Predicting rose production using machine learning: a comparison between phenological and autoregressive models.

Predicción de la producción de rosas mediante aprendizaje automático: una comparación entre modelos fenológicos y autorregresivos

Lina Constanza Caicedo Arroyave [1]; Wilmer Dario Pineda Rios [2]; Carlos Isaac Zainea Maya [3].

Información de los autores

1. Candidata a Magister en Estadística Aplicada - Universidad Santo Tomas. linacaicedo@usantotomas.edu.com
2. Director -Doctor en Estadística- Docente Universidad Santo Tomas. wilmerpineda@usta.edu.co
3. Codirector –Magister en Mátematicas, Docente Universidad Santo Tomas. carloszainea@usta.edu.co

Received: xx xx 2025. Accepted: xx xx 20xx.

ABSTRACT

This research compares the effectiveness of two modeling approaches for predicting rose production: one based on traditional phenological counts and another purely autoregressive, using classical Machine Learning (ML) algorithms as a low-cost alternative. Due to the limited availability of historical data, synthetic datasets were generated to preserve the seasonal and cyclical patterns of Colombia's floriculture sector. MLP, LSTM, and XGBoost models were evaluated under a reproducible experimental design, applying cross-validation and standard error metrics (MSE, R^2). Results indicate that the autoregressive XGBoost model achieved the best performance (R^2=0.82), outperforming models based on phenological information (R^2=0.809). These findings demonstrate that production history provides a stronger predictive signal than manual field counts, reducing dependence on subjective and labor-intensive procedures. The study offers a replicable predictive framework that enhances production planning and strengthens the competitiveness of the floriculture industry through efficient use of existing data.

Key words: production forecasting, floriculture, *machine learning*, XGBoost, time series, synthetic data.

RESUMEN

Esta investigación compara la efectividad de dos enfoques de modelado para la predicción de la producción de rosas: uno basado en conteos fenológicos tradicionales y otro puramente autorregresivo, empleando algoritmos clásicos de Machine Learning (ML) como alternativa de bajo costo. Debido a la limitada disponibilidad de datos históricos, se generaron datos sintéticos que preservan la estacionalidad y los patrones cíclicos del sector floricultor colombiano. Se evaluaron modelos MLP, LSTM y XGBoost bajo un diseño experimental reproducible, aplicando validación cruzada y métricas de error (MSE, R^2). Los resultados muestran que el enfoque autorregresivo con XGBoost alcanzó el mejor desempeño (R^2=0.82), superando a los modelos basados en información fenológica (R^2=0.809). Esto evidencia que la historia productiva contiene una señal predictiva más robusta que los conteos manuales, permitiendo prescindir de procesos costosos y subjetivos. El estudio aporta un modelo predictivo replicable que optimiza la

planeación productiva del sector floricultor, fortaleciendo su competitividad mediante el uso eficiente de los datos disponibles.

Palabras clave: predicción de producción, floricultura, *machine learning*, XGBoost, series temporales, datos sintéticos.

INTRODUCCIÓN

La industria floricultora en Colombia es una de las de mayor importancia a nivel mundial, además el país se ha consolidado como el segundo mayor exportador de flores después de Países Bajos (International Fresh Produce Association, 2025; ProColombia, 2025; Asocolflores 2023, Trendeconomy, 2023; 2023 B., R., & G., 2019; Asocolflores, s. f.). Este lugar de privilegio se sostiene gracias a procesos logísticos especializados y a una planeación productiva que requiere gran precisión (Rodríguez, Barrero, Calderón, & Cardoso, 2025; Piza, 2023, pp. 34–36). Uno de los puntos clave en estos procesos es la proyección de la producción, ya que de una buena estimación depende tanto el control de los costos como la eficiencia en la cadena de suministro y la competitividad del sector (Mordor Intelligence, 2025; Encalada Ruiz & Rivadeneira Morales, 2020, pp. 58, 63, 72).

De manera tradicional, las proyecciones de cosecha se basan en la observación de cantidad de flores en los invernaderos a través de conteos de los estados fenológicos de las plantas, buscando anticipar la producción con varias semanas de antelación (Contrera Mercado, 2024, pp. 22–23; Mora Quintero, 2019; Rodríguez & Flórez, 2006). Una técnica empleada consiste en tomar en cuenta los estados fenológicos más representativos de la etapa reproductiva los cuales son arroz, arveja, garbanzo y rayando color (Perilla Garzón, 2019).

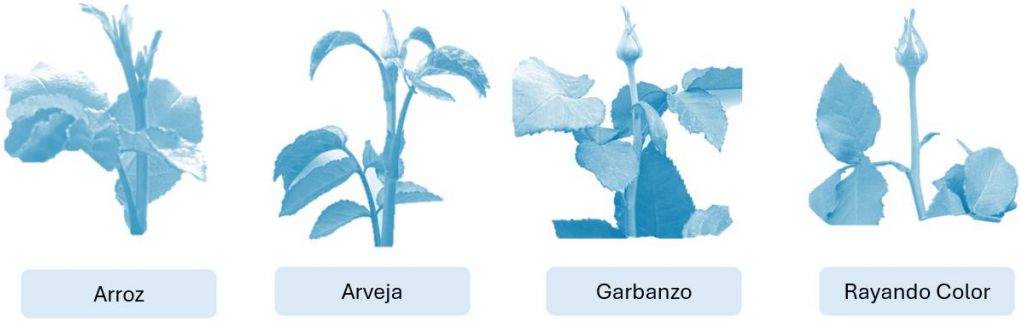


Figure 1. Representación de los estados fenológicos de la rosa en la etapa reproductiva.

Nota-1. Elaboración propia con base en Castro & Palomar (2022) y Perilla Garzón (2019).
Nota-2. Relación entre los estados de desarrollo del botón floral y la proyección de producción en el sector floricultor. Los botones en estado arroz corresponden a la producción estimada para cuatro semanas, los de estado arveja a tres semanas, los de estado garbanzo a dos semanas y los de estado rayando color a una semana antes de la cosecha.

Si bien esta técnica ha sido el estándar, su principal limitación radica en la subjetividad inherente al proceso de conteo y, por tanto, en su precisión. Los modelos que se generan a partir de ello son puntuales y experimentales, y se basan en un ciclo productivo medido en campo, no en datos históricos (Castro & Palomar, 2022). Esta condición metodológica introduce una variabilidad no cuantificada en las estimaciones, lo que debilita el rigor estadístico de los pronósticos. De hecho, como referente proxi Herrera et al. (2024) muestran que el conteo manual presenta una variabilidad significativamente mayor (CV = 42.69 %) frente a otros métodos automatizados basados en UAVs (CV = 36.08 %), lo que evidencia la menor precisión y el mayor error inherente a los métodos tradicionales.

Estas limitaciones se traducen en errores de pronóstico que resultan en pérdidas económicas considerables para el sector, estimadas en millones de dólares anualmente ya que generan ineficiencias logísticas. La magnitud del impacto es particularmente relevante si se considera que Colombia exportó US$ 2.385 millones en productos de la floricultura durante 2024 (Corficolombiana, 2025). Ya sea por excedentes de producción que no pueden ser comercializados o por la incapacidad de satisfacer pedidos comprometidos, la rentabilidad y la competitividad del sector se ven afectadas.

Ante este panorama, surge la necesidad de establecer metodologías más sólidas que permitan mejorar la fiabilidad de las proyecciones y apoyar la toma de decisiones en la planeación productiva. En este sentido, las metodologías de machine learning aparecen como una alternativa, ofreciendo mayor precisión en los pronósticos de cosecha (Liu et al., 2024, pp. 150–155; Mejía & Páez, 2023, pp. 12–13, 55–59; Albán Bautista & Zabala Chico, 2022, pp. 11–19).

La predicción del rendimiento de cultivos es un campo en constante evolución. Las investigaciones de vanguardia a nivel internacional se orientan hacia la automatización de la recolección de datos mediante vehículos aéreos no tripulados (UAVs) y la aplicación de modelos de visión por computador (e.g., YOLOv5, YOLOv8) para el conteo y clasificación de botones florales (Lai et al., 2025; Herrera et al., 2024). Estas tecnologías prometen, a largo plazo, eliminar la subjetividad inherente al conteo manual. Sin embargo, la adopción de estas tecnologías en el sector floricultor colombiano enfrenta barreras significativas de implementación a corto y mediano plazo, incluyendo altos costos de inversión y la necesidad de personal con competencias técnicas avanzadas (Mantilla, Mejía, & Tascón, 2025; Kumari et al., 2018). La realidad operativa de una gran mayoría de fincas productoras en el país sigue dependiendo, y seguirá dependiendo en el futuro previsible, de los métodos de conteo manual como principal fuente de datos para

la planificación (Carangui et al., 2024; Bahuguna et al., 2020). Ante este escenario, surge una brecha de conocimiento de carácter práctico y aplicado, dado que los datos de conteos manuales son y seguirán siendo la norma operativa, existe una clara ausencia de estudios rigurosos que determinen cuáles son las metodologías analíticas más potentes para extraer el máximo valor predictivo de esta información actualmente disponible. Por lo tanto, este trabajo adopta un enfoque metodológico más clásico y accesible, ya que está limitado por la naturaleza de los datos existentes (Araújo et al., 2023).

