Revista Electrónica Formación y Calidad Educativa
ISSN1390-9010
RefCalE
refcale@uleam.edu.ec
Uleam
Vicerrectorado Académico

DOI: https://doi.org/10.56124/refcale.v13i3.005

Análisis predictivo del rendimiento agrícola del cacao en el cantón Balao usando enfoques de aprendizaje profundo

Análisis predictivo del rendimiento agrícola del cacao

Autores:

Autor [1] Gabriel Jesús Vega Delgado

Autor [2] Yasmany Fernández Fernández

DIRECCIÓN PARA CORRESPONDENCIA:

Dirección: Av. La Paz 402 y callejón sin nombre

Nombres: Gabriel Jesús Vega Delgado

Correo: gvega@est.unibe.edu.ec

Telf.: 0987826252

Fecha de recepción: agosto 05, 2025

Fecha de aceptación: diciembre 20, 2025

[1] Gabriel Jesús Vega Delgado: Ingeniero en Tecnologías de la Información. Universidad Iberoamericana del Ecuador, UNIBE, gvega@est.unibe.edu.ec, https://orcid.org/0009-0007-5967-5570, Guayas, Ecuador.

[2]Yasmany Fernández Fernández: Ingeniero en Informática. Postgrado en Matemática Aplicada. Docente en Posgrados de la Universidad Iberoamericana del Ecuador, UNIBE, y Docente de la Universidad Politécnica Estatal de Carchi, UPEC, yfernandez@doc.unibe.edu.ec, yfernandezf@upec.edu.ec, https://orcid.org/0000-0002-9530-4028, Tulcán, Ecuador.

RESUMEN

La problemática abordada en esta investigación fue la necesidad de optimizar los procesos de cuidado del cultivo de cacao CCN-51 en el cantón Balao, provincia del Guayas, frente a la alta variabilidad de los fenómenos meteorológicos. Por ello, se planteó desarrollar un enfoque predictivo para las producciones agrícolas del cacao CCN-51 mediante la implementación de redes neuronales LSTM, con el objetivo de predecir los rendimientos agrícolas a partir de datos multifuentes. El estudio optó un enfoque cuantitativo, con un diseño no experimental-longitudinal. La selección de muestra se basó en el método no probabilístico por conveniencia y se trabajó con un total de 384 huertas de cacao. El proceso de trabajo comenzó con la recopilación de información de diversas fuentes, análisis de suelo, índices de vegetación, datos históricos y registros de producción. Obtenido el dataset se realizó el proceso EDA para normalizar las variables, se aplicó el método VIF para determinar las variables predictoras a utilizar y se desarrolló un modelo utilizando varias configuraciones de LookBack y Horizon acompañadas de métodos de evaluación. Los resultados obtenidos permitieron determinar que las redes LSTM procesan mucho mejor variables multifuentes que los métodos tradicionales, obteniendo un error RMSE aceptable, pero el modelo no captaba de buena forma los patrones de las variables presentando un coeficiente de determinación bajo, esto debido a la heterogeneidad de los datos. La investigación recalca el potencial de las redes LSTM pero también expone lo fundamental que es un dataset de calidad y con datos homogéneos.

PALABRAS CLAVES: Rendimiento agrícola; Aprendizaje profundo; Redes temporales LSTM; Predicción de cultivos; Series temporales.

Revista Electrónica Formación y Calidad Educativa
ISSN1390-9010
RefCalE
refcale@uleam.edu.ec
Uleam
Vicerrectorado Académico

Predictive analysis of cacao crop yield in Balao canton using deep learning approaches

ABSTRACT

The problem addressed in this research was the need to optimize the care processes of the CCN-51 cacao crop in Balao canton, Guayas province, due to the high variability of meteorological phenomena. Therefore, it was proposed to develop a predictive approach for the CCN-51 cacao agricultural productions through the implementation of LSTM neural networks, aiming to predict agricultural yields using multisource data. The study adopted a quantitative methodological approach, based on a non-experimental longitudinal design, since data expressed over a period of time was used without variable manipulation. The sample selection was based on a non-probabilistic convenience method, working with a total of 384 cacao plantations, which were selected for having the required information and proximity and availability to participate in the research. The work process began with data collection from various sources, soil analysis, vegetation indexes, historical data, and production records. Once the dataset was obtained, an exploratory data analysis (EDA) was performed to normalize variables, the variance inflation factor (VIF) method was applied to determine predictor variables, and a model was developed using several LookBack and Horizon configurations accompanied by evaluation methods. The results showed that LSTM networks process multisource variables much better than traditional methods, achieving an acceptable RMSE error; however, the model did not adequately capture the variable patterns, presenting a low coefficient of determination due to data heterogeneity. The research highlights the potential of LSTM networks while also emphasizing the critical importance of a high-quality and homogeneous dataset.

KEYWORDS: Agricultural Yield; Deep Learning; LSTM-based Temporal Neural Networks; Crop Yield Prediction; Multivariate Time Series.

INTRODUCCIÓN

El Theobroma cacao L. o cacao, conocido comúnmente, es uno de los cultivos más importantes en el Ecuador. La relevancia de este fruto radica en su aportación económica, social y cultural al país. Las variantes existentes son la de aroma fino y CCN-51 (Colección Castro Naranjal 51), esta última desarrollada en 1960 utilizando cruces controlados y mejoramiento genético. En las últimas décadas la variante CCN-51 se ha convertido en la más cultivada debido a que presenta resistencia a plagas/enfermedades, mejor adaptación a los diferentes climas del país y una alta y rápida producción (Jaimez et al., 2022). El Cantón Balao, en la provincia del Guayas, es una de las zonas donde la variedad CCN-51 tiene una representación relevante en la economía de los pequeños y medianos productores, de acuerdo con Sornoza y colaboradores (2022) los productores de esta zona han optado por esta variante como una alternativa ante las variaciones identificadas en el mercado internacional.

Pese a las ventajas, el cultivo de cacao CCN-51 actualmente enfrenta factores negativos concernientes a las variaciones de temperaturas en el ambiente, la rápida degradación del suelo, recurrencia a enfermedades como monilla (Moniliophthora roreri), escoba de bruja (Moniliophthora perniciosa), distintos tipos de musgos y mal de machete (Ceratocystis fimbriata), y malas prácticas agrícolas de los productores (Dos Santos Pereira et al., 2024). Todos estos factores perjudican el rendimiento de la plantación dificultando la proyección de producción a futuro.

Tradicionalmente, la estimación del rendimiento del cacao y de otros cultivos era a través de promedios basados en datos históricos o en modelos lineales simples. Estas técnicas eran eficientes en escenarios sin una alta complejidad o interacción de factores, pero en la actualidad la interacción y rápida variación de factores agroclimáticos, edáficos y fenológicos, sumado a aspectos como disponibilidad de nutrientes, técnicas de poda, control fitosanitario y prácticas de fertilización, han dejado obsoletos estos métodos.

En la actualidad, el auge de la tecnología permite acceder a datos emitidos por sensores agrícolas e información satelital, que sumado a los datos históricos se pueden desarrollar análisis predictivos sofisticados y con un menor margen de error, con la capacidad de procesar una gran cantidad de información en tiempo real. Deep Learning o aprendizaje profundo es una de las herramientas más utilizadas en la predicción del sector agrícola. Esta tecnología detecta y aprende patrones dentro de una gran cantidad de datos a través de algoritmos. Un claro ejemplo de esto son las redes neuronales

convolucionales (CNN), redes neuronales recurrentes (RNN) y las redes de memoria a largo plazo (LSTM), que mientras más grande o representativa sea la muestra, más preciso es el modelo predictivo.

