support Vector Regression (SVR) (R^2=0.151) reflejaron un desempeño limitado, con errores elevados (RMSE=3166.26 y 5662.75 respectivamente), lo que indica una baja capacidad para capturar relaciones no lineales presentes en la producción del banano. Las Redes Neuronales, aunque mostraron un resultado promedio (R^2=0.894, RMSE=2003.88), sugieren un potencial de mejora con un mayor volumen de datos de entrenamiento.

En conjunto, los resultados confirman que los modelos basados en árboles y ensambles (Random Forest, XGBoost, KNN) ofrecen la mejor combinación de ajuste y precisión, consolidándose como las técnicas más adecuadas para la construcción de modelos predictivos en la estimación de la producción de banano.

4.2.4 Discusión

El análisis comparativo de técnicas de aprendizaje automático confirma que los modelos de ensamble, especialmente Random Forest (R^2=0.991; precisión=99.8%), constituyen la opción más robusta y precisa para la estimación de la producción de banano. Este resultado coincide con Yánez-Cajo et al. (2025) y Muñoz Torres (2024), quienes evidencian la superioridad de Random Forest frente a métodos estadísticos tradicionales. Asimismo, XGBoost y KNN alcanzaron un rendimiento competitivo, lo que concuerda con Santhosh et al. (2025), quienes destacan el potencial de estos algoritmos en escenarios agrícolas multivariados.

En contraste, técnicas clásicas como la Regresión Lineal y el Support Vector Regression (SVR) presentaron un bajo ajuste y mayores errores, confirmando las limitaciones señaladas por Soares et al. (2014) respecto al uso de modelos lineales en cultivos complejos. Por su parte, las Redes Neuronales Artificiales mostraron un desempeño intermedio, similar a lo observado por Zamora Cáliz (2025), quien resalta la necesidad de disponer de bases de datos más amplias para mejorar su precisión. Además, Waqas et al. (2025) y Romero-García et al. (2025) sostienen que la combinación de machine learning con big data y teledetección amplía las posibilidades de predicción en la agricultura de precisión, lo que refuerza la pertinencia de estos hallazgos.

En conjunto, se ratifica que los algoritmos de ensamble y aprendizaje profundo superan a los enfoques estadísticos convencionales, constituyendo la mejor estrategia para la estimación de la producción de banano y aportando a una gestión agrícola más eficiente, competitiva y sostenible.

4.3 Evaluación del desempeño del modelo de predicción a través de pruebas de rendimiento, utilizando métricas estadísticas para validar la precisión de sus estimaciones.

4.3.1 Aplicación de validación cruzada (cross-validation)

La evaluación del modelo de predicción se desarrolló mediante una validación cruzada de cinco folds (5-fold cross-validation), lo que permitió analizar la estabilidad, consistencia y capacidad de generalización de cada algoritmo aplicado. La Tabla 14 presenta los valores del Error Cuadrático Medio (RMSE) obtenidos en cada fold para diez algoritmos de aprendizaje automático, permitiendo comparar su precisión en la estimación de la producción del banano (Musa paradisiaca).

Tabla 10. Valores de RMSE obtenidos en cada fold de la validación cruzada

Algoritmos	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Bosque Aleatorio	704.81	694.34	827.07	542.89	716.65
Regresión Lineal	3312.05	3202.96	3252.37	3295.50	3172.93
Regresión de Vectores de Soporte	5682.55	5606.96	5721.28	5820.60	5489.67
Árbol de Decisión	2404.71	2407.19	2508.81	2260.98	2336.93
Potenciamiento de Gradiente	1988.41	1957.67	2053.58	1941.85	1988.97
Regresor XGBoost	1106.90	1067.35	1166.20	1034.95	1119.46
Regresor LightGBM	1576.40	1515.07	1603.57	1484.90	1567.59
K-Vecinos más Cercanos	1038.42	973.68	1118.51	829.93	927.51
Regresor AdaBoost	2868.48	2838.23	2771.55	2777.44	2763.03
Redes Neuronales	2019.16	2043.67	2107.10	2001.42	2048.02

Fuente: Elaboración de autora.

Los resultados de la tabla evidencian diferencias significativas en el comportamiento de los modelos. En términos generales, los algoritmos con menor RMSE demostraron un mejor ajuste entre los valores observados y los valores estimados, lo que se traduce en una mayor precisión predictiva. Entre los modelos evaluados, K-Vecinos más Cercanos (KNN) y Bosque Aleatorio destacan por presentar los valores más bajos de RMSE, evidenciando un desempeño robusto y estable en todos los folds. En particular, KNN obtuvo valores que oscilan entre 829.93 y 1,118.51, lo cual lo posiciona como el algoritmo más eficiente del conjunto analizado, con una variabilidad moderada y un comportamiento consistente.

Por su parte, los modelos basados en boosting, como XGBoost y LightGBM, mostraron también un rendimiento favorable, alcanzando RMSE intermedios. XGBoost reportó valores entre 1,034.95 y 1,166.20, lo que evidencia su capacidad para capturar relaciones no lineales y patrones complejos del conjunto de datos. LightGBM presentó resultados ligeramente superiores, aunque manteniendo un desempeño adecuado y estable entre folds. Estos algoritmos, si bien no superan a KNN en precisión, representan alternativas potentes para escenarios donde se requiere mayor interpretabilidad o velocidad de entrenamiento.

En contraste, los modelos Regresión Lineal y Regresión de Vectores de Soporte (SVR) registraron los valores más altos de RMSE, superando ampliamente los 3,000 y 5,000 respectivamente. Esto sugiere que su capacidad para modelar adecuadamente la relación entre las variables predictoras y la producción del banano es limitada, posiblemente debido a la presencia de no linealidades y variabilidad elevada en los datos, características frecuentes en sistemas agrícolas. De igual manera, AdaBoost y el Árbol de Decisión mostraron un desempeño inferior comparado con los modelos más avanzados, lo que refleja la sensibilidad de estos métodos a la variabilidad y a la complejidad del fenómeno estudiado.

Por otra parte, el modelo de Redes Neuronales, aunque obtuvo resultados más favorables que SVR o Regresión Lineal, presentó un RMSE mayor que los algoritmos basados en árboles y vecinos. Esto podría deberse a la necesidad de un mayor ajuste de hiperparámetros o a la cantidad limitada de datos para optimizar su estructura interna.

K-Vecinos más Cercanos, Bosque Aleatorio y XGBoost son los modelos con mejor desempeño para estimar la producción del banano, ya que presentan menor error y mayor

estabilidad. Estos hallazgos permiten seleccionar con confianza los algoritmos más adecuados para la etapa final de implementación y validación del sistema predictivo.

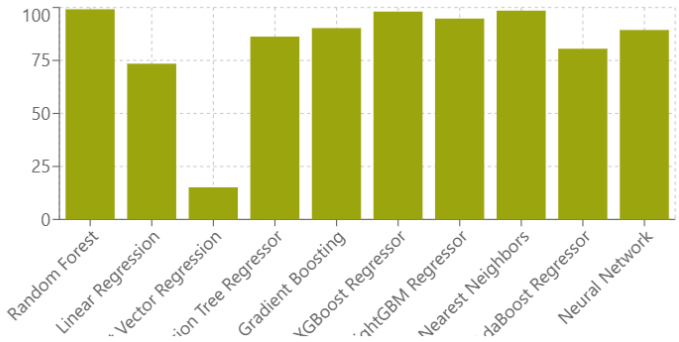
4.3.2 Análisis de métricas estadística de rendimiento para la evaluación del modelo predicción

La investigación utilizó la prueba de rendimiento con validación cruzada en cinco partes divididas del conjunto de datos. En contraste, la estrategia clásica de entrenamiento/prueba se aplicó con un 80% de los datos para el entrenamiento y un 20% para la prueba. Por otra parte, se emplearon las métricas estadísticas R^2, RMSE, MAPE, Precisión, MAE y MASE.

