TRABAJO DE GRADO
Opción Seminario-Diplomado.

Big Data en la Agricultura:
Predicción de cosechas y control de plagas

Tutor:
Juan Pablo Vélez Uribe

Estudiantes:
Sebastian Goez García
Wilson Alexander Orrego Arias
Julian Hurtado Ramírez

Corporación Universitaria Remington.
Facultad de Ingeniería.
Ingeniería de Sistemas.

Opción de Trabajo de grado Seminario.
15/2025

Tabla de Contenidos

Resumen

Este trabajo de investigación tiene como objetivo implementar los conocimientos del Big Data en el sector agrícola, abarcando áreas importantes como lo son el control de plagas, la predicción y la evolución de los cultivos.
El cambio climático ha causado un gran impacto en la agricultura que conocemos hoy en día, con ayuda de nuevas tecnologías ha logrado superar grandes retos, como la demanda de alimentos o escasez de recursos y la pérdida de cultivos por infestación de plagas. Gracias a los grandes avances en la tecnología se crean herramientas que ayudan con el análisis de los datos, estas cumplen un rol muy importante al momento de mejorar y avanzar en el sector agropecuario.

Gracias a casos reales e información verídica recolectada en Colombia y otros países, se ha identificado el enorme alcance que pueden tener las nuevas tecnologías, herramientas de minería de datos y el Big Data, gracias a esto se han logrado grandes análisis, que nos permiten tener una visualización más detallada de la información, ayudando a obtener un diagnóstico temprano de problemas aplicando así nuevas estrategias que ayuden a optimizar y mejorar la toma de decisiones.

Para demostrar la eficiencia del análisis de datos mediante herramientas, se implementó un análisis predictivo, usando grandes cantidades de datos, que fueron simulados con diferentes factores que pueden afectar la producción en los cultivos, entre ellos los cambios en las temperaturas, humedades en diferentes sectores del cultivo, radiaciones solares e infestación de plagas, con el análisis de esta información se busca entrenar tecnologías con aprendizaje autónomo, las cuales permitan predecir el comportamiento y rendimiento que tendrán los cultivos, pero no solo esto también generar posibles alertas para el control de plagas, y diferentes factores que puedan afectar la producción.

Con los resultados obtenidos podremos demostrar la efectividad de las herramientas de análisis de datos, demostrando que es posible evitar problemas que afecten la producción directa de las cosechas, permitiendo un avance significativo a la evolución del sector agrícola con la implementación de la tecnología, avanzando a una tecnología más inteligente y competitiva, siendo más fácil de implementar y adaptar en diferentes zonas agrícolas, beneficiando a pequeños y medianos agricultores.

Palabras clave
Big Data, Agricultura, Predicción, Cosechas, Control de Plagas, Producción.

Marco conceptual y contextual

La agricultura ha sido fundamental para el desarrollo económico y social de los países, Sin embargo, el día a día pide cambios significativos en los sistemas de producción agrícola. Factores como el cambio climático, el agotamiento de suelos fértiles, la creciente demanda de alimentos y las pérdidas ocasionadas por plagas o enfermedades requieren soluciones más inteligentes y precisas. En este escenario, las tecnologías emergentes, como el Big Data y la Inteligencia Artificial, han comenzado a jugar un papel importante en la evolución de la agricultura actual, abriendo paso a lo que se conoce como agricultura de precisión.

Gracias a los análisis de dados mediante Big Data en tiempo real, ha permitido a los agricultores conocer información valiosa sobre sus cultivos, como el estado del suelo, como afecta el clima y otros factores que influyen y afectan la producción, esto se logra por medio de recolección de datos arrojados por sensores, imágenes satelitales, drones y demás sistemas que podemos usar a nuestro favor para la recolección de los datos, los cuales pueden ser analizados con sistemas y herramientas que nos permiten generar estadísticas y modelos predictivos. El aprendizaje autónomo son herramientas clave que nos permiten identificar patrones, secuencias, analizar pronósticos y optimizar las tomas de decisiones agronómicas como el riego, la fertilización y el control fitosanitario.