Según la obra *"The Panoramic View of Ecuadorian Soil Nutrients (Deficit/Toxicity) from Different Climatic Regions and Their Possible Influence on the Metabolism of Important Crops"* se describe que Balao presenta condiciones agroecológicas óptimas para el cultivo de cacao CCN-51 (Mihai et al., 2023). Sin embargo, los agricultores tienen como problemática la planificación de las tumbas o cosechas debido a la carencia de alguna herramienta tecnológica que les permita tener una proyección bajo diferentes situaciones climáticas y varias condiciones de manejo o administración. Dicha carencia afecta también a la economía de la zona, pues los productores pierden la capacidad de negociar el producto con intermediarios.

La importancia de crear un modelo predictivo para los cultivos de cacao CNN-51 en el cantón Balao involucra integrar información climática, edafológica, fenológica y de administración agrícola. Adicional, permitiría a los productores complementar sus conocimientos y mejorar el proceso de toma de decisiones, despejando dudas ante los cambios climáticos.

La presente investigación se basa en la idea de que la agricultura es un procedimiento complejo que engloba factores en distintas escalas temporales y espaciales. La implementación de Deep Learning permite abarcar el fenómeno de estudio de una manera global, resolviendo la problemática con innovación tecnológica. Al plantear soluciones de este tipo ante un cultivo tropical como el cacao se encuentra la limitante de la recolección y sistematización de datos de calidad; pues en cantones como Balao, la información suele estar dispersa o carecer de algún modelo que estandarice la misma. Otra limitante es la carencia de capacitaciones técnicas a los productores o capataces, por lo que una solución con modelos predictivos involucra aspectos técnicos, sociales y educativos.

El trabajo de campo también se ve afectado por los sucesos dados en el cambio climático, hecho que afecta de manera directa la agricultura del país. Al adoptar una solución de este tipo permite reducir pérdidas

económicas y desarrollar estrategias con el objetivo de que el cultivo no se vea afectado gravemente. En base a (Bowen Quiroz & Medranda Cobeña, 2024), la Organización de las Naciones Unidas para la Alimentación y la Agricultura promueve la técnica de agricultura inteligente frente a las complicaciones del clima.

El objetivo central de la investigación es la predicción del rendimiento agrícola del cultivo de cacao CCN-51 en base a modelos de aprendizaje profundo. Con esto se definen fines específicos asociados al proceso de recolección y tratamiento del dataset, entrenar el modelo predictivo y evaluar el mismo. La hipótesis definida establece que un modelo predictivo basado en redes neuronales LSTM tiene la capacidad de procesar de mejor forma variables multifuentes, complejas y no lineales que en comparación a los procedimientos estadísticos tradicionales.

Entre los principales materiales a utilizar en la investigación están los registros climáticos históricos del cantón, análisis de suelos realizados por los productores, datos fenológicos del cacao CCN-51 y prácticas agrícolas registradas por observación y entrevistas. El dataset obtenido será refinado a través del proceso EDA, una vez limpios serán utilizados para el entrenamiento del modelo predictivo. La metodología aplicada será un enfoque cuantitativo acompañado de un diseño experimental aplicado, con el fin de evaluar el rendimiento del modelo con el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2).

Big Data en la agricultura

Big Data es un grupo de herramientas tecnológicas utilizadas para almacenar, administrar y analizar grandes volúmenes de datos, estén estos estructurados o no. Dentro del sector agrícola, el Big Data ofrece soluciones transformadoras frente a la situación que enfrentan los cultivos con respecto al cambio climático y la demanda que cada vez va en aumento. La principal actividad de Big Data dentro de la agricultura es analizar gran cantidad de datos, de diversas fuentes como sensores, registros históricos, imágenes satelitales, etc.; para así entender los procesos en los cultivos y mejorar la toma de decisiones hacia los mismos en varios posibles escenarios.

Cravero y colegas (2022) destacan que las fuentes de los datos en los cultivos son varios y pueden ser estructurados o no. Los investigadores identifican fuentes como sensores, los cuales miden en tiempo real la temperatura del aire, humedad, pH del suelo, etc. Otra fuente son las cámaras, los sistemas de GPS y bases de datos históricas. Sumado al

conocimiento adquirido por los productores a lo largo del tiempo, se puede fortalecer los algoritmos de Big Data.

La implementación de Big Data en el campo permite proyectar la producción del cultivo, la detección de plagas y mejorar las técnicas de cuidado (Cravero, Pardo, Sepúlveda, et al., 2022). Este tipo de tecnología también facilita la gestión y un mejor aprovechamiento de los recursos naturales, brindando sugerencias personalizadas para cada caso. Pese a los beneficios, también se encuentran limitantes. La primera son las diversas fuentes de datos y la calidad de estos. Adicional la transformación y normalización del dataset es un proceso complejo debida a que se necesita integrar datos de varias plataformas.

Series temporales en predicciones agrícolas

Las series temporales optimizan el proceso de toma de decisiones basados en predicciones espacio-temporales. Esta técnica se trata de un conjunto de información recopilada de manera secuencial en determinados lapsos de tiempo con la finalidad de observar el comportamiento de un determinado fenómeno. La predicción de series temporales en los cultivos abarca el análisis de variables recopiladas en un determinado tiempo. Estas variables son la temperatura, la humedad, la radiación solar, los índices de vegetación, etc., obtenidos de varias fuentes.

El uso más común de las series temporales es la predicción de la producción de los cultivos. Con el análisis de patrones, los algoritmos de Machine Learning estiman una producción, esta ventaja permite a los productores planificar el tiempo de tumba y el proceso de comercialización del producto. Kumar y compañía (2024), hacen hincapié en el desarrollo de modelos con un bajo margen de error con el objetivo de tener una excelente seguridad alimentaria.

La predicción de eventos climáticos permite a los agricultores desarrollar mejores calendarios de siembra y cosecha e implementar sistemas de riegos óptimos para el aprovechamiento del recurso hídrico. En la obra *"Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review"*, los investigadores determinan que la

velocidad en adquirir y procesar los datos de series temporales es de vital importancia (Cravero, Pardo, Sepúlveda, et al., 2022).

Las técnicas de Machine Learning utilizadas para el análisis de datos son las Redes Neuronales, Bosques Aleatorios, Máquinas de Vectores de Soporte y Árboles de Decisión. Las redes neuronales son las herramientas más óptimas para procesar y gestionar datos no lineales y complejos, esto convierte a esta técnica en un recurso valioso a la hora de analizar entornos dinámicos. Una tendencia observada últimamente es la combinación de varias técnicas de Machine Learning con la finalidad de mejorar la precisión de las predicciones.

Redes neuronales recurrentes (LSTM)

Las redes Long Short-Term Memory (LSTM), ha nacido como una herramienta necesaria en la predicción de series temporales complejas. Este tipo de red neuronal se caracteriza por modelar dependencias a largo plazo basada en datos por secuencia, esta característica la hace indispensable en las proyecciones agrícolas pues los datos que se manejan evolucionan y no tienen patrones lineales.

Citando a (Nurraharjo et al., 2024), se determina que las redes LSTM son utilizadas para la predicción de variables críticas, en su trabajo menciona que en la actualidad se han combinado las LSTM con dispositivos IoT para captar datos en tiempo real y poder realizar las respectivas predicciones con el objetivo de optimizar aspectos como el riego y la aplicación de abono.

Las redes LSTM sobrepasan las limitantes de las redes convolucionales tradicionales al implementar accesos de entrada, olvido y salida que gestionan y optimizan el flujo de la información, pudiendo conservar o desechar información en el transcurso de todo el flujo. De acuerdo con Del Coco, Leo y Carcagni (2024) esta arquitectura es relevante en el campo agrícola pues permite captar la influencia de fenómenos meteorológicos pasados en el presente y futuros en los sembríos.

