Predicción del rendimiento de cultivos agrícolas en los cinco corregimientos de la ciudad de Medellín, utilizando modelos de *Machine Learning*

Alba Miriam Gómez Arango

Director: Edison Valencia Diaz
Escuela de Ciencias Aplicadas e Ingeniería.
evalenci@eafit.edu.co
Docente del Área de Computación y Analítica de Datos

Co - director: Juan Fernando Zuluaga Orrego
Coordinador Nacional Territorial
Organización de Agricultura Familiar para la Alimentación y la Agricultura
Juan.zuluagaorrego@fao.org
319 207 5459

UNIVERSIDAD EAFIT
ESCUELA DE ADMINISTRACIÓN
MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA
MEDELLIN
2024

1 Resumen

En un contexto global donde la agricultura y la producción de alimentos desempeñan un papel crucial para la seguridad alimentaria, el empleo y la sostenibilidad, este estudio se centra en predecir rendimiento de los cultivos agrícolas presentes en los cinco corregimientos de Medellín. El objetivo principal consiste en diseñar un modelo de predicción para nueve cultivos locales mediante el uso de técnicas de aprendizaje automático (*Machine Learning*).

Medellín se distingue por su diversidad de cultivos, que incluyen una agricultura periurbana caracterizada por microparcelas productivas distribuidas en varios cultivos de tipo chagra. Estas prácticas agrícolas tradicionales son llevadas a cabo por una población de agricultores envejecida. La precisión en la predicción del rendimiento se vuelve esencial, ya que una parte significativa de la producción se destina al autoconsumo, con un enfoque de subsistencia. Sin embargo, también se comercializan excedentes, lo que impacta directamente en la seguridad alimentaria de la comunidad local.

Los resultados destacan la efectividad de los modelos de aprendizaje automático, en particular, los modelos de Boosting como PCA Random Forest y PCA XGB Boosting, en la predicción de los cultivos objeto de estudio. Estos modelos muestran la capacidad para capturar las relaciones entre las variables y la heterogeneidad presente en la producción del territorio. Sin embargo, se han identificado oportunidades de mejora relacionadas con la reducción del error de los modelos, las cuales pueden abordarse mediante la recopilación continua de datos y el apoyo técnico brindado a los agricultores. Esto no solo aumentará la disponibilidad de datos, sino que también contribuirá a perfeccionar el modelo y a comprender el comportamiento del rendimiento en los cultivos analizados, facilitando la toma de decisiones en el sector agrícola del municipio de Medellín.

Este proyecto representa una herramienta valiosa tanto para profesionales del sector agropecuario como para las instituciones encargadas de la planificación y desarrollo agrícola. Así mismo, ofrece un enfoque innovador al análisis de datos del sector, aprovechando las ventajas de la ciencia de datos. A través de estas técnicas, se abren oportunidades para establecer estrategias, planes y proyectos que contribuyan a la planificación de siembras, la gestión de las zonas productivas del municipio y el fortalecimiento de la seguridad alimentaria local.

Palabras clave: *Seguridad Alimentaria; Producción agrícola, rendimiento de cultivos, Machine Learning, regresión.*