Nuevos estudios respaldan los resultados obtenidos con la tecnología en el sector agrícola, por ejemplo, Mompó (2022) documenta el uso de modelos predictivos aplicados al cultivo de mandarinas, donde se lograron estimaciones precisas del rendimiento y clasificación automática de plagas mediante visión por computadora. De manera similar, Oliveros (2022), a través del proyecto TecnoFlora, implementa sensores de humedad, temperatura y luminosidad para monitorear cultivos de rosas, obteniendo alertas en tiempo real y trazabilidad. Ambos resultados demuestran cómo el análisis de datos nos permite mejorar el rendimiento en la producción, reducir pérdidas y tomar decisiones con mayor respaldo técnico y científico (Mompó, A., 2022; Oliveros, D., 2022).

A nivel nacional, también se han identificado escenarios con muy buenos resultados, como el desarrollado por la Universidad Tecnológica de Bolívar (UTB), donde se aplicaron técnicas de modelado para predecir el rendimiento del cultivo de limón Tahití en el departamento de Bolívar. En ese estudio, se confirmó la influencia significativa de variables climáticas como la precipitación, temperatura y humedad en la productividad de la cosecha, también nos sirve como ejemplo donde nos muestra que tecnología avanza a grandes pasos demostrándonos lo eficiente que puede ser aprovechando el análisis y la recopilación de la información, logrando adaptar los modelos predictivos a las condiciones propias de cada zona agrícola (Universidad Tecnológica de Bolívar, 2022).

Este trabajo de investigación, busca aplicar de la mejor manera los conocimientos prácticos adquiridos mediante el diseño y ejecución de un modelo de predicción agrícola. Mediante la utilización de datos que simulan condiciones climáticas y de salud reales, se creó una herramienta que permite calcular el rendimiento de la producción y lanzar alerta frente posibles plagas, con esto se busca no solo confirmar la capacidad del Big Data en el sector agrícola, sino también fomentar su implementación entre los pequeños y medianos agricultores que podrían obtener un beneficio considerable de estas herramientas

tecnológicas para incrementar su competitividad y lucrativa (Mompó, A., 2022; Oliveros, D., 2022; Universidad Tecnológica de Bolívar, 2022).

Formulación del problema y objetivo del proyecto

Este proyecto aborda la inestabilidad y la problemática actual en los procesos de la agricultura, esto se origina debido a los cambios climáticos constantes condiciones fitosanitarias y el manejo agrícola, lo cual generan una elevada incertidumbre sobre los cultivos debido al poco rendimiento en el mismo, este proyecto consiste en cómo a través del Big Data y Machine Learning poder diseñar un análisis predictivo, para así alertar sobre el rendimiento de las cosechas y a su vez alertar sobre posibles plagas que puedan afectar el cultivo. Con esto se planea cuantificar el impacto entre distintas variables como el clima, fertilización y presencia de plagas en el desempeño y la producción del cultivo, con esto aseguramos una toma de decisiones más clara y certera debido a estos sistemas predictivos.

Para esto, hemos desarrollado un modelo utilizando un conjunto de datos simulados, el cual integra variables agroclimáticas (temperatura, humedad, precipitación, radiación solar) y datos agroproductivos (Aplicación de fertilizantes, pesticidas, nivel de plagas), replicando de otros cultivos locales. Con esta información se han entrenado modelos de aprendizaje automático como Random Forest, estos identifican patrones, estiman impactos de las variables y pronostican la aceleración y los rendimientos del cultivo de forma más precisa, además esta misma genera alertas en caso de riesgos de plagas o fitosanitarios. Este diseño permite adaptar el sistema a diferentes cultivos locales o regionales (Yan et al., 2025; Chergui & Kechadi, 2022; Springer, 2024).