Al integrar las redes LSTM con sensores o dispositivos IoT los datos serán continuos y en tiempo real, mejorando las predicciones. Sin embargo, las LSTM en el campo agrícola también presentan falencias como la demanda de un equipo computacional de alta gama y la capacidad de poder gestionar datos faltantes o ruidosos (Sun et al., 2023).

Integración de datos de varias fuentes

La integración de datos captados de diversas fuentes es un recurso crítico y esencial para el desarrollo de un sistema predictivo de sembríos. Los datos generados en la agricultura son de tipo estructurado, semiestructurado y no estructurados, debido a que proceden de sensores, imágenes satelitales, bases de datos, etc. La combinación y depuración correcta de la data obtenida permite una visión global del medio agrícola.

Los Data Lakes y herramientas en la nube permiten una rápida consolidación y almacenamiento de una gran cantidad de información. Estas tecnologías permiten el almacenaje de datos en tiempo en real y a la vez la ejecución de procesos de Machine Learning que necesitan de datos de diversas fuentes para modelar un determinado suceso (Assimakopoulos et al., 2024).

La integración de los datos es fundamental para la proyección de precios de los frutos, generando de esa manera modelos que representan la variabilidad del mercado. Adicional, se pueden incorporar análisis de sentimientos para observar el comportamiento de las personas hacia un determinado fruto. Pese a los beneficios, el proceso de integración enfrenta la necesidad de una correcta sincronización de los datos tiempo-espacio, sin esta el modelo proyectaría información errónea.

Refinación de datos

La refinación o limpieza de datos es un paso importante en la utilización de la información Generalmente los datos presentan errores, valores faltantes y ruido, aplicando el refinamiento de datos se evita el sesgo de la información y se mejora la proyección y entrenamiento de los algoritmos de Machine Learning (Del Coco et al., 2024).

El proceso EDA es la técnica para entender un dataset , con este método se pueden corregir valores atípicos, imputar datos faltantes, eliminar

datos duplicados y estandarizar las variables. Adicional, nos permite determinar datos que son pocos confiables para el entrenamiento del modelo predictivo. Refinar datos es fundamental porque la data que se utilice determinará la capacidad que tendrá el modelo para detectar patrones.

La refinación o depuración de datos no solo mejora la calidad de los mismos, sino que facilita el entendimiento de estos y de como funciona el negocio. Adicional, un proceso transparente a la hora de manipular la información y tener la capacidad de explicar las proyecciones incrementa el nivel de confianza de los productores.

Ejemplos de modelos predictivos agrícolas

Del Coco y colegas (2024) exponen un modelo en función a una LSTM que proyecta temperatura, humedad y humedad del suelo con un error cuadrático medio de 2.35%. A su vez, los autores recopilas otro caso de éxito implementado por Gao en el cuál se desarrolla una red bidireccional LSTM con el objetivo de optimizar la predicción de la conductividad eléctrica y humedad del suelo en cultivos de frutas cítricas.

Por otra parte, Sun y colegas (2023) destacan la integración de algoritmos Support Vector Machines (SVM), Random Forest y redes neuronales con la finalidad de proyectar rendimientos en base a datos históricos, condiciones climáticas y propiedades del suelo. Adicional, los autores mencionan que la implementación de técnicas como análisis discriminante lineal (LDA) mejoran el proceso de entrada de datos para algoritmos como PSO-SVM.

Las redes convolucionales también han aportado en la detección de plagas y enfermedades. Al implementar esta tecnología se puede procesar imágenes, como en el caso de Mohanty, proyecto en el cual se detectaron 26 enfermedades en cultivos y la precisión de identificación fue de 99.35% (Assimakopoulos et al., 2024). Pese a aquello, también se ha fortalecido esta detección con el acompañamiento de técnicas como Few-Shot Learning para optimizar el proceso de identificación en un grupo de datos complejos.

MATERIALES Y MÉTODOS

La presente investigación posee un enfoque cuantitativo debido a que implica la recolección, análisis y procesamiento de información numérica asociada a variables agroclimáticas, edáficas y productivas de las huertas de cacao del cantón Balao. Al utilizar un método cuantificado se pudo implementar aprendizaje automático y observar la estrecha relación que comparten las variables predictoras y el rendimiento agrícola.

El tipo de investigación es explicativa-aplicada, porque identificó la correlación existente entre los distintos tipos de variables utilizadas. Adicional, se generó una solución práctica para la problemática estudiada. En cuanto al diseño de investigación se abarcó no experimental-longitudinal. Es diseño no experimental porque se procesaron datos históricos obtenidos en campo y no se manipularon las variables de estudio. Longitudinal porque se analizó el comportamiento de las variables en un determinado intervalo de tiempo.

El contexto del trabajo radica en buscar y desarrollar una solución para que los productores y determinados organismos de gestión agrícola, o que demuestren interés en este campo, puedan tener una herramienta que optimice el proceso de toma de decisiones sobre los cultivos de cacao CCN-51. Se empleó la metodología CRISP-DM para el proceso de análisis predictivo. La solución tecnológica se basa en una red neuronal recurrente LSTM y modelos de regresión lineal y ARIMA.

En la primera etapa de la metodología se comprendió el negocio definiendo y entendiendo la necesidad que existe de anticipar la producción cacaotera. De acuerdo con (Murillo Martínez & Cano Lara, 2025), comprender los objetivos, las necesidades y el entorno del negocio es un paso fundamental para implementar exitosamente un proceso o herramienta. Se identificó las variables a examinar, la variable dependiente de nuestro estudio es el rendimiento agrícola expresado en quintales por hectárea, y la variable de entrada sería la serie temporal por cada finca. Como segundo paso, se comprende los datos, se determina el tamaño de la muestra, y se procede a recolectar la información en un dataset. Posterior se realiza la

preparación de los datos, limpiándolos y manejando los respectivos valores atípicos.

En cuarto lugar, se implementaron técnicas de modelado predictivo. Se ejecutaron series temporales con varios LookBack y Horizon. La quinta fase fue la evaluación de los resultados para determinar cual era la combinación de variables optimizaba de mejor manera el rendimiento del modelo. Finalmente, se obtuvieron los entregables o resultados.

Las fuentes de donde se obtuvieron los datos fueron entrevistas con los productores del CCN-51, estaciones meteorológicas ubicadas en la localidad, análisis de suelo realizados por los agricultores, índices de vegetación (obtenidos de imágenes satelitales) y registros históricos de la producción del cacao obtenidos del INEC y de registros que llevan productores de la zona.

La población abarcada por la investigación son las unidades de producción cacaotera del cantón Balao, provincia del Guayas. La cantidad de cacahueras es considerablemente amplia y en la actualidad los registros en el GAD del cantón estaban desactualizados por los que se desconocía el tamaño de la población. Al desconocerse ese dato, se calculo una muestra de 384 cacahueras, seleccionando 16 fincas en cada uno de los recintos del cantón. Se empleó un método no probabilístico por conveniencias, el cuál radica en seleccionar a los participantes por su disponibilidad o proximidad. La razón por la cual se aplicó este método fue por la disponibilidad de información que presentaban las fincas seleccionadas.

El desarrollo del modelo predictivo se basó en la implementación de redes neuronales recurrentes de tipo LSTM. Adicional, se combinó un esquema de validación cruzada con series temporales utilizado para la evaluación del modelado. El proceso de evaluación consistió a través de dos métricas estadísticas, error cuadrático medio y coeficiente de determinación. Estas métricas se calcularon en la etapa de validación cruzada y en la evaluación final sobre el conjunto de datos de prueba.