Índice de Contenidos

Índice de Ilustraciones

Índice de tablas

Abreviaturas y acrónimos

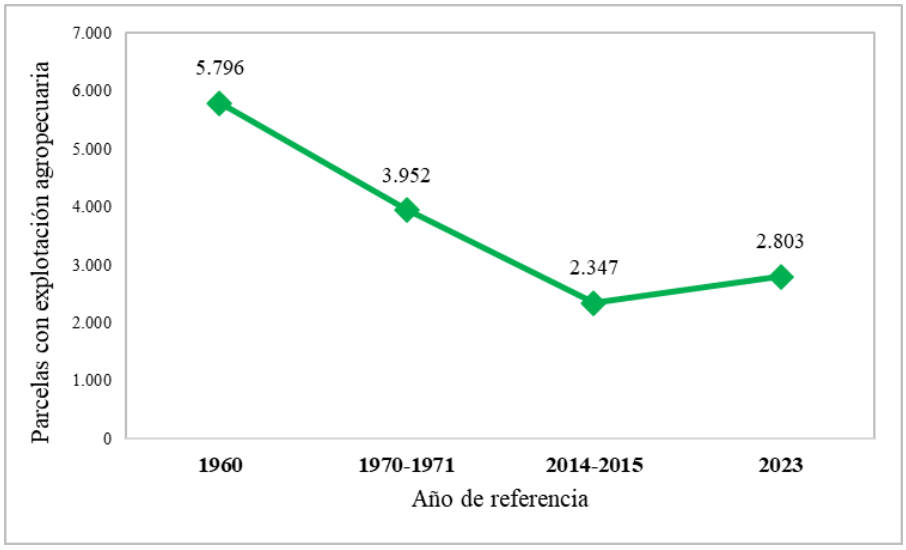
ANN	Red Neuronal Artificial *(por sus siglas en inglés)*
CEO	Centro de Estudios de Opinión
CNN	Redes Neuronales Convolucionales
COMER	Conexión Medellín Rural *(por sus siglas en inglés)*
CRISP-DM	Cross Industry Standard Process for Data Mining
DANE	Departamento Administrativo Nacional de Estadística
DNN	Redes Neuronales Profundas *(por sus siglas en inglés)*
FAO	Organización de las Naciones Unidas para la Alimentación y la Agricultura *(por sus siglas en inglés)*
MAE	Error Absoluto Medio *(por sus siglas en inglés)*
MedAE	Error Mediano Absoluto *(por sus siglas en inglés)*
MLR	Regresión Lineal Múltiple *(por sus siglas en inglés)*
NDVI	Vegetación de Diferencia Normalizada *(por sus siglas en inglés)*
ODS	Objetivos de Desarrollo Sostenible *(por sus siglas en inglés)*
PCA	Análisis de Componentes Principales *(por sus siglas en inglés)*
POT	Plan de Ordenamiento Territorial *(por sus siglas en inglés)*
RF	Bosque Aleatorio *(por sus siglas en inglés)*
RNN	Redes Neuronales Recurrentes *(por sus siglas en inglés)*
SD	Desviación Estándar *(por sus siglas en inglés)*
SVM	Máquina de Vectores de Soporte *(por sus siglas en inglés)*
TCN	Red Convolucional Temporal *(por sus siglas en inglés)*
UNFPA	Fondo de Población de las Naciones Unidas *(por sus siglas en inglés)*

2 Planteamiento del problema

La distribución poblacional en el municipio de Medellín ha provocado una disminución en las áreas de tierra destinadas a la producción de alimentos. Específicamente, las parcelas dedicadas a la agricultura campesina, familiar y comunitaria han mostrado una tendencia negativa y una disminución constante a lo largo del tiempo, según datos de censos agropecuarios realizados en el periodo de tiempo comprendido entre los años 1960 y 1970, así como información recopilada por el Centro de Estudios de Opinión (CEO) de la Universidad de Antioquia entre el año 2014 y el año 2015. (ver Figura 1)

En el año 2023, durante el proceso de Caracterización Rural llevado a cabo en el marco del Convenio de Cooperación Internacional entre el Municipio de Medellín y la FAO, se observó un aumento de aproximadamente 500 parcelas con explotación agropecuaria y forestal en las áreas corregimentales. Este aumento se atribuye en gran medida a la partición de espacios productivos en procesos hereditarios. Este cambio en la dinámica de producción señala la complejidad de los factores que influyen en la distribución de la tierra y la producción de alimentos en la región.

Figura 1. Cantidad de parcelas con explotación agropecuaria en los territorios rurales del Distrito de Medellín (1960 -2023)



Fuente: Tomado de DANE: Censos Agropecuarios 1960 y 1070/71 en (CEO & Alcaldía de Medellín, 2016)

Además de los cambios en las áreas de producción, se ha observado una disminución del 35% en la población rural dedicada a la agricultura en el período comprendido entre la caracterización del CEO

y la caracterización rural de 2023 en el Distrito. Esta cifra ha pasado de 11.978 personas entre los años 2014 y 2015 a 7.765 personas en 2023. La actual población productora de alimentos en las zonas rurales del Municipio se caracteriza por ser, en su mayoría, de edad avanzada, con un índice de envejecimiento de 2.9 y una edad promedio de los agricultores de 58 años.

Ante los cambios demográficos y en la producción agrícola, se hace necesario reevaluar el territorio y adaptarlo a una realidad en constante transformación donde se promueva la preservación de los espacios rurales como una oportunidad para el fomento de la producción de alimentos. Esto implica la generación de incentivos para la agricultura en estas áreas, la promoción de programas de formación y financiamiento, así como el acceso a tecnologías agrícolas modernas que optimicen la productividad sin agotar los recursos naturales y faciliten la transición de parcelas a las generaciones más jóvenes. No obstante, un desafío importante radica en la limitada disponibilidad de datos, lo que dificulta la toma de decisiones informadas en el sector agrícola del país y la ciudad. La información se torna esencial en la búsqueda de soluciones.