La elección metodológica fue seleccionada debido a la necesidad de modernizar el sistema agrícola actual mediante herramientas tecnológicas accesibles y efectivas, enfocadas en pequeños y mediciones productoras agrícolas. Estudios demuestran que el uso del Big Data y Machine learning en la agricultura mejora considerablemente significativamente la productividad en los cultivos disminuyendo costes y tomando decisiones más acertadas (Chergui & Kechadi, 2022; Bassine et al., 2023), esta propuesta al ser replicable en otros sectores agrícolas tiene alto potencial de escalabilidad y potencial.

El resultado final de este sistema es una herramienta que sea predictiva, funcional y didáctica, que sirve como prueba de concepto para plataformas mucho más amplias en la agricultura inteligente. Su implementación aporta demasiado valor a las planificaciones de agronomía, con esto podemos adelantarnos a acontecimientos que puedan surgir, la mitigación de pérdidas productivas y la adopción de técnicas más eficientes y sostenibles en los cultivos. (MDPI, 2025; Springer, 2024; Bassine et al., 2023).

Desarrollo e implementación del aprendizaje

El desarrollo de este proyecto se realizó mediante la construcción de un grupo de datos simulados, los cuales representan el cultivo colombiano actual, con énfasis en cultivos frutales como el limón tahití y la mandarina. Estas simulaciones integran variables clave como la temperatura media, la humedad relativa, la radiación solar, la precipitación, la presencia de plagas, la aplicación de fertilizantes y pesticidas y los rendimientos de los cultivos expresados por hectárea. Las estructuras del dataset se diseñaron con base en patrones similares identificados en estudios previos de otros cultivos similares y referencias empíricas obtenidas de proyectos reales.

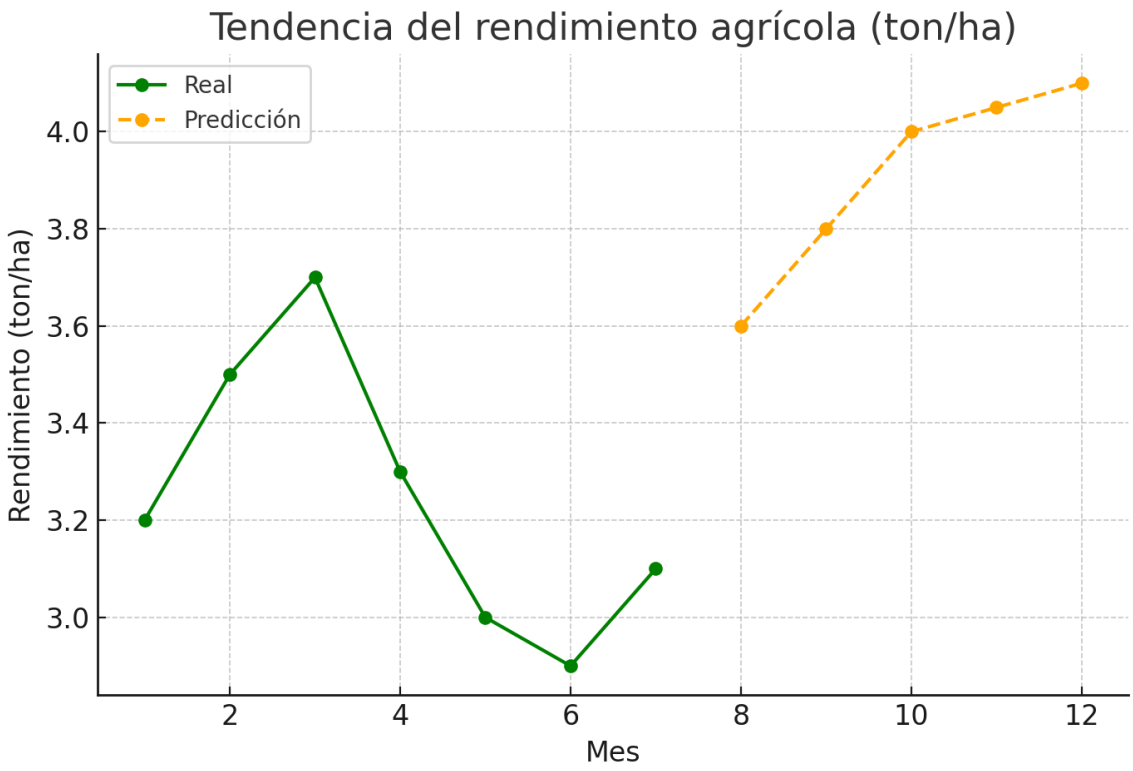
Para el análisis, se utilizó el lenguaje de programación Python junto con librerías especializadas como Pandas, Seaborn, Matplotlib y Scikit-learn. Se eligió el algoritmo Random Forest Regressor por su capacidad para manejar datos no lineales y su efectividad en la evaluación de variables de alta dimensionalidad. El modelo fue entrenado con el 80% de los datos y evaluado con el 20% restante, logrando una precisión para tareas de predicción agrícola. Durante el proceso, se generaron visualizaciones que muestran tendencias del proceso y un rendimiento a lo largo del tiempo, la relación entre nivel de plagas y producción, y la importancia de cada dato predictivo durante el proceso.