Los recursos destinados a la investigación son de carácter humano, técnico, financiero y temporal. En el ámbito humano, esta la participación de los investigadores y de los productores del campo. En recursos técnicos, se define el uso de computadoras, librerías y entornos de desarrollo; adicional de teléfonos para la grabación de audio de las entrevistas. Los recursos financieros destinados a la movilización en cada recinto. Por último, los

recursos temporales, los cuales hacen alusión al cronograma de ejecución del proyecto.

RESULTADOS Y DISCUSIÓN:

La aplicación de aprendizaje profundo en el campo de la agricultura permitió desarrollar una serie de pasos estructurados. El método de trabajo comenzó con el análisis de imágenes satelitales hasta el desarrollo del modelo, incluyo procesos de teledetección, estadística descriptiva, análisis de multicolinealidad y modelo de Deep Learning.

Análisis de imágenes satelitales

En la primera fase del proceso se obtuvieron y se procesaron imágenes satelitales originadas en el satélite Sentinel-2 del sistema de satélites Constelación Sentinel del programa Copernicus, estas imágenes se encuentran disponibles en Google Earth Engine (GEE). Este método permitió determinar los índices espectrales como NDVI, EVI, SAVI y NDWI. GEE ofrece una fácil manipulación de colecciones masivas de imágenes.

Figura 1. Función calcular índices

```
# Función para calcular índices
def calcular_indices(imagen):
    # Seleccionamos solo las bandas necesarias para evitar inconsistencias
    imagen = imagen.select(['B2', 'B3', 'B4', 'B8'])

    ndvi = imagen.normalizedDifference(['B8', 'B4']).rename('NDVI')
    evi = imagen.expression(
        '2.5 * ((NIR - RED) / (NIR + 6 * RED - 7.5 * BLUE + 1))', {
            'NIR': imagen.select('B8'),
            'RED': imagen.select('B4'),
            'BLUE': imagen.select('B2')
        }).rename('EVI')
    savi = imagen.expression(
        '((NIR - RED) / (NIR + RED + 0.5)) * 1.5', {
            'NIR': imagen.select('B8'),
            'RED': imagen.select('B4')
        }).rename('SAVI')
    ndwi = imagen.normalizedDifference(['B3', 'B8']).rename('NDWI')

    return imagen.addBands([ndvi, evi, savi, ndwi])
```

En primer lugar, se autenticó e inicializó el entorno de Google Earth Engine con la finalidad de acceder al repositorio de las imágenes satelitales y a funciones de reducción espacial y temporal que también ofrece. Para calcular los índices espectrales se utilizó las coordenadas de cada finca, dichos datos fueron definidos como objetos geométricos POINT. Posteriormente, se desarrolló e implementó la función CALCULAR_INDICES, esta se basó en el análisis de bandas espectrales como: B2 (Azul), B3 (verde), B4 (rojo) y B8 (infrarrojo cercano).

Citando a (De La A Salinas et al., 2025), la importancia de estos índices es la siguiente: NDVI (Normalized Difference Vegetation Index) permite conocer el vigor y la biomasa vegetal, EVI (Enchange Vegetation Index) utilizado en zonas que presentan un alto grado de vegetación y el NDVI sufre en ocasiones saturaciones, SAVI (Soil Adjusted Vegetation Index) tiene utilidad en casos agrícolas en los cuales el suelo desnudo y la vegetación varían por estaciones, y NDWI (Normalized Difference Water Index) es una métrica que ayuda a relacionar la condición hídrica de la siembra con su producción.

El lapso de tiempo analizado en las imágenes va desde enero de 2022 hasta junio de 2025. Sentinel-2 proporciona un promedio de 5 a 6 imágenes

al mes de una determinada ubicación. Por esa razón, el procedimiento es iterativo y promedia el índice. Adicional se incorporaron filtros de tiempo, espacio y nubosidad. Este último permitía descartar imágenes con un porcentaje de nubosidad mayor al 80%, esto con la finalidad de minimizar un sesgo en el cálculo de los índices por las diversas condiciones climáticas que se pueden presentar.

El resultado de este proceso fue un primer dataset con las variables Finca_Id, Fecha, y los respectivos índices. Al integrar GEE se redujo el tiempo de procesamiento, además de ahorrar tiempo en descarga masiva de imágenes y de poseer recursos en el equipo para un almacenamiento local. Adicional, al estandarizar la información por cada mes, se pudo realizar comparaciones entre las fincas en distintos intervalos de tiempo.

Proceso EDA

El proceso del Análisis Exploratorio de Datos (EDA), permite entender el comportamiento del dataset, con este proceso se puede identificar anomalías como valores atípicos y estandarizar la información de cada variable con la finalidad de evitar sesgos en el modelo predictivo. Del primer Dataset obtenido en el cálculo de los índices se incorporaron variables edafoclimáticas como ph del suelo, porcentaje de materia orgánica presente, precipitación, temperatura mínima y máxima, humedad relativa, evapotranspiración y radiación solar.

Para comprender el dataset se utilizó el comando shape, el cual presenta la dimensión o cantidad de registros y variables existentes en el conjunto de datos. Posterior, se realiza una visualización de las primeras filas con el objetivo de validar la estructura del archivo. Estos primeros pasos de verificación son fundamentales pues aseguran la congruencia entre el modelo y los supuestos metodológicos.

Figura 2. Dimensiones y validación de los datos

```
print("Dimensiones del dataset:", df.shape)
print("\nPrimeras filas:")
print(df.head())
```

```
Dimensiones del dataset: (16128, 33)

Primeras filas:
    Finca_Id      Fecha      Nombre_Finca   Tipo_cacao Cantón      Localidad  \
0  Finca_001 2022-01-01  Finca La Beatriz  Cacao CCN51  Balao  Cien Familias
1  Finca_001 2022-02-01  Finca La Beatriz  Cacao CCN51  Balao  Cien Familias
2  Finca_001 2022-03-01  Finca La Beatriz  Cacao CCN51  Balao  Cien Familias
3  Finca_001 2022-04-01  Finca La Beatriz  Cacao CCN51  Balao  Cien Familias
4  Finca_001 2022-05-01  Finca La Beatriz  Cacao CCN51  Balao  Cien Familias

    Latitud   Longitud  ph_suelo  Materia_organica_porc  ...  Horas_sol  \
0 -2.911044 -79.669952     7.007                  49.88  ...     358.12
1 -2.911044 -79.669952     7.462                  50.49  ...     345.13
2 -2.911044 -79.669952     7.415                  69.04  ...     357.26
3 -2.911044 -79.669952     7.242                  77.09  ...     302.90
4 -2.911044 -79.669952     7.275                  59.07  ...     289.92

   Radiacion_solar  Fertilizantes_kg_ha  Pesticidas_lt_ha  Cantidad_podas  \
0           498.20                  490               175               3
1           463.40                  495               190               3
2           462.90                  476               197               3
3           423.44                  487               182               3
4           404.75                  468               184               3

   Fenomeno_climatico_extremo    NDVI     EVI    SAVI    NDWI
0                           1  0.3565  0.4895  0.1954  0.7248
1                           1  0.0516  0.7898  0.1613  0.4786
2                           0  0.9339  0.5815  0.7073  0.9449
3                           0  0.2261  0.9342  0.0097  0.5543
4                           0  0.5898  0.2389  0.5456  0.5645
```

