Caracterización analítica de sustratos con coco y cascarilla de arroz en sistemas de cultivo sin suelo mediante sensores IoT y técnicas de analítica de datos en la sabana de Bogotá

Juan Sebastian López Galvis

Asesor

Fernando Luis Carrascal Porras

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización Ciencia y Analítica de Datos

2026

Agradecimientos

El autor expresa su sincero agradecimiento a CINTEL, Centro de Investigación y Desarrollo en Tecnologías de la Información y las Comunicaciones, por facilitar el acceso a los datos, la infraestructura de sensores IoT y el proyecto Agrotech como marco institucional para el desarrollo de este trabajo. Un reconocimiento especial al Director Ejecutivo Manuel Martínez y al Director Técnico Mario Castaño, cuyo respaldo institucional abre camino a este tipo de iniciativas de análisis aplicado que buscan generar conocimiento útil para los productores agrícolas de la región. Igualmente, el autor agradece a Piter Romero, Gerente Senior del Centro de Competencias Agrotech, y a Alfonso Díaz, Coordinador de Proyectos, por su acompañamiento técnico y disposición permanente durante el proceso de comprensión y análisis de los datos. Los resultados obtenidos son también un reflejo del compromiso de CINTEL con el desarrollo del sector agrícola colombiano y con la transferencia de conocimiento hacia quienes más lo necesitan.

Resumen

Los sustratos agrícolas determinan la necesidad de agua y nutrientes en sistemas de cultivo sin suelo, pero su comportamiento bajo condiciones de altitud en invernadero en Colombia ha sido poco documentado. Este trabajo caracterizó el comportamiento físico e hídrico del sustrato de coco y cascarilla de arroz en un cultivo de arándanos bajo invernadero en la Sabana de Bogotá a 2.600 msnm, mediante datos de una red de 21 sensores IoT y mediciones manuales realizadas durante 90 días.

Se construyó un dataset con 102 observaciones integrando humedad del suelo, temperatura, conductividad eléctrica y radiación fotosintéticamente activa, con registros manuales de tallos productivos y agua aplicada. Se entrenaron dos modelos de Random Forest, uno de clasificación del sustrato y uno de regresión de producción de tallos, validados con la estrategia Leave-One-Group-Out.

El coco retiene más del doble de humedad que la cascarilla y acumula cuatro veces más sales. El modelo de clasificación alcanzó exactitud perfecta en los 17 grupos evaluados, identificando el sustrato mediante una única regla basada en la conductividad eléctrica máxima diaria. El modelo de regresión mostró que la producción de tallos responde principalmente a la temperatura mínima del suelo y la radiación, más que al tipo de sustrato. Los resultados proveen recomendaciones técnicas para productores de la región en la selección y manejo de sustratos orgánicos en sistemas de cultivo sin suelo.

***Palabras clave*:** Sustrato orgánico, cultivo sin suelo, sensores IoT, conductividad eléctrica, humedad, Random Forest, clasificación, invernadero, Sabana de Bogotá.

Abstract

Agricultural substrates determine the water and nutrient requirements in soilless cultivation systems, but their behavior under high-altitude greenhouse conditions in Colombia has been poorly documented. This study characterized the physical and water behavior of a coconut coir and rice hull substrate in a greenhouse-grown blueberry crop in the Bogotá Savannah at 2,600 meters above sea level, using data from a network of 21 IoT sensors and manual measurements taken over 90 days.

A dataset of 102 observations was built integrating soil moisture, temperature, electrical conductivity, and photosynthetically active radiation with manual records of productive stems and daily water applied. Two Random Forest models were trained, a substrate classification model and a stem production regression model, both validated using a Leave-One-Group-Out strategy.

Coconut fiber retains more than twice the moisture of rice husk and accumulates four times more salts. The classification model achieved perfect accuracy across all 17 evaluated groups, identifying substrate type through a single rule based on maximum daily electrical conductivity. The regression model revealed that stem production responds primarily to minimum soil temperature and radiation rather than substrate type. The results provide technical recommendations for producers in the region on the selection and management of organic substrates in soilless cultivation systems.

Keywords**:** Organic substrate, soilless culture, IoT sensors, electrical conductivity, moisture, Random Forest, classification, greenhouse, Sabana de Bogotá.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La agricultura en cultivos sin suelo bajo invernadero ha tomado mayor importancia en los últimos años como alternativa productiva eficiente frente a las limitaciones de condiciones naturales del suelo convencional, especialmente en regiones con climas particulares como la Sabana de Bogotá. En estos cultivos, el sustrato cumple un papel importante en el manejo del agua y la retención de nutrientes para el cultivo, por lo que su selección y manejo adecuado son determinantes para el éxito del cultivo a desarrollar.

A pesar de su importancia, la caracterización técnica de los sustratos en condiciones locales bajo invernadero sigue siendo escasa. La mayoría de las recomendaciones disponibles provienen de contextos geográficos distintos o de forma empírica, lo que limita su aplicabilidad directa a las condiciones precisas de altitud, temperatura y radiación propias de la región de estudio.

El presente trabajo aborda esta brecha mediante la integración de dos fuentes de información complementarias: una red de sensores IoT que registra de manera continua variables de humedad, temperatura y conductividad eléctrica del suelo, y mediciones manuales de riego y producción realizadas a lo largo del ciclo de cultivo. A partir de estos datos, se desarrolla un análisis estadístico y temporal del comportamiento de dos sustratos orgánicos, coco y cascarilla de arroz en composiciones específicas que se mencionarán en el capítulo de metodología, y se construyen modelos de machine learning que permiten clasificar los sustratos e identificar las variables que mejor explican su desempeño.

Los resultados buscan generar recomendaciones técnicas concretas que orienten a productores de la región en la mejora de la selección y gestión de sustratos, aportando evidencia analítica local a una práctica agrícola en crecimiento.

Planteamiento del Problema

El crecimiento de los sistemas de cultivo sin suelo bajo invernadero en Colombia ha impulsado el uso de sustratos orgánicos como el coco y la cascarilla de arroz en estructuras de invernadero. Sin embargo, la adopción de estos materiales por parte de los productores se basa frecuentemente en criterios empíricos o en recomendaciones provenientes de contextos geográficos diferentes, sin contar con evidencia técnica generada en las condiciones específicas de la región. Raviv y Lieth (2008) señalan que las propiedades hídricas y físicas de un sustrato son altamente dependientes del entorno en el que opera, lo que hace necesario caracterizarlos bajo las condiciones reales del sitio de producción.

La Sabana de Bogotá presenta condiciones propias de altitud, aproximadamente 2.600 metros sobre el nivel del mar, temperatura moderada y alta variabilidad en radiación solar, que inciden directamente sobre la dinámica hídrica del sustrato y la demanda evaporativa del sistema. Bojacá, Gil y Cooman (2009) documentan que estas condiciones microclimáticas generan una alta variabilidad en el comportamiento de los cultivos bajo invernadero en esta región, lo que refuerza la necesidad de estudios locales que orienten las decisiones técnicas de los productores. Si bien estudios recientes como el de Monsalve Camacho et al. (2021) han avanzado en la caracterización fisicoquímica de sustratos en Colombia, el comportamiento dinámico de estos materiales bajo condiciones continuas de monitoreo con sensores IoT en la Sabana de Bogotá no ha sido documentado.

La conductividad eléctrica y la humedad del sustrato son variables ampliamente reconocidas como indicadores del estado hídrico y nutricional del sustrato del cultivo (Savvas & Gruda, 2018), la diferencias entre estos sustratos bajo condiciones locales de fertirrigación no ha sido documentado con el nivel de detalle que la tecnología de sensores IoT puede ofrecer. Este

ausencia de información técnica para la región genera incertidumbre en los productores al momento de seleccionar y gestionar adecuadamente el sustrato más apropiado para sus cultivos.

Bajo este contexto de incertidumbre surge la siguiente pregunta de investigación: *¿Qué patrones de comportamiento físico e hídrico presentan los sustratos de coco y cascarilla de arroz en cultivos bajo invernadero en la Sabana de Bogotá, y en qué medida las variables registradas por sensores IoT permiten caracterizarlos y distinguirlos mediante modelos de machine learning para orientar su selección y manejo eficiente?*

Dar respuesta a esta pregunta, mediante el análisis integrado de datos de sensores IoT complementado con datos de registro manual y el desarrollo de análisis de modelos de machine learning constituye el fin de este proyecto de investigación.

Justificación

La producción agrícola bajo invernadero en sistemas sin suelo representa una alternativa de alto potencial para la Sabana de Bogotá, donde las condiciones climáticas particulares exigen soluciones técnicas adaptadas al contexto local. La selección adecuada del sustrato es una decisión crítica en estos sistemas, dado que determina la disponibilidad hídrica y nutricional del cultivo para su ciclo de crecimiento y producción. Sin embargo, como señalan Raviv y Lieth (2008), las propiedades funcionales de los sustratos varían significativamente según las condiciones del entorno, lo que hace insuficiente extrapolar recomendaciones generadas en otros contextos geográficos.

La disponibilidad de tecnología de sensores IoT ofrece hoy una oportunidad sin precedentes para generar información técnica local de alta resolución temporal. Tzounis et al. (2017) destacan que el monitoreo continuo de variables de suelo mediante redes de sensores permite capturar dinámicas que no son observables mediante mediciones manuales esporádicas, abriendo la posibilidad de caracterizar el comportamiento de los sustratos con un nivel de detalle que antes no se había desarrollado para los productores de la región.

A esto se suma el aporte que representan los modelos de machine learning como herramienta de análisis. Sharma et al. (2021) establecen que la integración de datos de sensores agrícolas con algoritmos de aprendizaje automático permite identificar patrones y relaciones entre variables que enriquecen la comprensión del sistema de cultivo y apoyan la toma de decisiones técnicas fundamentadas. En ese sentido, el presente trabajo no solo genera conocimiento sobre el comportamiento físico e hídrico de dos sustratos orgánicos en condiciones locales, sino que propone una metodología replicable de monitoreo y análisis de datos que puede ser adoptada por otros productores e investigadores de la región.

Objetivos

Objetivo General

Caracterizar analíticamente el comportamiento físico e hídrico de los sustratos de coco y cascarilla de arroz en sistemas de cultivo sin suelo bajo invernadero en la Sabana de Bogotá, a partir del análisis de datos de campo y el desarrollo de modelos de machine learning, contribuyendo al conocimiento técnico sobre el manejo de sustratos orgánicos en la región.

Objetivos Específicos

Analizar estadística y temporalmente las variables registradas por los sensores IoT para identificar patrones de comportamiento, diferencias significativas entre sustratos y la dinámica de retención hídrica y respuesta de las variables ante eventos de riego y fertirrigación, determinando rangos operativos diferenciales entre el sustrato de coco y la cascarilla de arroz.

Desarrollar modelos de clasificación y regresión basados en Random Forest que, a partir de datos de sensores y mediciones de campo, permitan identificar el tipo de sustrato y explorar la relación entre las variables monitoreadas y la producción de tallos como indicador de respuesta del cultivo.

Formular recomendaciones técnicas basadas en los resultados preliminares del análisis y los modelos desarrollados, que orienten a productores de la región en la selección y gestión de sustratos orgánicos en sistemas de cultivo sin suelo bajo invernadero

Marco Teórico

Sistemas de Cultivo sin Suelo e Importancia del Sustrato

Los sistemas de cultivo sin suelo, también denominados sistemas hidropónicos, comprenden un conjunto de técnicas agrícolas en las que las plantas se desarrollan en ausencia del suelo mineral, utilizando en su lugar sustratos sólidos, soluciones nutritivas o la combinación de ambos. Según Raviv y Lieth (2008), estos sistemas permiten un control preciso de las variables nutricionales e hídricas del entorno radical, lo que los hace especialmente adecuados para la producción intensiva bajo invernadero.

En estos sistemas, el sustrato cumple una función estructural y funcional de primer orden: reemplaza al suelo como soporte mecánico de la planta y como medio de almacenamiento y distribución de agua y nutrientes. Urrestarazu (2004) establece que un sustrato agrícola debe reunir propiedades físicas, químicas y biológicas que garanticen la disponibilidad hídrica, la aireación radicular y la estabilidad química a lo largo del ciclo productivo. En el contexto del presente estudio, los sustratos de coco y cascarilla de arroz son evaluados bajo estas dimensiones mediante el monitoreo continuo con sensores IoT.

La creciente adopción de sustratos orgánicos en sistemas sin suelo responde, además, a criterios de sostenibilidad ambiental. Gruda (2019) señala que materiales como sutratos de coco y los subproductos agroindustriales, (entre ellos la cascarilla de arroz) representan alternativas como materiales no renovables ampliamente usados en horticultura, y que su caracterización funcional es indispensable para su validación técnica en condiciones locales de producción. En el contexto colombiano, Monsalve Camacho et al. (2021) documentan que la cascarilla de arroz es el sustrato más utilizado en sistemas de cultivo sin suelo en el país, lo que refuerza la pertinencia de su caracterización técnica comparativa frente a otros materiales como los sustratos de coco.

Propiedades Físicas e Hídricas de Sustratos Orgánicos

La caracterización de un sustrato se fundamenta en el análisis de sus propiedades físicas e hídricas, las cuales determinan su comportamiento ante el suministro de agua y la retención de nutrientes. De acuerdo con Barrett et al. (2016), las propiedades más relevantes incluyen la porosidad total, la capacidad de retención de agua, el espacio de aire y la densidad aparente. Estas propiedades condicionan directamente las variables que los sensores IoT registran en el presente estudio: humedad del suelo, temperatura y conductividad eléctrica.

La curva de retención hídrica describe la relación entre el contenido de agua en el sustrato y el potencial matricial, y varía significativamente según la estructura y composición del material (Raviv & Lieth, 2008). Un sustrato con alta capacidad de retención, como el coco, tenderá a mantener niveles más estables de humedad entre eventos de riego, mientras que uno con mayor macroporosidad, como la cascarilla de arroz, presentará drenajes más rápidos y mayor variabilidad en las lecturas del sensor. Esta diferencia es precisamente la que el análisis estadístico y temporal planteado en los objetivos del presente estudio busca cuantificar.

La conductividad eléctrica (CE) es una propiedad derivada de la concentración iónica en la solución del sustrato y constituye uno de los indicadores más utilizados para evaluar el estado nutricional y salino del medio de cultivo (Raviv & Lieth, 2008). Su monitoreo continuo mediante sensores permite detectar respuestas del sustrato ante eventos de fertirrigación, tal como se plantea en el objetivo de análisis de la dinámica hídrica del presente trabajo.

Coco y Cascarilla de Arroz Como Sustratos Orgánicos

El sustrato de coco utilizado en el presente estudio corresponde a una mezcla de turba y chip de coco. La turba es una enmienda orgánica de alta capacidad de retención hídrica y pH ácido, ampliamente utilizada en horticultura como componente de sustratos (Raviv & Lieth,

2008). El chip de coco es un subproducto del procesamiento del fruto de Cocos nucifera, de granulometría gruesa, que aporta aireación y estructura al sustrato. La combinación de ambos materiales genera un medio con alta retención hídrica proveniente de la turba y buena aireación radicular aportada por el chip, lo que explica los elevados valores de humedad registrados en este sustrato a lo largo del experimento.

La cascarilla de arroz utilizada para este estudio está compuesta por una parte cruda, otra tostada y virtua de pino patula, lo que la convierte en un material que presenta alta macroporosidad y baja capacidad de retención hídrica, características que la diferencian estructuralmente de la fibra de coco (Noguera et al., 2003). Monsalve Camacho et al. (2021) confirman este comportamiento en el contexto colombiano, señalando además que su baja capacidad de intercambio catiónico limita la retención de nutrientes en la solución del sustrato, lo que tiene implicaciones directas sobre la conductividad eléctrica registrada por los sensores. Su comportamiento hídrico es marcadamente distinto al del coco: drena el exceso de agua con mayor rapidez, lo que genera ciclos de humectación y desecación más pronunciados entre riegos. Noguera et al. (2003) establecen que la granulometría y la estructura interna del sustrato determinan en gran medida estas diferencias, las cuales deben ser evaluadas en las condiciones específicas del sitio de producción.

La comparación entre estos dos materiales bajo condiciones controladas de invernadero en la Sabana de Bogotá constituye el núcleo de caracterización del presente estudio, siendo las variables monitoreadas por los sensores IoT los indicadores cuantitativos de su comportamiento diferencial.

Dinámica Hídrica y Fertirrigación en Sistemas sin Suelo

La gestión del agua en sistemas de cultivo sin suelo es un proceso activo que involucra el suministro controlado de solución nutritiva y la respuesta del sustrato ante cada evento de riego o fertirrigación. Urrestarazu (2004) describe esta dinámica como un ciclo de humectación, retención y drenaje cuyas características dependen tanto de las propiedades físicas del sustrato como de la frecuencia e intensidad del riego aplicado. El monitoreo continuo de la humedad del sustrato permite reconstruir este ciclo a partir de las series temporales registradas por los sensores.

La conductividad eléctrica del sustrato responde de manera directa a los eventos de fertirrigación: aumenta tras el suministro de solución nutritiva y disminuye progresivamente por efecto de la absorción radicular y el drenaje. Savvas y Gruda (2018) señalan que el seguimiento de la CE en tiempo real es una herramienta clave para evaluar la eficiencia nutricional del sistema y detectar condiciones de salinidad que puedan afectar el desarrollo del cultivo. En este proyecto, el análisis de respuesta de la CE ante eventos de fertirrigación permite identificar rangos operativos diferenciales entre coco y cascarilla de arroz.

Urrestarazu (2004) destaca que la gestión del riego en invernadero debe considerar la heterogeneidad del sustrato dentro del mismo sistema, dado que distintos materiales presentan diferentes velocidades de absorción y tiempos de respuesta. Esta heterogeneidad es precisamente la que justifica el análisis temporal comparativo entre los 17 grupos del presente estudio, donde cada grupo corresponde a un sensor instalado en una planta individual con un tipo específico de sustrato.

Interacción Sustrato Ambiente en Condiciones de Invernadero

El comportamiento físico e hídrico de un sustrato no ocurre de manera aislada, sino en constante interacción con las condiciones ambientales del entorno. En sistemas de cultivo bajo invernadero, variables como la temperatura del ambiente, la humedad relativa y la radiación influyen sobre la demanda evaporativa del sistema y, por tanto, sobre la velocidad de secado del sustrato entre riegos (Fernández et al., 2010). Esta interacción es relevante para interpretar correctamente las series temporales de los sensores de suelo.

La radiación fotosintética, expresada en términos de densidad de flujo de fotones (PPFD, por sus siglas en inglés), es la variable ambiental con mayor incidencia sobre la demanda hídrica en sistemas de cultivo protegido, dado que regula la tasa de transpiración y la actividad fotosintética del cultivo (Fernández et al., 2010). En el presente estudio, la iluminación registrada por los sensores de ambiente en términos de luxes fue convertida a PPFD y constituye la única variable ambiental incluida en los modelos de machine learning, tras verificar que presenta una variación temporal significativamente mayor que las demás variables ambientales entre las fechas de muestreo, con un coeficiente de variación de 39,6 % frente a un máximo de 8,6 % en las restantes, lo que la identifica como la única variable ambiental con capacidad explicativa real para el modelado.

Las condiciones particulares de la Sabana de Bogotá, ubicada a aproximadamente 2.600 metros sobre el nivel del mar, confieren características específicas al ambiente de invernadero: temperaturas moderadas y relativamente estables a lo largo del año, alta variabilidad en radiación solar asociada a nubosidad frecuente, y baja presión barométrica. Bojacá, Gil y Cooman (2009), en un estudio realizado en invernaderos de esta región, documentan la alta variabilidad espacial y temporal de las condiciones microclimáticas como un factor determinante en la respuesta de los

cultivos. Estas condiciones locales refuerzan la pertinencia de realizar una caracterización de sustratos en este contexto geográfico específico.

Monitoreo con Sensores IoT en Agricultura

El IoT(siglas en Inglés) o Internet de las Cosas aplicado a la agricultura hace referencia a redes de dispositivos físicos equipados con sensores, capacidad de procesamiento y conectividad, que permiten la recolección automática y continua de datos del entorno de cultivo. Tzounis et al. (2017) identifica el monitoreo en tiempo real de variables de suelo y ambiente como una de las aplicaciones más consolidadas del IoT agrícola, con impacto directo en la toma de decisiones de riego, fertilización y control ambiental.

En el contexto de cultivos sin suelo, los sensores capacitivos de humedad, temperatura y conductividad eléctrica del suelo permiten registrar de manera continua el estado hídrico y salino del sustrato a lo largo del ciclo de vida del cultivo. Ferrández-Pastor et al. (2018) destacan que la frecuencia de muestreo (que en implementaciones típicas puede alcanzar decenas o centenares de registros por día por sensor) genera series temporales de alta resolución, cuyo procesamiento estadístico es indispensable para extraer información relevante sobre el comportamiento del sustrato.

La variedad y la cantidad de los datos capturados por los sensores IoT representan, al mismo tiempo, una oportunidad y un reto analítico. Sharma et al. (2021) señala que la integración de técnicas de ciencia de datos y machine learning con datos de sensores agrícolas es una tendencia consolidada que permite superar las limitaciones del análisis manual y extraer patrones que no son evidentes en la inspección directa de las series. En el presente estudio, los datos generados por los 17 sensores de suelo y los sensores de ambiente a lo largo de 90 días de

monitoreo constituyen la base para el análisis estadístico y el desarrollo de los modelos predictivos.

Ciencia de Datos y Random Forest en Contextos Agrícolas

Análisis Exploratorio y Estadístico de Datos de Sensores

El análisis exploratorio de datos (EDA) es el proceso mediante el cual se examinan las distribuciones, tendencias, patrones y relaciones presentes en un conjunto de datos antes de aplicar modelos formales. Tukey (1977), quien formalizó este enfoque, propone el uso de estadísticos descriptivos, visualizaciones y análisis de variabilidad como herramientas fundamentales para comprender la estructura de los datos. En el presente estudio, el EDA de las series temporales de los sensores IoT permite identificar ciclos diarios de humedad, temperatura y conductividad eléctrica, así como diferencias entre los sustratos evaluados.

Para la síntesis de series temporales de alta frecuencia (como las generadas por sensores con registros cada diez minutos) es práctica estándar, calcular estadísticos diarios que capturen el comportamiento de la variable a lo largo de las 24 horas: media, desviación estándar, mínimo, máximo y rango. Wilks (2011) establece que la desviación estándar y el rango son especialmente informativos en este contexto, pues reflejan la estabilidad o variabilidad del sistema durante los días, complementando así, la información que obtenemos con el promedio.

Modelos de Clasificación y Regresión con Random Forest

El modelo de Random Forest es una técnica de aprendizaje automático que funciona combinando varios árboles de decisión, cada uno de estos árboles se entrena con diferentes muestras aleatorias de los datos y con distintas variables, lo que permite, que al unir sus resultados se obtenga una predicción más precisa y confiable. Propuesto originalmente por Breiman (2001), este método reduce el sobreajuste propio de los árboles individuales y produce

estimaciones más estables e interpretables para la importancia relativa de las variables. Su aplicabilidad en datasets de tamaño moderado o pequeño, su tolerancia a variables de distinta naturaleza y su capacidad para modelar relaciones no lineales lo hacen especialmente adecuado para el dataset que tenemos como objeto de estudio.

En aplicaciones agrícolas, Random Forest ha demostrado un desempeño consistente para ejercicios de clasificación y de regresión. Jeong et al. (2016) lo emplean para predicción de rendimiento de cultivos a escala global, obteniendo resultados superiores a modelos de regresión lineal, y destacan la interpretabilidad de las importancias de variables como una de sus ventajas más relevantes para el análisis agronómico. Fernández-Delgado et al. (2014), en una comparación exhaustiva de algoritmos de clasificación, posicionan a Random Forest entre los métodos de mayor precisión en escenarios con datos reales de diversas disciplinas.

En el presente estudio se implementan dos configuraciones del algoritmo: un modelo de clasificación orientado a distinguir el tipo de sustrato a partir de las variables monitoreadas, y un modelo de regresión que relaciona dichas variables con la producción de tallos como indicador de la respuesta del cultivo. Ambos modelos son complementados con un árbol de decisión de profundidad controlada que facilita la visualización e interpretación de las reglas de clasificación en el documento final de la tesis.