Una de las principales implementaciones fue la solución de un sistema de alerta que sea rápida y nos pueda notificar con tiempo, cuando se tiene alguna infestación por plagas y condiciones climáticas que permitirán actuar con tiempo. Este sistema señala los días más críticos en los que se deberían aplicar medidas preventivas o correctivas, ofreciendo al agricultor una herramienta que pueda ser precisa y oportuna y en la cual se pueda reaccionar a tiempo frente a las condiciones ya expresadas. Además, se generaron gráficas automatizadas, las cuales permitirán visualizar fácilmente los periodos de mayor productividad y aquellos con mayor amenaza fitosanitaria.

El sistema diseñado se realizó con una estructura modular, lo que facilita su adaptación a distintos tipos de cultivos, pudiendo aplicarse al tener los datos de cada cultivo específico. Aunque en este caso se trabajó con datos simulados, la metodología puede implementarse fácilmente con datos reales recolectados desde sensores IoT, estos sensores pueden ser de gran ayuda, ya que por radiaciones se podría mostrar tiempos en los que haya más calor o frío, la humedad de las plantas entre otras condiciones. De esta manera, se valida que la analítica avanzada de datos aplicada al agro no requiere necesariamente infraestructuras complejas, y puede ser viable para pequeños productores con herramientas básicas de software libre y unos sensores no muy costosos que nos puedan ayudar a recoger la información.