Mientras la investigación de frontera busca reemplazar la fuente de datos, aqui el enfoque esta en optimizar el análisis de la fuente de datos existente. La presente investigación se posiciona deliberadamente para llenar este vacío, investigando cómo los algoritmos de ML pueden mejorar la precisión de los pronósticos utilizando los datos que los productores ya recolectan, ofreciendo una solución de alto impacto y baja barrera de entrada.

En particular, esta investigación se plantea como un estudio comparativo entre dos enfoques conceptuales para abordar el problema de la predicción de la producción. El propósito es evaluar de manera empírica la eficacia de dos estrategias de modelado: una basada en conteos fenológicos tradicionales y otra de carácter puramente autorregresivo. Ambas se orientan a la estimación de la producción de rosas en dos semanas consecutivas de interés, empleando algoritmos clásicos de Machine Learning (ML) como una alternativa de bajo costo y fácil implementación.

Para lograr este objetivo, se plantean dos hipótesis principales:

Hipótesis 1. Basada en conocimiento agronómico tradicional: se postula que la relación proporcional entre los conteos de estados fenológicos (arroz, arveja, garbanzo, rayando color) y la producción futura, un pilar del conocimiento agronómico tradicional, puede ser modelada con mayor precisión y robustez mediante algoritmos de aprendizaje automático (ML), superando las limitaciones de los métodos de estimación lineal simple.

Hipótesis 2. Basada en series de tiempo: se plantea la hipótesis alternativa de que un modelo puramente autorregresivo, que utiliza únicamente los registros históricos de producción como predictores, puede lograr una precisión de pronóstico superior a los modelos basados en datos fenológicos. Esto implicaría que las dependencias temporales inherentes a la serie de tiempo de producción contienen una señal predictiva más fuerte que los conteos manuales exógenos.

La Hipótesis 1 se operacionaliza bajo la premisa de proporcionalidad entre los conteos fenológicos y la producción floral, donde los estados reproductivos de la rosa permiten anticipar la producción con un horizonte de cuatro semanas posteriores al conteo (Castro & Palomar, 2022; Cabrera Loja, 2021).

Table 1. Hipótesis de proporcionalidad entre los conteos fenológicos y la producción floral.

Estado reproductivo de la rosa	Porcentaje del conteo	Porcentaje de producción de las siguientes cuatro semanas
Rayando Color (Rc)	$p1 = \frac{\#\ Rc}{\#\ (Rc+G+Av+Ar)} * 100\%$	$P1 = \frac{producción\ a\ \mathbf{1\ semana}}{producción\ total\ a\ 1,2,3,4\ semanas} * 100\%$
Garbanzo (G)	$p2 = \frac{\#\ G}{\#\ (Rc+G+Av+Ar)} * 100\%$	$P2 = \frac{producción\ a\ \mathbf{2\ semanas}}{producción\ total\ a\ 1,2,3,4\ semanas} * 100\%$
Arveja (Av)	$p3 = \frac{\#\ Ar}{\#\ (Rc+G+Av+Ar)} * 100\%$	$P3 = \frac{producción\ a\ \mathbf{3\ semanas}}{producción\ total\ a\ 1,2,3,4\ semanas} * 100\%$
Arroz (Ar)	$p4 = \frac{\#\ Ar}{\#\ (Rc+G+Av+Ar)} * 100\%$	$P4 = \frac{producción\ a\ \mathbf{4\ semanas}}{producción\ total\ a\ 1,2,3,4\ semanas} * 100\%$

Donde #: número de tallos en el estado indicado (conteo)

Para trabajar con esta hipótesis, se trabajó con proporciones de producción organizadas en conjuntos de cuatro semanas continuas, lo que permite analizar la distribución relativa de la producción semanal dentro de cada grupo.

Table 2. Ejemplo de producción total en cuatro semanas

Semana	Producción
1	100
2	50
3	200
4	80
5	120
6	30
…	…

Los grupos se forman desplazando la ventana de cuatro semanas de manera continua:

- Grupo 1: semanas 1–4 (100, 50, 200, 80).
- Grupo 2: semanas 2–5 (50, 200, 80, 120).

Table 3. Ejemplo de procentaje de producción en cuatro semanas

Grupo	$P1$	$P2$	$P3$	$P4$
1	23%	12%	46%	19%
2	11%	44%	18%	27%

Teniendo en cuenta la tabla 1 Ejemplo: P1 = 100 / (100+50+200+80) = 23%

Este enfoque permite modelar la relación proporcional entre los estados fenológicos observados y los volúmenes de producción posteriores, proporcionando la base para el entrenamiento y validación de los modelos de machine learning.

MATERIALES Y METODOS

Simulación de datos de producción y coherencia metodológica

Simulación de producción de tallos. La investigación enfrentó la limitación típica de la ciencia de datos aplicada a la industria: la escasa disponibilidad de registros históricos de producción detallados y continuos. Dada la naturaleza reservada del sector floricultor, se optó por generar datos sintéticos, una estrategia metodológica que permite reproducir las propiedades estadísticas y temporales del fenómeno real, creando un entorno controlado para comparar de forma rigurosa distintas arquitecturas de modelos predictivos.

Base de datos proxy relevante. Como base para la simulación se utilizó el conjunto de datos de demanda semanal de la tesis de Calderón (2005). Aunque corresponde a la demanda y no a la producción directa, se considera el proxy público más adecuado, ya que en un sector exportador como la floricultura, la planificación de la producción responde directamente a la demanda proyectada del mercado. Estos datos reflejan la estacionalidad crítica —como los picos de San Valentín y el Día de la Madre— y los patrones cíclicos que determinan los cronogramas de siembra, poda y cosecha. Específicamente, se utilizaron los anexos 4.6 (correspondiente a 12 variedades) y 4.7 (las primeras 11 variedades) del trabajo de Calderón (2005), derivados del presupuesto de producción anual de la empresa Melody Flowers Ltda. Para alinear estos datos de demanda con el ciclo biológico de la producción, se aplicó un desfase temporal de 9 semanas (Calderón, 2005, p. 43).

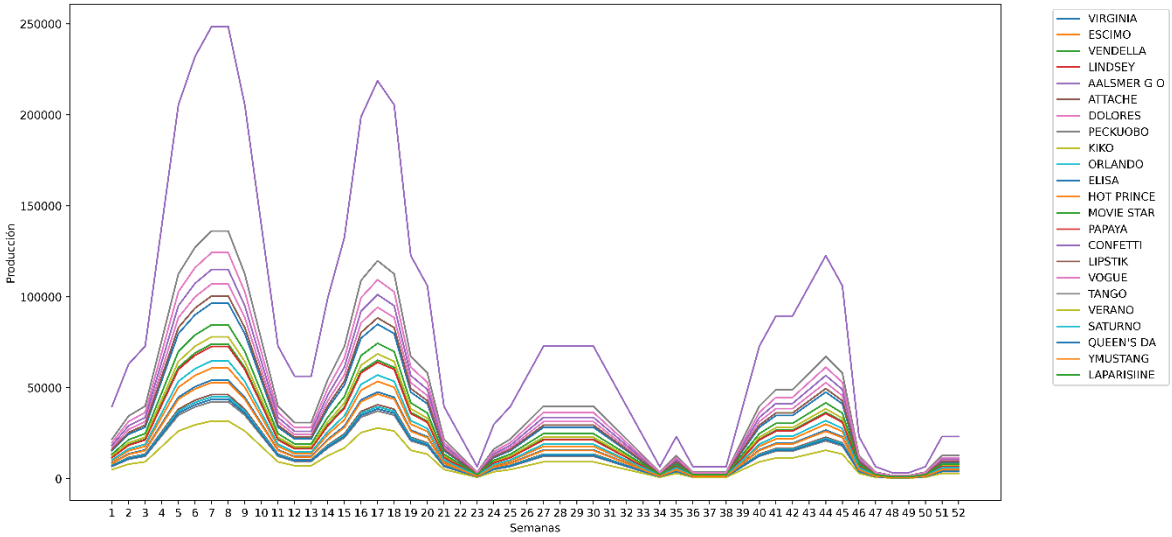


Figure 2. Curvas generadas a partir de los 23 conjuntos de datos presentados en los apéndices 4.6 y 4.7 del trabajo de Calderón (2005), que se emplearon como referencia para la simulación de nuevos conjuntos de datos.