Validación Cruzada en Datasets de Tamaño Reducido: Leave-One-Group-Out

La validación cruzada es una técnica que se usa para evaluar qué tan bien un modelo de machine learning puede funcionar con datos nuevos, especialmente cuando no se cuenta con mucha información, como ocurre en este trabajo de investigación. En lugar de una única partición para el entrenamiento de prueba, que bien, puede ser inestable con pocos datos, la validación cruzada se usa para entrenar y evalúar el modelo en múltiples subconjuntos del

dataset, promediando los resultados para obtener una estimación más confiable del desempeño (Kohavi, 1995).

Cuando los datos se forman en una estructura de grupos, tal como ocurre en la data que tenemos en el presente estudio, donde cada uno de los 17 grupos es observado en múltiples fechas, la variante denominada Leave-One-Group-Out (LOGO) es la más apropiada para la validación. Esto funciona de la siguiente manera: para cada iteración, un grupo completo con toda la información de una fecha, se excluye del entrenamiento y se utiliza exclusivamente para la evaluación, repitiéndose este proceso hasta que cada grupo haya actuado una vez como el conjunto de prueba (Roberts et al., 2017). Esto nos garantiza que el modelo sea evaluado sobre grupos que nunca ha visto durante el entrenamiento, simulando su aplicación a nuevas plantas o nuevos sustratos no contemplados en el experimento incial.

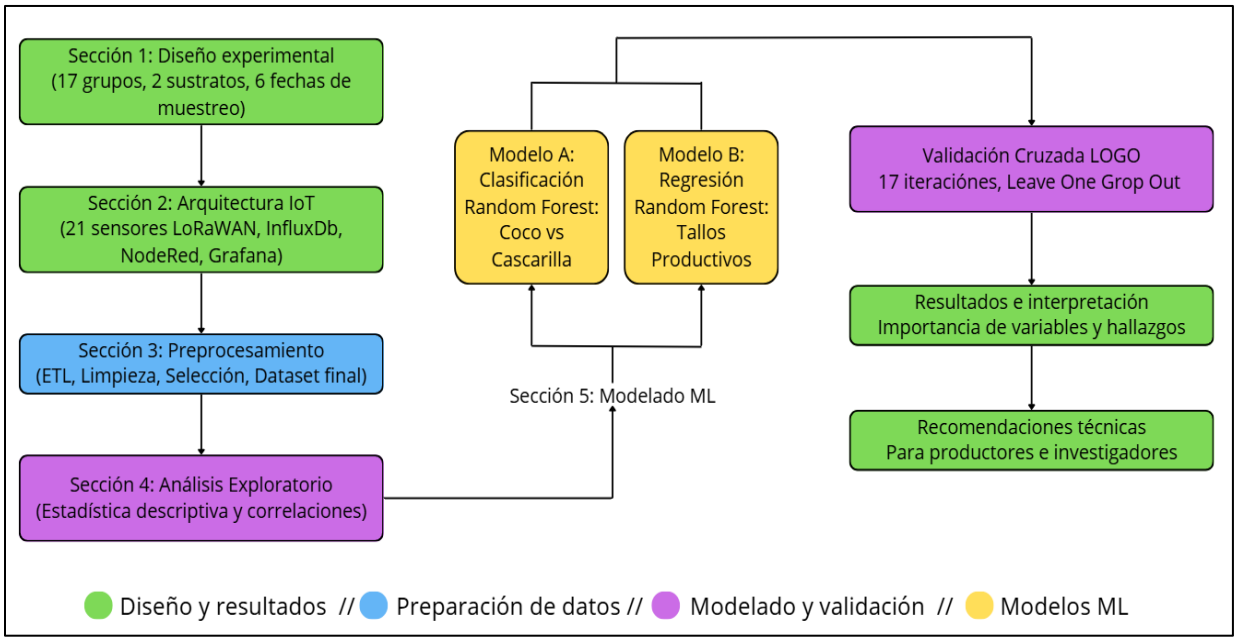
Hastie, Tibshirani y Friedman (2009) enfatizan que ignorar la estructura de grupos en la validación (realizando particiones aleatorias de filas) introduce una fuga de información que sobreestima artificialmente el desempeño del modelo. En el contexto del presente estudio, esto ocurriría si observaciones del mismo grupo aparecieran tanto en entrenamiento como en prueba, lo que permitiría al modelo memorizar patrones específicos de ese grupo en lugar de aprender características generalizables del sustrato.

Metodología

La Figura 1 presenta el diagrama de flujo general de la metodología aplicada en el estudio.

Figura 1

Diagrama de Flujo General de la Metodología del Estudio



Nota. Las etapas en verde corresponden al diseño experimental y los resultados finales; las etapas en azul a la preparación y construcción del dataset; las etapas en morado al análisis exploratorio y la validación del modelo; y las etapas en amarillo a los dos modelos de machine learning desarrollados.

Descripción General del Estudio

El presente estudio corresponde a una investigación de carácter descriptivo y cuantitativo, orientada a la caracterización del comportamiento físico e hídrico de dos sustratos orgánicos: uno que es una mezcla de turba al 70% y chip de coco al 30%, denominada en todo el documento

como “sustrato de coco o coco” y una mezcla de cascarilla de arroz cruda al 60%, cascarilla de arroz tostada al 28% y viruta de pino patula al 12%, denominada en todo el documento como “cascarilla de arroz”, ambas utilizadas en un sistema de cultivo sin suelo bajo invernadero, esto quiere decir que están sembradas cada una en bolsas individuales, las bolsas usadas son de polietileno de un solo uso, con doble filtro UV, calibre 6, color negro, de 40 cm de diámetro y se usaron 2 vólumnes, una de 60 litros y otra de 30 litros de capacidad. La investigación se desarrolló en un invernadero ubicado en la Sabana de Bogotá, aproximadamente a 2.600 metros sobre el nivel del mar, dentro del marco de un proyecto de investigación tecnológico del Centro de Investigación y Desarrollo en Tecnologías de la Información y las Comunicaciones pudiendo usar la sigla CINTEL, en el centro de competencias AGROTECH orientado al sector agropecuario.

Este estudio abarca el período comprendido entre el 1 de enero y el 31 de marzo de 2026, correspondiente a una etapa de crecimiento vegetativo inicial del cultivo de arándano variedad Emerald, sembradas el 31 de octubre de 2025, utilzando plántulas genéticamente idénticas. Al momento de la siembra, las plantas contaban aproximadamente con 60 días desde su siembra, por lo que los datos capturados representan el comportamiento de los sustratos durante la fase de establecimiento y primer desarrollo del cultivo, no desde la siembra de las plántulas.

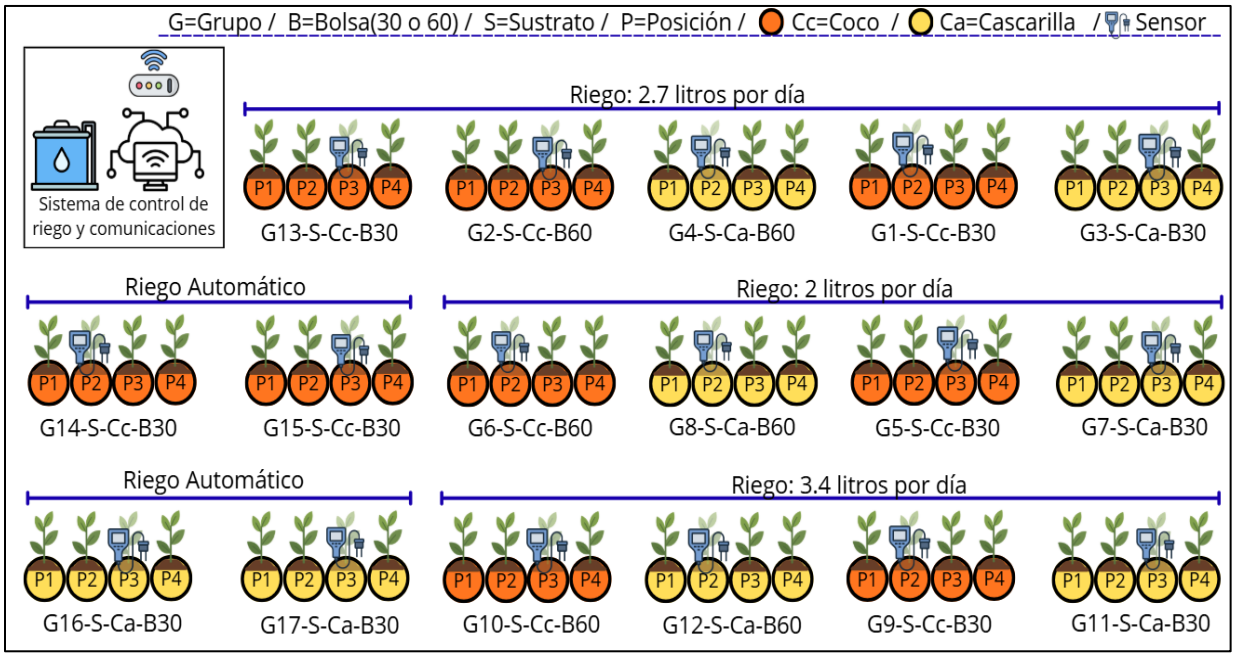
Diseño Experimental

El sistema de monitoreo se estructuró sobre 17 grupos de estudio, cada uno conformado por 4 plantas, para un total de 68 plantas. Cada grupo cuenta con un único sensor de suelo instalado en una de las cuatro plantas, lo que permite el monitoreo continuo de las variables físicas del sustrato en esa posición específica. La Figura 2 presenta un plano de distribución

espacial de los grupos en el invernadero, identificando el tipo de sustrato, la capacidad de bolsa, la modalidad de riego asignada a cada uno y la ubicación del sensor en cada grupo, y la Tabla 1 muestra la distribución de los grupos según el tipo de sustrato y capacidad de la bolsa.

Figura 2

Distribución Espacial de los 17 Grupos en el Invernadero.



Nota. Esquema planimétrico del invernadero con la ubicación de los 17 grupos de estudio, diferenciados por tipo de sustrato (coco y cascarilla de arroz), capacidad de bolsa (30 L y 60 L) y modalidad de riego asignada.

Tabla 1

Distribución de Grupos de Estudio por Tipo de Sustrato y Capacidad de Bolsa

Sustrato	Bolsa 30 L	Bolsa 60 L	Total grupos
Coco	6 grupos	3 grupos	9 grupos
Cascarilla de arroz	5 grupos	3 grupos	8 grupos

Total	11 grupos	6 grupos	17 grupos

Nota. Distribución de los 17 grupos de estudio según el tipo de sustrato (coco y cascarilla de arroz) y la capacidad de bolsa (30 L y 60 L), con los totales por fila y columna.

Para el sistema de riego se utilizó agua natural proveniente del nacimeinto de la montaña, de esta manera, se implementó el estudio bajo cuatro modalidades diferenciadas, lo que introduce una variable de manejo hídrico relevante para el análisis comparativo de los sustratos:

- Riego fijo de 2,0 L/día: aplicado a 4 grupos (G5, G6, G7 y G8).
- Riego fijo de 2,7 L/día: aplicado a 5 grupos (G1, G2, G3, G4 y G13).
- Riego fijo de 3,4 L/día: aplicado a 4 grupos (G9, G10, G11 y G12).
- Riego automático por histéresis de humedad del suelo: aplicado a 4 grupos, de los cuales 3 (G15, G16 y G17) operaron bajo este sistema durante todo el período y 1 (G14) cambió a riego fijo de 1,6 L/día a partir del 17 de marzo de 2026.

Variables Monitoreadas

Las variables del estudio provienen de dos fuentes complementarias: la red de sensores IoT, que registra datos de manera automática cada 10 minutos durante los 90 días del período de estudio, y las mediciones manuales realizadas en seis fechas específicas de muestreo. La Tabla 2 presenta el resumen de todas las variables monitoreadas, organizadas por fuente y tipo de sensor. La segunda fuente son las mediciones manuales, realizadas en 6 fechas de muestreo distribuidas entre el 19 de enero y el 30 de marzo de 2026, que registran el número de tallos productivos y la cantidad de agua recibida por planta en cada grupo.

Tabla 2

Variables Monitoreadas por Fuente de Datos en el Sistema de Cultivo

Fuente	Variable	Unidad	Frecuencia
Sensor de suelo	Humedad del sustrato	%	Cada 10 min
Sensor de suelo	Temperatura del suelo	°C	Cada 10 min
Sensor de suelo	Conductividad eléctrica	µS/cm	Cada 10 min
Sensor de ambiente	Temperatura ambiente	°C	Cada 10 min
Sensor de ambiente	Humedad relativa	%	Cada 10 min
Sensor de ambiente	CO_2	ppm	Cada 10 min
Sensor de ambiente	Iluminación / PPFD	Lux/ µmol m^{-2} s^{-1}	Cada 10 min
Sensor de ambiente	Presión barométrica	hPa	Cada 10 min
Sensor de agua	pH del agua de riego	0-14	Cada 10 min
Medición manual	Tallos productivos	Unidades	6 fechas
Medición manual	Cantidad de agua	L/día	6 fechas

Nota. Las variables de suelo fueron registradas por 17 sensores individuales, uno por grupo de estudio. Las variables de ambiente fueron registradas por 3 sensores compartidos entre todos los grupos y 1 sensor de pH del agua de riego. La radiación fue originalmente registrada en lux y convertida a densidad de flujo de fotones fotosintéticamente activos (PPFD) mediante el factor de conversión 0,0185. Las variables manuales corresponden a mediciones realizadas directamente en campo por el equipo investigador en las seis fechas de muestreo.

Enfoque Analítico

La integración de ambas fuentes de datos constituye el núcleo metodológico del estudio. Dado que la frecuencia temporal de los datos de los sensores, con aproximadamente 140 registros diarios por grupo, es significativamente mayor que la de las mediciones manuales con un registro cada 15 días. Se aplicó un proceso de agregación estadística que sintetiza el comportamiento diario de cada variable de sensor en estadísticos representativos: media, desviación estándar, mínimo, máximo y rango. Este proceso permite alinear ambas fuentes en un único dataset analítico de 102 observaciones para 17 grupos con 6 fechas de muestreo, sobre el cual se desarrolla el análisis estadístico y los modelos de machine learning descritos en las siguientes secciones.

Arquitectura de la Red IoT Para el Monitoreo de Sustratos

La caracterización de sustratos en cultivos sin suelo requiere un monitoreo continuo y de alta resolución temporal de sus propiedades físicas y químicas. Para ello, en CINTEL se implementó una arquitectura de red IoT basada en la tecnología LoRaWAN, la cual ofrece bajo consumo energético, largo alcance y robustez en entornos agrícolas. Esta red permite la adquisición automática de datos cada 10 minutos (parámetro configurable para los sensores), superando significativamente la frecuencia de las mediciones manuales tradicionales cada 15 días y facilitando el posterior alineamiento de datos para el análisis y el modelado de machine learning.

La red IoT está compuesta por un total de 21 sensores marca Milesight distribuidos de la siguiente manera:

17 sensores de suelo EM500-SMTC, encargados de medir humedad, temperatura y conductividad eléctrica (EC) del sustrato.

3 sensores ambientales (EM500-CO2, EM300-TH y EM500-LGT), que registran temperatura, humedad relativa, CO2, iluminación (lux) y presión atmosférica del entorno.

1 sensor de agua (RIKA RK500-22), destinado al monitoreo del pH del agua de riego.

La Figura 3 ilustra la instalación física representativa de un sensor de suelo dentro de una bolsa de cultivo, mostrando su posición en relación con las cuatro plantas del grupo.

Figura 3

Fotografía de la Instalación Física de un Sensor de Suelo en Bolsa de Cultivo.

Nota. Fotografía de campo que muestra la instalación de un sensor de suelo dentro de una bolsa de cultivo, con las cuatro plantas del grupo visibles en su contexto de invernadero. El sensor se instaló a 5 cm de la superficie de forma vertical para todos casos.

Todos los sensores operan como dispositivos end-devices de clase A en LoRaWAN y transmiten sus paquetes de datos de forma periódica al gateway central. Este gateway actúa como concentrador, consolida la información recibida y la reenvía a través de protocolo TCP/IP hacia la infraestructura de almacenamiento y procesamiento en un servidor local.

El flujo de datos continúa a través de las siguientes capas de software:

InfluxDB: Base de datos de series temporales optimizada para el almacenamiento eficiente de datos de alta frecuencia.

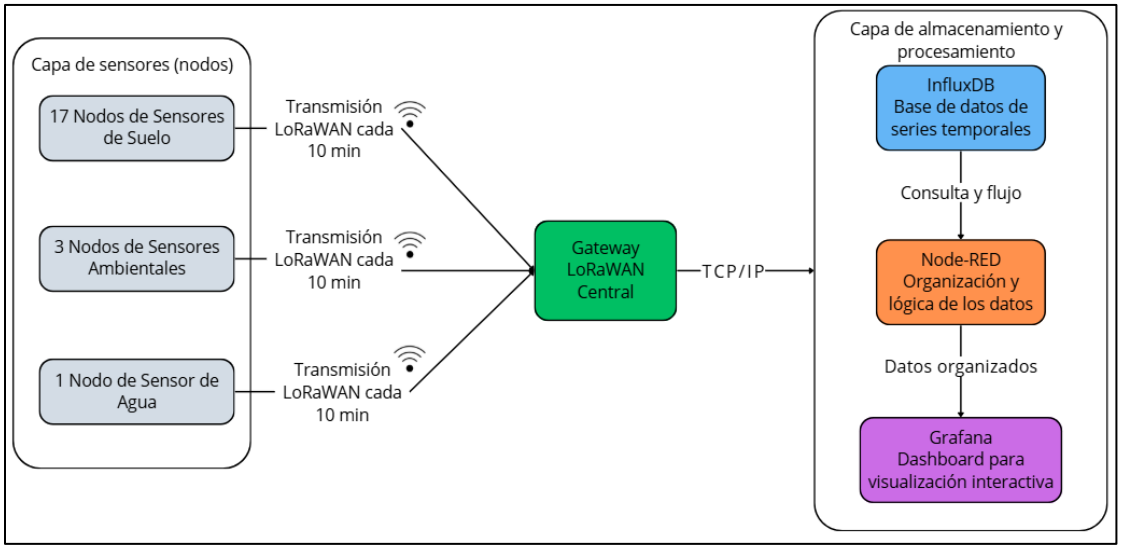
Node-RED: Plataforma de programación visual de bajo código que se encarga de la organización, filtrado, validación y control lógico de los flujos de datos.

Grafana: Herramienta de visualización que genera un dashboard interactivo en tiempo real, permitiendo al usuario consultar tendencias, alarmas y métricas clave de los sustratos de manera clara y accesible.

La Figura 4 ilustra esquemáticamente el flujo completo de la información desde los sensores hasta la interfaz de usuario.

Figura 4

Diagrama de la Arquitectura de la red IoT LoRaWAN Implementada Para el Monitoreo Continuo de Sustratos.



Nota. Diagrama de arquitectura por capas de la red IoT LoRaWAN implementada, que ilustra el flujo de datos desde los sensores end-devices hasta la interfaz de visualización en Grafana, pasando por el gateway central, InfluxDB y Node-RED.

La Tabla 3 presenta un resumen de los componentes principales de la red.

Tabla 3

Componentes de la Arquitectura IoT LoRaWAN

Componente	Cantidad	Función principal	Protocolo de comunicación
Sensores de suelo	17	Humedad, temperatura, EC	LoRaWAN (cada 10 min)
Sensores ambientales	3	Temperatura, humedad relativa, CO2, Iluminación (lux) y presión atmosférica	LoRaWAN (cada 10 min)
Sensor de agua	1	pH	LoRaWAN (cada 10 min)
Gateway central	1	Recepción y consolidación de paquetes	LoRaWAN → TCP/IP
InfluxDB	1	Almacenamiento de series temporales	TCP/IP
Node-RED	1	Organización, filtrado y control lógico	Interno
Grafana	1	Visualización y dashboard interactivo	Interno

Nota. Listado de los componentes de la red IoT LoRaWAN implementada, con su cantidad, función principal y protocolo de comunicación utilizado en cada capa de la arquitectura.

Esta arquitectura garantiza escalabilidad, redundancia y facilidad de mantenimiento, además de proporcionar la resolución temporal de 10 minutos que resulta fundamental para el preprocesamiento y el alineamiento con las mediciones manuales complementarias descritas en la sección de construcción del dataset para modelado.

Recolección y Preprocesamiento

Carga y Unión Inicial de Datos

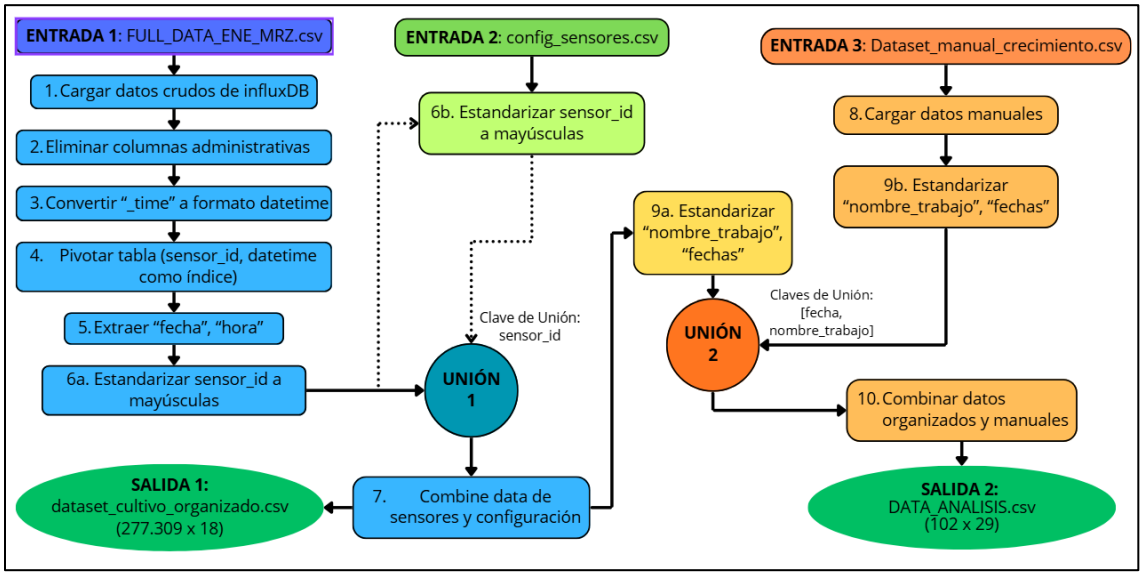
Los datos crudos provenientes de los sensores fueron extraídos desde la base de datos InfluxDB en formato CSV. Dado que estos registros se encontraban identificados únicamente mediante el campo sensor_id, se realizó un proceso de integración con el archivo auxiliar (config_sensores.csv) donde se establece la equivalencia entre el sensor_id y su nombre de trabajo para el proyecto, utilizando dicha variable como clave de unión. Este procedimiento permitió enriquecer cada observación con información contextual relevante: nombre del sensor, tipo de sustrato, tamaño de bolsa, posición del sensor dentro del grupo y grupo de estudio al que pertenece. El resultado de esta etapa se consolidó en el archivo dataset_cultivo_organizado.csv, que constituye la base de datos cruda enriquecida para todos los análisis posteriores.

Posteriormente se realizó un segundo proceso de integración incorporando los datos recolectados manualmente de cada 15 días desde el archivo Dataset_manual_crecimiento.csv. Las claves de unión utilizadas fueron la fecha de registro y el identificador del grupo (nombre_trabajo). Durante esta fase se mantuvieron diferenciadas las variables de posición provenientes de cada fuente (la del sensor y la del registro manual) con el propósito de preservar la trazabilidad completa de la información e incluir tanto las plantas instrumentadas con sensor como las demás plantas del grupo.

El flujo completo de integración, incluyendo las fuentes de entrada, las claves de unión y los archivos generados en cada etapa, se ilustra en la Figura 5. El procedimiento detallado de implementación se describe en el Apéndice A: Pseudocódigo de carga y unión de datos.

Figura 5

Diagrama de Flujo de la Integración de Fuentes de Datos



Nota. Diagrama de flujo que muestra las tres fuentes de datos de entrada, registros crudos de InfluxDB, archivo de configuración de sensores y dataset manual de crecimiento, las claves de unión utilizadas en cada operación de integración y los archivos generados en cada etapa.