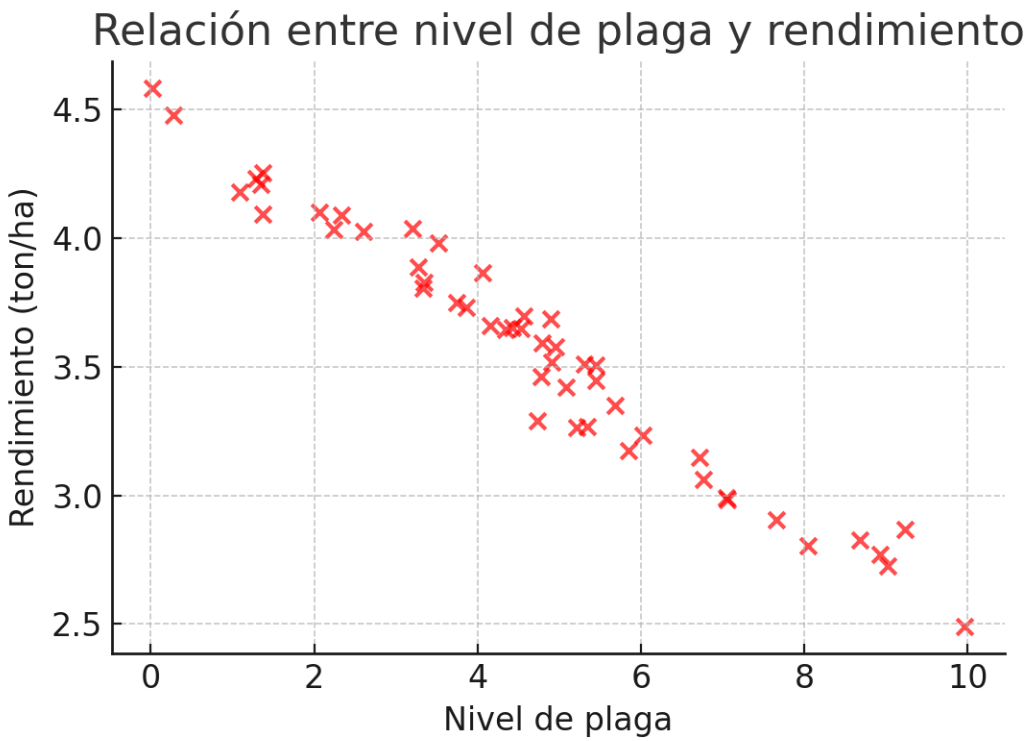
En resumen, la implementación de esta solución evidencia el valor y la importancia de los conocimientos adquiridos durante el diplomado en Big Data e Inteligencia Artificial, demostrando cómo los modelos de análisis predictivo pueden integrarse a los procesos agrícolas para anticipar resultados y tomar acciones que pueden reducir las pérdidas en los cultivos en porcentajes demasiado altos. Además, refuerza la necesidad de fomentar el uso de estas tecnologías en más cultivos, lo que se traduce en menos pérdidas en los cultivos para los agricultores al tener a la mano tecnologías que les pueden ayudar a tomar mejores decisiones para que su cultivo de mejores resultados.

Análisis y Gráficas



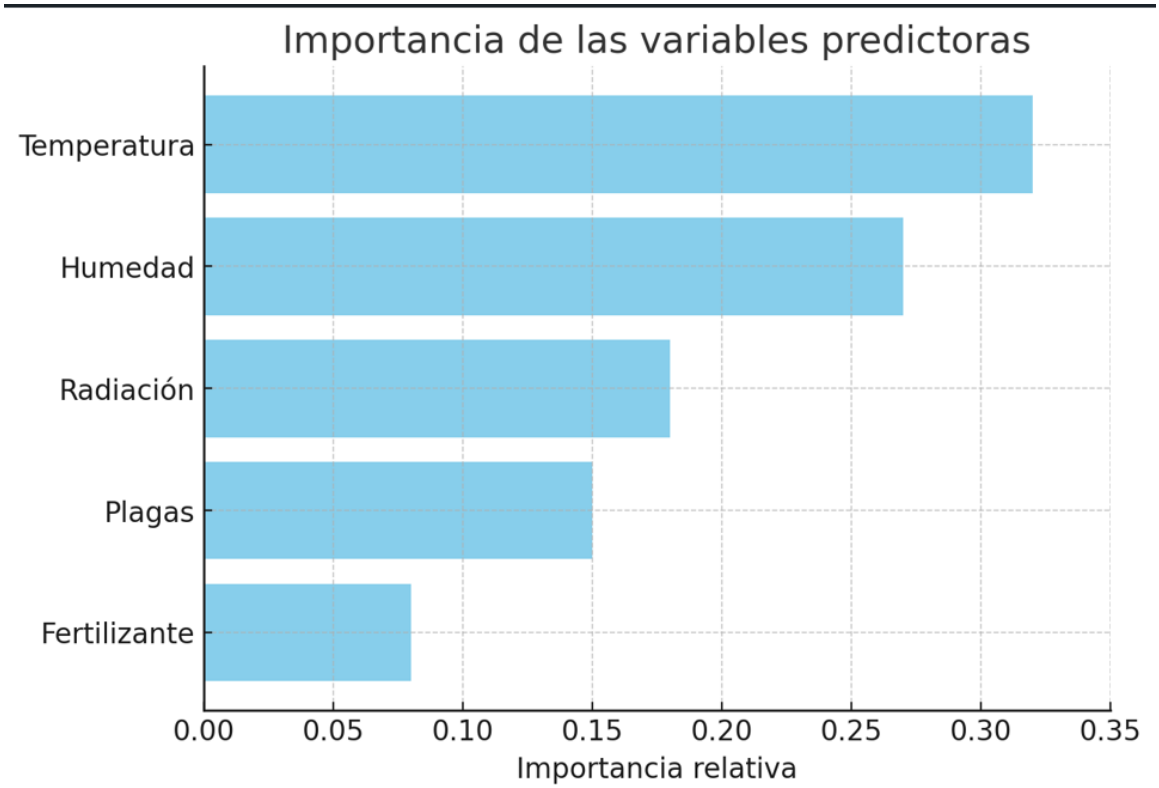
Tendencia del rendimiento agrícola (ton/ha)