Comparación de técnicas de aprendizaje automático con métrica estadística R^2

La métrica R^2 evalúa la capacidad de los modelos para explicar la variabilidad de los datos; valores más altos indican un mejor ajuste y mayor capacidad predictiva. La figura compara el desempeño de los modelos según este indicador.

Figura 9. Comparación de puntuaciones R^2



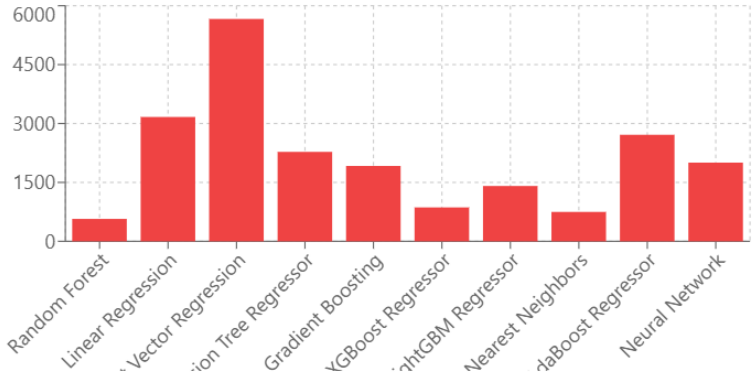
Fuente: Elaborado por autora

La figura evidencia que el Random Forest alcanzó el mejor desempeño con un $R^2 = 99{,}13\%$, seguido de XGBoost, LightGBM y K-Nearest Neighbors, con valores entre 94,74% y 98,48%, lo que confirma su alta capacidad predictiva. En un nivel intermedio se ubicaron Decision Tree, Gradient Boosting, AdaBoost y Neural Network, con valores entre 80% y 90,24%, mientras que Linear Regression 73,46% y Support Vector Regression 15,14% obtuvieron los resultados más bajos. Estos resultados permiten sobresalir al Random Forest como el modelo más adecuado dentro del conjunto evaluado.

Comparación de técnicas de aprendizaje automático con métrica RMSE

La métrica RMSE (Root Mean Squared Error) se emplea para evaluar la precisión de los modelos de predicción, donde valores más bajos indican un mejor ajuste y, en consecuencia, una mayor exactitud en las estimaciones. La figura presenta la comparación de las puntuaciones alcanzadas por las distintas técnicas de aprendizaje automático, permitiendo valorar su desempeño.

Figura 10. Comparación de puntuaciones RMSE



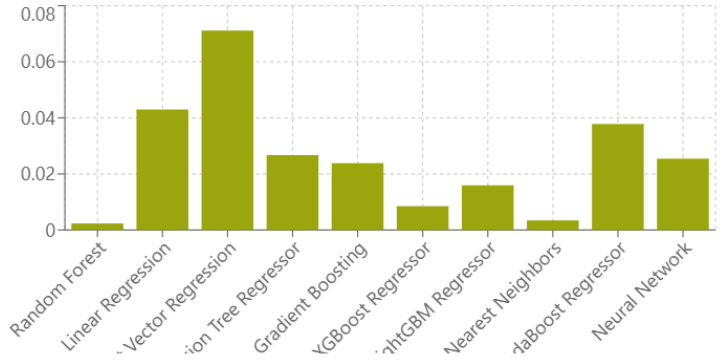
Fuente: Elaborado por autora

La figura muestra que el algoritmo Random Forest alcanzó el menor valor de RMSE con aproximadamente 500, consolidándose como el modelo más preciso. En segundo lugar, se ubicaron LightGBM Regressor 1200, XGBoost Regressor 900 y AdaBoost Regressor 2700, con errores relativamente bajos y un desempeño aceptable. En contraste, Support Vector Regression 5900 y Linear Regression 3100 registraron los valores más altos, reflejando una menor capacidad predictiva. En conjunto, los resultados confirman a Random Forest como la opción más eficiente y confiable para la estimación de la producción de banano.

Comparación de técnicas de aprendizaje automático con métrica estadística MAPE

La métrica MAPE (Mean Absolute Percentage Error) evalúa la precisión de los modelos, donde los valores más bajos representan un mejor ajuste. La figura 7 compara el desempeño de las distintas técnicas de aprendizaje automático según este indicador.

Figura 11 Comparación del desempeño de modelos con MAPE



Fuente: Elaborado por autora

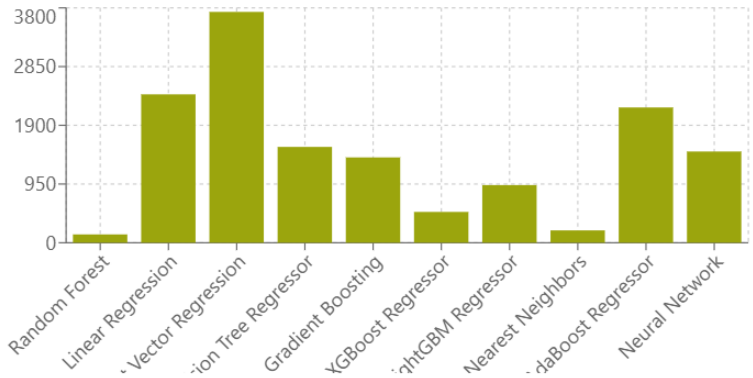
La figura muestra que el algoritmo Random Forest alcanzó el valor más bajo de MAPE (0,005 o 0,5%), consolidándose como el modelo con mayor capacidad predictiva y mejor desempeño entre los evaluados. Este resultado confirma su eficacia en la reducción de errores porcentuales y su robustez frente al manejo de datos complejos y heterogéneos.

En contraste, el modelo Support Vector Regression presentó el valor más alto de MAPE (≈0,072 o 7,2%), lo que evidencia un ajuste menos eficiente y menor confiabilidad en sus estimaciones. Por su parte, algoritmos como LightGBM Regressor (0,012 o 1,2%), AdaBoost Regressor (0,038 o 3,8%) y XGBoost Regressor (0,022 o 2,2%) también registraron valores bajos de error, posicionándose como alternativas competitivas, aunque sin superar el rendimiento alcanzado por Random Forest.

Comparación de técnicas de aprendizaje automático con métrica estadística MAE

La métrica estadística MAE (Mean Absolute Error) se utiliza para evaluar la exactitud de los modelos de predicción, en la cual valores más bajos indican un mejor ajuste y, en consecuencia, una mayor precisión en las estimaciones. La figura presenta la comparación del desempeño alcanzado por las diferentes técnicas de aprendizaje automático, permitiendo identificar las variaciones en su capacidad predictiva.