La Tabla 4 resume las dimensiones y características de los datasets generados en esta fase.

Tabla 4

Estructura y Dimensiones de los Datasets Generados en la Fase de Integración

Dataset generado	Filas	Columnas	Descripción
dataset_cultivo_organizado.csv	277.309	18	Data cruda enriquecida con metadata de sensores

DATASET_ANALISIS.csv	102	29	Dataset final agregado para modelado

Nota. El DATASET_ANALISIS contiene 29 columnas en total, de las cuales 2 son identificadores del grupo (datetime y nombre_trabajo), 1 fue excluida del análisis (inicio de producción) y 1 variable de radiación fue descartada por valores sistemáticamente nulos (radiacion_min). Cada modelo utiliza 24 de las 25 columnas restantes como variables predictoras, dado que la variable objetivo no actúa simultáneamente como predictora.

Limpieza y Preparación de los Datos

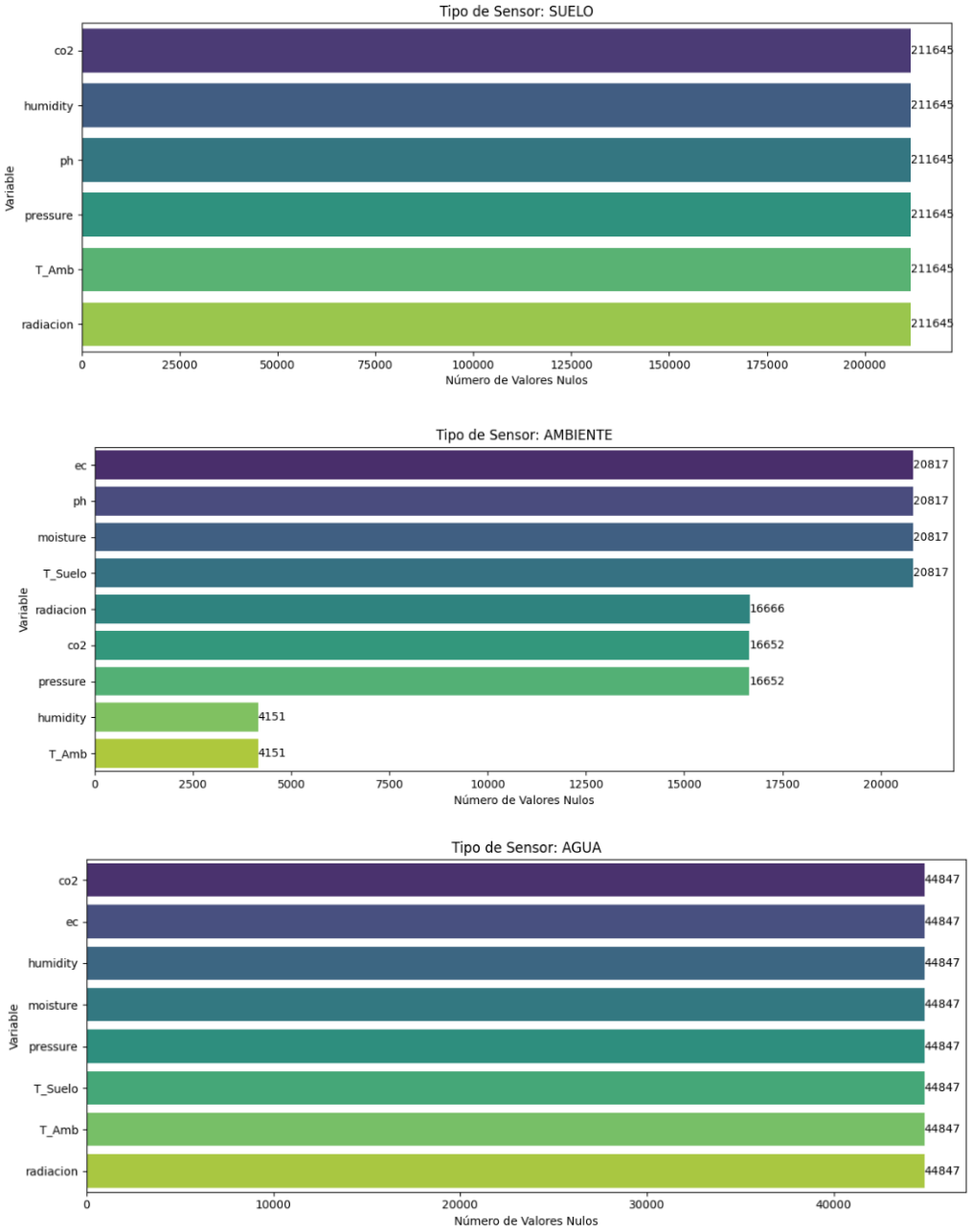
El proceso de limpieza se aplicó sobre el archivo dataset_cultivo_organizado.csv y tuvo como propósito garantizar la integridad, consistencia y calidad de los datos antes de cualquier análisis o modelado. Cada decisión de transformación se fundamentó en criterios estadísticos o en conocimiento del dominio agrícola, priorizando la conservación de la información sobre su eliminación.

Separación de Variables de Temperatura. La variable de temperatura registrada por los sensores combinaba en una misma columna mediciones provenientes de dos entornos distintos: el suelo y el ambiente. Dado que estas dos magnitudes tienen comportamientos físicos y rangos de valores diferentes, se separaron en dos columnas independientes (T_Suelo y T_Amb) asignando cada valor según el tipo de sensor que lo originó. La columna original fue eliminada para evitar redundancia. Los valores nulos resultantes en cada columna se preservaron como nulos estructurales, ya que representan la ausencia legítima de medición: un sensor de suelo no registra temperatura ambiente y viceversa.

Tratamiento de Valores Faltantes. Los valores faltantes se abordaron con estrategias diferenciadas según la naturaleza de cada variable. Como parte del protocolo estándar de preprocesamiento, se revisó la presencia de valores faltantes en todas las variables de sensor mediante el conteo de nulos agrupado por sensor y por variable. El análisis reveló que las series temporales de los sensores de suelo y de ambiente presentaron continuidad prácticamente completa a lo largo del período de estudio, resultado consistente con la alta confiabilidad de la arquitectura LoRaWAN implementada. Para los valores faltantes residuales detectados en variables dispersas (pH, radiación, humedad relativa, CO_2 y presión, cuya cobertura es estructuralmente menor por corresponder a sensores únicos) se aplicó imputación por mediana calculada individualmente por sensor. En variables de suelo con nulos estructurales en los extremos de las series, se aplicó forward-fill y backward-fill como medida de respaldo. En ambos casos, si un sensor no reportó históricamente ningún valor para una variable, los nulos se mantuvieron como estructurales para preservar la integridad semántica del dataset.

Figura 6

Recuento de Valores Nulos por Variable Antes del Tratamiento de Datos Faltantes, Desagregado por Tipo de Sensor

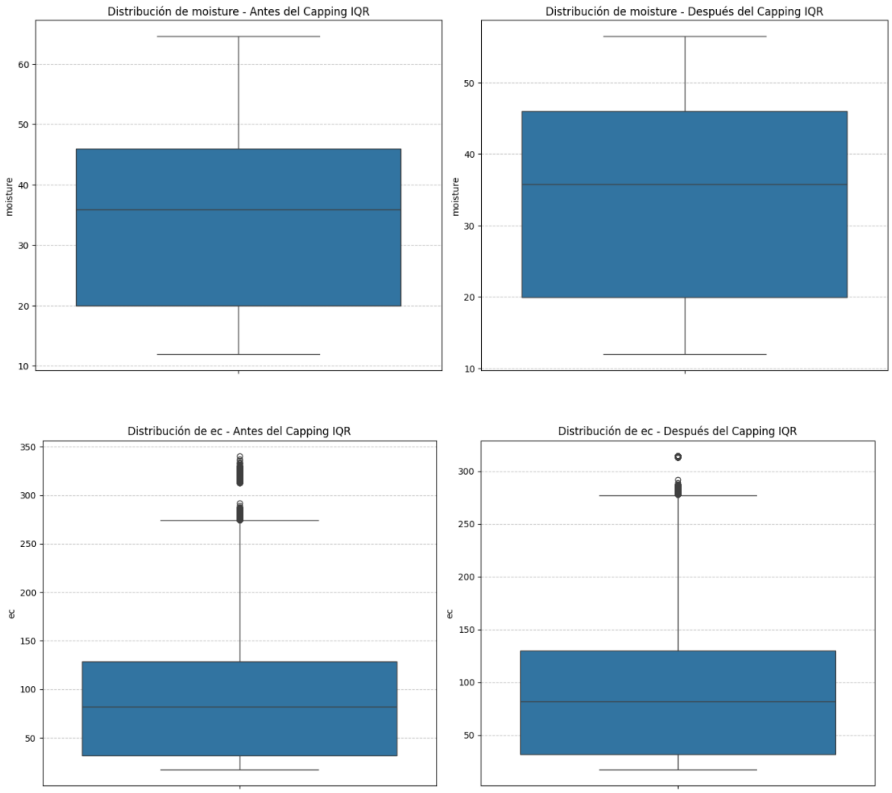


Nota. Gráfico de barras horizontales con el recuento de valores nulos detectados en cada variable del dataset antes del tratamiento de datos faltantes, diferenciando entre nulos estructurales propios de variables con cobertura parcial y nulos en series continuas de sensores de suelo.

Identificación y Corrección de Valores Atípicos. La detección de valores atípicos se hizo utilizando el método del rango intercuartílico (IQR) este se aplicó por separado a cada sensor, con el fin de tener en cuenta las características propias de cada dispositivo y no mezclar comportamientos distintos. Los valores identificados fuera de los límites calculados fueron acotados (técnica conocida como capping) en lugar de eliminados, lo que evita la pérdida de registros y mantiene la estructura temporal de las series. Adicionalmente, se aplicaron restricciones basadas en el conocimiento del dominio: la humedad del sustrato y la humedad relativa del ambiente se limitaron al rango físicamente válido de 0 % a 100 %, y la conductividad eléctrica se restringió a valores no negativos. Estas restricciones corrigen lecturas que, aunque no sean outliers estadísticos, son físicamente imposibles en el contexto del experimento. La Figura 7 presenta los diagramas de caja comparativos de humedad del sustrato y conductividad eléctrica antes y después de la aplicación del método IQR, ilustrando el efecto del tratamiento sobre la distribución de ambas variables**.**

Figura 7

Boxplots Comparativos de Humedad del Sustrato y Conductividad Eléctrica Antes y Después del Tratamiento de Valores Atípicos Mediante Capping IQR



Nota. Diagramas de caja comparativos de la humedad del sustrato y la conductividad eléctrica antes y después de la aplicación del método de rango intercuartílico (IQR) para la corrección de valores atípicos, mostrando el efecto del tratamiento sobre la distribución de ambas variables.

Eliminación de Duplicados y Estandarización de Etiquetas. Se identificaron y eliminaron registros duplicados para garantizar que cada observación fuera única. Adicionalmente, las variables categóricas (tipo de sustrato, tamaño de bolsa y tipo de sensor) fueron estandarizadas convirtiendo sus valores a mayúsculas y eliminando espacios en blanco,

con el fin de evitar que variaciones de formato generaran categorías artificialmente distintas durante el análisis.

Conversión de Unidades de Iluminación. La variable de iluminación, registrada originalmente en lux, fue convertida a radiación fotosintéticamente activa expresada en micromoles por metro cuadrado por segundo (μmol m^{-2} s^{-1}), utilizando el factor de conversión 0,0185. Esta transformación es estándar en investigación agrícola y permite expresar la variable en las unidades físicamente relevantes para el análisis del efecto de la radiación sobre el sistema de cultivo (Fernández et al., 2010).

El resultado de todas estas etapas es un dataset limpio, consistente y listo para la fase de integración y modelado, cuyo resumen de intervenciones se presenta en la Tabla 5. El procedimiento completo de implementación de cada etapa de limpieza se describe en el Apéndice B Pseudocódigo de limpieza y preparación de datos.

Tabla 5

Resumen de Intervenciones de Limpieza Aplicadas al Dataset

Aspecto	Criterio de detección	Acción realizada
Temperatura duplicada	Dos entornos en una columna	Separación en T_Suelo y T_Amb
Valores faltantes en series continuas	Brechas en humedad y EC	Interpolación lineal por sensor
Valores faltantes en variables dispersas	pH, radiación, CO_2, presión	Imputación por mediana por sensor
Valores atípicos estadísticos	Método IQR por sensor	Capping en límites calculados

Aspecto	Criterio de detección	Acción realizada
Valores físicamente imposibles	Conocimiento del dominio	Restricción de rango por variable
Registros duplicados	Filas idénticas	Eliminación directa
Inconsistencias categóricas	Variaciones de formato	Estandarización a mayúsculas
Unidades de iluminación	Lux → PPFD	Conversión × 0,0185

Nota. Resumen de las ocho intervenciones de limpieza aplicadas al dataset organizado, con el criterio de detección utilizado en cada caso y la acción ejecutada para corregir o tratar el problema identificado.

Selección de Variables Ambientales

La red IoT implementada registra seis variables de ambiente de manera continua: temperatura ambiente (T_Amb), humedad relativa, CO_2, radiación (convertida a PPFD), presión barométrica y pH del agua de riego. Sin embargo, dado que estas variables son registradas por sensores compartidos entre todos los grupos (a diferencia de los sensores de suelo, que son individuales por grupo), su valor es idéntico para los 17 grupos en una misma fecha. Esta característica implica que su capacidad de discriminar entre sustratos dentro de una misma fecha de muestreo es nula: el árbol de decisión no podría utilizar estas variables para separar coco de cascarilla si ambos comparten exactamente el mismo valor ambiental en el mismo momento.

Por esta razón, la inclusión de variables ambientales en el dataset de modelado se justifica únicamente si presentan variación real entre las seis fechas de muestreo. Una variable ambiental con baja variabilidad entre fechas no aporta información explicativa al modelo y solo incrementa la dimensionalidad del dataset sin beneficio analítico.

Variación Temporal entre Fechas de Muestreo. Para identificar qué variables ambientales presentan variación temporal relevante para el modelado, se realizó el cálculo para el coeficiente de variación (CV) de cada una entre las seis fechas de muestreo. Este criterio aplica exclusivamente a las variables ambientales, dado que su valor es idéntico para todos los grupos en una misma fecha al provenir de sensores compartidos y no individuales por grupo. A diferencia de las variables de suelo, donde cada sensor produce lecturas independientes y la variación entre grupos es la información discriminante, las variables ambientales solo pueden aportar información al modelo si presentan variación entre las fechas de muestreo. El CV, que muestra la desviación estándar como un porcentaje de la media, permite comparar que tanta variabilidad tienen distintas variables, siendo en este caso el indicador adecuado para evaluar si una variable compartida captura condiciones temporales diferenciadas a lo largo del periodo de estudio.

Los resultados se presentan en la Tabla 6 y muestran una brecha marcada entre las variables evaluadas: la radiación fotosintéticamente activa presenta un CV de 39,6 %, mientras que todas las demás variables no superan el 8,6 %. Esta separación natural en los datos identifica a la radiación como la única variable ambiental con variación temporal suficiente para aportar información discriminante al modelo. Las cinco variables restantes fueron excluidas al presentar valores agrupados en un rango de 0,1 % a 8,6 %, indicando comportamiento prácticamente constante a lo largo del período de estudio. El procedimiento completo de cálculo se describe en el Apéndice C: Pseudocódigo de selección de variables ambientales.

Tabla 6

Coeficiente de Variación de las Variables Ambientales Entre las Seis Fechas de Muestreo

Variable	CV (%)	Decisión
Radiación (PPFD)	39,6 %	Incluir
T_Amb	8,6 %	Descartar
Humedad relativa	5,1 %	Descartar
CO_2	3,2 %	Descartar
pH	7,0 %	Descartar
Presión barométrica	0,1 %	Descartar

Nota. Coeficiente de variación calculado para cada una de las seis variables ambientales monitoreadas entre las seis fechas de muestreo, con la decisión de inclusión o descarte para el dataset de modelado.

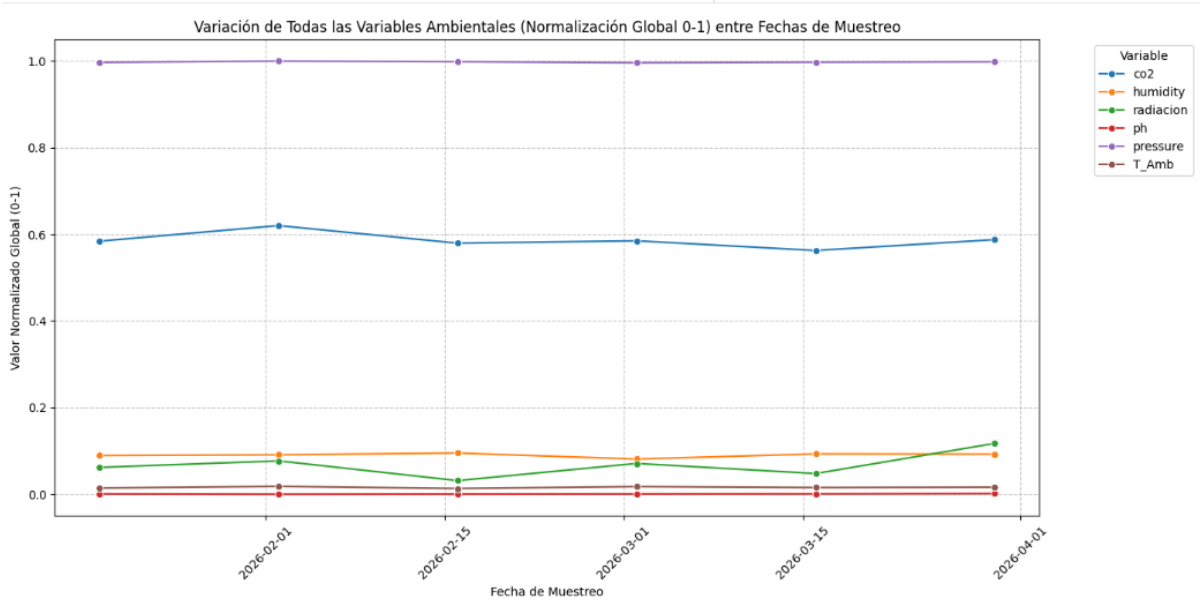
Interpretación de Resultados. La radiación fotosintéticamente activa es la única variable que presenta una variación temporal diferente a las demás, con un CV del 39,6 % frente a un máximo del 8,6 % en las restantes cinco variables, esta brecha natural en los datos la identifica como la única variable ambiental con variación temporal suficiente para aportar información al modelado. El resultado es coherente con las condiciones climáticas de la Sabana de Bogotá, donde la alta variabilidad en nubosidad genera diferencias significativas en la radiación incidente entre fechas, mientras que la temperatura, la humedad relativa y la presión barométrica permanecen relativamente estables a lo largo del año debido a la altitud constante y el régimen climático de la región (Bojacá et al., 2009).

La presión barométrica presenta el CV más bajo (0,1 %), lo que era esperable dado que su valor depende fundamentalmente de la altitud, la cual no varía entre fechas. El CO_2 y la humedad relativa muestran variaciones menores al 6 %, consistentes con un ambiente de invernadero con ventilación natural. El pH del agua de riego, aunque registrado por el sensor de agua, presentó un CV del 7,0 % y fue igualmente descartado.

La Figura 8 permite visualizar este resultado de manera directa: mientras las cinco variables descartadas se representan como líneas prácticamente horizontales a lo largo de las seis fechas, la radiación muestra variaciones de pendiente entre cada fecha de muestreo, reflejando la alta variabilidad en la nubosidad característica de la Sabana de Bogotá durante el período de estudio.

Figura 8

Variación Normalizada de Variables Ambientales



Nota. Gráfico de líneas con la variación normalizada (escala 0-1) de las seis variables ambientales monitoreadas a lo largo de las seis fechas de muestreo, que permite comparar visualmente la estabilidad de cada variable en el período de estudio.

Resultado de la Selección. De las seis variables ambientales monitoreadas, únicamente la radiación fotosintéticamente activa (PPFD) fue seleccionada para integrarse al dataset de modelado. Las cinco variables restantes fueron excluidas por no presentar variación temporal suficiente entre las fechas de muestreo, condición que las hace prescindibles para el análisis diferencial entre sustratos. Esta decisión reduce la dimensionalidad del dataset sin pérdida de información relevante y mejora la interpretabilidad de los modelos desarrollados en la sección de desarrollo del modelo de machine learning.

Construcción del Dataset para Modelado

A partir del dataset_cultivo_organizado.csv limpio y transformado, se construyó el dataset final denominado DATASET_ANALISIS, que constituye el insumo principal para el análisis estadístico y el desarrollo de los modelos de machine learning. Este proceso implicó cuatro operaciones secuenciales: filtrado por fechas de muestreo, agregación estadística de variables de sensor, integración de variables manuales y la de ambiente (radiación), y ajuste final de tipos de datos.

Filtrado por Fechas de Muestreo. Dado que las mediciones manuales se realizaron en seis fechas específicas, el dataset de sensores, que abarca 90 días continuos de monitoreo, se filtró para conservar únicamente los registros correspondientes a esas seis fechas. Este criterio establece la prioridad de las fechas manuales sobre la serie completa de sensor, garantizando que cada fila del dataset final tenga correspondencia con una observación manual validada en campo. Las fechas de muestreo utilizadas se presentan en la Tabla 7.

Tabla 7

Fechas de Muestreo Utilizadas como Referencia Para la Construcción del Dataset

Fecha	Días desde inicio	Variables manuales registradas
19/01/2026	0	Tallos productivos, cantidad de agua (L)
02/02/2026	14	Tallos productivos, cantidad de agua (L)
16/02/2026	28	Tallos productivos, cantidad de agua (L)
02/03/2026	42	Tallos productivos, cantidad de agua (L)
16/03/2026	56	Tallos productivos, cantidad de agua (L)
30/03/2026	70	Tallos productivos, cantidad de agua (L)

Nota. Listado de las seis fechas de muestreo manual utilizadas como referencia para la construcción del dataset, con los días transcurridos desde el inicio del estudio y las variables manuales registradas en cada fecha.

Agregación Estadística de Variables de Sensor. Cada variable de sensor registra aproximadamente 140 lecturas por día para cada grupo. Utilizar estos registros individuales como filas independientes del dataset de modelado constituiría una pseudoreplicación, ya que el valor manual asociado (número de tallos productivos o cantidad de agua) sería idéntico para todas las lecturas del mismo grupo en la misma fecha. Para evitar este sesgo, se calcularon estadísticos diarios que sintetizan el comportamiento de cada variable de sensor a lo largo de las 24 horas de cada fecha de muestreo. Para las tres variables de suelo (humedad del sustrato, temperatura del suelo y conductividad eléctrica) se calcularon cinco estadísticos: media, desviación estándar, mínimo, máximo y rango. Esto generó un total de 15 variables de sensor por

observación, que capturan tanto el nivel promedio como la variabilidad y la amplitud de respuesta del sustrato durante el día.

Integración de Variables Manuales y de Radiación. Las variables manuales (número de tallos productivos y cantidad de agua (L)) se incorporaron al dataset mediante una unión por fecha de muestreo y nombre del grupo. Adicionalmente, los estadísticos diarios de radiación (PPFD) calculados en la sección de selección de variable sambientales (media, desviación estándar, mínimo, máximo y rango) se integraron como variables ambientales mediante una unión por fecha, resultando en 4 variables adicionales de radiación.

La variable inicio de producción, presente en el dataset manual, fue excluida del dataset de modelado por dos razones: su tipo de dato original es categórico sin una codificación numérica directa, y su contenido (indicar si en esa fecha el grupo entró o no a producción) no corresponde a una variable predictora del comportamiento del sustrato sino a un evento discreto del cultivo, cuyo análisis excede el alcance del presente estudio.

La variable radiacion_min fue excluida del modelo dado que su valor es sistemáticamente cero en todos los grupos y fechas, correspondiendo a las horas nocturnas de cada día de muestreo, por lo que no aporta información discriminativa.

Ajuste de Tipos de Datos y Estructura Final. Una vez integradas todas las fuentes, se realizaron ajustes de tipos de datos para garantizar la compatibilidad con los algoritmos de machine learning: la columna Capacidad de bolsa se convirtió a tipo texto para tratarla como variable categórica, y la columna Cantidad de agua (L) se procesó para corregir el uso de coma como separador decimal, convirtiéndola a tipo numérico. La columna datetime se reordenó al inicio del dataset para facilitar su identificación, y la variable Posicion en el grupo se estandarizó como tipo texto.