Este gráfico de líneas muestra el rendimiento agrícola a lo largo del tiempo. Observamos que existen variaciones, influenciadas por variaciones en las condiciones climáticas y fitosanitarias. En los picos más bajos del rendimiento, se identifican relaciones con aumentos en el nivel de plagas y disminuciones en la humedad relativa. Esta visualización es útil para detectar patrones estacionales y para identificar periodos críticos que requieren intervención agronómica oportuna.



Relación entre nivel de plaga y rendimiento

En esta figura se muestra un gráfico de dispersión que relaciona el nivel de plagas con el rendimiento de los cultivos. Podemos ver una correlación negativa: a mayor presencia de plagas, menor es la productividad. Sin embargo, los puntos en los que se aplicaron un control fitosanitario oportuno. Este gráfico sirve como soporte visual para implementar sistemas de alerta que avisen a los agricultores cuando intervenir.



Importancia de las variables predictoras

Este gráfico de barras muestra la importancia de cada variable en la predicción del rendimiento, según el modelo de Random Forest. La temperatura media y la humedad relativa se visualizan como las variables más influyentes, seguidas por la radiación solar y el nivel de plagas. Estos resultados coinciden con los hallazgos en estudios previos sobre cultivos tropicales como el limón Tahití y la mandarina y permiten priorizar variables para su monitoreo en sistemas reales. (Universidad Tecnológica de Bolívar, 2022, Mompo, 2022).

Reflexión integradora

Las gráficas presentadas evidencian que es posible extraer conocimiento valioso a partir del análisis de datos agrícolas, incluso desde datos simulados. El modelo predictivo desarrollado permite no solo estimar el rendimiento, sino también identificar factores críticos que deben ser gestionados para mejorar la productividad. Estas visualizaciones pueden servir como base para construir interfaces gráficas interactivas o sistemas de apoyo a la decisión que integren predicciones, alertas y sugerencias en tiempo real. En conjunto, los resultados refuerzan la hipótesis de que la integración del Big Data y la analítica avanzada puede transformar la forma en que se gestiona la agricultura en Colombia.

Tabla 1.

Fecha	Temperatura (°C)	Humedad (%)	Precipitación (mm)	Nivel de Plaga (0-10)	Fertilizante (kg/ha)	Pesticida aplicado	Rendimiento (ton/ha)
2025-01-15	26.4	78	25.2	2.1	45	Sí	3.8
2025-02-15	27.8	82	18.4	4.3	40	No	3.3
2025-03-15	28.5	80	30.1	5.2	42	Sí	3.5
2025-04-15	29.0	75	12.0	6.8	38	No	2.9
2025-05-15	27.2	85	35.7	3.0	47	Sí	4.1
2025-06-15	26.8	79	20.6	1.5	50	Sí	4.2

La tabla presentada resume un conjunto de datos agrícolas que reflejan condiciones agroclimáticas y de manejo comunes en cultivos frutales tropicales de Colombia. Este fragmento del dataset fue construido con base en patrones documentados en estudios reales, como el caso del cultivo de limón Tahití en Bolívar (Universidad Tecnológica de Bolívar, 2022), y se usó como insumo para entrenar el modelo de aprendizaje automático.