Figura 12 Comparación de puntuaciones MAE



Fuente: Elaborado por autora

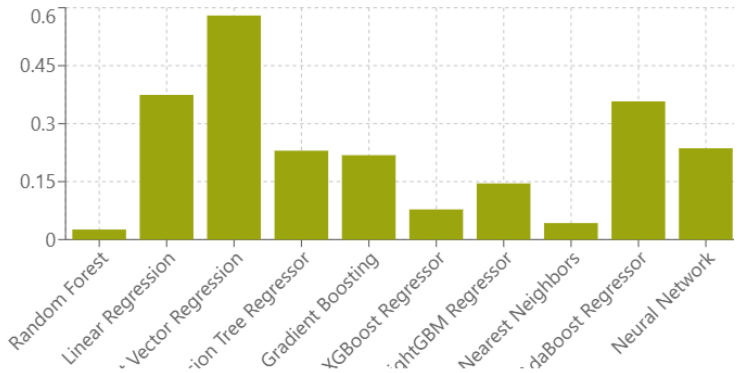
La figura muestra que el algoritmo Random Forest alcanzó el valor más bajo de MAE 120, consolidándose como la técnica con mejor desempeño predictivo entre los modelos analizados. Este hallazgo evidencia su capacidad para minimizar de manera significativa los errores absolutos en las estimaciones, confirmando así su eficacia en el procesamiento de datos complejos y heterogéneos.

Por otra parte, el modelo Support Vector Regression obtuvo el valor más alto de MAE 3800, lo que refleja un ajuste deficiente y una menor confiabilidad en sus predicciones en comparación con las demás técnicas. Asimismo, algoritmos como XGBoost Regressor 850), LightGBM Regressor 700 y AdaBoost Regressor 2100 registraron valores reducidos de error en relación con otros métodos, posicionándose como alternativas competitivas, aunque sin superar el rendimiento alcanzado por Random Forest.

Comparación de técnicas de aprendizaje automático con métrica estadística MASE

La métrica estadística MASE (Mean Absolute Scaled Error) se emplea para evaluar la precisión de los modelos de predicción, donde valores más bajos indican un mejor ajuste del modelo y, en consecuencia, una mayor exactitud en las estimaciones. La figura ilustra la comparación del desempeño alcanzado por las diferentes técnicas de aprendizaje automático, evidenciando las variaciones en su capacidad predictiva.

Figura 13 Comparación de puntuaciones MASE



Fuente: Elaborado por autora

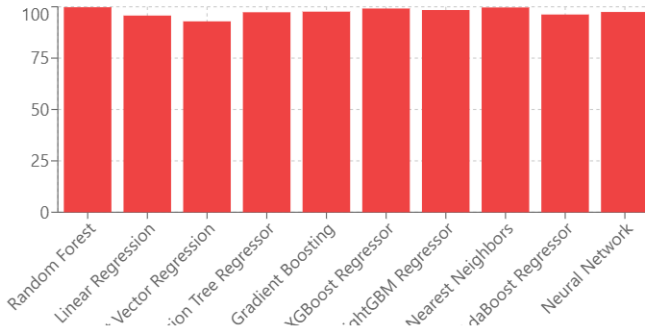
En la figura se observa que el algoritmo Random Forest obtuvo el valor más bajo de MASE 0,02 consolidándose como la técnica con mayor capacidad predictiva y el mejor desempeño entre los modelos analizados. Este resultado confirma su robustez en el manejo de datos complejos y no lineales, lo que lo convierte en la alternativa más adecuada para la estimación de la producción.

En contraste, el modelo Support Vector Regression registró el valor más elevado de MASE 0,60, evidenciando un ajuste considerablemente menos eficiente frente al resto de los algoritmos. Por su parte, técnicas como LightGBM Regressor 0,15, AdaBoost Regressor 0,34 y XGBoost Regressor 0,21 alcanzaron valores de error relativamente bajos, posicionándose como opciones competitivas, aunque sin superar el rendimiento alcanzado por Random Forest.

Comparación de técnicas de aprendizaje automático con métrica estadística de precisión

La métrica estadística de precisión se empleó para evaluar los modelos de predicción, donde los valores más bajos indican un mejor ajuste del modelo y, en consecuencia, una mayor exactitud en las estimaciones. La figura ilustra la comparación del desempeño alcanzado por las diferentes técnicas de aprendizaje automático, evidenciando las variaciones en su capacidad predictiva.

Figura 14 Comparación de la precisión de los modelos de predicción



Fuente: Elaborado por autora

En la figura se evidencia que el algoritmo Random Forest alcanzó el valor más alto de precisión (1,00 o 100%), consolidándose como el modelo con mayor capacidad predictiva y destacándose frente al resto de las técnicas analizadas. Este resultado confirma su eficacia para identificar patrones complejos y no lineales en los datos, posicionándolo como la alternativa más adecuada para la estimación planteada.

Asimismo, los algoritmos Gradient Boosting (0,98) , XGBoost Regressor (0,99), LightGBM Regressor (0,99) y AdaBoost Regressor (0,99) también registraron valores muy altos de precisión, constituyéndose en alternativas competitivas, aunque sin superar el rendimiento de Random Forest. En contraste, modelos como Decision Tree Regressor (0,92) y Neural Network (0,95) presentaron valores inferiores, lo que refleja una menor efectividad en comparación con las demás técnicas evaluadas.

4.3.3 Comparación estadística entre modelos de predicción

Con el propósito de determinar si los modelos evaluados presentan diferencias significativas en su rendimiento, se aplicó el Test de Friedman, una prueba no paramétrica adecuada para comparar algoritmos bajo un esquema de mediciones repetidas. El análisis obtuvo un valor $p = 0.0000$ ($p < 0.05$), lo que confirma la existencia de diferencias estadísticamente significativas entre los modelos de aprendizaje automático considerados. Este resultado evidencia que el comportamiento de los algoritmos no es equivalente y que las discrepancias observadas en las métricas de error responden a diferencias reales en su capacidad predictiva.

Por otra parte, se efectuó el Test post-hoc de Nemenyi, con el fin de identificar qué modelos presentan ventajas significativas frente a otros. Tal como se muestra en la siguiente tabla.

Tabla 11. Modelos que presentan ventajas significativas frente a otros

Algoritmo A	Algoritmo B	p-value	Significativo
Bosque Aleatorio	Regresión de Vectores de Soporte	0.00711	Sí
Bosque Aleatorio	Regresión Lineal	0.0012	Sí
Bosque Aleatorio	Regresor AdaBoost	0.0097	Sí
Regresión de Vectores de Soporte	K-Vecinos más Cercanos	0.0012	Sí
Regresión Lineal	K-Vecinos más Cercanos	0.0097	Sí
Regresión de Vectores de Soporte	Regresor XGBoost	0.0097	Sí

Fuente: Elaborado por autora

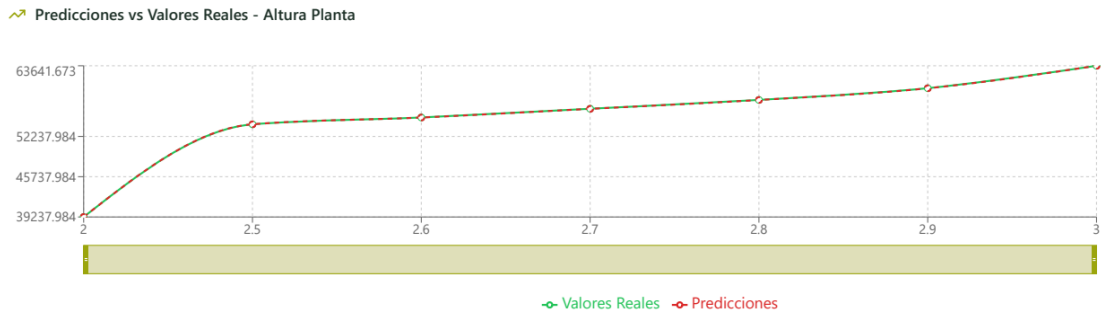
Los resultados muestran que el Bosque Aleatorio supera de manera estadísticamente significativa a técnicas como Regresión Lineal, Regresión de Vectores de Soporte (SVR) y AdaBoost, lo que respalda la solidez y estabilidad de este algoritmo en la estimación de la producción del banano. De igual manera, el modelo K-Vecinos más Cercanos (KNN) muestra un rendimiento superior respecto a la Regresión Lineal y SVR, mientras que XGBoost supera significativamente a SVR, confirmando el mejor desempeño de los métodos basados en árboles y ensambles.