El dataset final DATASET_ANALISIS contiene 102 observaciones y 29 variables, con cero valores nulos en todas las columnas. Su estructura completa se presenta en la Tabla 8 y su proceso de construcción se detalla en el Apéndice D: Pseudocódigo de construcción del dataset para modelado.

Tabla 8

Estructura del DATASET_ANALISIS

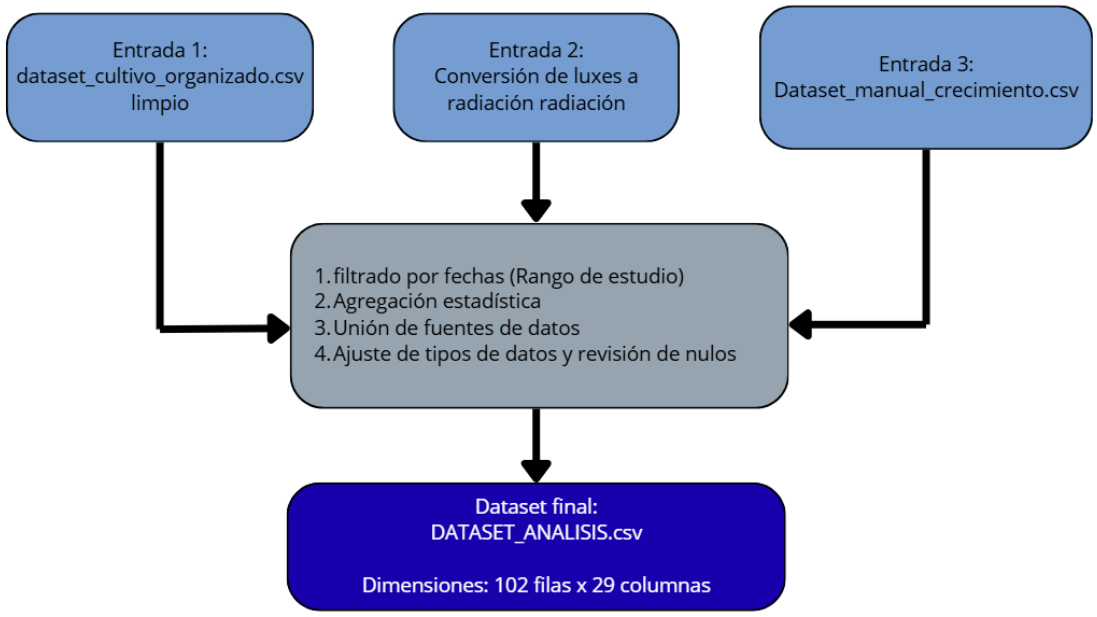
Grupo	Variables	Tipo	Origen
Identificación	datetime, nombre_trabajo	Fecha / texto	Sensor
Caracterización	Sustrato, Capacidad de bolsa, Posicion en el grupo, Tipo de riego	Texto	Configuración
Sensor suelo — humedad	moisture_mean, moisture_std, moisture_min, moisture_max, moisture_range	Numérico	Sensor (agregado)
Sensor suelo — temperatura	T_Suelo_mean, T_Suelo_std, T_Suelo_min, T_Suelo_max, T_Suelo_range	Numérico	Sensor (agregado)
Sensor suelo — CE	ec_mean, ec_std, ec_min, ec_max, ec_range	Numérico	Sensor (agregado)
Ambiente	radiacion_mean, radiacion_std, radiacion_max, radiacion_range	Numérico	Sensor ambiente (agregado)
Manual	numero de tallos productivos, Cantidad de agua (L)	Numérico	Medición manual

Nota. Estructura completa del DATASET_ANALISIS con las 29 variables que lo componen, organizadas por grupo funcional, con su tipo de dato y la fuente de origen de cada una.

La Figura 9 ilustra el flujo completo de construcción del dataset, desde las fuentes de entrada hasta el archivo final de 102 x 29.

Figura 9

Diagrama de Flujo de la Construcción del DATASET_ANALISIS.



Nota. Diagrama de flujo de la construcción del DATASET_ANALISIS, con tres nodos de entrada en la parte superior, dataset de sensores limpio, datos manuales y estadísticos de radiación, las operaciones aplicadas en cada etapa y el dataset resultante de 102 × 29 como salida.

El DATASET_ANALISIS en su forma final contiene 102 observaciones y 29 columnas. De estas, dos corresponden a identificadores del grupo (datetime y nombre_trabajo), una fue excluida del análisis por carecer de valor predictivo (inicio de producción), y una variable de radiación fue descartada por presentar valores sistemáticamente nulos (radiacion_min). Las 25 columnas restantes constituyen el conjunto de variables disponibles para el modelado, de las cuales cada modelo utiliza 24 como variables predictoras, según la configuración descrita en la sección de preparación final para el modelado, dado que la variable objetivo de cada modelo no actúa simultáneamente como predictora.

Preparación Final para el Modelado

A partir del DATASET_ANALISIS construido en la sección anterior, se realizó la preparación específica para el entrenamiento de los modelos de machine learning. Esta etapa comprende la definición de las configuraciones de entrada y salida para cada modelo, el tratamiento de variables categóricas y la definición de la estrategia de validación.

Configuración de los dos Modelos. El dataset unificado de 102 x 24 sirve como base para dos configuraciones de modelado independientes, que comparten las mismas variables de entrada pero difieren en su variable objetivo:

El Modelo A de clasificación tiene como objetivo determinar si el modelo puede identificar el tipo de sustrato, bien sea coco o cascarilla de arroz a partir de las variables registradas por los sensores y las mediciones manuales. La variable objetivo es Sustrato, codificada de forma binaria (Coco = 1, Cascarilla = 0). Las variables de entrada incluyen todas las variables de sensor, las de radiación, la cantidad de agua y el número de tallos productivos, además de las variables categóricas de caracterización del grupo.

El Modelo B de regresión tiene como objetivo predecir el número de tallos productivos a partir de las condiciones del sustrato registradas por los sensores, la radiación y la cantidad de agua recibida. La variable objetivo es numero de tallos productivos como valor numérico continuo. En este modelo, la variable Sustrato pasa a ser una variable de entrada, codificada numéricamente, dado que el tipo de sustrato es una condición del sistema que puede influir en la respuesta productiva. La Tabla 9 resume la configuración de variables y variable objetivo para cada modelo.

Tabla 9

Configuración de Variables y Variable Objetivo por Modelo

Variable	Modelo A (Clasificación)	Modelo B (Regresión)
Capacidad de bolsa	☑	☑
Posicion en el grupo	☑	☑
Tipo de riego	☑	☑
moisture_mean/std/min/max/range	☑	☑
T_Suelo_mean/std/min/max/range	☑	☑
ec_mean/std/min/max/range	☑	☑
radiacion_mean/std/max/range	☑	☑
Cantidad de agua (L)	☑	☑
numero de tallos productivos	☑ variable	objetivo
Sustrato	objetivo	☑ variable (codificada)
Total variables	24	24

Nota. Configuración de las variables predictoras y la variable objetivo para cada uno de los dos modelos de machine learning desarrollados, indicando qué variables actúan como entrada y cuál como salida en el Modelo A y en el Modelo B.

Codificación de Variables Categóricas. Las variables categóricas presentes en el dataset que son: Sustrato, Capacidad de bolsa, Posicion en el grupo y Tipo de riego, no pueden ser procesadas directamente por los algoritmos de machine learning en su forma textual. Para su

inclusión en los modelos se aplicó codificación mediante LabelEncoder, que asigna un valor numérico entero a cada categoría. Esta técnica es apropiada para Random Forest dado que el algoritmo no asume relaciones de orden entre los valores codificados, sino que los trata como identificadores discretos dentro de los splits de cada árbol.

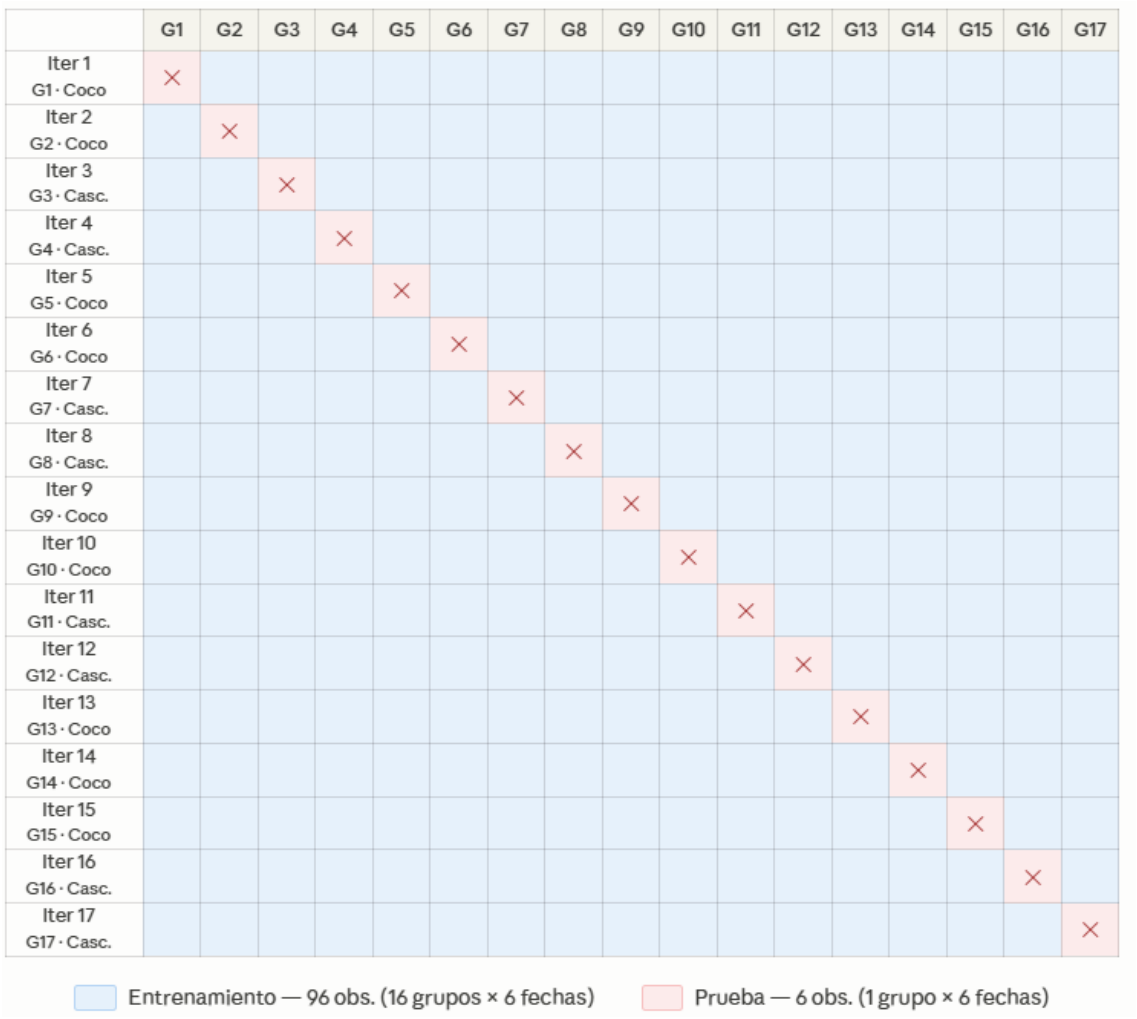
Estrategia de Validación Leave-One-Group-Out (LOGO). La elección de la estrategia de validación es una decisión metodológica crítica en este estudio, determinada por la estructura del diseño experimental. El dataset contiene 17 grupos observados cada uno en 6 fechas distintas, lo que genera una dependencia temporal entre las filas de un mismo grupo. Una división aleatoria de filas en conjuntos de entrenamiento y prueba permitiría que observaciones del mismo grupo aparecieran simultáneamente en ambos conjuntos, lo que provocaría una fuga de información: el modelo podría memorizar patrones específicos de ese grupo durante el entrenamiento y evaluarse sobre datos que ya conoce parcialmente, sobreestimando su capacidad de generalización (Roberts et al., 2017).

Para evitar este problema se implementó la validación Leave-One-Group-Out (LOGO). En cada iteración del proceso de validación, las seis observaciones de un grupo completo se excluyen del entrenamiento y se utilizan exclusivamente para la evaluación. Este proceso se repite 17 veces (una por grupo) hasta que cada grupo haya actuado una vez como conjunto de prueba. El modelo resultante es evaluado sobre grupos que nunca ha visto durante el entrenamiento, lo que simula de manera realista su aplicación a nuevas plantas o nuevos sustratos no contemplados en el experimento original (Hastie et al., 2009).

La Figura 10 ilustra esquemáticamente el funcionamiento de la validación LOGO aplicada al diseño del presente estudio.

Figura 10

Esquema de la Validación Leave-One-Group-Out (LOGO) Aplicada al Dataset de 17 Grupos por 6 Fechas



Nota. Esquema de la validación Leave-One-Group-Out (LOGO) aplicada al dataset de 17 grupos por 6 fechas, donde cada fila representa una iteración: el grupo excluido como conjunto de prueba se indica en rojo y los 16 grupos de entrenamiento en azul.

El esquema de la Figura 10. muestra como cada fila es una iteración donde el grupo marcado con x en rojo es excluido del entrenamiento y usado solo para prueba, mientras los 16

grupos restantes en azul forman el conjunto de entrenamiento. Las etiquetas de cada fila incluyen el número de iteración, el nombre del grupo excluido y su tipo de sustrato.

Métricas de Evaluación. Para el Modelo A de clasificación se utilizaron tres métricas complementarias: exactitud, F1-score ponderado y área bajo la curva ROC (ROC-AUC). La exactitud indica qué porcentaje de las predicciones del modelo son correctas, por su parte, el F1-score combina la precisión y la exhaustividad, teniendo en cuenta si las clases están balanceadas o no. Finalmente, el ROC-AUC mide qué tan bien el modelo puede diferenciar entre clases, sin depender de un umbral específico de decisión. Para el Modelo B de regresión se utilizaron el coeficiente de determinación R cuadrado y el error absoluto medio (MAE), expresado en número de tallos. El R cuadrado indica la proporción de varianza de la variable objetivo explicada por el modelo, mientras que el MAE cuantifica el error promedio de predicción en las unidades originales de la variable.

La selección de R^2 y MAE para el Modelo B responde a la naturaleza continua de la variable objetivo. El R^2 permite evaluar si el modelo captura patrones reales en los datos o simplemente predice la media, mientras que el MAE cuantifica el error en las unidades originales de la variable (número de tallos) facilitando la interpretación práctica del desempeño. Se descartó el error cuadrático medio (MSE) por su alta sensibilidad a valores extremos, condición especialmente problemática dado el comportamiento atípico del Grupo 10 identificado en el análisis. El error porcentual medio (MAPE) fue igualmente descartado por ser matemáticamente indefinido cuando el valor real es cero, condición presente en la primera fecha de muestreo donde ningún grupo había iniciado producción.

El procedimiento completo de preparación, configuración y validación de ambos modelos se describe en el Apéndice E: Pseudocódigo de preparación para el modelado.

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos constituye la fase previa al modelado y tiene como fin, entender la estructura, la distribución y el comportamiento de las variables del dataset DATASET_ANALISIS.csv antes de aplicar los modelos de machine learning. En el contexto del presente estudio, el EDA cumple una función doble: por un lado, permite verificar la calidad e integridad del dataset consolidado; por otro, genera evidencia descriptiva sobre el comportamiento diferencial de los sustratos de coco y cascarilla de arroz, que sustenta y complementa los resultados de los modelos desarrollados en la sección de desarrollo del modelo para machine learning.

Todo el análisis exploratorio se realizó en Google Colab utilizando las librerías pandas, matplotlib, seaborn y scipy. El procedimiento completo se describe en el Apéndice F: Pseudocódigo del análisis exploratorio de datos.

Estadística Descriptiva por Sustrato

El primer paso del EDA consistió en calcular los estadísticos descriptivos fundamentales (media, mediana, desviación estándar, mínimo, máximo y coeficiente de variación) para cada variable de sensor, desglosados por tipo de sustrato. Este análisis permite identificar diferencias en los niveles promedio y en la variabilidad de cada variable entre el sustrato coco y cascarilla de arroz, proporcionando una primera caracterización cuantitativa del comportamiento físico e hídrico de cada material.

Los estadísticos calculados para cada variable de sensor, desglosado por tipo de sustrato, se presentan en las Tablas 10 y 11 de la sección de resultados.

A diferencia de las variables ambientales, cuya selección se fundamentó en su variación entre fechas de muestreo (sección de selección de variables ambientales), las variables de suelo

se evalúan por su variación entre grupos, que es donde reside su capacidad discriminativa. En este caso, la diferencia entre coco y cascarilla (superiores al doble en humedad y 4,2 veces en conductividad eléctrica) justifican su inclusión, independientemente de su coeficiente de variación temporal.

Adicionalmente, se generaron diagramas de caja para las variables de mayor relevancia analítica, las cuales permiten visualizar de manera simultánea la distribución, la mediana, el rango intercuartílico y los valores atípicos residuales de cada variable según el tipo de sustrato.

Análisis Temporal de las Variables de Sensor

El comportamiento de las variables de suelo a lo largo de las seis fechas de muestreo permite identificar tendencias temporales y evaluar si los sustratos presentan patrones de evolución diferenciados a medida que avanza el crecimiento de cultivo. Este análisis es especialmente relevante para la conductividad eléctrica, cuyo comportamiento refleja la acumulación o lixiviación de sales en el sustrato con el paso del tiempo, y para la humedad, cuya evolución indica posibles cambios en la capacidad de retención hídrica del material (Savvas & Gruda, 2018). La Figura 12 presenta estas trayectorias para las seis fechas de muestreo.

Complementariamente, se analizó la variabilidad diaria de las variables de sensor utilizando los estadísticos de rango y desviación estándar calculados en la etapa de agregación. Estos estadísticos capturan la amplitud de oscilación de cada variable durante las 24 horas de cada fecha de muestreo y permiten evaluar si uno de los sustratos presenta un comportamiento más estable o más reactivo ante los eventos de riego y fertirrigación.

Análisis por Capacidad de Bolsa y Tipo de Riego

Dado que el diseño experimental incluye dos capacidades de bolsa (30 L y 60 L) y cuatro modalidades de riego, se evaluó si estas variables de diseño introducen diferencias sistemáticas

en las variables de sensor que pudieran confundirse con el efecto del tipo de sustrato. Este análisis es clave para interpretar bien los resultados de los modelos, si factores como la capacidad de bolsa o el tipo de riego explican una parte importante de la variabilidad de los datos de los sensores, entonces deben ser considerados como covariables en el análisis y no tratarse solo como variables adicionales sin mayor peso. Los resultados de este análisis se presentan en las Figuras 14 y 15.

Análisis de Correlación entre Variables

Para identificar las relaciones directas entre las variables del dataset y evaluar cuáles presentan mayor asociación con las variables objetivo — tipo de sustrato y número de tallos productivos —, se calculó la matriz de correlación de Pearson sobre el conjunto completo de variables numéricas. Este análisis permite orientar la interpretación de los resultados de importancia de variables generados por los modelos de Random Forest y anticipar qué variables tienen mayor poder discriminativo antes del modelado.

Complementariamente, se calcularon las correlaciones individuales de cada variable con la variable objetivo del Modelo A — Sustrato codificado numéricamente — ordenadas por valor absoluto descendente. Este análisis univariado permite identificar las variables con mayor capacidad de separación entre sustratos de forma independiente al modelo.

Los resultados de ambos análisis se presentan en las Figuras 16 y 17 de la sección de resultados.

Análisis Comparativo de la Variable Manual: Tallos Productivos

El número de tallos productivos es la variable manual de mayor relevancia analítica, ya que constituye la variable objetivo del Modelo B y un indicador de la respuesta del cultivo ante las condiciones del sustrato. Su análisis exploratorio permite evaluar la distribución de esta

variable, identificar diferencias entre sustratos y verificar que su variabilidad es suficiente para que el modelo de regresión pueda aprender patrones significativos.

Se analizó su distribución por tipo de sustrato, por capacidad de bolsa y por fecha de muestreo. Particular atención se prestó a los valores de cero tallos productivos registrados en la primera fecha de muestreo (19 de enero de 2026), que son biológicamente válidos y coherentes con la etapa fenológica del cultivo: con aproximadamente 80 días desde la siembra, las plantas se encontraban aún en fase de establecimiento radicular y crecimiento vegetativo, sin haber iniciado la emisión de tallos productivos. La distribución y evolución temporal de los tallos productivos obtenidos se presentan en las Figuras 18 y 19 de la sección de resultados

Desarrollo del Modelo de Machine Learning

Con base en el DATASET_ANALISIS preparado en la Sección de preparación final para el modelado, se desarrollaron dos modelos de machine learning complementarios utilizando el algoritmo Random Forest: un modelo de clasificación orientado a distinguir el tipo de sustrato a partir de las variables monitoreadas, y un modelo de regresión orientado a predecir el número de tallos productivos como indicador de la respuesta del cultivo ante las condiciones del sustrato. Ambos modelos fueron implementados en Google Colab utilizando la librería scikit-learn. El procedimiento completo se describe en el Apéndice G: Pseudocódigo del desarrollo de los modelos de machine learning.

Selección y Justificación del Algoritmo

El algoritmo Random Forest fue seleccionado como método principal por su adecuación a las características específicas del dataset y del problema de investigación. Breiman (2001), quien propuso formalmente el algoritmo, explica que Random Forest logra reducir el sobreajuste propio de los árboles de decisión individuales al combinar múltiples árboles entrenados sobre

submuestras aleatorias de los datos y subconjuntos aleatorios de variables, lo que permite obtener resultados mas estables y generalizables a nuevos datos.

En el contexto del presente estudio, su elección se justifica por cuatro razones específicas. Primera, su robustez ante datasets de tamaño moderado: con 102 observaciones y 24 variables predictoras por modelo, Random Forest produce estimaciones estables sin requerir grandes volúmenes de datos, a diferencia de algoritmos como redes neuronales que demandan mayor cantidad de registros para un entrenamiento confiable (Fernández-Delgado et al., 2014). Segunda, su capacidad para modelar relaciones no lineales entre variables de sensor y tipo de sustrato, que no son capturables mediante regresión logística o modelos lineales simples. Tercera, su tolerancia a variables de distinta naturaleza (numéricas y categóricas) dentro del mismo modelo. Cuarta y más relevante para la tesis, la generación de métricas de importancia de variables que permiten identificar cuáles propiedades del sensor caracterizan mejor el comportamiento diferencial de los sustratos, respondiendo directamente al objetivo general del estudio.

Complementariamente, se entrenó un árbol de decisión individual de profundidad controlada con el propósito exclusivo de generar una representación visual interpretable de las reglas de clasificación. El análisis de interpretabilidad de ambos modelos, incluyendo la importancia relativa de cada variable en la clasificación del sustrato y en la predicción de tallos productivos, se presenta en la sección de comparación de importancia de variables entre modelos en los resultados

Modelo A Clasificación del Tipo de Sustrato

El Modelo A tiene como objetivo determinar si es posible identificar el tipo de sustrato (coco o cascarilla de arroz) a partir exclusivamente de las variables registradas por los sensores y

las mediciones manuales, sin necesidad de una etiqueta externa. Una exactitud alta en este modelo constituye evidencia de que los dos sustratos presentan comportamientos físicos e hídricos suficientemente diferenciados como para ser distinguidos automáticamente, lo que representa en sí mismo un resultado de caracterización.

La variable objetivo Sustrato fue codificada de forma binaria (Coco = 1, Cascarilla = 0). El modelo fue configurado con los hiperparámetros que se presentan en la Tabla 10, seleccionados para controlar el sobreajuste en un dataset de tamaño reducido. Los hiperparámetros del modelo fueron definidos mediante una configuración razonada basada en las características del dataset y en los criterios documentados en la literatura para conjuntos de datos de tamaño reducido (Hastie et al., 2009; Breiman, 2001). Con el fin de verificar que los resultados no dependen de la configuración específica elegida, se realizó un análisis de sensibilidad evaluando cinco combinaciones de hiperparámetros mediante validación LOGO: profundidad máxima de 3, 5 y 7 niveles, sin límite de profundidad, y variaciones en el número mínimo de muestras por hoja. Todas las configuraciones evaluadas alcanzaron una exactitud de 1,000, confirmando que la separabilidad entre sustratos es robusta e independiente de los hiperparámetros seleccionados. Este resultado respalda que la distinción entre coco y cascarilla de arroz está determinada por las propiedades físicas de los sustratos capturadas por los sensores, y no por el ajuste fino del algoritmo. Los valores seleccionados para el modelo final se presentan en la Tabla 10 junto con su justificación individual.