En el contexto de los desafíos mencionados, uno de los enfoques más prometedores para impulsar mejoras significativas es la aplicación estratégica de la ciencia de datos. Este método, respaldado por herramientas y técnicas avanzadas, permite analizar conjuntos de datos, identificar patrones, predecir tendencias y tomar decisiones informadas. Esta herramienta resulta efectiva al transformar la información limitada en conocimientos valiosos, orientando estrategias para impulsar la producción de alimentos y mejorar la calidad de vida de quienes habitan en las zonas rurales y urbanas por igual. Al emplear esta disciplina de manera estratégica, no solo se fortalece la sostenibilidad agrícola, sino que se genera oportunidades para el desarrollo económico local y reducción en las presiones que impulsan la migración hacia entornos urbanos.

3 Justificación

El uso de la ciencia de datos como herramienta de predicción del rendimiento de cultivos agrícolas representa una contribución significativa en el ámbito de la agricultura, especialmente en un contexto global marcado por desafíos como el crecimiento de la población, la degradación de los recursos naturales y el cambio climático. La necesidad de asegurar la disponibilidad de alimentos en cantidades suficientes para una población en constante aumento exige un cambio de enfoque en la gestión de la producción agrícola (FAO, 2018).

La predicción del rendimiento de los cultivos agrícolas contribuye a la seguridad alimentaria en el sentido que permite planificar de manera más eficiente la producción mejorando la disponibilidad de alimentos y optimizando el uso de recursos. Además, desde una perspectiva económica, contribuye a la reducción de riesgos, mejora la rentabilidad y estabiliza los precios en el mercado. En términos ambientales, promueve prácticas agrícolas sostenibles y respalda la conservación del entorno. Tradicionalmente, la evaluación del rendimiento en cultivos se ha llevado a cabo de manera retrospectiva de corto plazo, esperando hasta la cosecha para obtener datos concretos. Sin embargo, la ciencia de datos permite un enfoque predictivo, anticipando el rendimiento de los cultivos mediante modelos basados en una diversidad de variables, desde condiciones climáticas hasta características del suelo (FAO, 2022).

El rendimiento de los cultivos es un fenómeno multifactorial, influenciado por elementos como las condiciones climáticas, la calidad del suelo y el acceso al agua, entre otros. En este contexto, la ciencia de datos emerge como una herramienta invaluable al posibilitar la integración de diversas variables en modelos predictivos. Dada la complejidad inherente de estos factores, se requiere un enfoque analítico avanzado para identificar patrones y relaciones que de otra manera serían difíciles de comprender (FAO, 2022).

En el caso específico de los corregimientos de Medellín, la aplicación de la ciencia de datos no solo aporta a la mejora en la eficiencia de la producción de alimentos, sino que también proporciona a los agricultores y a nivel sectorial información para la toma de decisiones. Este proyecto busca aplicar metodologías de ciencia de datos para prever el rendimiento de los cultivos en esta región específica. Los resultados de esta investigación pueden influir directamente en la toma de decisiones de políticas agrícolas, aumentando la sostenibilidad y la eficacia de las prácticas agrícolas locales.

4 Objetivos

4.1 Objetivo general

Diseñar un modelo de predicción del rendimiento de los cultivos en las áreas rurales cercanas a la ciudad de Medellín mediante la aplicación de técnicas de *Machine Learning*.

4.2 Objetivos específicos

1. Realizar un análisis exploratorio y preprocesamiento del conjunto de datos para la implementación de los modelos de *Machine Learning* seleccionados.
2. Desarrollar modelos de predicción utilizando el conjunto de datos con el fin de abordar el problema de pronóstico de rendimientos de cultivos en el Distrito de Medellín.
3. Evaluar los resultados de los modelos seleccionados y comparar las medidas de calidad.

5 Estado del arte y Marco teórico

5.1 La producción de Alimentos en el mundo

Con el crecimiento de la población, la degradación de los recursos naturales, el cambio climático, los avances tecnológicos, la distribución de ingresos y los cambios en la manera de concebir las dietas alimenticias, la producción de alimentos en el mundo debe afrontar unos retos e incertidumbres de cara a la garantía de hambre cero, del derecho humano a la alimentación y la búsqueda armónica del cumplimiento de los demás objetivos de desarrollo sostenible propuestos en la Agenda 2030 de las Naciones Unidas (FAO, 2023)

Paradójicamente, la producción agrícola se ve limitada en el mundo debido a la pérdida de la calidad de los recursos naturales, especialmente los suelos y el agua. Esta degradación en muchos sentidos, está provocada por la misma actividad productiva del sector que generan un gran impacto ambiental a través de la generación de gases de efecto invernadero, vertimiento de agroquímicos a los afluentes, incremento de la huella hídrica y de carbono, pérdida de biodiversidad y pérdida del hábitat de especies que contribuyen en diferente medida al incremento de la temperatura global y alteraciones climáticas que conducen a sequías prolongadas e inundaciones constantes, entre otras. En este sentido, se identifica la necesidad de cambios estructurales en la manera como se han venido desarrollando las actividades, toda vez que el ritmo de producción y consumo podrían derivar en inseguridad alimentaria persistente y un crecimiento económico insostenible (FAO, 2018).