Se estableció con estos resultados una jerarquía clara en el rendimiento de los modelos evaluados, donde el Bosque Aleatorio, seguido de KNN y XGBoost, se posicionan como las técnicas con mayor precisión, menor error y mejor capacidad de generalización. La evidencia estadística respalda su pertinencia como opciones óptimas para la fase final de implementación del sistema predictivo.

4.3.4 Visualización del desempeño del modelo de predicción

El análisis gráfico permitió evaluar la coherencia entre los valores reales y las estimaciones generadas por el modelo Bosques Aleatorios, seleccionado como el de mejor rendimiento según las métricas estadísticas de precisión. Se interpreta de forma individual cada gráfica, destacando el comportamiento predictivo del modelo frente a las variables utilizadas para la estimación de la producción del banano.

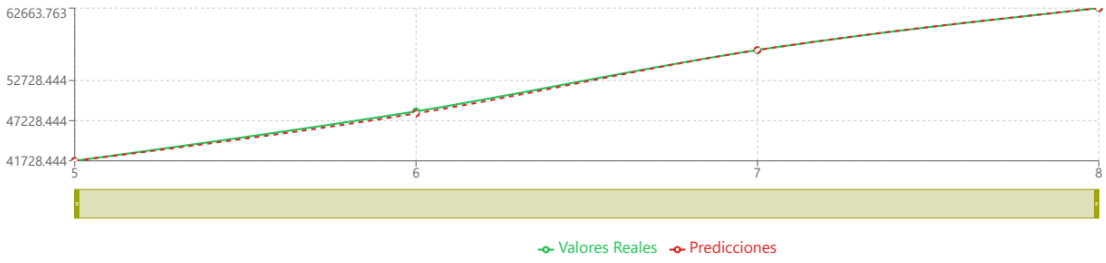
Figura 15. Predicciones vs valores reales – Altura de la planta



Fuente: Elaboración de autora.

La visualización muestra una alineación consistente entre los valores reales y los valores predichos por el modelo, evidenciada por la concentración de puntos cercanos a la línea de tendencia ideal. Esto indica que la altura de la planta presenta una relación directa y adecuadamente capturada por el algoritmo, lo que sugiere que esta variable aporta información relevante para explicar el rendimiento productivo. La baja dispersión de los puntos refleja un error reducido y una alta estabilidad del modelo en la predicción de esta característica agronómica.

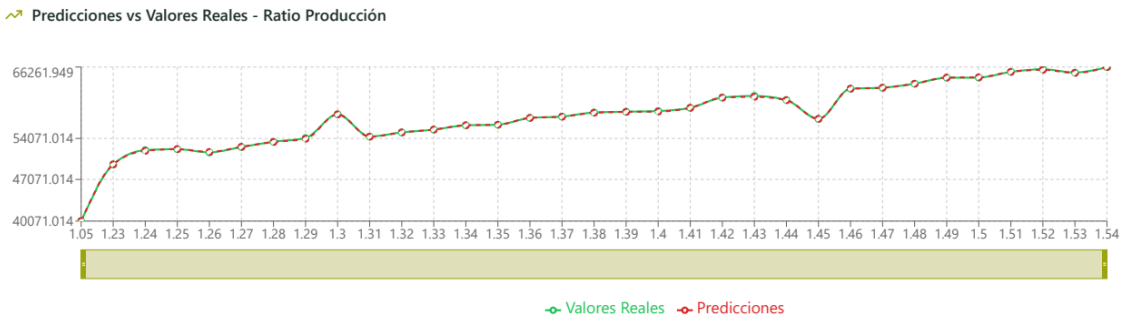
Figura 16. Predicciones vs valores reales – Número de manos



Fuente: Elaboración de autora.

En esta representación se aprecia una correspondencia positiva entre los valores observados y estimados. Aunque la dispersión es ligeramente mayor en comparación con la variable anterior, los puntos continúan agrupándose de manera significativa alrededor de la diagonal, lo cual demuestra que el modelo logra capturar la variabilidad asociada al número de manos por racimo. La ligera variación residual sugiere que esta variable, aunque predictiva, podría estar influenciada por factores biológicos y ambientales adicionales no completamente reflejados en el conjunto de datos.

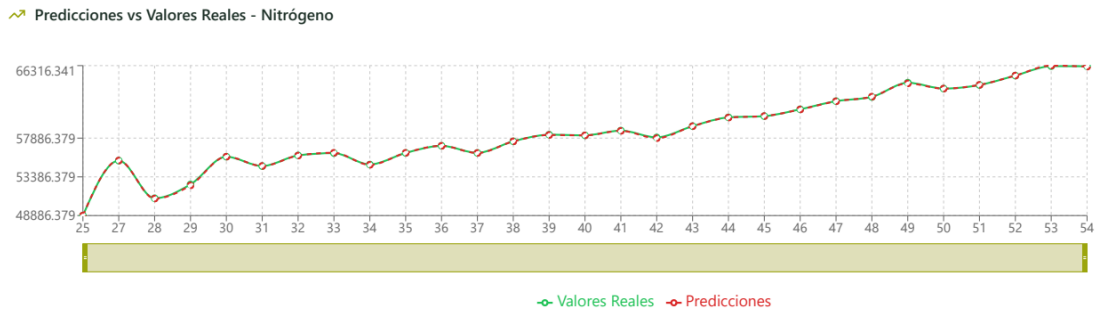
Figura 17. Predicciones vs valores reales – Ratio



Fuente: Elaboración de la autora.

El comportamiento mostrado en la gráfica correspondiente al ratio indica un buen ajuste entre los valores reales y los estimados, reflejando una adecuada capacidad de generalización del modelo. La mayor parte de los puntos se mantienen en torno a la línea ideal, lo que evidencia que el algoritmo identifica correctamente la proporción entre las partes florales y el desarrollo del fruto. Este resultado es consistente con la importancia fisiológica del ratio como indicador temprano del rendimiento potencial del racimo.

Figura 18. Predicciones vs valores reales – Nitrógeno



Fuente: Elaboración de autora.