Tabla 10

Hiperparámetros del Random Forest para el Modelo A (clasificación)

Hiperparámetro	Valor	Justificación
n_estimators	200	Número de árboles suficiente para estabilizar las estimaciones
max_depth	5	Limita la profundidad para evitar sobreajuste con pocas observaciones
min_samples_leaf	3	Exige mínimo 3 muestras por hoja para evitar nodos muy específicos
class_weight	balanced	Compensa el leve desbalance entre 9 grupos coco y 8 cascarilla
random_state	42	Garantiza reproducibilidad de los resultados

Nota. Hiperparámetros seleccionados para el Random Forest del Modelo A de clasificación, con el valor asignado a cada parámetro y la justificación metodológica de su elección.

Con esta configuración se entrenó el modelo sobre el dataset completo y se evaluó mediante validación LOGO, las métricas de evaluación utilizadas fueron exactitud, F1-score ponderado y ROC-AUC, los resultados de clasificación, las métricas por iteración y la importancia de variables se presentan en la sección de resultados.

Complementariamente, se entrenó un árbol de decisión individual con profundidad máxima de 4 niveles (valor menor al del Random Forest principal), con el propósito exclusivo de generar una representación visual interpretable de las reglas de clasificación aprendidas, su representación gráfica se presenta en la Figura 21 de la sección de resultados.

Modelo B Regresión de Tallos Productivos

El Modelo B tiene como objetivo predecir el número de tallos productivos a partir de las condiciones del sustrato registradas por los sensores, la radiación y la cantidad de agua recibida. En este modelo, el tipo de sustrato se incorpora como variable de entrada codificada numéricamente, dado que es una condición del sistema con potencial influencia sobre la respuesta productiva del cultivo.

La variable objetivo numero de tallos productivos presenta un coeficiente de variación global del 94,7 %, lo que indica una alta dispersión en los datos y, por tanto, una señal suficiente para que el modelo de regresión pueda aprender patrones asociados a las condiciones del sustrato. Los hiperparámetros del modelo se presentan en la Tabla 11.

Tabla 11

Hiperparámetros del Random Forest para el Modelo B (regresión)

Hiperparámetro	Valor	Justificación
n_estimators	200	Consistente con Modelo A para comparabilidad
max_depth	5	Control de sobreajuste en dataset pequeño
min_samples_leaf	3	Garantiza generalización mínima por nodo
random_state	42	Reproducibilidad de resultados

Nota. Hiperparámetros seleccionados para el Random Forest del Modelo B de regresión, con el valor asignado a cada parámetro y la justificación de su elección en coherencia con el Modelo A.

Las métricas de evaluación utilizadas fueron el coeficiente de determinación R^2 y el error absoluto medio (MAE) expresado en número de tallos, calculadas mediante validación LOGO sobre las 17 iteraciones. La selección de R^2 y MAE para el Modelo B responde a la naturaleza

continua de la variable objetivo. El R^2 permite evaluar si el modelo captura patrones reales en los datos o simplemente predice la media, mientras que el MAE cuantifica el error en las unidades originales de la variable (número de tallos) facilitando la interpretación práctica del desempeño. Se descartó el error cuadrático medio (MSE) por su sensibilidad a valores extremos, y el error porcentual medio (MAPE) por ser indefinido cuando el valor real es cero, condición presente en la primera fecha de muestreo donde ningún grupo había iniciado producción.

Los resultados de predicción del Modelo B, incluyendo la importancia de variables y la dispersión entre valores observados y predichos, se presentan en las Figuras 22 y 23 de la sección de resultados.

Consideraciones Éticas y Limitaciones Metodológicas

Contexto Institucional y Propiedad de los Datos

El presente trabajo de grado se desarrolló en articulación con el Centro de Investigación y Desarrollo en Tecnologías de la Información y las Comunicaciones(CINTEL), entidad sin ánimo de lucro dedicada al fortalecimiento de capacidades tecnológicas y a la promoción de la transformación digital en diferentes sectores de la economía colombiana. El autor hace parte del equipo del centro de competencias Agrotech de CINTEL en calidad de contratista, desde el cual se ejecutan proyectos tecnológicos orientados al sector agropecuario.

La infraestructura experimental sobre la que se fundamenta esta investigación (el cultivo de arándano variedad Emerald bajo invernadero en la Sabana de Bogotá, la red de sensores IoT, el sistema de almacenamiento y procesamiento de datos, y la totalidad de los datos sensoriales y manuales recolectados) son propiedad de CINTEL y forman parte de un proyecto de investigación interno en curso. El uso de estos recursos para el desarrollo del presente trabajo de grado fue autorizado por la institución, bajo el entendimiento de que el análisis se enfocaría exclusivamente en la caracterización analítica de los sustratos de coco y cascarilla de arroz, sin comprometer los objetivos estratégicos ni la confidencialidad de los resultados operativos del proyecto institucional.

En coherencia con los principios éticos y de confidencialidad establecidos en el Código de Ética de CINTEL, el tratamiento de la información se realizó exclusivamente con fines académicos. No se divulgan en este documento datos crudos de configuración técnica de la red, registros operativos específicos ni resultados que puedan comprometer los procesos internos del proyecto de investigación. Los análisis desarrollados se orientan exclusivamente a la generación de conocimiento técnico sobre el comportamiento físico e hídrico de los sustratos, sin incluir

información sobre la operación del sistema productivo ni datos que permitan identificar configuraciones técnicas propietarias del proyecto institucional.

Alcance y Delimitación del Estudio

Una delimitación central del presente trabajo es su enfoque en el sustrato como objeto de análisis, en lugar del cultivo o la planta. Aunque el experimento se realiza sobre un cultivo de arándano variedad Emerald (especie de alto valor comercial en la producción hortícola colombiana), el análisis de su desarrollo biológico, sus índices de producción, su comportamiento fenológico o su respuesta agronómica a las condiciones de cultivo no forma parte del alcance de esta investigación. Esta delimitación responde a dos razones: primero, el estudio del comportamiento productivo del arándano bajo las condiciones específicas del proyecto es objeto de análisis interno de CINTEL; segundo, el objetivo del presente trabajo es la caracterización del sustrato como medio de cultivo, para lo cual la variable de tallos productivos se utiliza únicamente como indicador de la respuesta del sistema, no como variable de producción agrícola en sí misma.

Es importante señalar que el período de monitoreo corresponde a una etapa específica del ciclo del cultivo (fase de establecimiento y crecimiento vegetativo inicial) y no abarca el ciclo productivo completo desde la siembra. Los resultados obtenidos, incluyendo los valores de conductividad eléctrica, humedad y producción de tallos, son representativos de esta etapa particular y podrían presentar variaciones en fases posteriores del cultivo cuando la demanda hídrica y nutricional de la planta sea diferente.

En consecuencia, las recomendaciones técnicas generados en este trabajo se dirigen a orientar decisiones sobre selección y manejo de sustratos, y no deben interpretarse como recomendaciones sobre el manejo agronómico del cultivo de arándano.

Limitaciones Metodológicas

Este estudio tiene cuatro limitaciones metodológicas que es importante tener en cuenta al momento de interpretar los resultados y analizar qué tan aplicables pueden ser en otros contextos.

La primera es el tamaño del dataset de modelado. Con 102 observaciones generadas a partir de 17 grupos y 6 fechas de muestreo, el dataset es de tamaño reducido para los estándares del aprendizaje automático. Esta limitación es inherente al diseño experimental de campo, en el que las mediciones manuales (que determinan el número de fechas de muestreo) no pueden realizarse con mayor frecuencia sin interferir con el desarrollo normal del cultivo. La validación Leave-One-Group-Out fue seleccionada precisamente para maximizar el uso de los datos disponibles y obtener estimaciones de desempeño tan confiables como sea posible bajo estas condiciones.

La segunda es la especificidad geográfica y de especie. Los resultados fueron obtenidos bajo condiciones particulares de altitud, temperatura, radiación y manejo hídrico propias de la Sabana de Bogotá y de un cultivo de arándano variedad Emerald. Su transferibilidad directa a otros contextos geográficos, otras especies o variedades, o diferentes sistemas de riego no está garantizada y requeriría validación experimental adicional.

La tercera es la heterogeneidad de las modalidades de riego. Los cuatro tipos de riego implementados en el experimento (tres volúmenes fijos con distinta cantidad de agua diaria y cuatro automáticos por histéresis) introducen una fuente de variación adicional que no siempre puede separarse completamente del efecto del tipo de sustrato en el análisis. Aunque el tipo de riego fue incluido como variable en los modelos, su interacción con el sustrato en condiciones de mayor homogeneidad de riego podría arrojar resultados diferentes a los aquí obtenidos.

La cuarta limitación es la especificidad de la etapa fenológica evaluada. El período de monitoreo comprende exclusivamente la fase de establecimiento y crecimiento vegetativo inicial del cultivo, iniciado el 31 de octubre de 2025. El comportamiento de los sustratos en fases posteriores (floración, fructificación y producción plena) puede diferir del documentado en este estudio, dado que la demanda hídrica y la dinámica de absorción de nutrientes varían significativamente a lo largo del ciclo productivo del arándano.

La posible influencia del sensor sobre la planta instrumentada fue evaluada comparando la producción de tallos entre la planta con sensor y las tres plantas restantes de cada grupo. El análisis no reveló diferencias estadísticamente significativas ($t = -0,052$, $p = 0,959$), por lo que este factor no constituye una limitación activa del estudio.

Resultados

Análisis Exploratorio de Datos

Estadística Descriptiva por Sustrato

El primer análisis consistió en comparar cómo se comportaron los dos sustratos durante el período de estudio en las variables medidas por los sensores. Los resultados muestran diferencias muy marcadas entre el coco y la cascarilla de arroz en dos de las tres variables de suelo, tal como se presenta en las Tabla 12.

Tabla 12

Estadística Descriptiva de las Variables de Sensor de Suelo Desagregada por Tipo de Sustrato

Variable	Sustrato	Media	Mediana	Desv. Est.	Mínimo	Máximo	CV (%)
Humedad del sustrato media (%)	Coco	45,35	45,66	4,91	35,90	54,24	10,8
	Cascarilla	19,66	19,95	2,85	14,65	24,34	14,5
Temperatura del suelo media (°C)	Coco	19,47	19,64	1,07	16,84	21,17	5,5
	Cascarilla	20,01	19,99	1,30	16,61	22,59	6,5
Conductividad eléctrica media	Coco	137,71	128,44	47,87	72,04	277,36	34,8
	Cascarilla	33,14	32,89	6,61	21,34	48,99	19,9

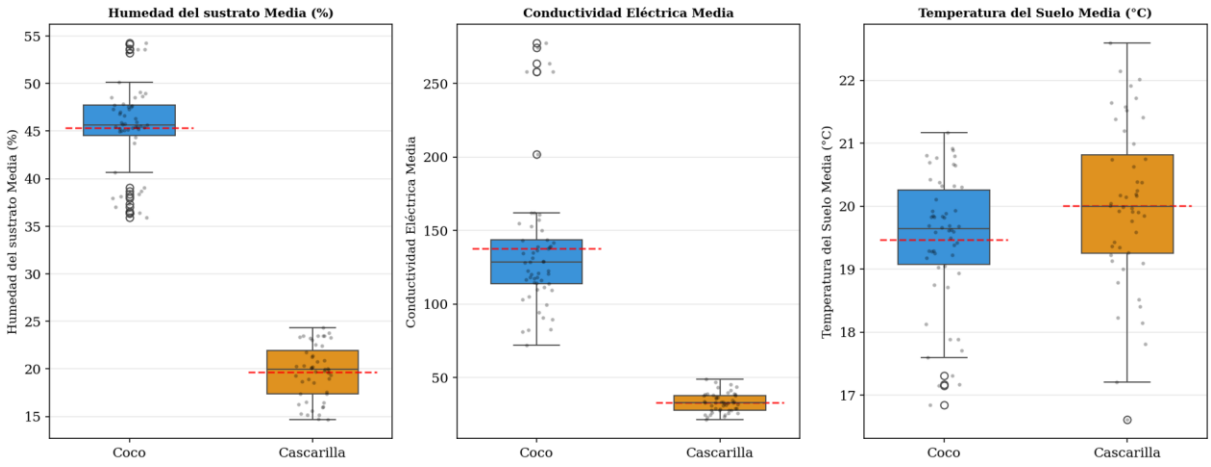
Nota. Estadísticos descriptivos, media, mediana, desviación estándar, mínimo, máximo y coeficiente de variación de las variables de sensor de suelo desagregados por tipo de sustrato.

La diferencia más evidente es en la humedad como podemos ver en la figura 11, el sustrato de coco mantuvo en promedio un nivel de humedad del 45,35 %, mientras que la cascarilla solo alcanzó el 19,66 %. Para ser mas claros, el coco retuvo más del doble de agua que la cascarilla en las mismas condiciones de riego. Esto no es un dato menor: significa que si un productor riega ambos sustratos con la misma cantidad de agua, el coco la conserva la humedad mucho más tiempo mientras que la cascarilla la pierde rápidamente.

La Figura 11 muestra visualmente esta separación, el diagrama de cajas de los dos sustratos no se tocan en ningún punto, lo que confirma que la diferencia es consistente y no hay casos intermedios.

Figura 11

Comparación de Variables de Suelo por Tipo de Sustrato



Nota. Diagramas de caja de la humedad del sustrato media, la conductividad eléctrica media y la temperatura del suelo media, comparados entre el sustrato de coco y la cascarilla de arroz, con los valores individuales de cada observación superpuestos.

La segunda diferencia importante está en la conductividad eléctrica, que mide qué tan cargada de sales y nutrientes está el agua dentro del sustrato. El coco registró en promedio un

valor de 137,71, aproximadamente cuatro veces mayor que el de la cascarilla 33,14. Esto ocurre porque el coco retiene los nutrientes del abono mucho más que la cascarilla, que los elimina más rápidamente con el agua de drenaje. Para un productor esto tiene una implicación directa: con coco hay mayor acumulación de sales entre riegos, lo que puede requerir ajustes en la fertilización para evitar salinidad excesiva o ajustes en la aplicación del riego.

La temperatura del suelo, en cambio, fue prácticamente igual en ambos sustratos, 19,47 °C en coco y 20,01 °C en cascarilla. Esta pequeña diferencia confirma que la temperatura del suelo depende principalmente del ambiente del invernadero, no del tipo de sustrato.

La radiación registró valores idénticos para ambos sustratos, lo cual era esperado ya que es una variable del ambiente del invernadero y afecta a todas las plantas por igual.

Respecto a la producción de tallos, la cascarilla generó en promedio más tallos nuevos que el coco 5,02 vs 3,43 durante el período de estudio como se puede ver en la Tabla 13. Sin embargo esta diferencia viene acompañada de alta variabilidad, algunos de los grupos de cascarilla produjeron muy pocos tallos y otros muchos, lo que indica que otros factores además del sustrato influyen en la producción.

Tabla 13

Estadística Descriptiva de la Radiación (PPFD) y Variables Manuales Desagregada por Tipo de Sustrato

Variable	Sustrato	Media	Mediana	Desv. Est.	Mínimo	Máximo	CV (%)
Radiación media (µmol m^{-2} s^{-1})	Coco	55,46	54,54	20,26	28,26	92,59	36,5
	Cascarilla	55,46	54,54	20,28	28,26	92,59	36,6

Tallos productivos	Coco	3,43	3,00	3,33	0	14	97,3
	Cascarilla	5,02	4,00	4,44	0	21	88,5
Cantidad de agua recibida (L/día)	Coco	2,21	2,70	1,06	0,10	3,40	48,0
	Cascarilla	2,67	2,70	0,99	0,66	6,25	37,2

Nota. Estadísticos descriptivos, media, mediana, desviación estándar, mínimo, máximo y coeficiente de variación de la radiación fotosintéticamente activa (PPFD), el número de tallos productivos y la cantidad de agua recibida, desagregados por tipo de sustrato.

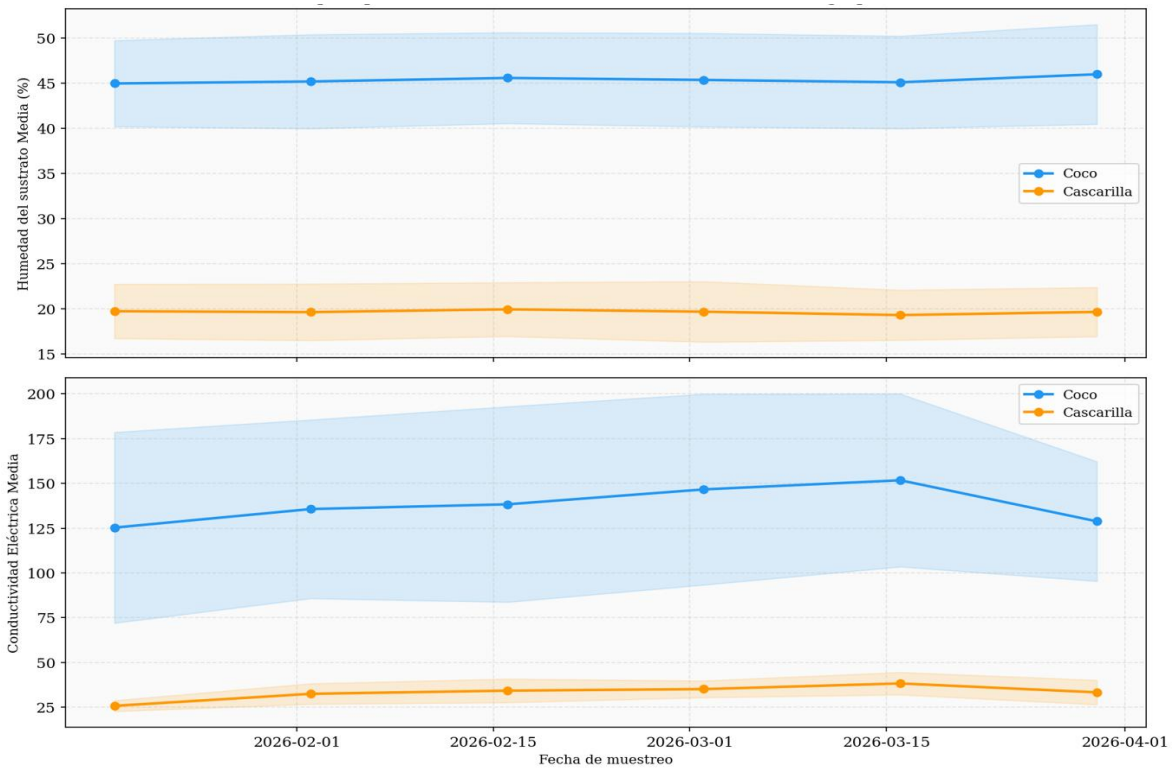
Análisis Temporal de las Variables de Sensor

Este análisis busca responder una pregunta sencilla: ¿las diferencias entre coco y cascarilla se mantienen en el tiempo o cambian a medida que avanza el cultivo?

La Figura 12 muestra la evolución de la humedad y la conductividad eléctrica a lo largo de las seis fechas de muestreo. El resultado es claro: la diferencia entre los dos sustratos se mantuvo estable durante todo el período sin que las curvas se acercaran en ningún momento. Esto es relevante porque confirma que las propiedades físicas de cada sustrato son estables y no se homogenizan con el tiempo bajo las condiciones del experimento.

Figura 12

Evolución Temporal de la Humedad del Sustrato Media y la Conductividad Eléctrica Media por Tipo de Sustrato a lo Largo de las Seis Fechas de Muestreo



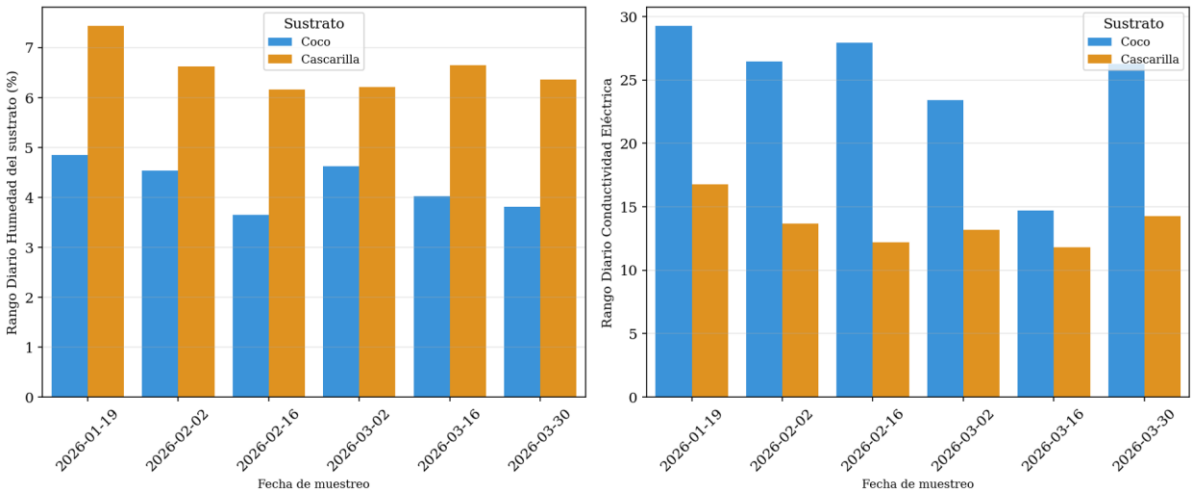
Nota. Gráfico de líneas con la evolución temporal de la humedad del sustrato media y la conductividad eléctrica media para cada tipo de sustrato a lo largo de las seis fechas de muestreo, con bandas sombreadas que representan ±1 desviación estándar entre grupos del mismo sustrato.

Para la Figura 13 se analiza algo más específico, cuánto varía la humedad de cada sustrato a lo largo de un mismo día. En ese sentido, la cascarilla mostró oscilaciones diarias más grandes que el coco en todas las fechas, esto se explica por su estructura física ya que la cascarilla absorbe el agua del riego rápidamente pero también la pierde con mayor velocidad, generando ciclos pronunciados de humedad alta justo después del riego y humedad baja antes del siguiente ciclo. Por el contrario, el coco al retener mejor el agua, mantiene niveles más estables

durante el día. En general para un productor esto significa que con cascarilla el momento del riego es más crítico si se atrasa, la planta puede experimentar estrés hídrico.

Figura 13

Variabilidad Diaria de la Humedad del Sustrato y la Conductividad Eléctrica por Tipo de Sustrato y Fecha de Muestreo



Nota. Gráfico de barras agrupadas con el rango diario, diferencia entre el valor máximo y mínimo del día de la humedad del sustrato y la conductividad eléctrica por tipo de sustrato y fecha de muestreo, como indicador de la variabilidad intradía de cada variable.

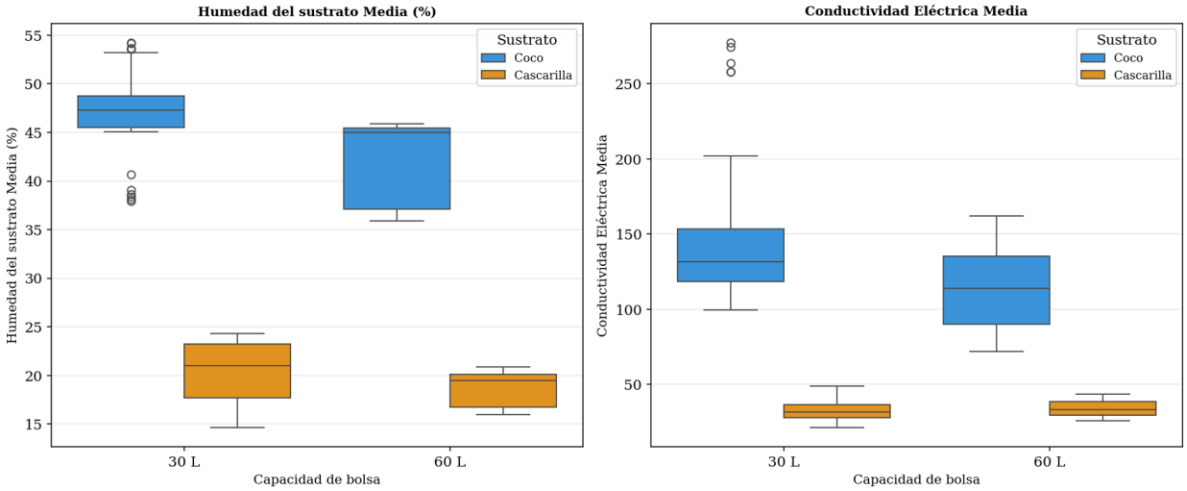
Análisis por Capacidad de Bolsa y Tipo de Riego

Ahora, se busca verificar que las diferencias observadas entre sustratos no se deban en realidad al tamaño de la bolsa o a la cantidad de agua aplicada.