Proceso de generación sistemático y reproducible. Cada uno de los 23 conjuntos de datos base (23 variedades) se utilizó como punto de partida para la creación de 300 nuevas series de datos sintéticos, con el fin de generar una diversidad de escenarios representativos. El proceso de simulación se estructuró de la siguiente manera: En primer lugar para cada una de las 52 semanas del periodo de estudio, se generaron 300 valores aleatorios mediante muestreo uniforme, delimitado por los valores mínimo y máximo correspondientes a la semana de interés en los datos originales. En segundo lugar para garantizar la reproducibilidad, se empleó una semilla aleatoria fija (42). Por último cada valor generado se promedió con la media semanal observada en los datos originales, lo que permitió preservar la tendencia central de los datos simulados y conservar la coherencia con el comportamiento de la variable. La coherencia metodológica principal de este estudio reside en el diseño de un flujo de trabajo (pipeline) analítico completo y riguroso que se aplica consistentemente a los datos generados, permitiendo una comparación justa de los dos enfoques de modelado planteados en las hipótesis.

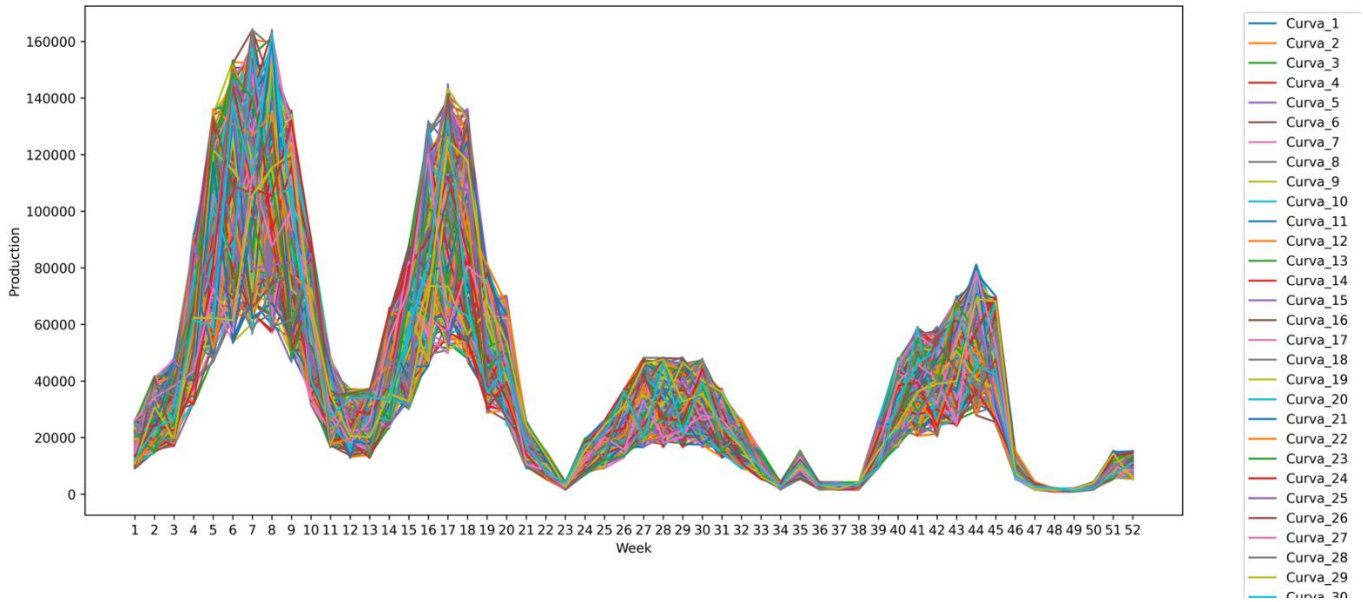


Figure 3. 300 curvas simuladas a través de software Python.

Técnicas de aprendizaje automático

Para la modelación y predicción se aplicaron diversas técnicas de aprendizaje automático, utilizando validación cruzada de cinco particiones para evaluar la estabilidad y generalización de los modelos. Los datos se dividieron en 80 % para entrenamiento y 20 % para validación, garantizando la reproducibilidad mediante una semilla aleatoria (random_state=42). El desempeño se evaluó con las métricas Error Cuadrático Medio (MSE) y Coeficiente de Determinación (R^2), calculadas por variable y de forma global. Todas las implementaciones se realizaron en Python, empleando scikit-learn, pandas, numpy y las librerías específicas de cada algoritmo.

Perceptrón Multicapa (MLP). Desde el campo de las redes neuronales artificiales, se utilizó un modelo de predicción basado en perceptrón multicapa (MLP) (Rumelhart & Hinton, 1986; Géron, 2022; Vállez Enano & Espinosa Aranda, s. f.). El cual está compuesto por *neuronas* que se organizan en capas, y cada neurona recibe información, que transforma y la envía hacia adelante (proceso *feedforward*), sin retrocesos. Cada neurona realiza una operación matemática en la que combina las entradas con 'pesos' (que indican qué tan importante es cada entrada), les suma un ajuste llamado sesgo, y luego aplica una función que decide cómo debe responder. Esta transformación se puede expresar como: $y = f(\sum w_i x_i + b)$, donde los x_i son las entradas, w_i los pesos, b el sesgo, y f(·) una función que introduce flexibilidad al modelo. A diferencia de otros métodos estadísticos aquí no se requiere suponer

una forma concreta de la relación entre las variables de entrada y de salida, ya que los MLP aprenden patrones complejos de los datos por sí solos, lo que los hace muy útiles para tareas donde las relaciones no son evidentes o no se pueden describir fácilmente con ecuaciones simples. Para la predicción de los porcentajes de estados fenológicos, se desarrollaron cuatro modelos con distintas combinaciones de variables de entrada y salida, como se indica en la tabla 4.

Table 4. Configuración de variables para modelado.

Configuración	Variables Entrada	Variables Salida
1	P1, P4	P2, P3
2	Semana, P1, P4	P2, P3
3	P1, P2	P3, P4
4	Semana P1, P2	P3, P4

Pi= Porcentaje de la producción i semanas adelante de la semana de conteo, Semana=Semana de conteo. Relacionar con lo presentado en la tabla 1.

Para la modelación, se implementaron diferentes combinaciones de capas ocultas, definidas mediante pruebas exploratorias que incluyeron arquitecturas de dos y tres capas, a partir de los cuales se obtuvieron (20, 80, 20), (200, 200, 200), (80, 80) y (250, 250, 250), respectivamente para cada configuración de la tabla 4. El entrenamiento de cada modelo se limitó a un máximo de 1.000 iteraciones, utilizando la función de activación ReLU, el optimizador Adam y la función de pérdida *mean_squared_error*. Previamente, las variables de entrada y salida fueron normalizadas mediante *StandardScaler*, y las predicciones se reescalaron a su dominio original para facilitar su interpretación. La evaluación del desempeño se realizó mediante validación cruzada de cinco particiones (*KFold*, n_splits = 5, random_state = 42), calculando las métricas de error cuadrático medio (MSE) y coeficiente de determinación (R^2), tanto de forma global como individual para cada variable de salida.

Extreme Gradient Boosting (XGBoost). También se aplicó el algoritmo Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016; Carrasco, Bueno & Montero, s. f.; XGBoost developers, 2022), una técnica de aprendizaje automático reconocida por su alta eficiencia computacional y su capacidad para modelar relaciones complejas no lineales. XGBoost es de naturaleza no paramétrica y permite la captura de estructuras de datos sin imponer una forma funcional específica. El algoritmo construye un conjunto de árboles de decisión de forma secuencial, donde cada nuevo árbol corrige los errores residuales de los árboles anteriores. La

optimización se realiza por gradientes de una función objetivo diferenciable, que combina una pérdida cuadrática y un término de regularización, controlando el sobreajuste y mejorando la capacidad de generalización del modelo.

Se desarrollaron cuatro modelos XGBoost utilizando las mismas combinaciones de variables de entrada y salida que se definieron para los modelos de MLP (tabla 4). Para la regresión multiobjetivo, se empleó *MultiOutputRegressor*, que permitió entrenar un modelo XGBoost independiente para cada variable dependiente. Los hiperparámetros del XGBRegressor se ajustaron según la configuración de cada modelo: el número de árboles (*n_estimators*) fue de 150, 300, 150 y 200 para cada una de las cuatro configuraciones, respectivamente. La tasa de aprendizaje (*learning_rate*) se mantuvo constante en 0.1 y la profundidad máxima (*max_depth*) en 5 en todas las configuraciones.

RNN (Redes Neuronales Recurrentes) – LSTM. Se utilizó un enfoque predictivo basado en redes neuronales recurrentes, específicamente del tipo Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997; James et al., 2023), diseñadas para capturar dependencias temporales en datos secuenciales mediante la incorporación de mecanismos de memoria a largo y corto plazo. Considerando factores propios de las áreas de proyección del sector floricultor, se evaluaron dos configuraciones de entrada y salida como las indicadas en la tabla 5.

Table 5. Configuración de variables para modelado.