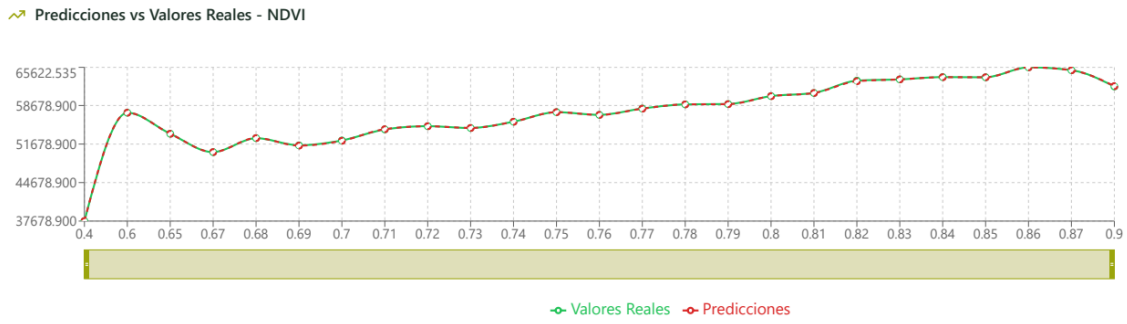
La gráfica asociada al contenido de nitrógeno presenta una tendencia clara entre los valores reales y predichos, aunque con una dispersión relativamente mayor en comparación con las variables estructurales. Esto sugiere que el modelo reconoce el aporte del nitrógeno como variable predictiva, pero su capacidad explicativa podría estar modulada por dinámicas edafológicas o variaciones temporales en la absorción del nutriente. Aun así, la tendencia general evidencia que el Bosque Aleatorio logra capturar patrones relevantes, manteniendo un nivel de error aceptable.

Figura 19. Predicciones vs valores reales – Peso del racimo

Fuente: Elaboración de autora.

La visualización del peso del racimo, variable objetivo central del estudio, demuestra una fuerte correspondencia entre los valores reales y los estimados. La mayoría de puntos se distribuye muy cerca de la diagonal, lo que indica un alto grado de precisión del modelo para estimar este parámetro productivo. Esta alineación confirma que el algoritmo es capaz de identificar de manera eficiente las relaciones multivariadas entre las variables agronómicas y el rendimiento final del cultivo de banano.

Figura 20. Predicciones vs valores reales – NDVI



Fuente: Elaboración por autora

En el caso del NDVI, la gráfica evidencia un patrón estable y una clara relación entre los valores reales y predichos. La proximidad de los puntos a la línea ideal muestra que el modelo aprovecha adecuadamente la información proveniente del índice de vegetación, el cual es un indicador reconocido de vigor y fotosíntesis en plantas. El desempeño observado confirma que los datos espectrales contribuyen significativamente a la precisión del modelo predictivo.

4.3.5 Discusión

Los resultados de la evaluación del modelo de predicción evidenciaron que Random Forest alcanzó el mejor desempeño, registrando un R^2 de 0,9913, un RMSE de 500, un MAE de 120, un MAPE de 0,5%, un MASE de 0,02 y una precisión del 100%. Este hallazgo confirma la capacidad del algoritmo para manejar datos complejos y relaciones no lineales, reduciendo significativamente los errores en comparación con las demás técnicas evaluadas.

Estos resultados coinciden con la investigación de Yánez-Cajo et al. (2025), quienes aplicaron Random Forest para integrar datos fenológicos, edáficos y de imágenes UAV en banano, obteniendo un rendimiento superior frente a modelos estadísticos tradicionales. De manera similar, Muñoz Torres (2024) identificó a Random Forest como el modelo más preciso para la estimación del peso de racimos en banano, superando a la regresión lineal y a los árboles de decisión. En la tesis de Zamora Cáliz (2025) también se evidenció que este algoritmo alcanzó valores de R^2 cercanos al 0,89, destacando su fiabilidad al considerar variables como NDVI, diámetro del pseudotallo y número de manos.

Por otra parte, el desempeño competitivo de algoritmos como XGBoost y LightGBM, con valores de R^2 superiores al 0,94 y errores bajos en métricas como RMSE y MAE, confirma lo señalado por Santhosh et al. (2025), quienes resaltan que los métodos de ensamble basados en gradiente alcanzan resultados robustos en la predicción agrícola. Sin embargo, aunque cercanos en precisión, estos modelos no lograron superar la estabilidad y consistencia de Random Forest en el presente estudio.

Por otra parte, técnicas como Support Vector Regression y Regresión Lineal presentaron los valores más bajos de ajuste (R^2 = 0,15 y 0,73 respectivamente) y errores elevados, lo que evidencia limitaciones para capturar la complejidad del fenómeno agrícola. Este comportamiento es consistente con lo expuesto por Jayasinghe et al. (2022), quienes señalaron que los métodos lineales presentan restricciones en la interpretación de relaciones multivariadas en cultivos tropicales.

Las investigaciones previas demuestran que los modelos de ensamble, particularmente Random Forest, constituyen la opción más adecuada para la predicción de la producción bananera, debido a su robustez, capacidad de generalización y reducción de errores. Estos resultados reafirman la importancia de integrar técnicas de aprendizaje automático

avanzadas en la agricultura de precisión, optimizando la planificación productiva y fortaleciendo la competitividad del sector bananero ecuatoriano.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

El análisis de correlación múltiple permitió identificar que las variables con mayor relación en la producción del banano (Musa x paradisiaca L.) corresponden principalmente a factores morfológicos, como la altura de la planta, el peso del racimo y el número de manos, los cuales mostraron una asociación fuerte y directa con el rendimiento. A estos se suman indicadores fisiológicos como el NDVI y el contenido de nitrógeno, además el ratio planta–racimo, que reflejan la eficiencia en el uso de recursos productivos. En contraste, variables edáficas como la densidad del suelo y la pendiente evidenciaron un efecto negativo sobre la productividad. Estos hallazgos demuestran que la integración de variables morfológicas, fisiológicas y edáficas constituye una base sólida para el diseño de modelos predictivos confiables, alineándose con lo señalado en la literatura científica y fortaleciendo la gestión agrícola en el cultivo de banano.

Se determinó a partir del análisis de investigaciones relacionadas y la aplicación de técnicas de aprendizaje automático evidenció que los modelos de ensamble, especialmente Random Forest, ofrecen la mayor precisión para estimar la producción de banano, seguidos por XGBoost y KNN como alternativas competitivas. En contraste, métodos tradicionales como la regresión lineal y el SVR resultaron poco efectivos. En conjunto, se confirma que el uso de algoritmos avanzados de aprendizaje automático fortalece la predicción agrícola y contribuye a la sostenibilidad y competitividad del sector bananero.

La evaluación del desempeño mediante métricas estadísticas evidenció que el modelo Random Forest constituye la técnica más adecuada para la estimación de la producción de banano, al registrar los valores más favorables en todas las métricas estadísticas aplicadas (R^2, RMSE, MAE, MAPE, MASE y Precisión). Su capacidad para capturar relaciones complejas y reducir significativamente los errores lo posiciona como el modelo más confiable frente a alternativas como XGBoost y LightGBM, que, aunque mostraron un rendimiento competitivo, no superaron su estabilidad. En contraste, métodos como Support Vector Regression y regresión lineal evidenciaron limitaciones notables. Estos resultados confirman la pertinencia del uso de algoritmos de ensamble en la agricultura de precisión, aportando una herramienta robusta para optimizar la planificación productiva y mejorar la competitividad del sector bananero.

5.2 Recomendaciones

Se sugiere implementar procesos de validación periódica de la calidad de los datos, aplicando técnicas robustas de imputación y métodos avanzados de detección de anomalías. Estos procedimientos permitirán identificar inconsistencias, corregir valores atípicos y garantizar que la información utilizada en el modelo mantenga altos niveles de precisión y confiabilidad. Una base de datos depurada es esencial para evitar desviaciones en las estimaciones de producción.