Para la Figura 14 comparamos los dos tamaños de bolsa, la de 30 Litros y la de 60 Litros por cada sustrato. La separación entre coco y cascarilla se mantiene en ambos tamaños, lo que confirma que la diferencia es del sustrato y no del volumen o capacidad de la bolsa.

Figura 14

Comparación de Variables de Suelo por Capacidad de Bolsa y Tipo de Sustrato

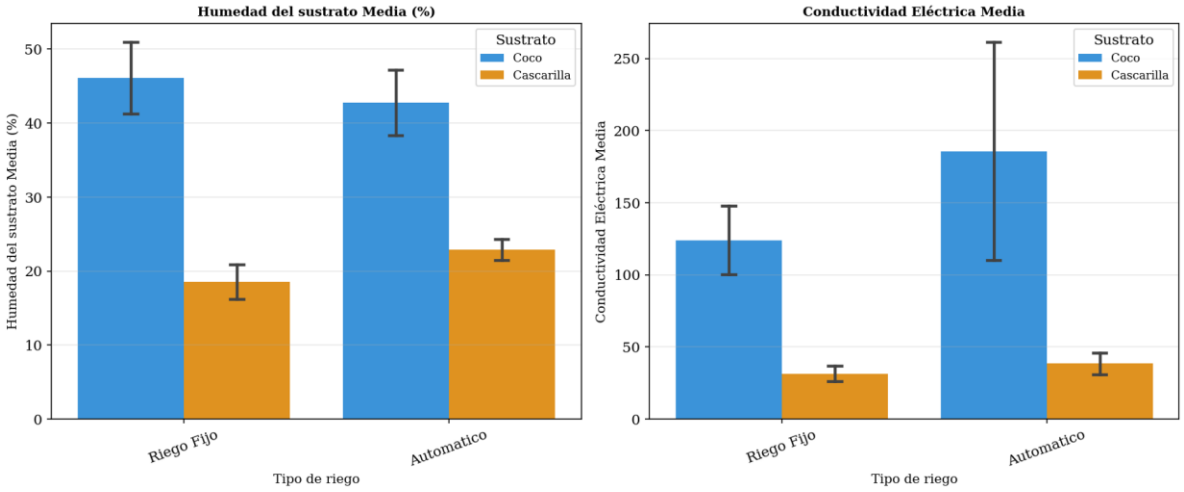


Nota. Diagramas de caja de la humedad del sustrato media y la conductividad eléctrica media desagregados por capacidad de bolsa (30 L y 60 L) y tipo de sustrato, para evaluar la interacción entre el volumen de contenedor y el comportamiento hídrico del sustrato.

La Figura 15 hace lo mismo con los tipos de riego. En todas las variaciones, el coco siempre mostró mayor humedad y mayor conductividad que la cascarilla. Esto refuerza que las diferencias son propias de cada sustrato y no dependen de cuánta agua se aplique.

Figura 15

Valores Medios de Humedad del Sustrato y Conductividad Eléctrica por Tipo de Riego y Sustrato



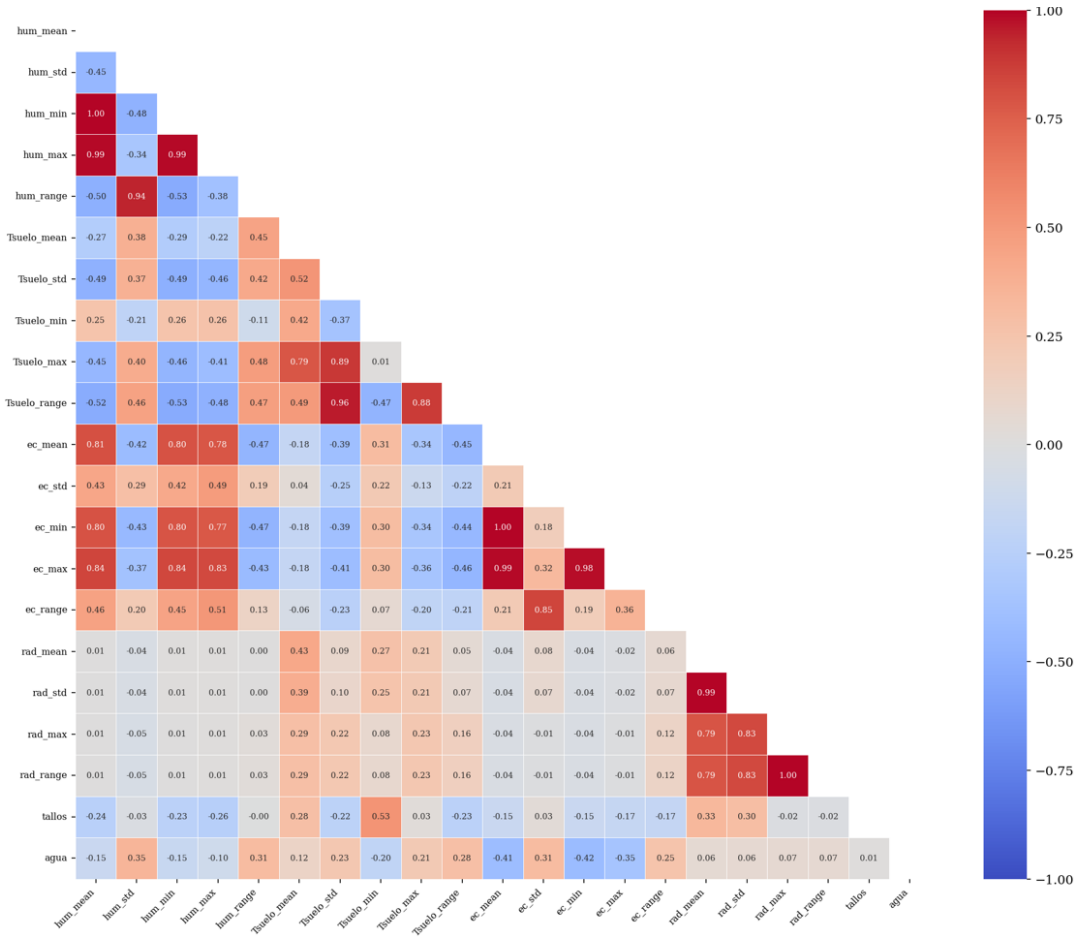
Nota. Gráfico de barras con valores medios para la humedad del sustrato y la conductividad eléctrica según el tipo de riego aplicado (Riego Bajo, Riego Estándar, Riego Alto y Automático) y el tipo de sustrato, con barras de error que representan ±1 desviación estándar.

Análisis de Correlación entre Variables

La Figura 16 muestra cómo se relacionan todas las variables entre sí. Sin abordar en detalles técnicos, el punto más importante es que las variables de humedad y conductividad eléctrica forman grupos muy cohesionados, cuando una sube las demás del mismo grupo también suben, esto indica que medir una de ellas da información sobre las demás.

Figura 16

Mapa de Calor de la Matriz de Correlación de Pearson Entre Todas las Variables Numéricas del DATASET_ANALISIS



Nota. Mapa de calor triangular de la matriz de correlación de Pearson entre todas las variables numéricas del DATASET_ANALISIS, con escala de color que va de azul (correlación negativa) a rojo (correlación positiva) y valores numéricos anotados en cada celda.

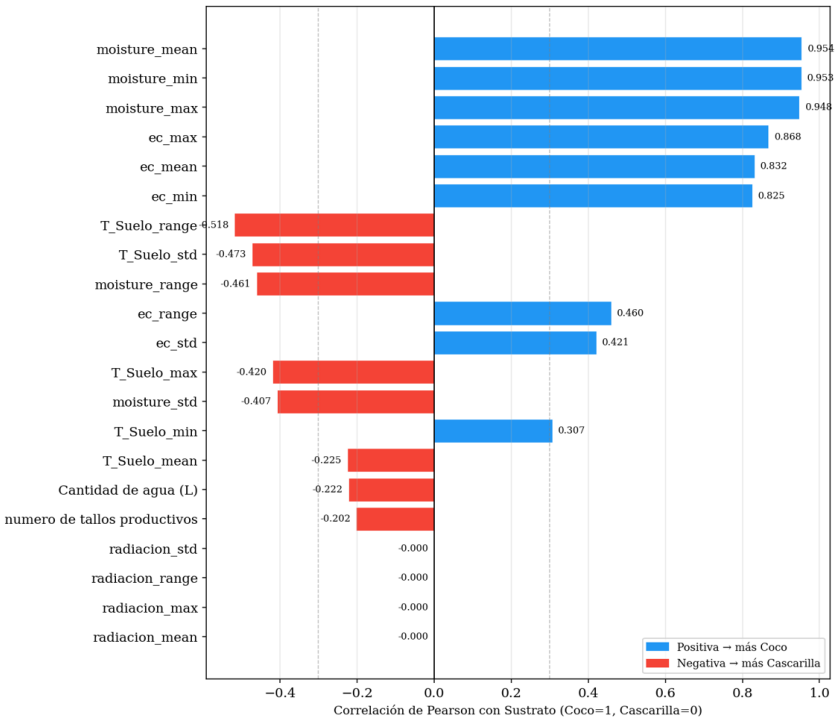
La Figura 17 responde una pregunta más directa: ¿qué variable, por sí sola, distingue mejor el coco de la cascarilla? Las tres medidas de humedad encabezan la lista con valores de correlación superiores a 0,948, son prácticamente perfectos, luego siguen las medidas de

conductividad eléctrica. Ahora, en el extremo opuesto, la radiación tiene correlación de cero con el sustrato, confirmando que no sirve para distinguir entre materiales dentro de una misma fecha.

Lo importante de esta figura es que deja ver que con una sola variable de humedad ya se distingue casi perfectamente los dos sustratos, el modelo de clasificación que se construye en la siguiente sección debería funcionar muy bien, y así fue.

Figura 17

Correlación de Cada Variable con la Variable Sustrato, Ordenadas por Valor Absoluto Descendente



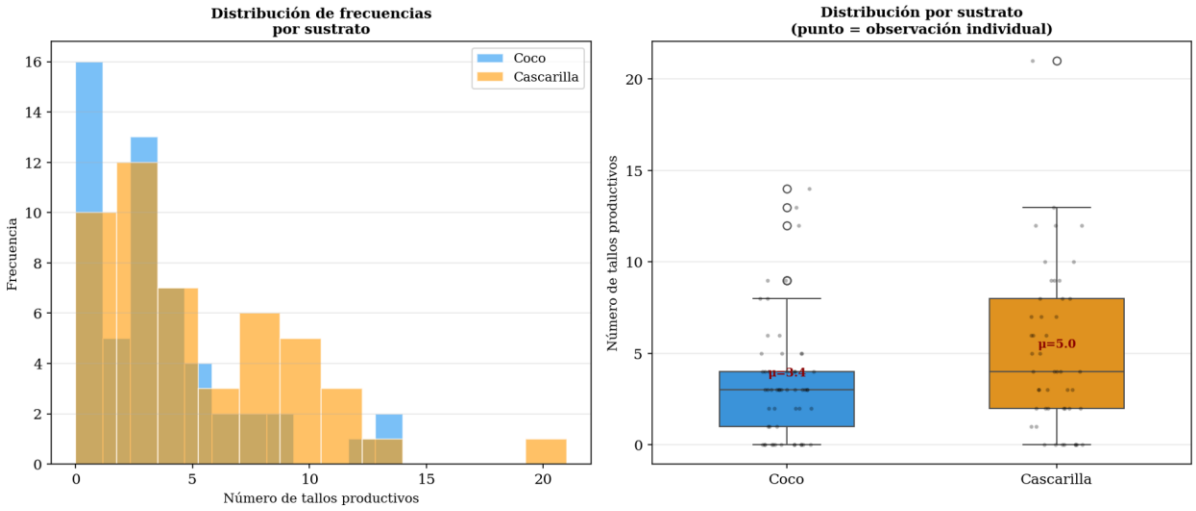
Nota. Gráfico de barras horizontales con la correlación de Pearson de cada variable con el tipo de sustrato codificado numéricamente (Coco = 1, Cascarilla = 0), ordenadas por valor absoluto descendente y diferenciadas por color según la dirección de la asociación.

Análisis de Tallos Productivos

La Figura 18 muestra cómo se distribuye la producción de tallos nuevos en cada sustrato. Ambos tienen alta variabilidad con grupos que produjeron muy pocos tallos y otros que produjeron muchos, incluso dentro del mismo tipo de sustrato. Esto indica que el sustrato no es el único factor que determina la producción: la cantidad de agua, la temperatura y la radiación también influyen.

Figura 18

Distribución del Número de Tallos Productivos por Tipo de Sustrato



Nota. Panel doble con el histograma de frecuencias del número de tallos productivos por tipo de sustrato (panel izquierdo) y los diagramas de caja correspondientes con los valores individuales superpuestos y la media indicada (panel derecho).

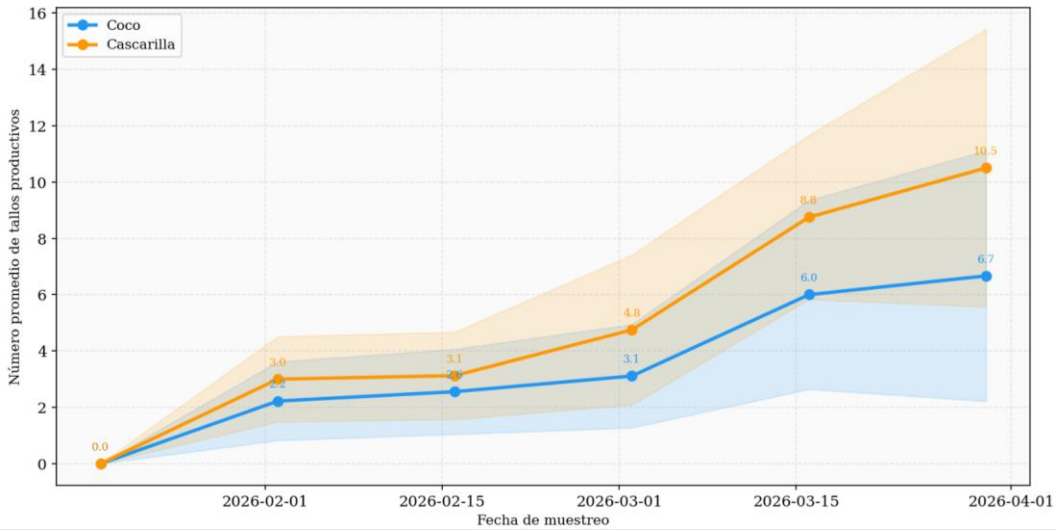
Un dato importante a resaltar es que, en la primera fecha de muestreo, el 19 de enero de 2026, todos los grupos presentaron cero tallos productivos. En ese momento las plantas llevaban aproximadamente 80 días desde su siembra y aún estaban en fase de establecimiento, es posible que sus raíces se estaban desarrollando y aún no habían iniciado la producción de tallos nuevos,

también es importante aclarar que el registro de tallos productivos al ser de forma manual, puede presentar errores humanos en el conteo.

La Figura 19 muestra cómo evolucionó la producción de tallos a lo largo del tiempo. La cascarilla superó consistentemente al coco en las cuatro últimas fechas de muestreo. Sin embargo, la alta dispersión entre grupos del mismo sustrato que se ve representada por las bandas sombreadas de la figura, indica que esta diferencia no es uniforme, para algunos grupos de coco se produjeron más tallos que algunos de cascarilla, y viceversa. Podemos decir que el sustrato si influye en la producción, pero no es el único factor determinante.

Figura 19

Evolución Temporal del Número Promedio de Tallos Productivos por Tipo de Sustrato



Nota. Gráfico de líneas con la evolución del promedio de tallos productivos por tipo de sustrato a lo largo de las seis fechas de muestreo, con bandas sombreadas de ±1 desviación estándar que ilustran la dispersión entre grupos del mismo sustrato en cada fecha.

Análisis Comparativo de la Planta con Sensor vs Planta sin Sensor

Una pregunta válida antes de confiar en los resultados es si el sensor mismo por su posición, dimensión y/o material afecta a la planta donde está instalado. Si la presencia del dispositivo dentro del sustrato alterara las condiciones de humedad o nutrientes alrededor de esa planta, los datos registrados no serían representativos y las comparaciones con las demás plantas del grupo no serían válidas.

Para responder esto se comparó la producción de tallos de la planta con sensor frente al promedio de las otras tres plantas del mismo grupo, en los 17 grupos y las seis fechas de muestreo. El resultado es contundente: prácticamente no hay diferencia. La planta con sensor produjo en promedio 4,12 tallos y las plantas sin sensor produjeron 4,14 tallos, una diferencia de apenas 0,02 tallos que estadísticamente es indistinguible del azar. Al separar el análisis por tipo de sustrato el resultado se mantiene igual, ni en coco ni en cascarilla la presencia del sensor marcó una diferencia real en la producción de la planta.

En términos simples, el sensor no afecta a la planta y puede instalarse dentro del sustrato sin alterar el comportamiento productivo del cultivo, lo que respalda tanto la validez del monitoreo como la confiabilidad de los datos utilizados en todo el análisis.

Resultados de los Modelos de Machine Learning

Modelo A Clasificación del Tipo de Sustrato

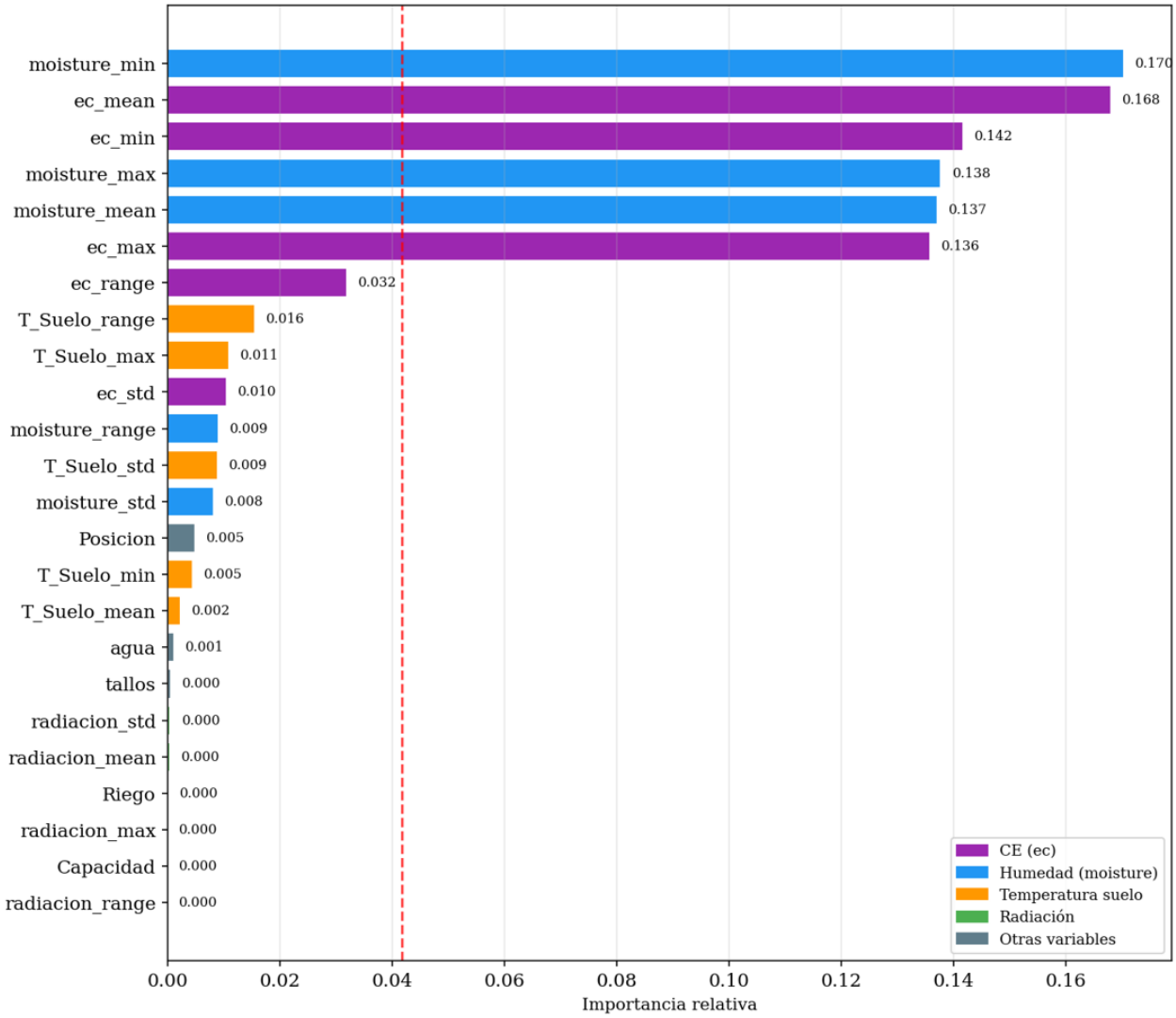
El Modelo A clasificó correctamente el tipo de sustrato en la totalidad de los 17 grupos evaluados, con exactitud de 1,000 y F1-score de 1,000 para cada iteración de la validación LOGO. Esto significa que el modelo identificó sin error si un sensor estaba en un sustrato coco o cascarilla, incluso sin haber visto ese grupo durante el entrenamiento. Los números indican que no hubo un solo grupo para el cual el modelo se equivocara.

La exactitud de 1,000 nos dice que todas las predicciones fueron correctas. El F1-score de 1,000 confirma que este resultado no se debe a un desbalance y que el modelo acierta muy bien de igual forma con coco que con cascarilla. La métrica ROC-AUC no se pudo calcular porque en la validación LOGO cada grupo pertenece a un único sustrato, por lo que el conjunto de prueba contiene una sola clase para cada iteración, esta condición hace que la métrica matemáticamente sea indefinida, sin embargo esto no afecta la interpretación del desempeño.

La Figura 20 muestra qué variables usó el modelo para tomar esas decisiones. Las seis más importantes concentran el 89,1 % de la información utilizada y son exclusivamente medidas de humedad del suelo y conductividad eléctrica, tenemos humedad minima de 0,170,EC promedio de 0,168, EC mínima de 0,142, humedad máxima de 0,138, humedad promedio de 0,137 y EC máxima de 0,136. Variables como la radiación, tipo de riego, capacidad de bolsa y tallos productivos obtuvieron una importancia de 0, lo que nos indica, que una vez conocidas la humedad y la CE, ninguna otra variable aporta información adicional para distinguir los sustratos.