Configuración	Variables Entrada	Variables Salida
5	Semana, Pd1, Pd2, Pd3, Pd4, Pd5, Pd6	Pd7, Pd8, Pd9, Pd10
6	Pd1, Pd2, Pd3, Pd4, Pd5, Pd6	Pd7, Pd8, Pd9, Pd10

Pd1 a Pd6= Producción en cada una de las seis semanas previas a las semanas de interes, Pd7 a Pd10=Producción de cada una de las futuras cuatro semanas de interes. Semana= Semana previa a la primera semana con el registro de producción. (Ejemplo: Si Pd1 corresponde a la producción de la semana 30, entonces Semana=29.)

En todos los casos, los datos fueron previamente normalizados entre 0 y 1 mediante *MinMaxScaler* y reorganizados en una estructura tridimensional compatible con el modelo LSTM (15.591 muestras, un paso de tiempo hacia adelante y siete características en el caso de las variables de entrada de la configuración 5 y seis caracteristicas en el caso de las variables de entrada de la configuración 6).
Los modelos se estructuraron con una única capa LSTM, variando el número de neuronas de acuerdo con la configuración obteniendo 80 neuronas para la

configuración 5 y 100 neuronas para la configuración 6. Esta capa fue seguida por una capa densa de salida con cuatro nodos, correspondiente a las variables de salida (tabla 5) que representan la producción proyectada para las cuatro semanas siguientes. Se evaluaron diferentes configuraciones de entrenamiento, variando el número de épocas y optando así por 50, el tamaño de lote 32 y la función de activación ReLU.
La red se entrenó utilizando el optimizador *Adam* y la función de pérdida *mean squared error(MSE)*. La implementación se realizó en Python utilizando las librerías TensorFlow y Keras para la construcción y entrenamiento del modelo.

XGBoots. Como complemento metodológico para la predicción de valores secuenciales de producción, se empleó el algoritmo XGBoost. El conjunto de entrada y salida se construyó a partir de las configuraciones descritas en la Tabla 5, utilizando ventanas temporales deslizantes y un enfoque de regresión multivariable implementado mediante *MultiOutputRegressor*. Se evaluaron de manera estructurada distintas combinaciones de tasa de aprendizaje (*learning_rate* = 0.1, 0.05 y 0.01), número de estimadores (*n_estimators* entre 50 y 600) y profundidad máxima del árbol (*max_depth* entre 2 y 5), las cuales fueron comparadas mediante validación cruzada. Como resultado, se seleccionaron dos configuraciones finales con 500 y 600 árboles, respectivamente, manteniendo una tasa de aprendizaje de 0,1 y una profundidad máxima de 5 en ambos casos, por presentar el mejor desempeño predictivo.

Calidad y rigor estadístico

Para garantizar la calidad y el rigor del estudio, se implementó un marco metodológico estructurado que abarca desde la generación de los datos hasta la evaluación final de los modelos. El rigor no se basa en que los datos sintéticos repliquen perfectamente la realidad, sino en la reproducibilidad, objetividad y validez interna del diseño experimental comparativo. Los principales elementos que aseguran este rigor son: Generación de datos sistemática y reproducible: se crearon 300 series de datos sintéticos utilizando una semilla aleatoria fija (42), lo que permite verificar y replicar los resultados. Validación estadística: se aplicó la prueba Kolmogorov-Smirnov (K-S) para confirmar que las distribuciones de los datos simulados no difirieran significativamente de las series base, garantizando su coherencia estadística. Rigor en el modelado y la evaluación: los datos se dividieron de manera estandarizada en 80 % para entrenamiento y 20 % para prueba, usando una semilla fija, y se aplicó validación cruzada de cinco particiones (5-fold) para evitar el sobreajuste. Métricas objetivas**:** el desempeño de los modelos se evaluó mediante el Error Cuadrático Medio (MSE) y el Coeficiente de

Determinación (R^2), lo que permitió una comparación directa y cuantitativa de los resultados.

Validez de los datos y consideraciones éticas

La validez de los datos en un estudio de modelado debe evaluarse según el propósito de la investigación. En este caso, centrado en la comparación metodológica de arquitecturas de aprendizaje automático, la validez de los datos sintéticos se define por su capacidad para reproducir las características estructurales y dinámicas del fenómeno analizado, más que por replicar fielmente la producción real actual.

Validez estructural y propósito. En tanto a la validez estructural, aunque los datos base de Calderón (2005) representan la demanda, reflejan la estacionalidad y los ciclos que caracterizan la planificación productiva en la floricultura de exportación. Al conservar la tendencia central y los rangos de variación del conjunto original, los datos sintéticos resultan estructuralmente válidos como un entorno de prueba realista para los modelos. Por otra parte en cuanto a la validez para el propósito(Fit-for-Purpose) el conjunto de datos fue diseñado específicamente para cumplir el objetivo del estudio, evaluar cuál de los enfoques de modelado —fenológico o autorregresivo— ofrece mayor eficacia en la predicción de series temporales con estas características.

Consideraciones sobre la protección de datos (habeas data)

Se revisaron los marcos normativos aplicables, incluida la Ley de Habeas Data (Ley 1581 de 2012); no obstante, se concluyó que esta regulación no aplica al presente estudio por las siguientes razones:

Naturaleza de los datos: la Ley de Habeas Data protege información personal de individuos identificables, mientras que los datos empleados son de carácter comercial y agregado. Fuente pública: los datos base provienen de una tesis de maestría de acceso público, no de una base privada con información sensible. Datos sintéticos: el conjunto final fue generado artificialmente, sin contener información real de personas o entidades, por lo que no implica riesgos de privacidad. La principal consideración ética del estudio fue mantener la transparencia y reproducibilidad del proceso de generación de datos, garantizada mediante el uso de algoritmos explícitos y semillas aleatorias fijas.

RESULTADOS Y DISCUSIÓN

Caracterización y validación de datos simulados

Los conjuntos de datos simulados (figura 4B) presentan una producción entre 758 y 164.336 tallos, los cuales se encuentran dentro del intervalo de producción de los datos de referencia (figura 4A). Además, se identifican comportamientos similares en la forma de distribución, ya que se observa un sesgo a la derecha y una menor dispersión entre el valor mínimo y la mediana en ambos casos. Los datos simulados incluyen un conjunto de valores atípicos que se mantienen dentro del mismo rango de los datos originales.

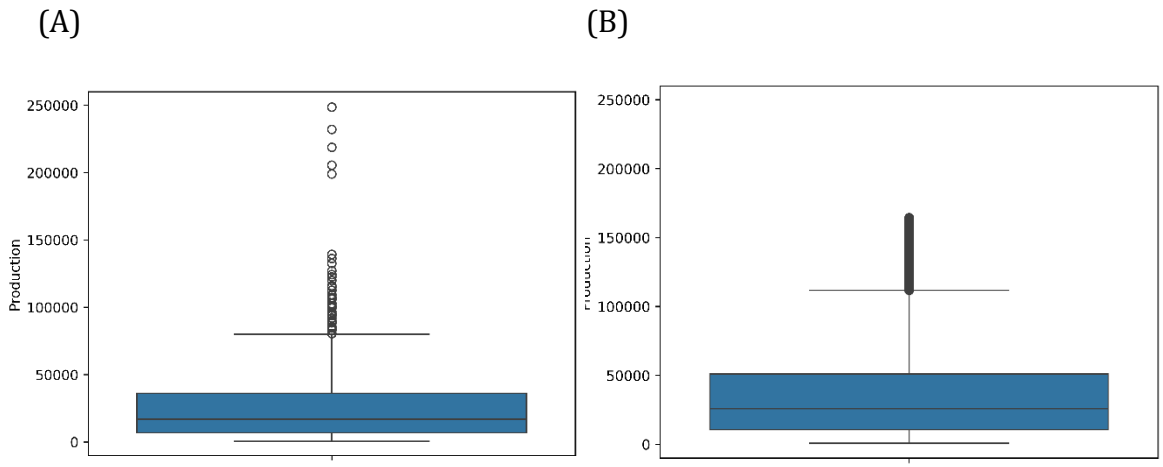


Figure 4. Boxplot de la producción de tallos (A) total datos de referencia (B) total datos simulados.

Aunque la producción simulada contiene valores distantes (Tabla 6) a los de referencia, se presenta una desviación similar en ambos conjuntos de datos y una distribución comparable a la descrita anteriormente, lo que permite identificar que, en primera instancia, los resultados no difieren drásticamente.