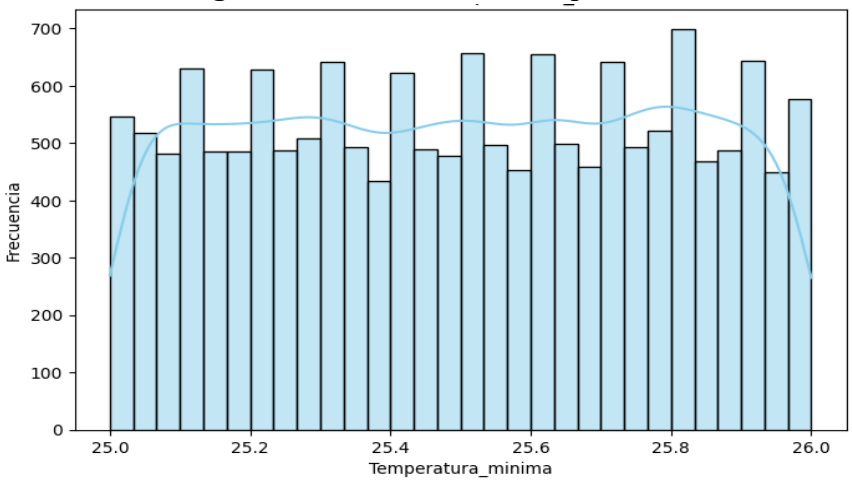
Figura 3. Análisis de datos duplicados o nulos

```
duplicados = df.duplicated().sum()
print(f"\nNúmero de filas duplicadas: {duplicados}")
if duplicados > 0:
    df = df.drop_duplicates()
    print("Duplicados eliminados.")


print("\nValores nulos por columna:")
print(df.isnull().sum())


faltantes = (df.isnull().sum() / len(df)) * 100
print("\nPorcentaje de valores nulos por columna:")
print(faltantes)
```

Luego, se procedió a realizar un análisis en la información para determinar si existen datos duplicados o valores nulos. El análisis reflejo que no existen datos duplicados ni valores nulos, descartando el riesgo para la consistencia de los resultados. Se implementó un análisis descriptivo con el comando describe(), este proporciona un resumen de las métricas estadísticas (media, desviación estándar, mínimo, máximo y cuartiles). En los resultados obtenidos se visualiza que variables como la radiación solar presenta una desviación estándar de 28.9 lo que permite denotar que existe una variabilidad en la cantidad de radiación recibida en los sembríos. Por otra parte, la variable de porcentaje de materia orgánica tiene una heterogeneidad moderada con una desviación estándar de 15.9.

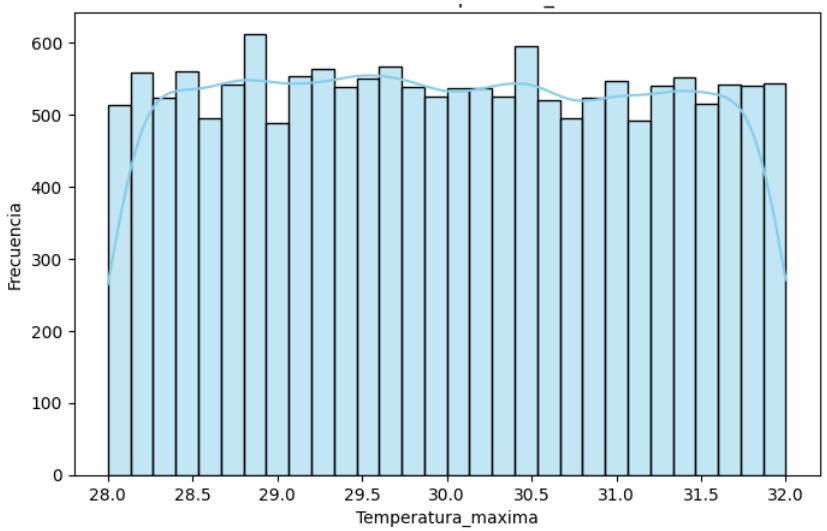
Adicional, se generaron histogramas para las variables de interés con el objetivo de identificar comportamientos normales, sesgados o multimodales. Los resultados revelaron que la mayoría de las variables presentan comportamientos multimodales, excepto las variables relacionadas a las temperaturas y evapotranspiración que presentan distribuciones más simétricas.

Figura 4. Distribución de temperatura mínima



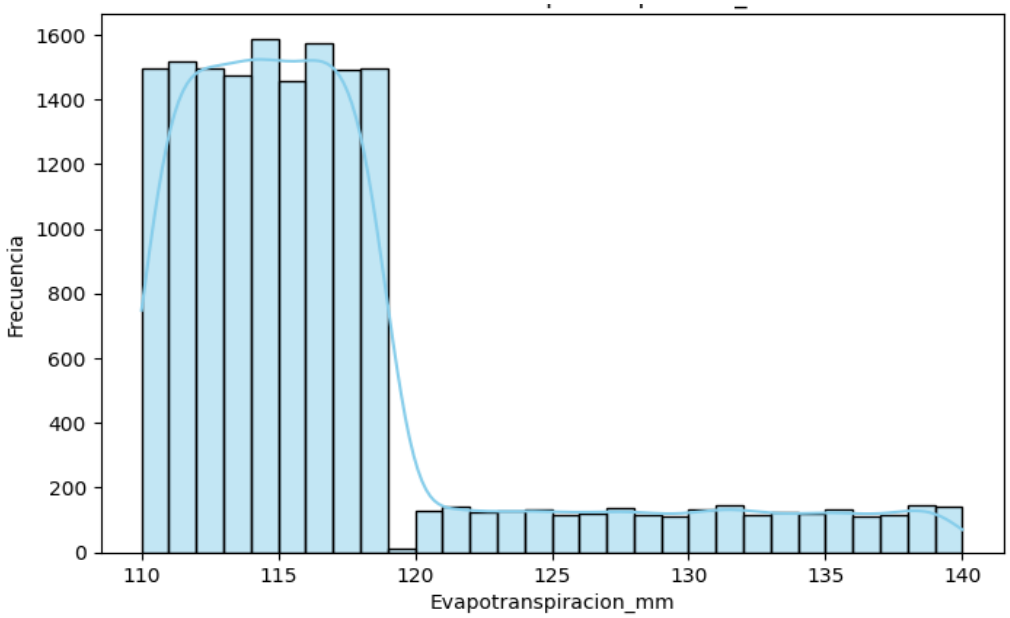
Nota: Presenta una distribución casi normal, con una ligera asimetría.

Figura 5. Distribución de temperatura máxima



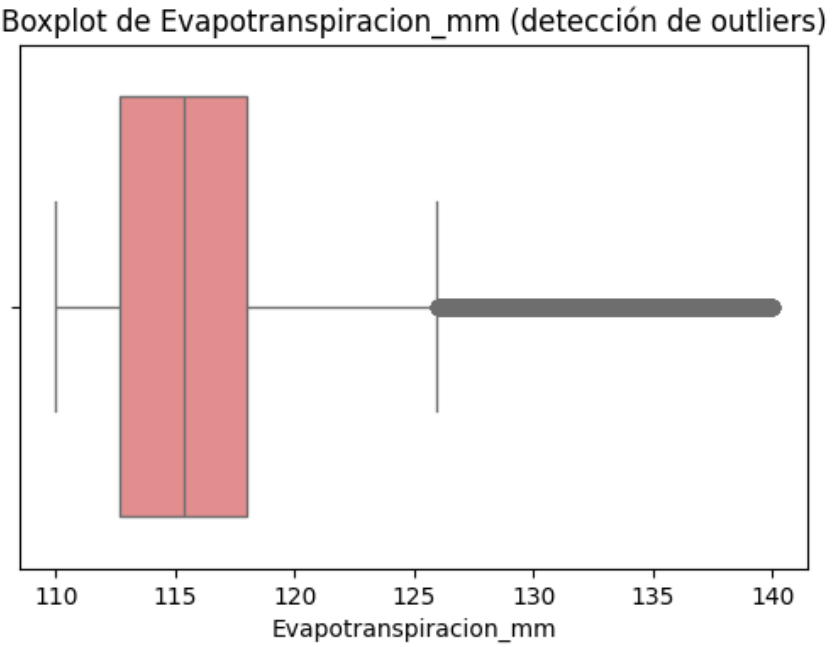
Nota: Presenta distribución normal, bastante simétrica.