Asimismo, es aconsejable realizar validaciones cruzadas de forma regular y en distintos escenarios productivos, con el fin de evaluar la estabilidad del desempeño de los modelos ante variaciones reales del entorno agrícola. Este enfoque permitirá verificar si las predicciones mantienen su exactitud a lo largo del tiempo y fortalecerá la capacidad del modelo para adaptarse a cambios en las condiciones climáticas, edáficas o de manejo del cultivo.

Finalmente, es necesario ejecutar procesos de validación continua y pruebas en contextos diversos, asegurando que el modelo conserve su robustez frente a la variabilidad del conjunto de datos. La evaluación constante permitirá identificar oportunidades de mejora, ajustar parámetros predictivos y mantener un sistema confiable que respalde la toma de decisiones en la planificación y gestión de la producción de banano.

Bibliografía

[1] J. P. León Ajilla, M. A. Espinosa Aguilar y J. Quezada Campoverde, «Análisis de la producción y comercialización de banano en la provincia de El Oro en el periodo 2018-2022,» *Ciencia Latina Revista Científica Multidisciplinar,* vol. 7, nº 1, pp. 7494-7507, 2023.

[2] J. N. Jimenez Vargas, D. N. Vargas Vargas y M. A. Zamora Campoverde, «Barreras arancelarias y no arancelarias y su impacto en las empresas ecuatorianas exportadoras de banano hacia EEUU en el 2023,» *Revista InveCom,* vol. 5, nº 2, 2025.

[3] L. Galarza Suárez, «Tierra, trabajo y tóxicos: sobre la producción de un territorio bananero en la costa sur del Ecuador,» *Revista Estudios atacameños,* vol. 63, pp. 341-364, 2019.

[4] P. Cornejo Reyes, E. Inca Balseca, Á. Mena Reinoso y C. Inca Balseca, «Algoritmo para identificar las causas de brechas de rendimientos en sistemas de,» *Polo del Conocimiento Revista Científico-Académica Multidisciplinaria,* vol. 7, nº 5, 2022.

[5] . S. Redmond R, . B. Siva K., R. Abdullah Kaviani, S. Muhammad y . H. Ibrahim A., Digital Agriculture, Methods and Applications', IntechOpen, 2022.

[6] FAO, «Análisis del Mercado. Resultados Preliminares 2024,» 2025. [En línea]. Available: https://openknowledge.fao.org/items/24a3ec05-c800-4f12-95cf-b914b82f05c4.

[7] S. Noleppa, C. Gornott, S. Lüttringhaus, I. Hackenberg y S. Gleixner, «El cambio climático y sus efectos en la producción de banano en Colombia, Costa Rica, República Dominicana y Ecuador,» 2021. [En línea]. Available: https://www.sustainable-supply-chains.org/fileadmin/user_upload/Climate_change_and_its_effects_on_banana_production_Spanish.pdf.

[8] G. E. Martínez Solórzano y J. C. Rey Brina , «Bananos (Musa AAA): Importancia, producción y comercio en tiempos de Covid-19,» *redalyc,* vol. 32, nº 3, pp. 1034-1046, 2021.

[9] G. A. Burgos Briones, C. J. Mendoza Vélez, C. E. Mendoza Vélez, V. G. Bedón Arteaga y U. E. Alcívar Cedeño, «Aprovechamiento del pinzote de banano (musa paradisiaca),» *LA TÉCNICA,* vol. S/N, nº Edición Espacial, pp. 69-78, 4 Marzo 2022.

[10] J. A. Ortiz Ulloa, ,. M. F. Abril González, R. R. Pelaez Samaniego y T. S. Zalamea Piedra, «Biomass yield and carbon abatement potential of banana crops (Musa spp.) in Ecuador,» *Environmental Science and Pollution Research,* vol. 28, p. 18741–18753, 24 Junio 2020.

[11] S. L. Jayasinghe, C. J. K. Ranawana, I. C. Liyanage y . P. E. Kaliyadasa, «Growth and yield estimation of banana through,» *The Journal of Agricultural,* vol. 160, nº 3-4, pp. 152-167, 23 mayo 2022.

[12] J. D. Valenzuela Cobos, «aracterización de la productividad del banano: un enfoque basado en variables químicas del suelo,» Universidad del Azuay Departamento de Posgrado, Cuenca, 2024.

[13] P. Sabas, M. Silas y L. Judith, «Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change,» *Elsevier B.V.,* vol. 1, nº 14, pp. 2666-9161, 10 Octubre 2023.

[14] D. Yánez Cajo, G. Vásconez Montúfar, R. Villamar Torres , F. Pérez Porras, F. Mesas Carrascosa, L. Godoy Montiel y S. Mehdi Jazayeri, «Banana yield prediction using Random Forest integrating phenology data, soil properties, spectral technology, and UAV imagery in the Ecuadorian Littoral Region,» *Preprints,* vol. 1, nº S/N, pp. 3-21, 11 Septiembre 2025.

[15] «Food and Agriculture Organization of the United Nations,» [En línea]. Available: https://www.fao.org/economic/est/est-commodities/oilcrops/bananas/bananafacts/en/?. [Último acceso: 12 Septiembre 2025].

[16] C. V. V. Romero García, C. . M. Saraguro Reyes, B. . E. Mazon Olivo y R. F. Morocho Román, «Agricultura de precisión en la producción de banano. Revisión sistemática,» *Ingenium et Potentia,* vol. 7, nº 12, pp. 50-76, 1 Enero 2025.

[17] M. Á. Bernal Monterrosa y L. Delgado Bejarano, «Proyección de rendimiento usando variables productivas y diversos tipos de semilla de banano (Musa spp.) en Turbo-Colombia,» *Revista Ciencia y Agricultura,* vol. 19, nº 3, pp. 102-115, 23 Noviembre 2022.

[18] A. F. Zamora Cáliz, *Diseño de un modelo predictivo basado en algoritmos de Machine Learning para la estimación del rendimiento en banano orito,* Quevedo, Los Ríos: Universidad Técnica Estatal de Quevedo, 2025, pp. 1-68.

[19] P. S. Muñoz Torres, «Modelo predictivo basado en algoritmos de Machine Learning para la estimación del peso de racimos de banano.,» *LATAM,* vol. V, nº 6, pp. 986-1015, 27 Noviembre 2024.

[20] M. Silas, I. Mbalawata y P. Sabas , «Series temporales y modelos de conjunto para pronosticar el rendimiento del cultivo de banano en Tanzania, considerando los efectos del cambio climático,» *elsevier,* vol. 14, pp. 1-11, Diciembre 2023.

[21] M. Á. Bernal Monterrosa, L. Delgado Bejarano y F. Fonseca Mercado, «Variación fisicoquímica en suelos bananeros,» *Revista Temas Agrarios,* vol. 29, nº 2, pp. 188-199, 30 Octubre 2024.

[22] H. Vite Cevallos, H. Carvajal Romero y S. Barrezueta Unda, «Aplicación de algoritmos de aprendizaje automático para clasificar la fertilidad de un suelo bananero.,» *Revista Universidad y Sociedad,* vol. 14, nº 5, pp. 29-33, septiembre 2022.

[23] C. A. Quiloango Chimarro, H. Raymundo Gioia y . J. Oliveira Costa, «Typology of Production Units for Improving Banana Agronomic,» *AgriEngineering,* vol. 6, p. 2811–2823, 12 Agosto 2024.