Table 6. Medidas estadísticas de datos de referencia y datos simulados

Conjunto de datos	Total de datos	Promedio	Desviación estándar	Mínimo	P_{25}	P_{50}	P_{75}	Máximo
Original /Referencia	1196	26598	30493	423	6787	16923	36122	248568
Simulado	15600	36196.8	34161.2	758	10593.5	25765	51051.5	164336

Para determinar si los datos simulados siguen la misma distribución que los datos de referencia, se realizó la prueba Kolmogorov–Smirnov, comparando cada uno de los 23 conjuntos de datos originales con cada uno de los 300 conjuntos de datos simulados. Esto con el propósito de verificar si cada curva simulada sigue la misma distribución que al menos una de las curvas originales. En la figura 5 se observa la distribución del porcentaje de veces que cada conjunto de datos simulado presentó una distribución estadísticamente equivalente a los conjuntos de datos originales, con p-value < 0,05 o incluso p-value < 0,01. Encontrando, que los porcentajes estuvieron por encima del 0 %, lo que indica que cada conjunto simulado presentó una distribución estadísticamente equivalente, al menos respecto a uno de los conjuntos de datos de referencia.

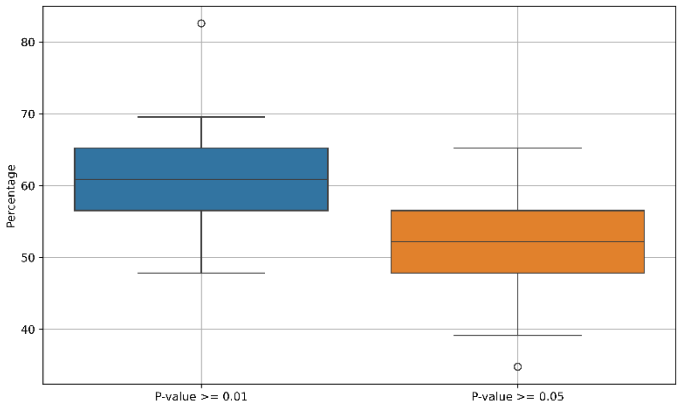


Figure 5. Distribución del porcentaje de curvas originales a las que una curva simulada es estadisticamente igual.

Con el fin de evaluar la coherencia temporal entre la serie original de producción y las series simuladas, se calculó la Función de Correlación Cruzada (CCF) para distintos rezagos temporales, con el propósito de analizar el grado de alineación entre ambas series. Los resultados muestran picos positivos altos y

estadísticamente significativos en el rezago cero, indicando una fuerte sincronía entre la serie original y las simuladas. Asimismo, se observan picos secundarios coherentes con la periodicidad estacional, lo que evidencia la conservación de la estructura estacional anual. En conjunto, la CCF confirma que las series simuladas reproducen el comportamiento temporal dominante de la serie original sin introducir desfases sistemáticos, respaldando la idoneidad de la estrategia de simulación adoptada.

Figure 5. Distribución del porcentaje de curvas originales a las que una curva simulada es estadisticamente igual.

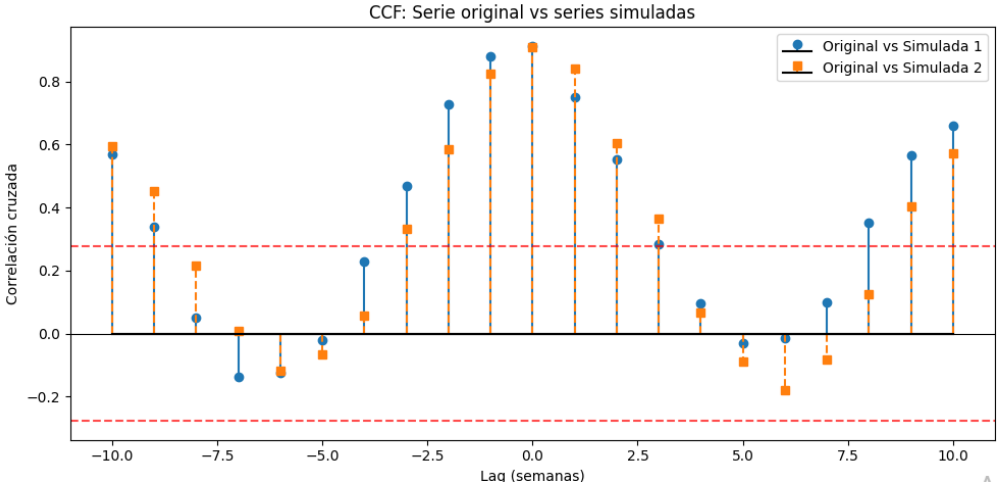


Figure 6. Función de correlación cruzada (CCF) entre la serie original y las series simuladas de producción.

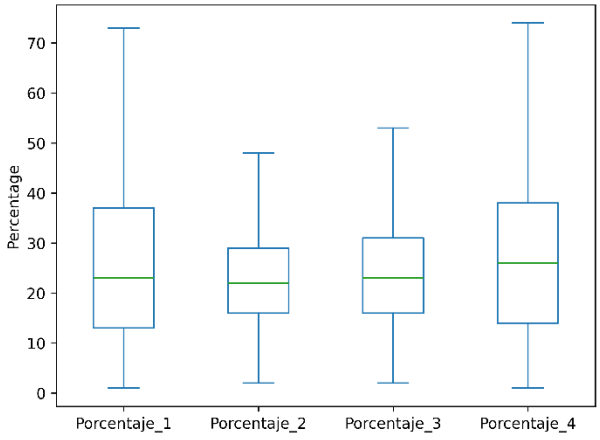
Evaluación del diseño experimental comparativo

En esta sección se presentan los resultados del diseño experimental comparativo, diseñado para evaluar las dos hipótesis de modelado propuestas. Primero, se analizó el rendimiento de los modelos que busca potenciar la hipótesis fenológica tradicional (configuraciones 1-4), donde los porcentajes de producción actúan como proxies de los conteos manuales. Posteriormente, se evaluó el rendimiento del enfoque alternativo basado en series temporales (configuraciones 5-6), que es agnóstico al conocimiento fenológico y se basa únicamente en la historia de la producción. La comparación directa del mejor modelo de cada enfoque permitirá responder a la pregunta de investigación central.

Evaluación del enfoque basado en la hipótesis fenológica (hipótesis 1)

La distribución de los porcentajes de estados fenologicos se muestra en la figura 6, en donde el porcentaje de la producción a una y cuatro semanas tiene mayor

rango de dispersión con máximos que no superan el 80% mientras, los porcentajes de la producción a dos y tres semanas son menos dispersos y no superan el 60%, además de presentar sesgos a la derecha principalmente los dos porcentajes inicialmente en mención. Sin embargo, esto no implica que los porcentajes se relación siempre igual en cada conjunto de semanas, como se observa en la tabla 7.



Porcentaje i (Pi) : De acuerdo con lo indicado en la tabla 1.

Figure 7. Distribución de los porcentajes de producción en cada uno de los momentos de conteo (relacionado con la información de la Tabla 1), calculados a partir de los datos simulados.

Table 7. Representación de la obtención de porcentajes de producción

Id conjunto de datos	Semana de Conteo	Producción a una semana	Producción a dos semanas	Producción a tres semanas	Producción a cuatro semanas	Total	Porcentaje_1	Porcentaje_2	Porcentaje_3	Porcentaje_4
0	52	18552	33944	35728	64695	152919	12	22	23	42
1	1	33944	35728	64695	84636	219003	15	16	30	39
2	2	35728	64695	84636	118045	303104	12	21	28	39
3	3	64695	84636	118045	103893	371269	17	23	32	28
4	4	84636	118045	103893	153135	459709	18	26	23	33
...	...	...	...	...	...	...	...	...	...	...
15592	44	42302	6268	2790	2076	53436	79	12	5	4
5593	45	6268	2790	2076	1179	12313	51	23	17	10

15594	46	2790	2076	1179	2193	8238	34	25	14	27
15595	47	2076	1179	2193	6939	12387	17	10	18	56
15596	48	1179	2193	6939	11143	21454	5	10	32	52

La producción se obtuvo de los datos simulados.

En el marco de la experiencia práctica del sector floricultor, en este estudio la propuesta inicial se orientó a estimar los porcentajes de producción como en la configuración 1 (tabla 4). Sin embargo, con el propósito de mejorar el desempeño, se incorporó la variable 'semana del conteo', conformando la 'configuración 2', que mostró un mejor ajuste tanto en MSE como en R^2. Posteriormente, se replanteó la lógica de variables de entrada y salida, dado que depende de conteos a una distancia de cuatro semanas podía generar inconsistencias en la predicción debido a variaciones propias del manejo y factores ambientales en periodos prolongados. En consecuencia, se propuso utilizar los porcentajes como en la 'configuración 3' y la 'configuración 4'.

Lo analizado hasta el momento se abordó mediante MLP; no obstante, se decidió contrastar los resultados con otra metodología para evaluar posibles mejoras en los ajustes de los pronósticos, seleccionando XGBoost. Este último ha demostrado un desempeño superior frente a las redes neuronales, alcanzando mayores niveles de precisión y menores errores en distintas aplicaciones debido posiblemente a la presencia de datos tabulares(Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022).