Figura 6. Distribución de evapotranspiración



Nota: Presenta una ligera asimetría hacia la derecha (sesgo positivo).

Figura 7. Detección de outliers



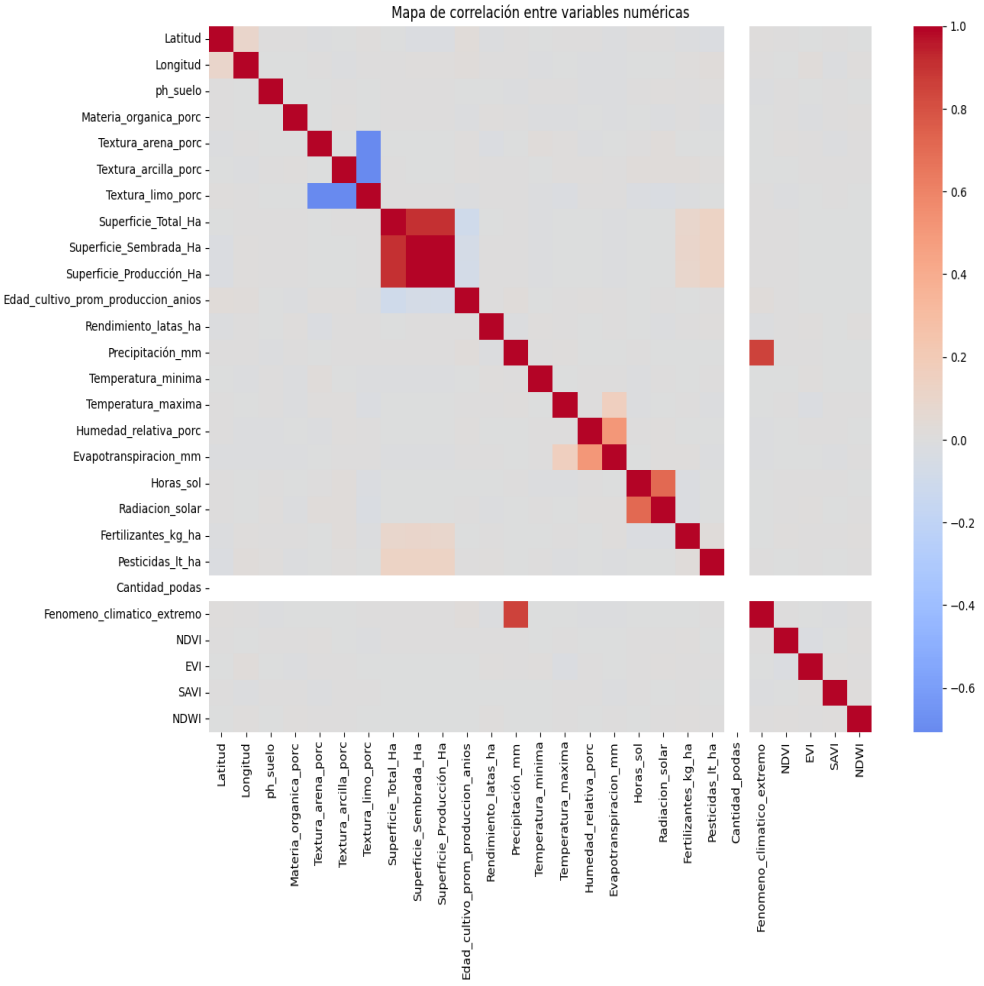
Posterior, se implementó el uso de boxplots con la finalidad de encontrar outliers. Este paso es importante porque al implementar outliers en redes neuronales puede haber distorsión de los parámetros. Se analizaron las variables de interés y se determinó que la única variable que tiene valores atípicos es Evapotranspiración. Se recalca que no todos los outliers deben eliminarse, pues son mediciones correctas y suprimirlas restaría representatividad en el caso de estudio.

Adicional a los boxplots, se incorporaron dos métodos complementarios para la detección de valores atípicos: el Z-score y el método del rango intercuartílico (IQR). El resultado de ambos métodos confirmo la existencia de valores atípicos en la variable Evapotranspiración, sin embargo, no se eliminaron estos datos, sino que se transformaron logarítmicamente. Esta transformación redujo la asimetría de la variable y mejoró el comportamiento estadístico de la misma, de forma que se pudo utilizar en el modelo predictivo.

Para el tratamiento de las otras variables se implementó la técnica StandardScaler. De acuerdo con (Setyo Priyambudi & Sulistyo Nugroho,

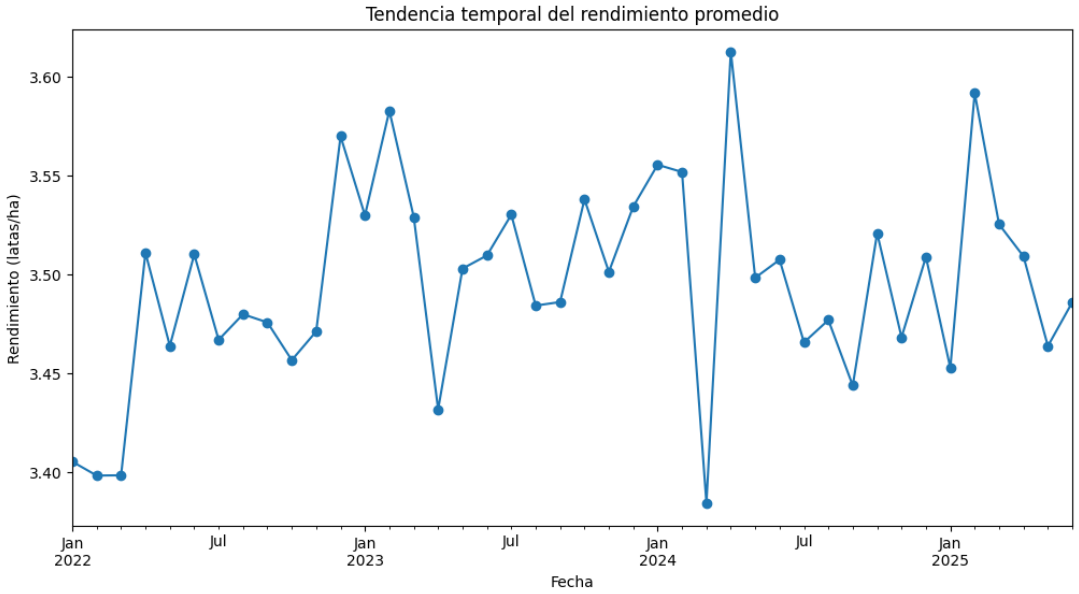
2024), StandarScaler centra todos los datos en la media de los mismos y los va escalando de acuerdo a la desviación estándar, obteniendo como resultado una media igual a cero y una desviación igual a uno. La técnica fue elegida porque los datos presentaban una distribución aproximadamente normal. La finalidad fue mejorar el proceso y velocidad de aprendizaje de la red LSTM.

Figura 8. Mapa de calor (Correlación entre variables)



Se generó un mapa de calor para examinar las relaciones lineales entre las variables. Como resultado se observa que los índices espectrales obtenidos del procesamiento de las imágenes satelitales presentan correlaciones elevadas entre sí, mientras que variables climáticas como la precipitación, temperatura y evapotranspiración presentan correlaciones de mediana intensidad. La importancia de este análisis es descartar redundancias e identificar las variables que aportarán al modelo predictivo.