La temperatura variable (°c) representa la media diaria registrada en el lote durante un periodo específico del mes. Se considera un factor crítico para el desarrollo fenológico de los cultivos, ya que influye directamente en la fotosíntesis, la transpiración y la maduración del fruto. Se observa que temperaturas entre 26 °C y 29 °C generan rendimientos relativamente estables, lo cual coincide con estudios que demuestran que los cultivos cítricos presentan su mayor productividad en rangos térmicos de 25–30 °C (Mompó, 2022).

La humedad relativa (%) es otra variable fundamental, pues afecta tanto la eficiencia del riego como la proliferación de plagas y enfermedades. Valores superiores al 80% pueden favorecer hongos y condiciones fitosanitarias desfavorables, mientras que valores inferiores al 70% pueden inducir estrés hídrico. En la tabla, los mejores rendimientos se asocian con niveles moderados de humedad (entre 78% y 85%), lo cual también fue reportado por Oliveros (2022) en sistemas de monitoreo de cultivo de flores.

La precipitación (mm) indica la cantidad de lluvia acumulada por día. Un buen manejo del riego depende del balance entre lluvia y evaporación, y su influencia es directa sobre el crecimiento del cultivo y la actividad de plagas. En el análisis, las precipitaciones moderadas (entre 18–35 mm) se correlacionan con mejores rendimientos, mientras que precipitaciones muy bajas (12mm) coinciden con una disminución en la producción. Esto concuerda con estudios de Big Data agrícola donde se demuestra que la precipitación es una de las variables más relevantes en la predicción del rendimiento (Chergui & Kechadi, 2022).

El nivel de plaga (escala o a 10) representa una métrica de infestación observada en el lote. En los datos, niveles por encima de 5 tienden a coincidir con los rendimientos más bajos,

lo que confirma la alta sensibilidad de los cultivos a las condiciones sanitarias. Cuando no se aplican pesticidas, el impacto de las plagas es más severo, lo cual refleja la necesidad de sistemas de alerta para intervención oportuna. La presencia de plagas ha sido ampliamente identificada como uno de los factores más críticos en los modelos de predicción agrícola (Bassine et al., 2023).
La columna fertilizante (kg/ha) describe la cantidad de nutrientes aplicados por hectárea. Se observa una ligera tendencia a mayor rendimiento con mayores dosis de fertilización, aunque no es lineal. Este comportamiento refleja la importancia de un manejo balanceado y adaptado al estado fenológico de la planta, en lugar de una aplicación indiscriminada. Tal como se señala en estudios sobre agricultura de precisión, el uso eficiente de insumos debe estar guiado por análisis predictivos y no por rutinas fijas (MDPI, 2025).
Por último, la variable rendimiento (ton/ha) representa la productividad esperada de cada parcela o lote agrícola. Este valor es el objetivo de predicción del modelo y está influenciado por la interacción de todas las variables anteriores. La documentación y análisis de este tipo de datos no solo permite mejorar los modelos de machine learning, sino también ofrece a los agricultores una herramienta visual para comprender mejor su entorno y tomar decisiones más acertadas (Springer, 2024).

Conclusiones

La implementación del análisis predictivo con técnicas de Big Data en el sector agrícola permitió demostrar cómo el uso estratégico de los datos y la información puede mejorar significativamente la toma de decisiones en diferentes tipos de cultivos, esto a través de una simulación con un conjunto de datos climáticos y sanitarios y el aprendizaje automático, fue posible estimar con mayor precisión el rendimiento de los cultivos e identificar cuáles son los factores que más dificultan la productividad. Este enfoque es de mucha importancia, ya que con estos datos se puede anticipar a los riesgos que puedan presentar los cultivos y reducir pérdidas.