Los modelos XGBoost superaron a los modelos MLP en todas las configuraciones (Tabla 8), ya que presentaron mayores coeficientes de determinación (R^2) y menores errores cuadráticos medios (MSE). La configuración 4 obtuvo el mayor R^2 conjunto (0.835 en entrenamiento y 0.809 en prueba). A nivel de variables individuales, el R^2 máximo fue de 0.908 en entrenamiento y 0.892 en prueba para P4, por lo que esta configuración resultó óptima cuando se prioriza la precisión en una única salida. Por su parte, la configuración 2 alcanzó el menor MSE en prueba (18.32), con un R^2 conjunto de 0.824 (entrenamiento) y 0.79 (prueba), lo que representó el mejor equilibrio para predecir simultáneamente P2 y P3 con error reducido, aunque con un R^2 inferior al de la configuración 4. Un patrón similar se observó con MLP. En todos los casos, se notó que la inclusión de la variable semana mejoró de forma consistente el ajuste de los modelos.

Table 8. Resultados de predicción de porcentajes con MLP y XGBoost.

Metodología	Configuracion	Variables de Entrada	Variables de Salida	MSE				R^2			
				Train	Test	Train Conjunto	Test Conjunto	Train	Test	Train Conjunto	Test Conjunto
MLP	1	P1; P4	P2; P3	43.4 43.31	43.91 43.79	43.36	43.85	0.588 0.581	0.583 0.576	0.584	0.579
	2	Semana; P1; P4	P2; P3	21.29 21.33	22.94 22.98	21.31	22.96	0.798 0.794	0.782 0.777	0.796	0.779
	3	P1; P2	P3; P4	54.92 55.0	55.36 55.54	54.96	55.45	0.468 0.794	0.464 0.791	0.631	0.628
	4	Semana; P1; P2	P3; P4	27.63 28.02	29.82 30.19	27.82	30.01	0.733 0.895	0.711 0.887	0.814	0.799
XGBoost	1	P1; P4	P2; P3	42.46 42.52	44.12 43.97	42.49	44.05	0.597 0.588	0.581 0.574	0.592	0.577
	2	Semana; P1; P4	P2; P3	18.34 18.29	21.95 21.83	18.32	21.89	0.826 0.823	0.791 0.789	0.824	0.79
	3	P1; P2	P3; P4	53.88 54.13	55.45 55.97	54.01	55.71	0.478 0.797	0.469 0.789	0.637	0.626
	4	Semana; P1; P2	P3; P4	24.58 24.57	28.3 28.86	24.57	28.58	0.762 0.908	0.726 0.892	0.835	0.809

MLP= perceptrón multicapa , MSE= Error Cuadratico Medio, XGBoost= , Pi= Porcentaje de la producción i semanas adelante de la semana de conteo, Semana=Semana previa de conteo. Configuración: representa la combinación de valores de entrada y de salida.

En resumen, los resultados de este primer enfoque demuestran que es posible modelar la relación fenológica con un grado de precisión considerable, validando parcialmente la Hipótesis 1. El modelo XGBoost en la configuración 4, que utiliza la semana y los porcentajes de producción de las semanas 1 y 2 para predecir los de las semanas 3 y 4, representa el rendimiento óptimo alcanzable bajo el paradigma de modernización del conocimiento agronómico tradicional, logrando un R^2 en prueba de 0.809. Este resultado establece el punto de referencia contra el cual se evaluará el enfoque alternativo.

Evaluación del enfoque agnóstico de series temporales (hipótesis 2)

Aunque no fue el enfoque central inicial, se exploró un método alternativo para responder a la pregunta de investigación desde una perspectiva puramente estadística, poniendo a prueba la Hipótesis 2. Este enfoque prescinde por completo de los porcentajes basados en conteos (la hipótesis fenológica) y trata la producción como una serie temporal pura, donde el pasado de la serie predice su futuro. Para ello, se implementaron modelos LSTM y XGBoost utilizando

únicamente registros de producción de semanas anteriores como variables de entrada (configuraciones 5 y 6).

Como se observa en la tabla 9, el modelo XGBoost demostró un rendimiento superior en comparación con el modelo LSTM, presentando un mejor coeficiente de determinación (R^2) y un menor error cuadrático medio (MSE) en ambas configuraciones (5 y 6). Específicamente, la configuración 5 del modelo XGBoost se destacó por obtener el mayor R^2 conjunto (0.94 en entrenamiento y 0.82 en prueba). A nivel de predicción de variables individuales, esta misma configuración alcanzó el R^2 más alto, con valores de 0.94 en entrenamiento y 0.82 en prueba para las cuatro variables de salida. Esta configuración es óptima cuando se busca la máxima precisión en la predicción. También se observó que la inclusión de la variable de entrada 'semana' fue fundamental para mejorar significativamente el rendimiento en ambos modelos.

Table 9. Resultados de predicción de la producción semanal con LSTM y XGBoost.

Metodología	Configuración	Variables de Entrada	Variables de Salida	MSE				R^2			
				Train	Test	Train Conjunto	Test Conjunto	Train	Test	Train Conjunto	Test Conjunto
LSTM	5	Semana Pd1, Pd2, Pd3, Pd4, Pd5, Pd6	Pd7, Pd8, Pd9, Pd10.	225248212.6 216241660.8 213945960.3 214384000.5	228916160.1 220506365.3 217021466.6 216892099.7	217454958.6	220834022.9	0.807 0.815 0.816 0.816	0.804 0.811 0.814 0.814	0.814	0.811
	6	Pd1, Pd2, Pd3, Pd4, Pd5, Pd6	Pd7, Pd8, Pd9, Pd10.	334362513.4 397650048.1 452522025.7 495530948.6	343372954.3 407302323.8 462448055.8 506203323.3	420016383.9	429831664.3	0.713 0.651 0.612 0.575	0.705 0.651 0.603 0.566	0.639	0.631
XGBoost	5	Semana Pd1, Pd2, Pd3, Pd4, Pd5, Pd6	Pd7, Pd8, Pd9, Pd10.	65522098.1 67364158.6 70046081.5 70841019.5	214397804.7 209938578.7 207773098.6 206457639.3	68443339.4	209641780.3	0.944 0.942 0.939 0.939	0.816 0.819 0.822 0.822	0.941	0.820
	6	Pd1, Pd2, Pd3, Pd4, Pd5, Pd6	Pd7, Pd8, Pd9, Pd10.	95487326.1 113463078.2 122879734.9 123622774.3	321358176.1 351379914.3 358774034.0 360406425.8	113863228.4	347979637.5	0.918 0.902 0.895 0.893	0.724 0.698 0.692 0.691	0.902	0.701

Pd_1 a Pd_6= Producción en cada una de las seis semanas previas a las semanas de interes, Pd_7 a Pd_10=Producción de cada una de las futuras cuatro semanas de interes. Semana= Semana previa a la primera semana con el registro de producción.

Para contextualizar estos resultados, se tomó como referencia el estudio de Mejía Giraldo y Páez Barreto (2023). En su investigación, se trabajó con varios modelos, entre ellos el modelo SARIMAX, que obtuvo un R^2 de 0.81, y Gradient Boosting, con un R^2 de 0.59. En contraste, los modelos XGBoost evidenciaron un ajuste mejor, con

valores de R^2 de 0.82, lo cual indica la superioridad de este enfoque en la precisión de los pronósticos.

Evaluación comparativa La comparación de los resultados de ambos enfoques es concluyente. El mejor modelo del enfoque fenológico (XGBoost, config. 4) alcanzó un R^2 de 0.809, mientras que el del enfoque de series temporales (XGBoost, config. 5) obtuvo un R^2 de 0.820 en la predicción directa de la producción semanal. Este resultado demuestra que, aunque los modelos de Machine Learning capturan información útil de los datos fenológicos, la historia de la producción —por su autocorrelación y patrones temporales— ofrece una señal predictiva más fuerte. El enfoque autorregresivo resulta más preciso y simple, al eliminar la dependencia del conteo manual. En síntesis, los resultados confirman la superioridad del algoritmo XGBoost y validan una estrategia de modelado más eficaz para este problema.

Alcances y limitaciones del estudio

Es fundamental interpretar los hallazgos de esta investigación dentro de su alcance metodológico específico y reconocer sus limitaciones inherentes, las cuales fueron decisiones de diseño tomadas para garantizar la validez interna del estudio ante restricciones del mundo real. En tanto a los alcances: el estudio realizó una evaluación comparativa de distintas arquitecturas de aprendizaje automático para predecir series temporales con patrones dinámicos y estacionales de la producción florícola. Su aporte principal fue identificar un *pipeline* analítico y un modelo XGBoost en modo autorregresivo como la opción más eficaz en un entorno controlado, ofreciendo un motor predictivo validado aplicable con datos reales del sector. En cuanto a las limitaciones, las principales se relacionan con el uso de datos sintéticos, lo cual restringe la validez externa aunque permite una comparación rigurosa bajo condiciones controladas. Adicionalmente, la simulación de los datos se realizó mediante un esquema independiente por semana, lo que permitió preservar la estacionalidad anual y los rangos empíricos de variabilidad observados en la producción, pero implicó una simplificación de la dependencia temporal de corto plazo al no reproducir explícitamente posibles interrelaciones intra-anuales entre semanas consecutivas. En consecuencia, los resultados asociados a la Hipótesis 2 deben interpretarse considerando que la dinámica temporal representada prioriza el comportamiento estacional global del sistema productivo. Asimismo, la ausencia de variables exógenas —como factores climáticos, nutricionales o de manejo— respondió a la dificultad de simular dichas interacciones y a la limitada disponibilidad de datos públicos.