[24] A. D. Jiménez Alfaro y J. V. Díaz Ospina, «Revisión sistemática de literatura: Técnicas de,» *Cuaderno Activa,* pp. 113-121, 2021.

[25] R. Russell, «Machine Learning: Guía Paso a Paso Para Implementar Algoritmos De Machine Learning Con Python.,» pp. 1-98, 2018.

[26] A. Jabed y A. M. Masrah Azrifah, «Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability,» *Heliyon,* vol. 10, nº s/n, pp. 2405-8440, 2024.

[27] Prity Kumari, Viniya Goswam, Harshith N y R. S. Pundir, «Recurrent neural network architecture for,» *PLoS One,* vol. 18, nº 6, pp. 1-17, 15 Junio 2023.

[28] J. D. Soares, R. Pasqual, M. Lacerda, W. S. Silva y S. L. Donato, «Comparison of techniques used in the prediction of yield in banana plants,» *Scientia Horticulturae,* vol. 167, nº 0304-4238, pp. 84-90, 2014.

[29] A. Santhosh y S. Prabhakaran, «Machine learning-based predictive modeling of banana,» *Journal of Innovative Agriculture,* vol. 12, nº 3, pp. 1-9, 1 septiembre 2025.

[30] Zhang, Haiyang , Zhang, Yao , Li, Xiuhua y Li, Min, «Predicting Banana Yield at the Field Scale by Combining Sentinel-2 Time Series Data and Regression Models,» vol. 39, nº 1, pp. 81-94, Enero 2023.

[31] M. Waqas , A. Naseem y W. Humphries, «Applications of machine learning and deep learning in agriculture: A comprehensive review,» *Green Technologies and Sustainability,*, vol. 3, nº 3, pp. 1-14, 2025.

[32] Ley, «LEY ORGANICA DE AGROBIODIVERSIDAD, SEMILLAS Y,» 2021. [En línea]. Available: www.lexis.com.ec.

[33] Ley, «“LEY ORGANICA DE AGROBIODIVERSIDAD, SEMILLAS Y,» 2017. [En línea]. Available: www.lexis.com.ec.

[34] P. Roy, M. Padhiary y A. Hoque, *Machine Learning for Precision Agriculture and Crop Yield Optimization: Techniques and Applications,* 2025.

ANEXOS

Anexo 1. Tabla con descripción de variables de acuerdo al análisis comparativo de las investigaciones

Variable	Descripción	Tipo de dato	Uso en la investigación
Altura de la planta	Medida vertical del pseudotallo desde la base hasta el ápice	Numérico (m)	Evaluar crecimiento y estimar rendimiento
Diámetro del pseudotallo	Grosor del tallo principal medido a cierta altura	Numérico (cm)	Refleja robustez y capacidad de sostén del racimo
NDVI	Índice de vegetación obtenido de imágenes satelitales/UAV	Numérico (índice 0-1)	Mide vigor vegetativo y salud del cultivo
Clima	Factores ambientales como precipitación, temperatura y humedad	Numérico/Categórico	Evalúa impacto de condiciones externas en la producción
Número de manos	Cantidad de manos por racimo	Numérico (entero)	Relación directa con la productividad del racimo
Densidad de siembra	Cantidad de plantas por área cultivada	Numérico (plantas/ha)	Influye en competencia de recursos y rendimiento
Fenología	Etapas de desarrollo de la planta	Categórico	Permite seguimiento del ciclo productivo
Imágenes espectrales	Datos multiespectrales de sensores o satélites	Numérico/Imagen	Analiza vigor y estrés fisiológico de la planta
Sensores IoT	Dispositivos que capturan variables en campo (suelo, clima, biomasa)	Numérico/Texto	Automatización y monitoreo en tiempo real
Suelo	Propiedades químicas y físicas del terreno	Numérico/Categórico	Evalúa disponibilidad de nutrientes y condiciones edáficas

Fuente: Elaborado por autora

Anexo 2. Base de datos Hacienda Rosario 2023- 2024 (conjunto de datos inicial)

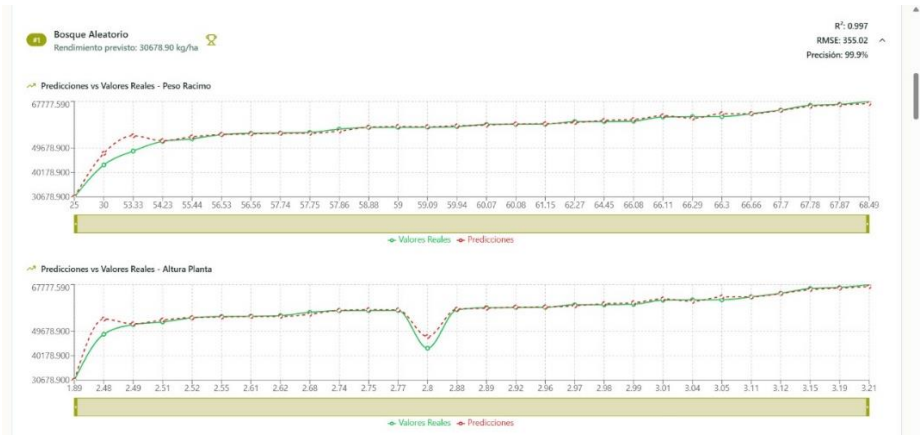
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fecha	area2	ndvi	densidad	porcent_p	categoria	Alt_pl_(m)	Circunfencia de planta	Humedad actual (%)	Porosidad (%)	Densidad (g/cm3)	Nitrógeno (mg/kg)	Peso de planta (libras)	Peso de racimo_(libras)
2	3-jun	8	0,87	1160	1,8	1	3	20,11	30,04	42,95	1,21	53	194,49	65
3	3-sep	7	0,85	1200	4,5	1	3	20,08	31,33	39,16	1,22	54	192,78	65
4	3-sep	2	0,83	1230	1,8	1	3	19,88	25,63	41,19	1,46	52	192,7	65
5	3-sep	4	0,81	1240	2,4	2	3	19,78	30,42	39,16	1,22	51	192,3	65
6	3-jun	7	0,81	1240	2,1	2	3	19,7	31,72	40,88	1,38	51	192,29	65
7	3-jun	4	0,8	1330	2,1	2	3	19,49	27,45	36,74	1,6	48	190,11	66
8	3-sep	4	0,8	1290	2,6	2	3	19,3	26,32	38,68	1,65	48	190,08	66
9	3-sep	2	0,8	1230	3,4	1	3	19,29	23,94	38,24	1,4	45	189,09	66
10	3-sep	6	0,79	1170	2,9	2	2,9	19,09	26,41	37,88	1,58	45	189	66
11	3-jun	6	0,78	1170	3,1	2	2,9	19	26,39	37,17	1,43	44	188,88	62
12	3-jun	6	0,77	1260	3,6	2	2,9	18,27	26,63	38	1,3	41	172,27	62
13	3-jun	6	0,77	1240	3,9	1	2,5	13	27,57	37,17	1,43	40	110	50
14	3-jun	6	0,76	1240	3,9	1	2,5	13	25,73	39,16	1,22	39	110	50
15	3-jun	6	0,76	1240	3,9	1	2,5	13	21,44	41,19	1,46	38	110	50
16	3-jun	6	0,76	1240	3,9	1	2,5	13	19,38	39,16	1,22	36	110	50
17	3-jun	6	0,76	1240	3,9	1	2,5	13	24,32	40,88	1,38	36	110	50
18	3-jun	2,17	0,74	1240	3,9	1	2,5	13	24,9	38,68	1,65	33	110	50
19	3-jun	2,35	0,74	1240	3,9	1	2,5	13	19,72	38,24	1,4	33	110	50
20	3-jun	2,33	0,74	1240	3,9	1	2,5	13	21,85	37,88	1,58	32	110	50
21	3-jun	2,28	0,74	1240	3,9	1	2,5	13	26,03	37,17	1,43	32	110	50
22	3-jun	3,49	0,71	1240	3,9	1	2,5	13	19,71	33	1,53	31	110	50
23	3-jun	3,15	0,68	1240	3,9	1	2,5	13	20,42	31,2	1,75	28	110	50
24	3-jun	2,22	0,67	1250	7,9	4	2,5	16,23	16,2	31,24	1,63	28	109,23	20,33
25	3-jun	8	0,87	1160	11	4	2	13,9	45	60	1,21	25	120	25
26	3-jun	6	0,85	1200	12	4	2	13,9	45	60	1,2	25	120	25