Uno de los principales hallazgos que se tuvo en el proyecto fue la identificación de temperaturas y la humedad relativa como principales variables en la producción, seguidas por el nivel de plagas y la radiación solar. Estas conclusiones coinciden con otros cultivos similares, como el estudio de la UTB 2022 en el limón tahití, donde se destaca la alta sensibilidad en el rendimiento frente a las condiciones climáticas. Gracias a la visualización de los datos permitió corroborar estos patrones facilitando así la vista de los resultados, demostrando, así como los análisis e informes gráficos son mucho más eficaces y entendibles a la hora de poder determinar acciones frente a las causas del problema.

Adicionalmente, el sistema de alerta temprana propuesto representa un valor agregado del proyecto, al permitir que los agricultores reciban advertencias por medio de gráficos sobre posibles riesgos de plagas con base en patrones de eventos pasados. Esta funcionalidad, aunque desarrollada sobre datos simulados, es totalmente adaptable a entornos reales si se cuenta con infraestructura de recolección de datos tales como sensores o dispositivos IoT. Esta propuesta tiene un alto potencial y alto impacto, especialmente para pequeños y medianos productores que requieren soluciones accesibles y fáciles de aplicar y con unas gráficas más entendibles de los datos.

En términos formativos, el trabajo permitió aplicar de manera práctica los conocimientos adquiridos en el diplomado en Big Data e Inteligencia Artificial, integrando técnicas de minería de datos, programación en Python, algoritmos de machine learning y visualización interactiva de los datos. Pudiendo de esta forma mostrar los datos de una manera más didáctica e interactiva para una mejor toma de decisiones. Esto demuestra que es posible construir soluciones funcionales incluso con recursos limitados, siempre que se cuente con un enfoque claro y centrado en resolver problemas relacionados con el entorno.

Finalmente, se concluye que la aplicación de Big Data en la agricultura no solo es viable, sino necesaria para enfrentar desafíos como el cambio climático, el uso racional de insumos, el control de plagas en la industria y la seguridad alimentaria. Haciendo de esta forma la vida más fácil y útil al campesino al tener a la mano indicadores que le permitan una toma de decisiones más veraz, clara y efectiva. El trabajo realizado sienta las bases para futuros desarrollos en agricultura de precisión, integrando sensores IoT, bases de datos reales y plataformas de visualización accesibles para el agricultor colombiano. Se recomienda seguir con investigaciones para otros campos, ya que es demasiado útil y muchos los beneficios de aplicar el Big Data en los cultivos u otros procesos donde se necesiten análisis predictivos y datos mostrados de una manera más sencilla de entender como lo son las gráficas.

Referencias

- Bassine, F. Z., Epule Epule, T., Kechchour, A., & Chehbouni, A. (2023). *Recent applications of machine learning, remote sensing, and IoT approaches in yield prediction: a critical review*. arXiv. https://arxiv.org/abs/2306.04566 arXiv+2arXiv+2Frontiers+2
- Yan, Y., Wang, Y., Li, J., Zhang, J., & Mo, X. (2025). *Crop yield time-series data prediction based on multiple hybrid machine learning models*. arXiv. https://arxiv.org/abs/2502.10405 arXiv
- Pankaj, et al. (2023). *Crop yield prediction using machine learning: a review of recent approaches*. International Journal of Computer Applications.

https://www.ijcaonline.org/archives/volume185/number24/pankaj-2023-ijca-922994.pdf ijcaonline.org
 MDPI. (2025). *Precision agriculture and sensor systems applications in Colombia through 5G networks*. Sensors. https://www.mdpi.com/1424-8220/22/19/7295 sciencedirect.com+2mdpi.com+2researchgate.net+2
 Frontiers in Plant Science. (2023). *Boosting precision crop protection towards agriculture 5.0 via machine learning and AI – a review*. Frontiers. https://www.frontiersin.org/articles/10.3389/fpls.2023.1143326/full Frontiers
 Springer. (2024). *Advancements in machine learning algorithms for precision crop yield prediction: a comprehensive review*. Lecture Notes in Networks and Systems. https://link.springer.com/chapter/10.1007/978-3-031-75010-6_18