En síntesis, aunque los datos simulados fueron estadísticamente validados, no sustituyen la variabilidad real de un cultivo. El estudio se centró en validar un motor analítico robusto (XGBoost autorregresivo) bajo condiciones experimentales controladas.

Recomendaciones y trabajo futuro

Futuras investigaciones deberían combinar la robustez de XGBoost con modelos capaces de representar secuencias complejas e incorporar variables climáticas, nutricionales o de manejo, con el fin de mejorar la precisión predictiva. Así mismo, es pertinente validar la metodología en otros cultivos para evaluar su escalabilidad y ampliar los horizontes de predicción optimizando la configuración de los modelos. A partir de los hallazgos de esta investigación, se derivan una serie de recomendaciones específicas para el campo de la Estadística Aplicada en la agroindustria.

Recomendaciones para la praxis profesional en la agroindustria.

Reorientación estratégica de la fuente de datos: Se recomienda a los planificadores de producción del sector floricultor priorizar la recolección y mantenimiento de bases de datos históricas de producción semanal que sean limpias y consistentes. Este enfoque de bajo costo y alto impacto aprovecha los datos que las empresas ya generan. Adopción de algoritmos de ensamble para modelado: Se recomienda la adopción de XGBoost como una herramienta robusta, eficiente y de alta precisión para el desarrollo de modelos de pronóstico de producción. Simplificación operativa: Se sugiere reducir la dependencia del costoso y subjetivo proceso de conteo manual en campo, enfocando los esfuerzos en el análisis de datos históricos para una mayor eficiencia.

Recomendaciones para la investigación futura

Investigación de las causas de la superioridad del enfoque autorregresivo: Futuros estudios deben cuantificar el nivel de ruido introducido por el conteo manual o analizar la "inercia" y la autocorrelación del proceso productivo para entender por qué la señal de la serie temporal es dominante. Desarrollo de modelos híbridos: Habiendo establecido la fuerza del componente autorregresivo como línea base, se recomienda investigar modelos híbridos. Estos modelos podrían utilizar la predicción de la serie temporal como predictor principal e incorporar variables exógenas (como factores agroclimáticos y de manejo) para modelar las desviaciones o residuos, buscando mejoras marginales en la precisión. Exploración de técnicas de aumento de datos: Dada la dificultad para acceder a datos en el sector, se recomienda explorar técnicas avanzadas de generación de datos

sintéticos para series temporales, como las Redes generativas antagónicas (e.g., TimeGAN), para entrenar modelos más robustos y generalizables.

CONCLUSIONES

La presente investigación comparativa metodológica alcanzó sus objetivos al evaluar la eficacia de dos enfoques de Machine Learning para el pronóstico de la producción semanal de rosas, basándose en datos simulados que conservan la estacionalidad del sector.

Hallazgo original y conclusión metodológica principal

El estudio produce un hallazgo original y contraintuitivo con implicaciones directas para la industria floricultora colombiana:

Superioridad del enfoque autorregresivo: La conclusión es que un modelo autorregresivo, basado únicamente en la historia de la producción (series temporales), supera consistentemente a los modelos que incorporan información fenológica. Este resultado desafía la premisa de que los conteos manuales son la fuente de información predictiva más valiosa. Validación de la Hipótesis 2: El mejor modelo del enfoque agnóstico de series temporales (XGBoost, Configuración 5) logró un coeficiente de determinación (R^2) de 0.820 en la predicción directa de la producción de las cuatro semanas siguientes, utilizando como insumo los registros de las seis semanas anteriores. Este rendimiento supera el mejor resultado del enfoque fenológico ($R\text{^}2 = 0.809$), confirmando la superioridad de la señal predictiva contenida en la autocorrelación de los datos históricos, aunque se debe advertir que los resultados se interpretan considerando las restricciones derivadas del uso de datos sintéticos y la estructura temporal artificialmente reproducida mediante Block Bootstrapping. Aporte a la competitividad: La implementación de este enfoque de series temporales puede aportar un valor significativo al sector floricultor, mejorando la competitividad de las empresas y reduciendo las pérdidas económicas derivadas de errores de proyección, siempre bajo la consideración de las limitaciones de validez externa de los datos simulados y la exclusión de variables agroclimáticas exógenas. Además, en la medida en que sea posible obtener datos reales, los hallazgos podrían validarse empíricamente, confirmando que las tendencias observadas hasta el momento se mantienen bajo condiciones reales de producción.

Eficacia del algoritmo xgboost y eficiencia operacional

Algoritmo óptimo: Los modelos basados en XGBoost superaron de forma consistente a los modelos MLP y LSTM, lo que confirma que, para datos tabulares y series temporales cortas, los métodos de ensamble resultan más robustos y menos sensibles a la variabilidad. Simplificación operativa: Desde una perspectiva práctica, la aplicación de este método autorregresivo permite optimizar recursos operativos al prescindir de procedimientos manuales como el conteo de tallos, y concentrar el análisis en los registros históricos disponibles. Así, se logra no solo una mayor eficiencia en la estimación de la producción, sino también una reducción en los costos asociados a la recolección de datos. Contribución original: Este trabajo representa una de las primeras investigaciones rigurosas que aplica y compara sistemáticamente estas arquitecturas de ML para el problema específico de la predicción de producción de rosas, utilizando como punto de partida los datos derivados de los conteos fenológicos y simulaciones validadas, generando conocimiento original sobre cómo maximizar el valor de los datos disponibles en el sector.

Esta investigación evidencia el valor de abordar el problema del pronóstico productivo desde enfoques alternativos a los tradicionalmente dominantes en la literatura, como lo son las aproximaciones basadas en tecnologías de automatización de alto costo, que hacen uso de drones o visión artificial, y de los modelos fisiológicos sustentados en grados día o acumulación térmica —frecuentemente limitados a condiciones específicas y poco generalizables—, el enfoque propuesto demuestra que es posible extraer información predictiva relevante a partir de datos que ya forman parte de la operación cotidiana de las fincas. En este sentido, el estudio consolida una propuesta metodológica accesible, de bajo costo y coherente con la incertidumbre biológica real, que se adapta de manera efectiva a la realidad productiva del país y amplía el marco de soluciones disponibles para el sector floricultor, complementando y cuestionando los enfoques convencionales actualmente utilizados.

AGRADECIMIENTOS

A Dios, por dar guia a mi camino cada día y brindarme la sabiduría necesaria para llevar a cabo esta investigación. A mi esposo, por su incondicional apoyo y paciencia; a mi madre, por su esfuerzo y amor; a mi abuela y hermana, por sus valiosos consejos y ánimo permanente.
Agradezco a la profesora Edna Moreno, por motivarme a perseverar en el desarrollo de esta investigación y alentarme en el uso de datos sintéticos como

parte de la propuesta; al profesor Wilmer Pineda, por su orientación en la construcción y desarrollo global del estudio; y al profesor Isaac Zainea, por su valioso acompañamiento y guía en el diseño e implementación de los algoritmos de *machine learning* necesarios para esta investigación.

Conflicto de intereses: Los autores declaran que no existe ningún conflicto de intereses.

REFERENCIAS BIBLIOGRAFICAS

Albán Bautista, J. F., & Zabala Chico, D. M. (2022). *Desarrollo de un modelo de machine learning en la nube para mejorar la producción de una plantación de rosas* [Tesis de maestría, Universidad Politécnica Salesiana]. Repositorio Institucional UPS. https://dspace.ups.edu.ec/bitstream/123456789/24255/1/MSQ511.pdf

Araújo, S., Peres, R., Ramalho, J., Lidón, F., & Barata, J. (2023). Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy*. https://doi.org/10.3390/agronomy13122976.