Figura 20

Importancia de Variables del Modelo A (Random Forest Clasificación)



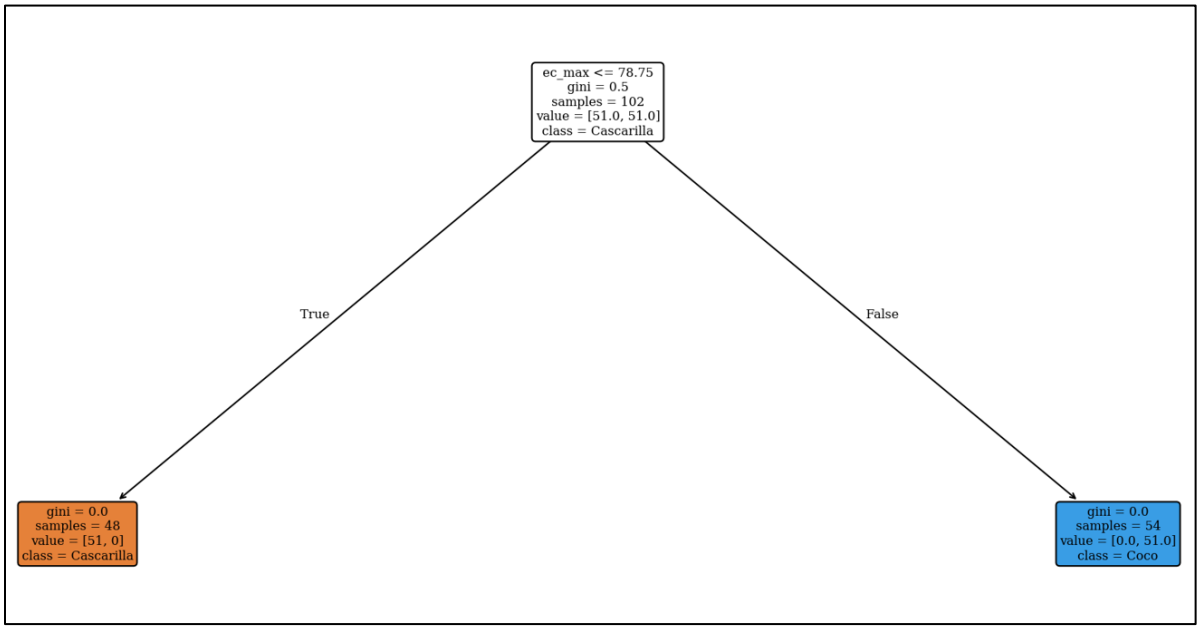
Nota. Gráfico de barras horizontales con la importancia relativa de cada variable en el Modelo A de Random Forest para la clasificación del tipo de sustrato, ordenadas de mayor a menor importancia y diferenciadas por color según el grupo de variable al que pertenecen.

La Figura 21 revela el mecanismo detrás de este resultado: el árbol de decisión clasificó perfectamente los 102 registros usando una sola regla, si la conductividad eléctrica máxima del

día supera 78,75, el sustrato es coco, si no lo supera, es cascarilla. Esta regla única con impureza Gini de 0,0 en ambos nodos finales significa que no hay ningún caso ambiguo en el dataset. Los dos sustratos están completamente separados por esta variable.

Figura 21

Árbol de Decisión Entrenado con Profundidad Máxima de 4 Niveles para la Clasificación de Sustratos (Coco vs. Cascarilla de Arroz)



Nota. Representación gráfica del árbol de decisión entrenado con profundidad máxima de cuatro niveles para la clasificación de sustratos, con nodos coloreados según la clase mayoritaria y etiquetas que muestran la variable de corte, el umbral, el índice de Gini y el número de muestras en cada nodo.

Modelo B Regresión para Tallos Productivos

El Modelo B intenta predecir cuántos tallos productivos generará una planta a partir de las variables del sensor. A diferencia del Modelo A, los resultados son heterogéneos, es decir que algunos grupos se predicen bien y otros no.

El R cuadrado promedio global obtenido es de -2,283, un valor negativo que en principio parece malo. Sin embargo este número está dominado por un único caso extremo, que es el Grupo 10 ya que obtuvo un R cuadrado de -41,774, lo que distorsiona el promedio general. Un R cuadrado negativo significa que para ese grupo específico el modelo predice peor que si simplemente usara el promedio de todos los datos como estimación. Excluyendo el Grupo 10, el R cuadrado resultante promedio es de 0,196, esto dice que el modelo explica aproximadamente el 20% de la variación en tallos productivos entre los grupos restantes. No es el mejor resultado, pero sí muestra que el modelo captura algo real.

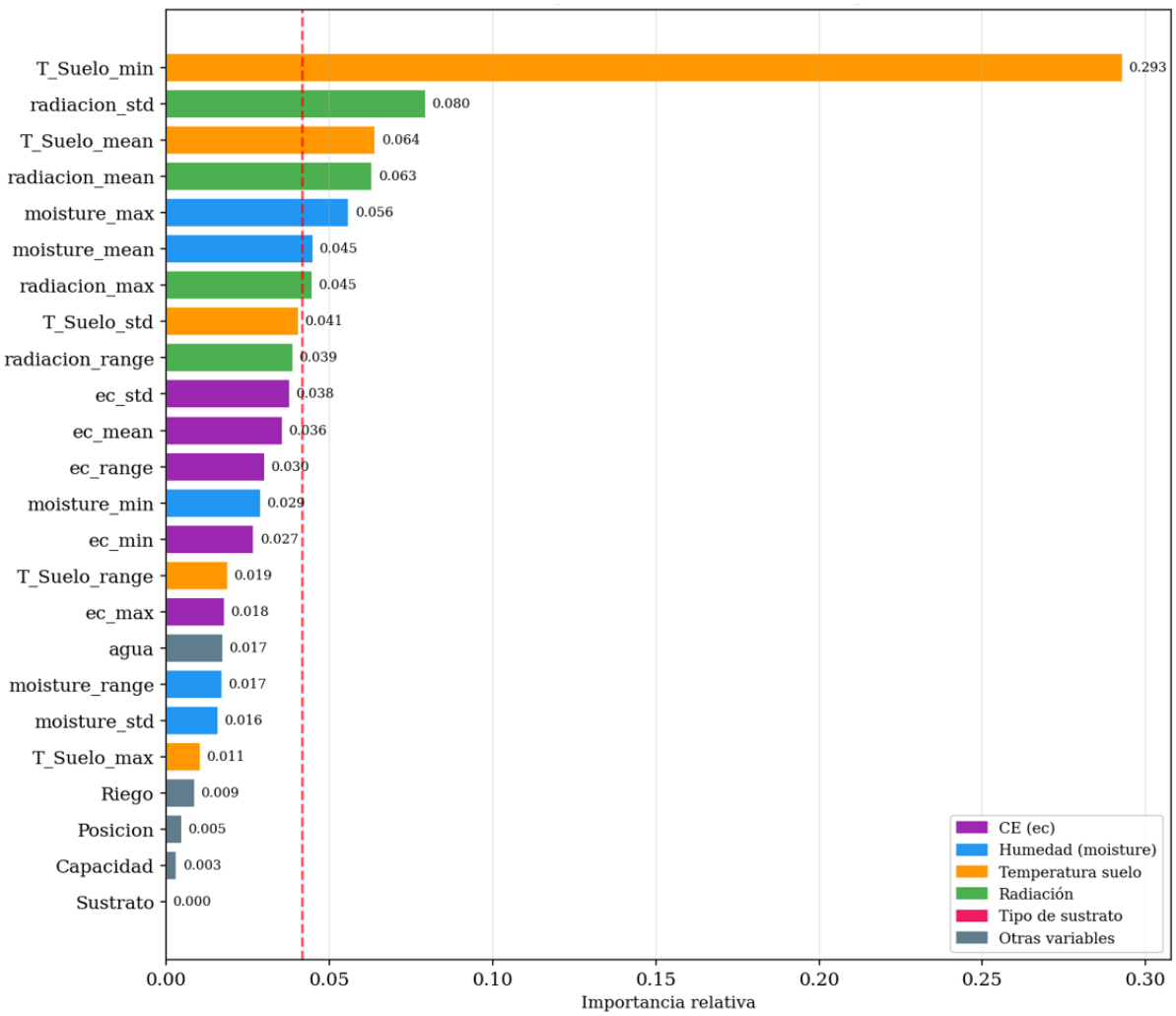
El valor de MAE obtenido de 2,26 tallos significa que en promedio el modelo se equivoca en 2,26 tallos al predecir la producción. Considerando que los valores observados van de 0 a 21 tallos, este error es moderado, y equivale a equivocarse en aproximadamente uno o dos tallos en la mayoría de los casos.

Seis grupos presentan un R cuadrado superior a 0,6, esto se entiende como una buena capacidad predictiva para esos grupos específicos, y son: Grupo 16 con 0,836, Grupo 15 con 0,810, Grupo 8 con 0,733, Grupo 7 con 0,719, Grupo 13 con 0,681 y Grupo 12 con 0,601. Para estos grupos el MAE oscila entre 0,68 y 2,22 tallos, el modelo así, predice con un margen de error de menos de 2 tallos en los mejores casos.

La Figura 22 nos muestra qué variables explican la producción de tallos, el hallazgo más revelador es que la temperatura mínima del suelo domina con una importancia de 0,293, más de tres veces por arriba de la segunda variable. Esto dice que los días con menor temperatura mínima coinciden con menor producción de tallos independientemente del tipo de sustrato, por ejemplo la variable Sustrato tiene importancia de 0, lo que confirma que el tipo de sustrato no aporta información adicional para predecir tallos.

Figura 22

Gráfico de Barras Horizontales con la Importancia de Variables del Modelo B



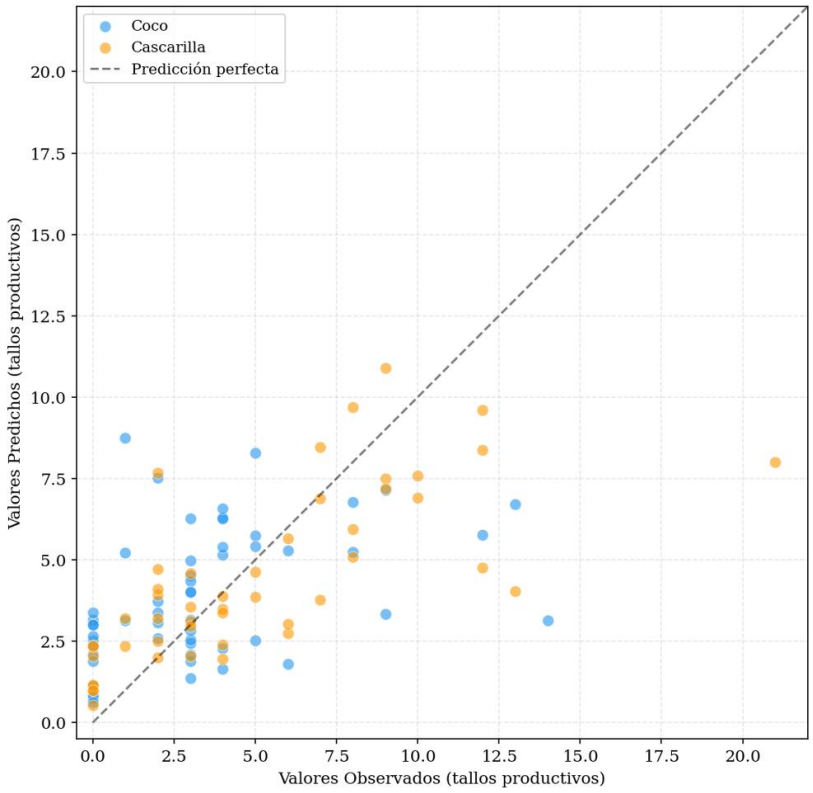
Nota. Gráfico de barras horizontales con la importancia relativa de cada variable en el Modelo B de Random Forest para la predicción del número de tallos productivos, ordenadas de mayor a menor importancia y diferenciadas por color según el grupo de variable.

La Figura 23 muestra la precisión del modelo, los puntos cercanos a la diagonal punteada son predicciones correctas y los alejados son errores. Se aprecia que el modelo predice bien en el rango bajo entre 0 y 3 tallos, pero sobreestima en valores medios y subestima en valores altos,

patrón típico de modelos de regresión que tienden a predecir hacia la media cuando hay alta variabilidad en los datos.

Figura 23

Gráfico de Dispersión de Valores Observados vs. Valores Predichos por el Modelo B para el Número de Tallos Productivos



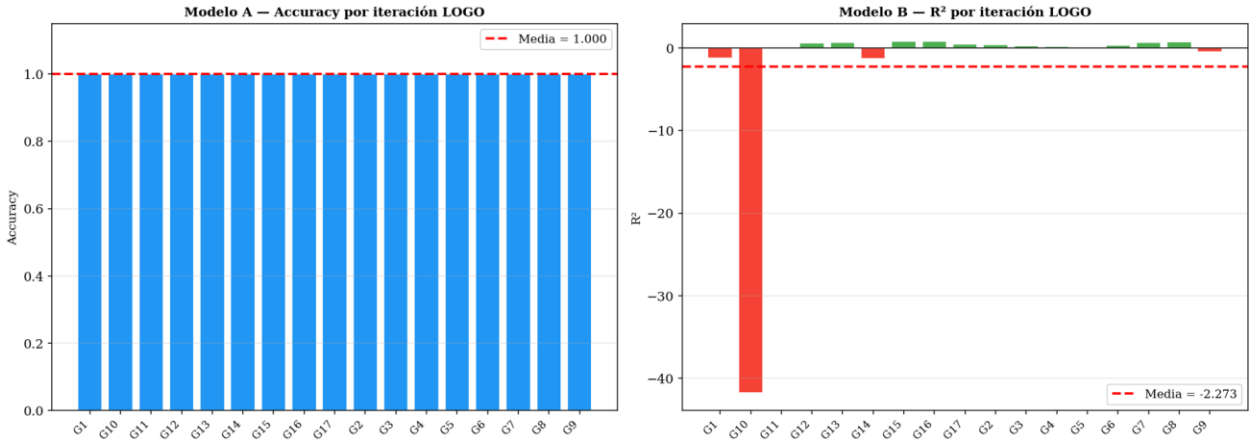
Nota. Gráfico de dispersión entre los valores observados y los valores predichos por el Modelo B para el número de tallos productivos, con puntos diferenciados por tipo de sustrato y una línea diagonal punteada que representa la predicción perfecta.

Estabilidad por Iteración LOGO

La Figura 24 permite ver el desempeño iteración por iteración, es decir, qué tan bien funcionó el modelo cada vez que se excluyó un grupo diferente y los resultados numéricos por grupo de ambos modelos se consolidan en la Tabla 14.

Figura 24

Distribución de las Métricas de Validación por Iteración LOGO



Nota. Panel doble con la distribución de las métricas de validación por iteración LOGO: exactitud del Modelo A (panel izquierdo) y R^2 del Modelo B (panel derecho), donde cada barra corresponde al grupo excluido en esa iteración y la línea horizontal indica el promedio global.

Resultados de Validación Leave-One-Group-Out para el Modelo A y el Modelo B

Tabla 10

Modelo A de Clasificación por Grupo

Grupo	Sustrato	Exactitud	F1 ponderado
Grupo 1	Coco	1,000	1,000
Grupo 2	Coco	1,000	1,000
Grupo 3	Cascarilla	1,000	1,000
Grupo 4	Cascarilla	1,000	1,000
Grupo 5	Coco	1,000	1,000
Grupo 6	Coco	1,000	1,000

Grupo	Sustrato	Exactitud	F1 ponderado
Grupo 7	Cascarilla	1,000	1,000
Grupo 8	Cascarilla	1,000	1,000
Grupo 9	Coco	1,000	1,000
Grupo 10	Coco	1,000	1,000
Grupo 11	Cascarilla	1,000	1,000
Grupo 12	Cascarilla	1,000	1,000
Grupo 13	Coco	1,000	1,000
Grupo 14	Coco	1,000	1,000
Grupo 15	Coco	1,000	1,000
Grupo 16	Cascarilla	1,000	1,000
Grupo 17	Cascarilla	1,000	1,000
Promedio		1,000 ± 0,000	1,000 ± 0,000

Tabla 11

Modelo B de Regresión por Grupo

Grupo	Sustrato	R^2	MAE (tallos)
Grupo 1	Coco	-1,209	2,77
Grupo 2	Coco	0,421	3,22
Grupo 3	Cascarilla	0,267	1,89
Grupo 4	Cascarilla	0,182	4,34
Grupo 5	Coco	0,135	1,27
Grupo 6	Coco	0,335	1,66

Grupo	Sustrato	R^2	MAE (tallos)
Grupo 7	Cascarilla	0,719	1,68
Grupo 8	Cascarilla	0,733	1,50
Grupo 9	Coco	-0,444	4,22
Grupo 10	Coco	-41,774	3,90
Grupo 11	Cascarilla	-0,087	3,48
Grupo 12	Cascarilla	0,601	2,22
Grupo 13	Coco	0,681	1,76
Grupo 14	Coco	-1,324	1,75
Grupo 15	Coco	0,810	0,68
Grupo 16	Cascarilla	0,836	1,11
Grupo 17	Cascarilla	0,473	0,94
Promedio total		-2,273 ± 10,200	2,26 ± 1,17
Promedio sin Grupo 10		0,196 ± 0,621	2,21 ± 1,19

Nota. Resultados de la validación Leave-One-Group-Out para ambos modelos: exactitud y F1-score ponderado por grupo para el Modelo A, y R^2 y MAE por grupo para el Modelo B, con los promedios globales al final.

Para el Modelo A, las 17 barras del panel izquierdo alcanzan exactitud de 1,0 sin excepción. Esto confirma que el resultado perfecto no depende de qué grupo se excluya, el modelo aprende la misma regla discriminante sin importar con cuáles 16 grupos se entrene, la separación entre sustratos es tan marcada que ningún grupo individual es crítico para que el modelo funcione.

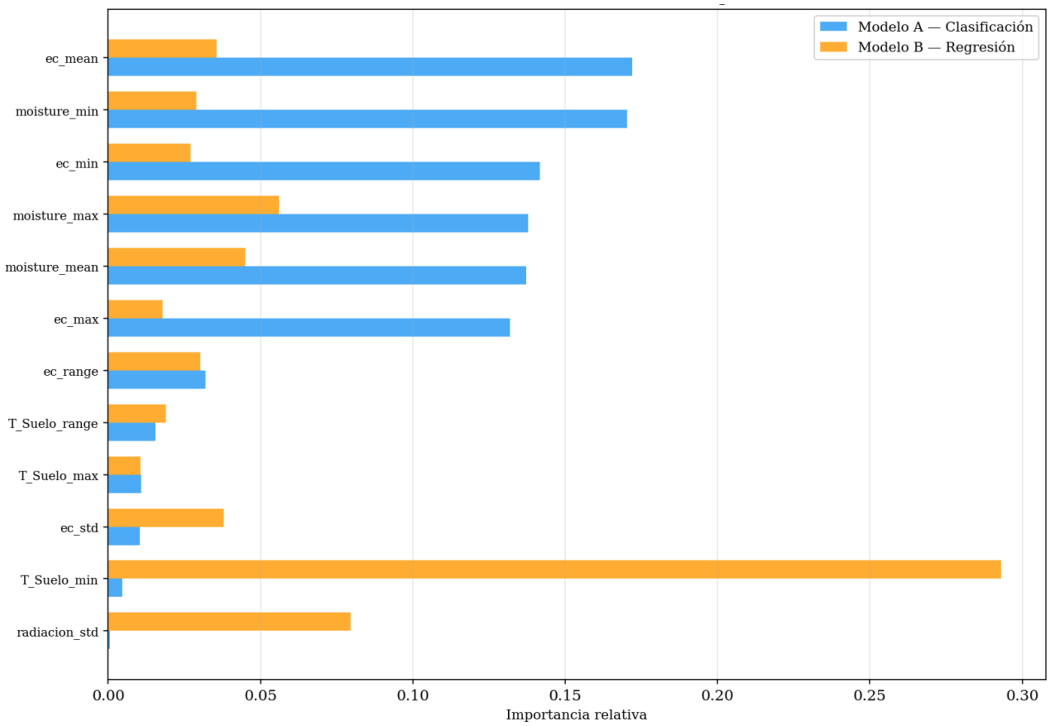
Ahora, para el Modelo B, el panel derecho muestra una imagen distinta. La barra del Grupo 10 cae para su R cuadrado hasta -41,774, dominando visualmente la figura. Las demás barras se distribuyen entre valores negativos moderados y positivos cercanos a 0,8. Esta variabilidad entre grupos indica que la producción de tallos no sigue un patrón único y predecible, algunos grupos responden de manera que el modelo puede aprender, y otros tienen comportamientos que no se generalizan desde el resto del experimento.

Comparación de Importancia de Variables entre Modelos

La Figura 25 pone en perspectiva la diferencia fundamental entre los dos modelos. En el Modelo A, las variables que más importan son la humedad y la CE, propiedades físicas estables del sustrato que el sensor mide con alta consistencia. Para el Modelo B, la variable más importante es la temperatura mínima del suelo, seguida de estadísticos de radiación.

Figura 25

Comparación de la Importancia de Variables entre el Modelo A (clasificación) y el Modelo B (regresión)



Nota. Gráfico de barras agrupadas con las importancias relativas de las variables más relevantes en el Modelo A (clasificación) y el Modelo B (regresión), que permite comparar visualmente qué variables son determinantes en cada objetivo de modelado.

Esta diferencia significa que, identificar el tipo de sustrato y predecir la producción de la planta dependen de variables completamente distintas. El sustrato se caracteriza por su humedad y sus sales. La producción de tallos responde a la temperatura y la luz. Un productor que quiera saber qué sustrato tiene, debe mirar la CE. Un productor que quiera saber cuándo esperar más tallos, debe mirar la temperatura mínima nocturna y la radiación del período.

Conclusiones

El presente trabajo caracterizó el comportamiento físico e hídrico de dos sustratos orgánicos, el de coco y la cascarilla de arroz, almacenados en bolsa y bajo condiciones de invernadero en la Sabana de Bogotá, mediante la integración de datos capturados por sensores IoT complementado con mediciones manuales periódicas, y el desarrollo de modelos de machine learning enfocados en su clasificación y análisis predictivo. Los resultados obtenidos permiten responder la pregunta de investigación y entregar conclusiones preliminares claras para la fase de crecimiento en el que se encuentra el cultivo y alineadas con cada objetivo específico. Es importante mencionar que los resultados presentados son representativos de la fase de establecimiento y crecimiento vegetativo inicial del cultivo de estudiado, ya que las plantas tenían aproximadamente 60 días desde la siembra al momento del monitoreo.

Sobre el análisis estadístico y temporal de las variables obtenidas por los sensores, los sustratos de coco y cascarilla de arroz exhiben comportamientos físicos e hídricos marcadamente diferenciados bajo las condiciones monitoreadas. La humedad media del sustrato de coco (45,35 %) es un poco mas del doble que el de la cascarilla de arroz (19,66 %), y la conductividad eléctrica media del coco (137,71) es casi cuatro veces más al de la cascarilla de arroz (33,14). Estas diferencias se mantienen constantes durante las seis fechas de muestreo manual sin convergencia temporal, lo que indica que las propiedades físicas e hídricas de cada sustrato no se homogenizan con el paso del tiempo bajo las condiciones del experimento.

La variabilidad diaria de la humedad del sustrato es sistemáticamente mayor en la cascarilla de arroz, confirmando su menor capacidad de retención hídrica y mayor velocidad de drenaje entre eventos de riego, también se observa que la temperatura del suelo no presenta diferencias relevantes entre ambos sustratos, lo que es coherente con su dependencia del

ambiente del invernadero, compartido por todos los grupos. De las variables ambientales monitoreadas por lo sensores, únicamente la radiación fotosintéticamente activa presentó variación temporal suficiente (CV = 39,6 %) para ser incorporada al análisis de modelado aunque su valor sea idéntico para todos los grupos en una misma fecha, su variación entre fechas la hace útil para el modelo de regresión.

Sobre los modelos de machine learning, el Modelo A de clasificación alcanzó una exactitud perfecta de 1,000 en las 17 iteraciones de la validación Leave-One-Group-Out, identificando correctamente el tipo de sustrato en todos los casos sin haber observado el grupo evaluado durante el entrenamiento. El árbol de decisión revela que esta clasificación se logra mediante una única regla: si la conductividad eléctrica máxima diaria supera el umbral de 78,75, el sustrato es coco; si no lo supera, es cascarilla. Este resultado tiene dos implicaciones: primero, que los dos sustratos son físicamente distinguibles de manera determinante a partir de los datos del sensor; y segundo, que la conductividad eléctrica máxima diaria es la propiedad más discriminante bajo las condiciones del experimento, por encima de la humedad.

El Modelo B de regresión presentó resultados heterogéneos con R^2 promedio de 0,196 excluyendo el grupo atípico (Grupo 10, R^2 = -41,774), y MAE promedio de 2,26 tallos. La variable más importante fue la temperatura mínima del suelo (T_Suelo_min, importancia = 0,293), seguida por estadísticos de radiación, lo que indica que la producción de tallos para el periodo de estudio responde positivamente a condiciones térmicas y de radiación fotosintética de cada fecha de muestreo, más que al tipo de sustrato en sí. Esta diferencia en la estructura de importancia de variables entre ambos modelos es el hallazgo más revelador e interesante del trabajo, porque las variables que caracterizan el sustrato como la CE y la humedad, son distintas de las que explican la respuesta productiva de la planta (temperatura mínima y radiación).