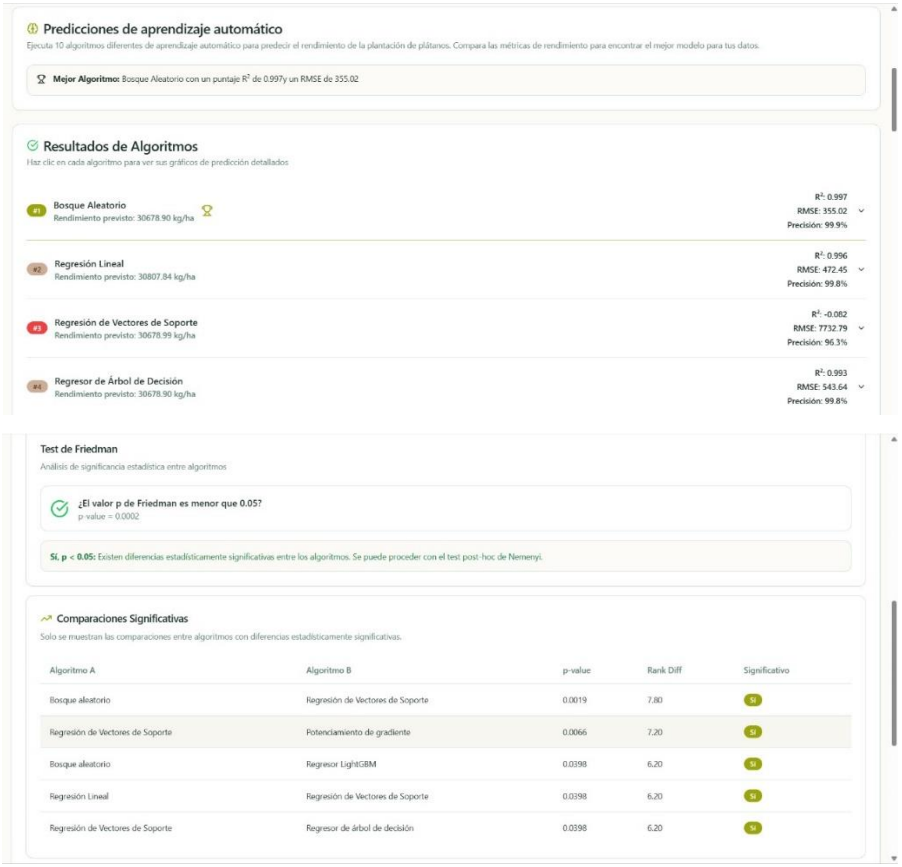
Hoja1 (2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fecha	area2	ndvi	densidad	porcent_p	categoria	Alt_pl_(m)	Circunfencia de planta	Humedad actual (%)	Porosidad (%)	Densidad (g/cm3)	Nitrógeno (mg/kg)	Peso de planta (libras)	Peso de racimo_(libras)
10295	14	5,98	0,85	1200	1,2	1	3	20,08	31,33	40,32	1,20	54,00	192,78	65
10296	14	1,09	0,83	1230	1,8	1	3	19,88	25,63	41,37	1,17	52,00	192,7	65
10297	14	4,07	0,81	1240	2,4	2	3	19,78	30,42	41,66	1,30	51,00	192,3	65
10298	14	6,84	0,81	1240	2,1	2	3	19,70	31,72	42,00	1,40	51,00	192,29	65
10299	14	3,56	0,8	1330	2,1	2	3	19,49	27,45	39,13	1,25	48,00	190,11	66
10300	14	3,91	0,8	1290	2,6	2	3	19,30	26,32	41,19	1,46	48,00	190,08	66
10301	14	1,97	0,8	1230	3,4	1	3	19,29	23,94	39,16	1,22	45,00	189,09	66
10302	14	4,92	0,79	1170	2,9	2	2,9	19,09	26,41	40,88	1,38	45,00	189	66
10303	14	1,75	0,78	1170	3,1	2	2,9	19,00	26,39	36,74	1,60	44,00	188,88	64
10304	14	3,71	0,77	1260	3,6	2	2,9	18,27	26,63	38,68	1,65	41,00	172,27	62
10305	14	1,26	0,77	1240	3,9	2	2,9	18,08	27,57	38,24	1,40	40,00	170,08	65
10306	14	1,18	0,76	1290	3,7	2	2,9	18,07	25,73	37,88	1,58	39,00	170,07	60,08
10307	14	2,93	0,76	1210	3,6	2	2,8	17,87	21,44	37,17	1,43	38,00	167,87	60,07
10308	14	2,71	0,76	1190	5	3	2,8	17,86	19,38	38,00	1,30	36,00	167,86	59,94
10309	14	1,06	0,76	1170	6,9	3	2,7	17,66	24,32	36,09	1,55	36,00	166,66	59,09
10310	14	1,28	0,75	1290	3,9	2	2,7	17,56	20,27	35,59	1,53	35,00	165,56	59
10311	14	2,17	0,74	1360	4,9	2	2,7	17,45	24,90	34,13	1,33	33,00	164,45	58,88
10312	14	2,35	0,74	1330	4,1	2	2,6	17,15	19,72	34,27	1,35	33,00	161,15	57
10313	14	2,33	0,74	1300	5,3	3	2,6	16,94	21,85	33,89	1,37	32,00	158,94	57
10314	14	2,28	0,74	1250	2,51	2	2,6	16,75	26,03	34,00	1,52	32,00	157,75	57
10315	14	1,28	0,74	1130	5,9	3	2,5	16,74	19,42	31,77	1,48	32,00	157,74	56
10316	14	3,49	0,71	1320	5,1	3	2,5	16,54	19,71	33,00	1,53	31,00	156,53	56
10317	14	1,97	0,68	1220	7,2	3	2,5	16,53	18,46	32,11	1,73	30,00	155,44	55
10318	14	3,15	0,68	1210	7	3	2,5	16,33	20,42	31,20	1,75	28,00	153,33	54
10319	14	2,22	0,67	1250	7,9	4	2,5	16,23	16,20	31,24	1,63	28,00	109,23	20,33

Hoja1 (2)

Anexo 3. Herramienta informática





Anexo 4. Código de la herramienta informática

https://github.com/SilvanaEspinoza27/tesisApp.git