Asocolflores. (2023, 23 de junio). *Asocolflores celebra 50 años impulsando las exportaciones, el trabajo formal y el desarrollo de las zonas rurales de Colombia*. Asocolflores.

https://asocolflores.org/es/asocolflores-celebra-50-anos-impulsando-las-exportaciones-el-trabajo-formal-y-el-desarrollo-de-las-zonas-rurales-de-colombia/

Asocolflores. (s. f.). *Sector floricultor*. Asocolflores. https://asocolflores.org/sector-floricultor/

B., A., R., O., & G., R. (2019). Current status of the OEE (overall equipment effectiveness) indicator in the operations of an export-type rose postharvest. *Cuadernos de Semilleros de investigación.* https://doi.org/10.33133/csi-7-2021-56

Bahuguna, S., Anchal, S., Chandel, A., Devi, M., Bhargava, B., & Kumar, A. (2020). Automated flower enumeration, a felicitous method developed for the floriculture industries. Flower and Ornamental Plants, 5(1), 51–60. https://doi.org/10.52547/flowerjournal.5.1.51

Cabrera Loja, J. (2021). *Modelos de predicción de producción basados en el Método de Grados Días de Desarrollo en tres variedades de Rosa Sp.* [Trabajo de grado, Universidad Central del Ecuador]. UCE. http://www.dspace.uce.edu.ec/handle/25000/25029

Calderón Novoa, F. (2005). *Modelo de optimización para la planeación de producción de un cultivo de rosas*. Uniandes. https://hdl.handle.net/1992/22705

Carangui, C., et al. (2024). Advanced algorithm for automated red rose counting using image processing techniques. 2024 IEEE Colombian Conference on Communications and Computing (COLCOM), 1–6. https://doi.org/10.1109/COLCOM62950.2024.10720327

Carrasco, R. A., Bueno, I., & Montero, J.-M. (s. f.). Boosting y el algoritmo XGBoost (Cap. 29). En *Fundamentos de ciencia de datos con R*. https://cdr-book.github.io/cap-boosting-xgboost.html

Castro Forero, M., & Palomar Rodríguez, X. (2022). *Generación de herramientas para la predicción de cosechas en variedad comercial de rosa de jardín basados en las acumulaciones de grados día y radiación acumulada (DLI)* [Trabajo de

especialización, Universidad Jorge Tadeo Lozano]. Repositorio Institucional Utadeo. https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/31187/Trabajo%20final%20Rosa%20de%20Jardin.pdf?sequence=3&isAllowed=y

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

Contrera Mercado, A. V. (2024). *Seguimiento del cultivo de rosas en la etapa productiva: calidad, cantidad y manejo ambiental en la finca Falcon Farms de Colombia S.A. (Suesca – Cundinamarca)* [Trabajo de grado, Universidad de Córdoba]. Repositorio Institucional Universidad de Córdoba. https://redcol.minciencias.gov.co/Record/UCORDOBA2_6d0db042ba17595e434e250440d6732a

Corficolombiana. Dirección de Análisis Sectorial y Sostenibilidad. (2025, febrero 14). Las flores en tiempos de San Valentín: La importancia estratégica de la floricultura colombiana. Informe Especial. https://investigaciones.corfi.com/documents/38211/0/250203.pdf/2e0cbee1-1520-1b40-c28f-6d7c39e20bc0

Encalada Ruiz, A. S., & Rivadeneira Morales, F. D. (2020). *Gestión de la cadena de abastecimiento y la eficiencia de los procesos en la florícola Rosas del Monte, S.A.* [Tesis de pregrado, Universidad Politécnica Estatal del Carchi]. Repositorio Institucional UPEC. https://repositorio.upec.edu.ec/server/api/core/bitstreams/53e69a22-cc23-4298-a5f4-7b8629a8bfde/content

Géron, A. (2023). *Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (3.ª ed.). O'Reilly Media.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint.* https://arxiv.org/abs/2207.08815

Herrera, D., Escudero-Villa, P., Cárdenas, E., Ortiz, M. y Varela-Aldás, J. (2024). Combinación de clasificación de imágenes y vehículos aéreos no tripulados para estimar el estado de rosas exploradoras. *AgriEngineering* , *6* (2), 1008-1021. https://doi.org/10.3390/agriengineering6020058

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

International Fresh Produce Association. (2025). *A look back at 2025: Floral industry report.* International Fresh Produce Association. https://www.freshproduce.com/siteassets/files/floral/a-look-back-2025_floral.pdf

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python* (Springer Texts in Statistics). Springer. https://doi.org/10.1007/978-3-031-38747-0

Kumari, S., Jeble, S., & Patil, Y. B. (2018). Barriers to technology adoption in agriculture-based industry and its integration into technology acceptance model. International Journal of Agricultural Resources, Governance and Ecology, 14(4), 338–351.https://doi.org/10.1504/IJARGE.2018.10017116

Lai, Q., Yang, Z., Su, W., Yan, C., Zhao, Q., Tan, Y., Que, Y., & Zheng, J. (2025). Enhancement of the prediction of the openness of fresh-cut roses with an improved YOLOv8s model validated by an automatic grading machine. *Frontiers in Plant Science, 16*, 1546503. https://doi.org/10.3389/fpls.2025.1546503

Liu, G., Zhong, K., Li, H., Chen, T., & Wang, Y. (2024). A state of art review on time series forecasting with machine learning for environmental parameters in agricultural greenhouses. *Information Processing in Agriculture, 11*(2), 143–162. https://doi.org/10.1016/j.inpa.2022.10.005

Mantilla, F., Mejía, G., & Tascón, D. (2025). *The role of industry 4.0 technologies in the export flower industry: Insights from a systematic literature review and surveys in emerging economies*. **Results in Engineering, 25**, 104507. https://doi.org/10.1016/j.rineng.2025.104507

Mejía Giraldo, Á., & Páez Barreto, C. (2023). *Modelos de pronóstico de producción de rosa Freedom: Un enfoque predictivo para mejorar el cumplimiento y reducir el desperdicio en la finca María Bonita* [Trabajo de grado, Universidad de La Sabana]. Repositorio Institucional Unisabana. https://hdl.handle.net/10818/60269

Mora Quintero, A. G. (2019). *Evaluación de herramientas de seguimiento fenológicos y curvas de desarrollo, para las mejoras en el cumplimiento de indicadores en la producción de cultivos de rosa* [Trabajo de grado, Universidad de los Llanos]. Repositorio Universidad de los Llanos. https://repositorio.unillanos.edu.co/entities/publication/1828d1b3-2cb6-41b1-8483-25b604543c46

Mordor Intelligence. (2025, 10 de septiembre). *Colombia floriculture market size & share analysis: Growth trends and forecast (2025–2030).* Mordor Intelligence. https://www.mordorintelligence.com/industry-reports/colombia-floriculture-market

Nicosia, G., Ojha, V. K., & La Malfa, E. (Eds.). (2020). *Machine learning, optimization, and data science* (pp. vi–vii). Springer. https://doi.org/10.1007/978-3-030-64583-0

Perilla Garzón, M. F. (2019). *Informe de pasantía universitaria: Estados fenológicos en la producción de rosas en Flores El Tandil* [Informe de pasantía, Universidad Nacional Abierta y a Distancia – UNAD]. Repositorio Institucional UNAD. https://repository.unad.edu.co/jspui/bitstream/10596/27387/1/%09mfperillag.pdf

Piza, J. P. (2023). *Proyecto de grado Juan Pablo Piza – Final* [Tesis de pregrado, Universidad Piloto de Colombia]. Repositorio Institucional UNIPILOTO. https://repository.unipiloto.edu.co/bitstream/handle/20.500.12277/13096/PROYECTO%20DE%20GRADO%20JUAN%20PABLO%20PIZA-FINAL.pdf?sequence=1&isAllowed=y

ProColombia. (2025, 12 de febrero). *Colombia conquista San Valentín: más de 60.000 toneladas de flores llevan el país de la belleza al mundo.* ProColombia. https://procolombia.co/sala-de-prensa/noticias/colombia-conquista-san-valentin-mas-de-60000-toneladas-de-flores-llevan-el-pais-de-la-belleza-al-mundo

Rodríguez, G., Barrero, M., Calderón, J., & Cardoso, E. (2025). Factores determinantes de la competitividad de las empresas floricultoras exportadoras en Colombia. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales.* https://doi.org/10.36390/telos272.10

Rodríguez, W. E., & Flórez, V. J. (2006). Comportamiento fenológico de tres variedades de rosas rojas en función de la acumulación de la temperatura. *Agronomía Colombiana, 24*(2), 247–257. https://www.redalyc.org/pdf/1803/180316239006.pdf

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion, 81,* 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Trendeconomy. (2023). *Cut flowers (commodity code 0603): International trade statistics.* Trendeconomy. https://trendeconomy.com/data/commodity_h2/0603

Vállez Enano, N., & Espinosa Aranda, J. L. (s. f.). Redes neuronales artificiales (Cap. 36). En *Fundamentos de ciencia de datos con R.* https://cdr-book.github.io/capNN.html

XGBoost developers. (2022). Demo for using xgboost with sklearn [Ejemplo de documentación]. En *XGBoost Documentation.* Recuperado el 22 de agosto de 2025, de https://xgboost.readthedocs.io/en/stable/python/examples/sklearn_parallel.html#sphx-glr-python-examples-sklearn-parallel-py