Sobre los resultados técnicos para los productores agrícolas, nos permiten exponer tres hallazgos técnicos concretos: El primero tiene que ver con la selección del sustrato en función del objetivo hídrico, de esta manera, el coco es el sustrato recomendado cuando se requiere alta estabilidad hídrica y menor frecuencia de riego, dado que mantiene rangos de humedad mas estables con baja variación diaria, este factor puede ser importante en regiones con acceso limitado al recurso hídrico. Por otra parte, la cascarilla de arroz es adecuada cuando se busca mayor aireación radicular y drenaje rápido, pero exige mayor frecuencia o volumen de riego para evitar estrés hídrico, dado que maneja un rango de humedad mayor y menos consistente.

En el segundo hallazgo podemos informar que utilizar un sensor para medir la conductividad eléctrica, es el instrumento más informativo para caracterizar y monitorear el estado de un sustrato, el umbral de 78,75 identificado por el árbol de decisión puede usarse como un indicador operativo: valores sostenidamente superiores señalan acumulación de sales, efecto propio para el sustrato de coco bajo el régimen de fertirrigación de este estudio. Para productores que manejen el sustrado de coco, el seguimiento de la CE máxima diaria permite detectar condiciones de salinidad que puedan requerir ajuste en la fertilización o en el volumen de lixiviación.

El tercer hallazgo nos presenta la respuesta productiva diferencial entre ambos sustratos. La cascarilla de arroz generó mayor producción promedio de tallos respecto al coco bajo las condiciones del experimento, sin embargo, dado que la temperatura mínima del suelo y la radiación fueron las variables más predictivas de la respuesta productiva y no el tipo de sustrato, se puede plantear que el manejo térmico y lumínico del invernadero puede tener mayor impacto sobre la producción que la elección del sustrato en sí misma, en condiciones de alta altitud como la de la Sabana de Bogotá.

En respuesta a la pregunta de investigación, los sustratos de coco y cascarilla de arroz presentan patrones de comportamiento físico e hídrico suficientemente diferenciados como para ser identificados con certeza mediante sensores IoT, con la conductividad eléctrica máxima diaria como variable principal. Los modelos de machine learning desarrollados demuestran que la caracterización basada en datos de sensor es viable, reproducible y metodológicamente sólida bajo condiciones de invernadero en la Sabana de Bogotá, aportando una base técnica objetiva para la toma de decisiones en sistemas de cultivo sin suelo en la región.

Recomendaciones

Las siguientes recomendaciones se derivan de los hallazgos del estudio y están dirigidas tanto a productores agrícolas de la región como a investigadores que busquen ampliar o replicar esta línea de trabajo.

Para los productores, la conductividad eléctrica máxima diaria del sustrato demostró ser el indicador más eficiente para diferenciar el comportamiento de los sustratos evaluados. Se recomienda incorporar sensores de conductividad eléctrica como herramienta de monitoreo rutinario en sistemas de cultivo sin suelo, priorizando el seguimiento del valor más alto del día sobre el promedio, ya que ese pico refleja el estado de retención de sales en el sustrato después de cada riego.

Para cultivos en los que la estabilidad hídrica sea prioritaria, el sustrato de coco es la opción recomendada bajo las condiciones de la Sabana de Bogotá, dado que mantiene niveles de humedad entre 35 % y 54 % con menor oscilación diaria. La cascarilla de arroz es adecuada cuando se requiere mayor aireación radicular, pero exige mayor atención en la frecuencia de riego para evitar períodos de estrés hídrico entre aplicaciones.

Dado que la temperatura mínima del suelo fue la variable más predictiva de la respuesta productiva, se recomienda prestar especial atención al manejo térmico del invernadero en los períodos nocturnos, particularmente en los meses de mayor nubosidad en la región, donde la caída de temperatura puede incidir negativamente en la actividad productiva independientemente del sustrato utilizado.

Para investigadores, se recomienda ampliar el período de muestreo manual a intervalos más frecuentes, quincenales en etapa de desarrollo y semanal en etapa de inicio de producción si las condiciones del experimento lo permiten, con el fin de incrementar el número de

observaciones disponibles para el modelado. El presente estudio demostró que con seis fechas de muestreo es posible construir un clasificador robusto, pero el modelo de regresión requeriría mayor volumen de datos para generalizar de manera más confiable entre grupos con comportamientos productivos atípicos como el observado en el Grupo 10.

Se recomienda explorar la incorporación de variables derivadas de la dinámica temporal del sensor, como la velocidad de caída de humedad entre riegos o la pendiente de recuperación post-riego como variables predictoras adicionales. Estas variables, calculables a partir del dataset de alta frecuencia disponible, podrían capturar propiedades hídricas del sustrato no representadas por los estadísticos diarios utilizados en este estudio.

Finalmente, se recomienda replicar la metodología desarrollada en otros contextos geográficos, especies de cultivo y combinaciones de sustratos, y poder contribuir a la consolidación de una base técnica regional para la caracterización de sustratos mediante sensores IoT y modelos de aprendizaje automático con el fin de evaluar la transferencia de los umbrales identificados, en particular el valor de la conductividad eléctrica medida puede variar respecto a otros estudios, factores como la ubicación y posición del sensor dentro del sustrato son determinantes para los valores que se generaron en este estudio.

Referencias Bibliográficas

Abad, M., Fornes, F., Carrión, C., Noguera, V., Noguera, P., Maquieira, Á., & Puchades, R. (2005). Physical properties of various coconut coir dusts compared to peat. HortScience, 40(7), 2138–2144.

Barrett, G. E., Alexander, P. D., Robinson, J. S., & Bragg, N. C. (2016). Achieving environmentally sustainable growing media for soilless plant cultivation systems. Scientia Horticulturae, 212, 220–234.

Bojacá, C. R., Gil, R., & Cooman, A. (2009). Use of geostatistical and crop growth modelling to assess the variability of greenhouse tomato yield. Biosystems Engineering, 103(3), 302–315.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Fernández, M. D., Bonachela, S., Orgaz, F., Thompson, R., López, J. C., Granados, M. R., & Fereres, E. (2010). Measurement and estimation of plastic greenhouse reference evapotranspiration in a Mediterranean climate. Irrigation Science, 28(6), 497–509.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15, 3133–3181.

Ferrández-Pastor, F. J., García-Chamizo, J. M., Nieto-Hidalgo, M., & Mora-Martínez, J. (2018). Precision agriculture design method using a distributed computing architecture on the internet of things context. Sensors, 18(6), 1731.

Gruda, N. (2019). Increasing sustainability of growing media constituents and stand-alone substrates in soilless culture systems. Agronomy, 9(6), 298.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2.ª ed.). Springer.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6), e0156571.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.

Monsalve Camacho, O. I., Henao Toro, M. C., & Gutiérrez Díaz, J. S. (2021). Caracterización de materiales con uso potencial como sustratos en sistemas de cultivo sin suelo. Ciencia y Tecnología Agropecuaria, 22(1).

Noguera, P., Abad, M., Puchades, R., Maquieira, A., & Noguera, V. (2003). Influence of particle size on physical and chemical properties of coconut coir dust as container medium. Communications in Soil Science and Plant Analysis, 34(3–4), 593–605.

Raviv, M., & Lieth, J. H. (2008). Soilless Culture: Theory and Practice. Elsevier.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929.

Savvas, D., & Gruda, N. (2018). Application of soilless culture technologies in the modern greenhouse industry. European Journal of Horticultural Science, 83(5), 280–293.

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive review. IEEE Access, 9, 4843–4873.

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

Tzounis, A., Katsoulas, N., Bartzanas, T., & Kittas, C. (2017). Internet of Things in agriculture, recent advances and future challenges. Biosystems Engineering, 164, 31–48.

Urrestarazu, M. (2004). Tratado de cultivo sin suelo (3.ª ed.). Mundi-Prensa.

Wilks, D. S. (2011). Statistical Methods in the Atmospheric Sciences (3.ª ed.). Academic Press.

Apéndices

Apéndice A

Pseudocódigo de Carga y Unión de Datos

ENTRADA: FULL_DATA_ENE_MRZ.csv, config_sensores.csv,

Dataset_manual_crecimiento.csv

1. Cargar datos crudos exportados desde InfluxDB

2. Eliminar columnas administrativas no relevantes

[result, table, _start, _stop, Unnamed]

3. Convertir columna de tiempo (_time) a formato datetime

4. Pivotar tabla: cada variable de sensor (_field) se convierte

en columna independiente

— Índices: datetime y sensor_id

5. Extraer columnas 'fecha' y 'hora' desde datetime

6. Estandarizar sensor_id a mayúsculas en datos de sensor

y en config_sensores.csv

7. Unir datos de sensores con config_sensores.csv

— Clave de unión: sensor_id

— Resultado: dataset_cultivo_organizado.csv (277.309 × 18)

8. Cargar Dataset_manual_crecimiento.csv

9. Estandarizar nombre_trabajo y fechas en ambos datasets

10. Unir dataset organizado con datos manuales

— Claves de unión: [fecha, nombre_trabajo]

— Resultado intermedio para análisis complementarios SALIDA:

dataset_cultivo_organizado.csv (277.309 × 18)

SALIDA: dataset_cultivo_organizado.csv

Apéndice B

Pseudocódigo de Limpieza y Preparación

ENTRADA: dataset_cultivo_organizado.csv

1. Separar temperatura por tipo de sensor:

  T_Suelo ← registros donde Tipo_sensor == 'SUELO'

  T_Amb ← registros donde Tipo_sensor == 'AMBIENTE'

  Eliminar columna 'temperature' original

  Preservar nulos resultantes como nulos estructurales

2. Ordenar dataset por [sensor_id, datetime]

3. Verificar valores faltantes por variable y sensor:

Para cada variable numérica:

    contar nulos agrupados por sensor_id

Si se detectan brechas en series continuas [moisture, ec, T_Suelo, T_Amb]:

    Aplicar interpolación lineal por sensor_id

→ En el presente estudio: continuidad completa confirmada, interpolación sin efecto neto

4. Imputación por mediana agrupada por sensor_id para:

  [ph, radiacion, humidity, co2, pressure]

  — Si sensor no tiene ningún registro para esa variable:

   mantener nulo estructural

5. Detección y corrección de outliers por sensor_id:

  Para cada variable numérica:

    Q1 = percentil 25 del sensor

    Q3 = percentil 75 del sensor

IQR = Q3 - Q1

límite_inf = Q1 - 1.5 × IQR

límite_sup = Q3 + 1.5 × IQR

Capping: valores fuera de límites se reemplazan

por el límite correspondiente

Restricciones de dominio adicionales:

moisture, humidity → rango válido [0, 100]

ec → mínimo 0 (no puede ser negativo)

6. Eliminar filas completamente duplicadas

7. Estandarizar variables categóricas:

[Tipo_bolsa, Tipo_sensor, Sustrato]

→ convertir a mayúsculas y eliminar espacios en blanco

→ reemplazar cadenas 'NAN' por valor nulo de pandas

8. Convertir iluminación a PPFD:

radiacion (µmol m^{-2} s^{-1}) = illumination (lux) × 0.0185

Eliminar columna 'illumination' original

SALIDA: dataset_cultivo_organizado.csv (limpio y transformado)

Apéndice C

Pseudocódigo de Selección de Variables Ambientales

ENTRADA: dataset_cultivo_organizado, fechas_muestreo (6 fechas)

1. Filtrar filas correspondientes a sensores de ambiente y agua
2. Conservar únicamente registros cuya fecha normalizada
   coincida con alguna de las 6 fechas de muestreo
3. Para cada variable ambiental [T_Amb, humedad_relativa,
   radiacion, ph, presion, co2]:
   a. Calcular la media diaria por fecha de muestreo
   b. Calcular el CV entre las 6 medias:
      CV = (desviación_estándar / media) × 100
   c. Si CV > 20%: marcar variable como INCLUIDA
      Si CV ≤ 20%: marcar variable como DESCARTADA
4. Conservar únicamente las variables marcadas como INCLUIDAS
5. Para cada variable incluida, calcular estadísticos diarios
   [media, std, min, max, rango] por fecha de muestreo

SALIDA: tabla de estadísticos diarios de variables seleccionadas
para integrar al dataset de modelado

Apéndice D

Pseudocódigo de Construcción del Dataset para Modelado

```
ENTRADA: dataset_cultivo_organizado.csv (limpio),
        Dataset_manual_crecimiento.csv,
        estadísticos diarios de radiación (de Apéndice C)
1. Identificar las 6 fechas de muestreo desde el dataset manual
  fechas_muestreo = fechas únicas del dataset manual
2. Filtrar dataset_cultivo_organizado por:
  — Tipo de sensor == 'SUELO'
  — Fecha normalizada contenida en fechas_muestreo
3. Para cada combinación [nombre_trabajo, fecha]:
  Calcular estadísticos diarios de cada variable de sensor:
  Para [moisture, T_Suelo, ec]:
    variable_mean  = media del día
    variable_std   = desviación estándar del día
    variable_min   = mínimo del día
    variable_max   = máximo del día
    variable_range = máximo - mínimo del día
  → Resultado: 15 features numéricas de sensor
4. Unir estadísticos de sensor con datos manuales:
  — Claves: [fecha, nombre_trabajo]
  — Conservar: numero_tallos_productivos,
            Cantidad_de_agua_L
```

— Excluir: inicio_de_produccion

5. Unir estadísticos de radiación:

— Clave: fecha

— Variables: radiacion_mean, radiacion_std,

radiacion_max, radiacion_range

Excluir: radiacion_min (valor sistemáticamente cero por lecturas nocturnas, sin aporte discriminativo)

→ Resultado: 4 variables adicionales adicionales

6. Ajustar tipos de datos:

Capacidad_de_bolsa → texto (categórica)

Posicion_en_el_grupo → texto (categórica)

Cantidad_de_agua_L → reemplazar coma por punto → numérico

datetime → conservar solo la fecha (sin hora)

7. Reordenar columnas:

[datetime, nombre_trabajo, Sustrato,

Capacidad_de_bolsa, Posicion_en_el_grupo,

Tipo_de_riego, variables_sensor (15),

tallos_productivos, Cantidad_agua,

variables_radiacion (5)]

8. Verificar integridad:

— Confirmar dimensiones: 102 filas × 29 columnas

— Confirmar cero valores nulos

9. Guardar como DATASET_ANALISIS.csv

SALIDA: DATASET_ANALISIS.csv (102 × 29, 0 nulos)

Apéndice E

Pseudocódigo de Preparación para el Modelado

```
ENTRADA: DATASET_ANALISIS.csv (102 × 29)

1. Codificar variables categóricas con LabelEncoder:
   Para [Sustrato, Capacidad_bolsa,
         Posicion_en_grupo, Tipo_riego]:
     Asignar valor numérico entero a cada categoría
     Guardar el codificador para interpretación posterior
2. Definir variables predictoras y objetivos por modelo:
   MODELO A (Clasificación):
     variables_A = todas las columnas excepto
                   [datetime, nombre_trabajo, Sustrato]
     objetivo_A = Sustrato codificado (Coco=1, Cascarilla=0)
   MODELO B (Regresión):
     variables_B = todas las columnas excepto
                   [datetime, nombre_trabajo,
                    numero_tallos_productivos]
     objetivo_B = numero_tallos_productivos
3. Definir grupos para validación LOGO:
   grupos = nombre_trabajo (17 valores únicos)
4. Configurar modelos Random Forest:
   RF_clasificador:
     n_estimators    = 200
```

max_depth = 5

min_samples_leaf = 3

class_weight = 'balanced'

random_state = 42

RF_regresor:

n_estimators = 200

max_depth = 5

min_samples_leaf = 3

random_state = 42

5. Ejecutar validación LOGO:

Para cada grupo g en [Grupo1 ... Grupo17]:

train = todas las filas donde nombre_trabajo ≠ g

test = todas las filas donde nombre_trabajo == g

Entrenar modelo con train

Evaluar modelo con test

Registrar métricas de la iteración

6. Calcular métricas promedio sobre las 17 iteraciones:

MODELO A:

accuracy_promedio ± desviación estándar

F1_ponderado_promedio ± desviación estándar

ROC_AUC_promedio ± desviación estándar

MODELO B:

R2_promedio ± desviación estándar

MAE_promedio ± desviación estándar (en tallos)

7. Entrenar modelo final con todo el dataset:

— Calcular importancia de variables

— Generar árbol de decisión de profundidad 4

para visualización e interpretación

SALIDA: métricas de validación por modelo,

importancia de variables,

árbol de decisión visualizable

Apéndice F

Pseudocódigo del Análisis Exploratorio de Datos

ENTRADA: DATASET_ANALISIS.csv (102 × 29)

1. Estadística descriptiva:

   Para cada variable numérica, agrupando por Sustrato:

      Calcular: media, mediana, std, min, max,

            CV = (std / media) × 100

   → Tablas 10 y 11

2. Boxplots comparativos por sustrato:

   Para [moisture_mean, ec_mean, T_Suelo_mean]:

      Generar boxplot con Sustrato en eje X

   → Figura 11

3. Análisis temporal:

   Para [moisture_mean, ec_mean, moisture_range, ec_range]:

      Agrupar por [datetime, Sustrato]

      Calcular media por grupo

      Generar gráfico de líneas con fechas en eje X

   → Figuras 12 y 13

4. Análisis por capacidad de bolsa y tipo de riego:

   Para [moisture_mean, ec_mean]:

      Boxplot cruzando Capacidad_bolsa × Sustrato

      Barplot cruzando Tipo_riego × Sustrato

   → Figuras 14 y 15

5. Matriz de correlación:

   Calcular correlación de Pearson entre

   todas las variables numéricas

   Generar heatmap con escala coolwarm

   → Figura 16

6. Correlación individual con variable objetivo:

   Codificar Sustrato como numérico (Coco=1, Cascarilla=0)

   Calcular correlación de cada variable con Sustrato_num

   Ordenar por valor absoluto descendente

   Generar gráfico de barras horizontales

   → Figura 17

7. Análisis de tallos productivos:

   Generar histograma y boxplot por sustrato

   Generar evolución temporal por sustrato

   → Figuras 18 y 19

SALIDA: Figuras 11 a 19, Tablas 10 y 11

Apéndice G

Pseudocódigo del Desarrollo de los Modelos de Machine Learning

```
ENTRADA: DATASET_ANALISIS.csv (102 × 29)

1. Codificar variables categóricas con LabelEncoder:
   [Sustrato, Capacidad_bolsa,
    Posicion_en_grupo, Tipo_riego]
   — Asignar valor numérico entero a cada categoría
   — Guardar codificador para interpretación posterior
   — Posicion_en_grupo: convertir a numérico con
     pd.to_numeric(..., errors='coerce') antes de codificar
2. Definir configuraciones por modelo:
   MODELO A (Clasificación):
      features_A = [Capacidad_enc, Posicion_enc, Riego_enc,
                    moisture_mean, moisture_std, moisture_min,
                    moisture_max, moisture_range,
                    T_Suelo_mean, T_Suelo_std, T_Suelo_min,
                    T_Suelo_max, T_Suelo_range,
                    ec_mean, ec_std, ec_min, ec_max, ec_range,
                    Cantidad_de_agua_L, radiacion_mean,
                    radiacion_std, radiacion_max, radiacion_range,
                    numero_tallos_productivos]
      objetivo_A = Sustrato_enc (Cascarilla=0, Coco=1)
      — Total: 24 variables predictoras
```

MODELO B (Regresión):

features_B = [Capacidad_enc, Posicion_enc, Riego_enc,
moisture_mean, moisture_std, moisture_min,
moisture_max, moisture_range,
T_Suelo_mean, T_Suelo_std, T_Suelo_min,
T_Suelo_max, T_Suelo_range,
ec_mean, ec_std, ec_min, ec_max, ec_range,
Cantidad_de_agua_L, radiacion_mean,
radiacion_std, radiacion_max, radiacion_range,
Sustrato_enc]

objetivo_B = numero_tallos_productivos

— Total: 24 variables predictoras

— Nota: radiacion_min excluida por ser sistemáticamente cero (lecturas nocturnas)

— Nota: inicio_de_produccion excluida por tipo object sin valor predictivo

grupos = nombre_trabajo (17 valores únicos)

3. Configurar modelos finales:

RF_clasificador (Modelo A):

n_estimators = 200

max_depth = 5

min_samples_leaf = 3

class_weight = 'balanced'

random_state = 42

RF_regresor (Modelo B):

n_estimators = 200

max_depth = 5

min_samples_leaf = 3

random_state = 42

DT_clasificador (árbol visual — Figura 21):

max_depth = 4

min_samples_leaf = 4

class_weight = 'balanced'

random_state = 42

4. Análisis de sensibilidad de hiperparámetros (Modelo A):

Evaluar 5 configuraciones mediante validación LOGO:

Configuración 1: max_depth=3, min_samples_leaf=3, n_estimators=100

Configuración 2: max_depth=5, min_samples_leaf=3, n_estimators=200

Configuración 3: max_depth=7, min_samples_leaf=3, n_estimators=200

Configuración 4: max_depth=5, min_samples_leaf=5, n_estimators=200

Configuración 5: max_depth=None, min_samples_leaf=1, n_estimators=200

Para cada configuración:

Instanciar RandomForestClassifier con class_weight='balanced'

Aplicar cross_val_score con cv=LOGO, scoring='accuracy'

Registrar exactitud promedio sobre 17 iteraciones

Resultado: todas las configuraciones → exactitud = 1,000

Conclusión: resultados robustos ante variación de hiperparámetros

5. Validación LOGO — Para cada modelo:

Inicializar LeaveOneGroupOut()

Para cada grupo g en [Grupo1 ... Grupo17]:

train = filas donde nombre_trabajo ≠ g

test = filas donde nombre_trabajo == g

Entrenar modelo con train

Predecir sobre test

Registrar métricas e identificador del grupo excluido

6. Calcular métricas promedio sobre las 17 iteraciones:

MODELO A:

exactitud = media(exactitud por iteración)

F1 = media(F1_ponderado por iteración)

Nota: ROC-AUC no calculable — cada grupo

pertenece a una sola clase; el conjunto

de prueba contiene una única clase

en cada iteración

MODELO B:

R^2 = media(R^2 por iteración)

MAE = media(MAE por iteración)

Nota: calcular también promedio excluyendo

Grupo 10 por ser valor atípico extremo

(R^2 = -41,774)

7. Entrenar modelos finales con todo el dataset:

   Ajustar RF_clasificador sobre (features_A, objetivo_A)

   Ajustar RF_regresor sobre (features_B, objetivo_B)

   Ajustar DT_clasificador sobre (features_A, objetivo_A)

8. Extraer importancia de variables:

   importancia_A = RF_clasificador.feature_importances_

   importancia_B = RF_regresor.feature_importances_

   Ordenar ambas de mayor a menor importancia relativa

9. Análisis comparativo sensor vs otras posiciones:

   Cargar Dataset_manual_crecimiento.csv (408 × 9)

   Identificar posición del sensor por grupo

   desde DATASET_ANALISIS (columna Posicion_en_grupo)

   Etiquetar cada fila: tiene_sensor = True/False

   Calcular media de tallos con sensor vs sin sensor

   Aplicar prueba t de Student (scipy.stats.ttest_ind)

   — con_sensor: n=102, media=4,12 tallos

   — sin_sensor: n=306, media=4,14 tallos

   — Resultado: t=-0,052, p=0,959

   — Conclusión: sin diferencia estadísticamente

    significativa (α=0,05)

10. Generar visualizaciones:

   — Importancia variables Modelo A → Figura 20

   — Árbol de decisión visual → Figura 21

— Importancia variables Modelo B → Figura 22

— Dispersión observado vs predicho → Figura 23

— Métricas por iteración LOGO → Figura 24

— Comparación importancias A vs B → Figura 25

SALIDA:

— Métricas de validación LOGO (Tabla 14)

— Análisis de sensibilidad de hiperparámetros

— Importancia de variables por modelo

— Árbol de decisión visualizable (Figura 21)

— Resultado análisis comparativo sensor

vs otras posiciones (t=-0,052, p=0,959)

— Figuras 20 a 25