Figura 9. Serie de tiempo de la producción agrícola promedio



El análisis temporal se represento mediante una serie de tiempo de la producción agrícola promedio. En la gráfica se observa una variabilidad moderada con crecientes interrumpidas por caídas pronunciadas, esto debido a las condiciones climáticas de aquellos de meses. Esta permite concluir, que el cultivo de cacao CCN-51 es muy sensible a factores ambientales y formas de cuidado. El análisis de la serie temporal permite determinar períodos de altas y baja producción, conocimiento fundamental para diseñar estrategias de intervención y sistemas de alertas en la agricultura.

Análisis VIF

El desarrollo de un modelo predictivo requiere pertinencia e independía en sus variables. Desarrollar un modelo con multicolinealidad va a hacer que los coeficientes de regresión sean inconsistentes y que el modelo no pueda generalizar. Se decidió utilizar el Factor de Inflación de la Varianza (VIF) por la cantidad de variables predictoras.

Para la implementación del modelo VIF se utilizó el paquete stasmodel en Python, el cual puede calcular el VIF de forma directa. Se comenzó definiendo las variables numéricas, dejando de lado variables categóricas. En segundo lugar, se realizó el ajuste de una regresión lineal para cada variable, utilizando el resto de variables como regresores. Como resultado se listo, de mayor a menor, las variables con el valor de VIF.

Figura 10. Resultados del análisis VIF

	Variable	VIF
3	Textura_arcilla_porc	inf
2	Textura_arena_porc	inf
4	Textura_limo_porc	inf
6	Superficie_Sembrada_Ha	1250.579562
7	Superficie_Producción_Ha	1205.229139
5	Superficie_Total_Ha	6.305503
9	Precipitación_mm	3.578967
19	Fenomeno_climatico_extremo	3.578069
14	Horas_sol	2.025482
15	Radiacion_solar	2.025409
13	Evapotranspiracion_mm_log	1.402990
12	Humedad_relativa_porc	1.366400
11	Temperatura_maxima	1.037805
17	Pesticidas_lt_ha	1.025789
8	Edad_cultivo_prom_produccion_anios	1.019388
16	Fertilizantes_kg_ha	1.013594
22	SAVI	1.001566
1	Materia_organica_porc	1.001231
21	EVI	1.001086
23	NDWI	1.001068
10	Temperatura_minima	1.001057
20	NDVI	1.000973
0	ph_suelo	1.000740
18	Cantidad_podas	0.000000

De acuerdo con los resultados, se observa una fuerte correlación entre las variables relacionadas a la textura del suelo, razón por la cual no se debe incluir a las tres en el modelo, solo bastaría con seleccionar una de ellas. También se debe considerar las variables relacionadas a la superficie, estas muestran valores muy altos y refleja una redundancia estadística casi total. Con el resto de las variables, presentan valores entre uno y tres, lo que significa que las hacen aceptables para el modelo.

Modelo predictivo

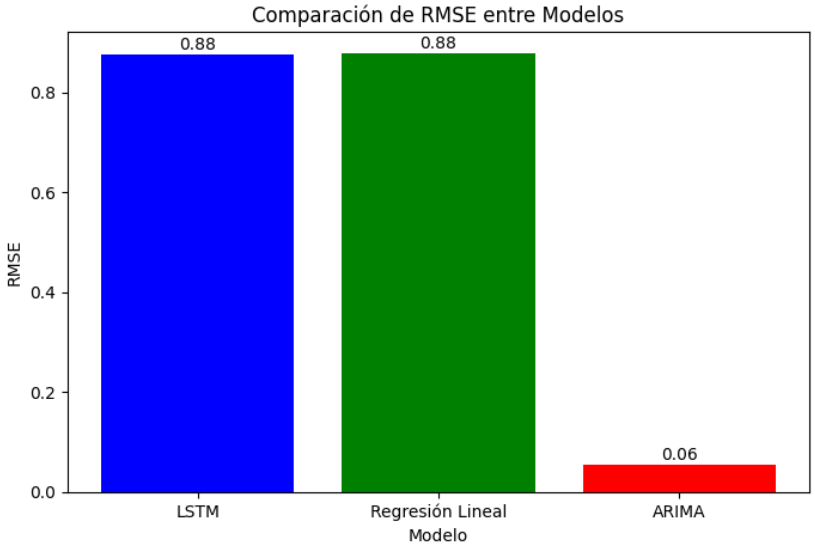
Se desarrollo el modelo trabajando con 18 variables predictivas, cada registro o fila en el dataset estaba identificada por un código y una determinada fecha. Se implementó un timestep, para utilizar información histórica dentro de cada tres mese para proyectar la producción agrícola del siguiente mes. De esa forma se organizó el dataset en tensores

tridimensionales, abarcando dimensiones relacionadas al número de muestras, horizonte temporal y número de variables predictoras.

Posterior a aquello se realizó la división del dataset para entrenamiento y prueba (80%-20%) de forma que se cuidara la secuencia cronológica de la información. Adicional, se implementó una validación cruzada con series temporales, con el objetivo de evaluar el desempeño del modelo LSTM, calculando el RMSE y el R^2. Los resultados obtenidos para RMSE (0.861) determinan que el resultado se considera aceptable, pero para el coeficiente de determinación promedio (-0.001) el resultado demuestra que modelo presenta limitaciones para determinar patrones relacionados a la dinámica temporal del rendimiento.

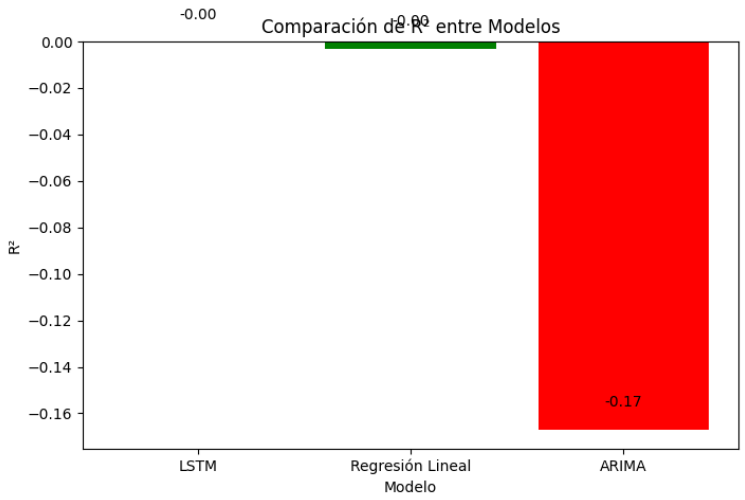
Posterior a aquello se implementó un modelo ARIMA y el modelo LSTM, donde el R^2 era negativo y el RMSE era valores bajos, por esa razón se determinó que el modelo tiene un comportamiento estable y coherente pero la heterogeneidad de las fincas y la variabilidad de la información por las fluctuaciones climáticas extremas generan dispersión en las series temporales.

Figura 11. Comparación de RMSE entre modelos



Nota: La comparación de los resultados permite denotar la capacidad de predicción.

Figura 12. Comparación de R2 entre modelos



Nota: Los resultados obtenidos por esta métrica permiten visualizar que el modelo no capta adecuadamente la dinámica temporal del rendimiento.

Discusión

El objetivo de la presente investigación es analizar la capacidad que poseen las redes LSTM para predecir el rendimiento agrícola, tomando como punto de referencia la hipótesis de que estas técnicas, dada su estructura para procesar secuencias temporales y relaciones no lineales complejas, podrían transformar de mejor manera las variables multifuentes agrícolas que los procedimientos estadísticos tradicionales. Sin embargo, los resultados obtenidos, con un R^2=-0.001 y un RMSE=0.861, no permiten validar la hipótesis planteada, sino que dejan observar que el modelo presenta una predicción baja. Al analizar el coeficiente de determinación se concluye que el modelo no logra capturar todas las variaciones existentes en el rendimiento agrícola.

Examinando el modelo y los resultados, se determina que el bajo desempeño se basa en una baja calidad y alta heterogeneidad de los datos de entrada. Investigaciones como las de Del Coco y colegas dictan que uno de los desafíos más grandes para Machine Learning es la escasez de datos fiables y de calidad. Las huertas no poseen atributos uniformes dejando a la vista una heterogeneidad en datos. Al intentar entrenar un modelo que generalice variables con tantos cambios bruscos se merma la capacidad de identificar y aprender patrones complejos. Assimakopoulos y colegas presentan un caso de estudio exitoso, en el cual se implementan herramientas de IA con variables de clima y topografía, implicando una

granularidad que en un set de datos heterogéneo de gran escala sería difícil de obtener.

Al estudiar huertas de cacao distantes y en diferentes zonas del cantón aumentan las diferencias, pues las condiciones edafoclimáticas varían de un lugar a otro. Al no poder sistematizar los datos, se limita la efectividad del modelo para captar variaciones espaciales relevantes. Por último, cabe recordar que las presentes condiciones meteorológicas, no solo en el cantón sino a nivel mundial, cambian drásticamente aumentando el ruido de datos. Todas estas restricciones si no son tratadas adecuadamente pueden provocar que el modelo predictivo identifique y aprenda "ruido" en vez de patrones. En conclusión, la capacidad de las LSTM para procesar información de diversas fuentes es dependiente de la calidad de la información obtenida.

CONCLUSIONES

La implementación de la metodología CRISP-DM en la investigación permitió desarrollar un algoritmo de trabajo que abarque fases de comprensión del negocio, recopilación de información, tratamiento de datos, desarrollo del modelo y evaluación.

Utilizar herramientas de Deep Learning como redes LSTM permitió observar el potencial que tiene esta para procesar variables multifuentes, complejas y no lineales en comparación a modelos estadísticos tradicionales. Pese aquello, obtener un coeficiente de determinación bajo fue por la calidad del dataset y la heterogeneidad de la información. Sugiriendo que en futuros procesos se mejore el sistema de recolección y normalización de la data.

Al normalizar las variables se pudo homogenizar las escalas y valores de las variables. Utilizando el método VIF se logró determinar las variables predictoras y desechar variables altamente colineales. Este paso demuestra que una selección adecuada es fundamental para tener un modelo fiable. Pese aquello, también se observó lo complicado que es trabajar con datos

heterogéneos, puesta esta afectó la capacidad del modelo para identificar y aprender patrones en vez de ruido.

El trabajo demuestra que las redes LSTM transforman de mejor forma las variables agrícolas, obteniendo resultados con errores aceptables, la causa de la deficiencia de la predicción sería la calidad del dataset. De igual forma, se observó que los modelos LSTM requieren de datos estandarizados con la finalidad de poder mejorar la precisión de las proyecciones.

REFERENCIAS BIBLIOGRÁFICAS

Assimakopoulos, F., Vassilakis, C., Margaris, D., Kotis, K., & Spiliotopoulos, D. (2024). Artificial Intelligence Tools for the Agriculture Value Chain: Status and Prospects. Electronics, 13(22), 4362. https://doi.org/10.3390/electronics13224362

Bowen Quiroz, G. A., & Medranda Cobeña, G. I. (2024). Impacto de los sistemas de información en la agricultura inteligente: Una revision general. Revista InGenio, 7(2), 117-136. https://doi.org/10.18779/ingenio.v7i2.824

Cravero, A., Pardo, S., Galeas, P., López Fenner, J., & Caniupán, M. (2022). Data Type and Data Sources for Agricultural Big Data and Machine Learning. Sustainability, 14(23), 16131. https://doi.org/10.3390/su142316131

Cravero, A., Pardo, S., Sepúlveda, S., & Muñoz, L. (2022). Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. Agronomy, 12(3), 748. https://doi.org/10.3390/agronomy12030748

De La A Salinas, L. D. R., Monserrate Rodríguez, J. P., Medina Robayo, A. I., & Tobar Cuesta, B. A. (2025). Uso de índices de vegetación para la detección del estrés hídrico en cultivos: Una revisión sistemática de estudios basados en teledetección. Polo del Conocimiento, 10(8), 791-813. https://doi.org/10.23857/pc.v10i8.10183

Del Coco, M., Leo, M., & Carcagnì, P. (2024). Machine Learning for Smart Irrigation in Agriculture: How Far along Are We? Information, 15(6), 306. https://doi.org/10.3390/info15060306

Dos Santos Pereira, J., Silva Santos, A., Martins Newman Luz, E. D., & Corrêa, R. X. (2024). Sources of resistance to witches' broom disease

in cacao (Theobroma cacao L.): Progress update and perspectives. Plant Breeding, 143(6), 798-809. https://doi.org/10.1111/pbr.13205

Jaimez, R. E., Barragan, L., Fernández Niño, M., Wessjohann, L. A., Cedeño Garcia, G., Sotomayor Cantos, I., & Arteaga, F. (2022). Theobroma cacao L. cultivar CCN 51: A comprehensive review on origin, genetics, sensory properties, production dynamics, and physiological aspects. PeerJ, 10, 1-23. https://doi.org/10.7717/peerj.12676

Kumar, R., Bhanu, M., Mendes Moreira, J., & Chandra, J. (2024). Spatio-Temporal Predictive Modeling Techniques for Different Domains: A Survey. ACM Computing Surveys, 57(2), 1-42. https://doi.org/10.1145/3696661

Mihai, R. A., Melo Heras, E. J., Terán Maza, V. A., Espinoza Caiza, I. A., Pinto Valdiviezo, E. A., & Catana, R. D. (2023). The Panoramic View of Ecuadorian Soil Nutrients (Deficit/Toxicity) from Different Climatic Regions and Their Possible Influence on the Metabolism of Important Crops. Toxics, 11(2), 123-144. https://doi.org/10.3390/toxics11020123

Murillo Martínez, G. Y., & Cano Lara, E. D. (2025). Desafios De Las Startups Del Ecuador Al Implementar Inteligencia Artificial En Su Gestión De Marketing: Desafios: Startups Del Ecuador E Inteligencia Artificial. REFCalE: Revista Electrónica Formación y Calidad Educativa, 13(1), 39-56. https://doi.org/10.56124/refcale.v13i1.003

Nurraharjo, E., Utami, E., Kusrini, & Ari Yuana, K. (2024). Hybrid LSTM-IoT in Agriculture: A Systematic Literature Review. 2024 International Conference on Information Technology and Computing (ICITCOM), 36-41. https://doi.org/10.1109/ICITCOM62788.2024.10762212

Setyo Priyambudi, Z., & Sulistyo Nugroho, Y. (2024). Which algorithm is better? An implementation of normalization to predict student performance. AIP Conference Proceedings, 2926(1), 020110. https://doi.org/10.1063/5.0182879

Sornoza Vélez, Lady, Valencia Carreño, L., Corozo Quiñónez, L., Sánchez Mora, F., Salas Macías, C., & Peña Monserrate, G. (2022). Recursos genéticos de cacao tipo Nacional en Ecuador: Una revisión sistemática. Revista Ciencia y Tecnología, 15(2), 31-44.

Sun, F., Meng, X., Zhang, Y., Wang, Y., Jiang, H., & Liu, P. (2023). Agricultural Product Price Forecasting Methods: A Review. Agriculture, 13(9), 1671. https://doi.org/10.3390/agriculture13091671