5.2 Rendimiento y/o productividad de los cultivos

El rendimiento de los cultivos, en términos generales, se refiere a la cantidad de producto comercial que se obtiene en una determinada área de superficie durante un ciclo de cultivo. Tradicionalmente, se solía esperar hasta la cosecha para calcular este dato considerando el total producido. Sin embargo, la disponibilidad de información oportuna y anticipada sobre la producción agrícola es beneficiosa para la gestión efectiva de actividades como la cosecha, el almacenamiento, los procesos de importación o exportación, el transporte y la comercialización de alimentos. También es esencial para definir políticas de precios, optimizar la eficiencia y rentabilidad de las inversiones, así como para minimizar riesgos económicos tanto para gobiernos, productores, compañías de seguros y otros actores sociales (Yildirim et al., 2022).

El rendimiento o la productividad de los cultivos agrícolas depende de muchos factores, entre ellos las condiciones climáticas, calidad del suelo, acceso al recurso hídrico, temperatura y precipitaciones. En este sentido, proponer modelos de predicción para este ítem requiere la combinación de la mayor cantidad de variables predictoras posibles de modo que se reduzca el error en las estimaciones y proyecciones productivas (Jhajharia & Mathur, 2022).

5.3 Técnicas de estimación de rendimiento de cultivos con algoritmos de aprendizaje automático

Desde el campo de la inteligencia artificial existen dos ramas fundamentales de estudio actual que han contribuido al desarrollo de nuevas técnicas precisas para la solución de problemas de regresión y clasificación en todo el mundo: el aprendizaje automático y el aprendizaje profundo.

En el contexto específico de la agricultura, el aprendizaje automático se convierte en una tecnología funcional para predecir el rendimiento de los cultivos. Esto se debe a su enfoque en estrategias de autoaprendizaje que permiten identificar patrones de asociación entre el rendimiento histórico anual de los cultivos y los datos relacionados con el rendimiento, lo que a su vez mejora las predicciones. En cambio, el aprendizaje profundo amplía su utilidad al permitir la inclusión de datos de teledetección, imágenes y análisis profundos de datos secuenciales (Sciforce, 2022).

A continuación, se detallarán los modelos, variables y técnicas de aprendizaje automático que se emplean en el análisis y la predicción del rendimiento de los cultivos agrícolas:

5.3.1 Algoritmo de regresión

5.3.1.1 Regresión lineal múltiple

Es un método estadístico que busca modelar la relación entre una variable continua y dos o más variables independientes mediante el ajuste de una ecuación lineal por mínimos cuadrados ordinarios. Existen limitaciones a considerar en el empleo de este tipo de modelos, tales como: Afectación en caso de inclusión de predictoras correlacionadas, se incorporan en el modelo todas las variables predictoras, incluso si no contienen información relevante y no puede ajustarse en los casos donde los registros sean inferiores a las variables a incluir en el modelo (IBM, 2022). Existen estrategias de mitigación a las limitaciones listadas anteriormente como los métodos de regularización, Lasso, Ridge y Elastic net.

- Regresión LASSO: Este método se conoce popularmente como L1 y penaliza la suma del valor absoluto de los coeficientes. Su objetivo es excluir los predictores no influyentes en el modelo (Amat, 2020).
- Regresión Ridge: Método de regularización L2 que penaliza la suma de los coeficientes elevados al cuadrado. Su objetivo es reducir de forma proporcional el valor de todos los coeficientes del modelo, pero sin que estos lleguen a cero. El grado de penalización está controlado por el hiper parámetro λ, donde medida que crece el su valor la penalización aumenta y el valor de los predictores disminuye. Su principal ventaja es la reducción de la varianza del modelo (Amat, 2020).
- Elastic net: Método de regularización que combina la penalización L2 y L2 que suele dar muy buenos resultados. El grado en que influye cada una de las penalizaciones está controlada por el hiper parámetro α. Y en los casos dónde exista cierta colinealidad entre varias características predictivas se escoger una o todas, de acuerdo con como haya sido parametrizado (Amat, 2020).

Con respecto a la aplicabilidad de estos métodos en modelos para la predicción del rendimiento en cultivos se listan a continuación algunos ejemplos encontrados en la literatura:

En el año 2022, Singh et al., realizaron un estudio de predicción del rendimiento del trigo a partir de datos históricos de alrededor de 20 años y datos meteorológicos recopilados en los distritos de

Amritsar, Bhatinda y Ludhiana en Punjab, India, En este emplearon modelos de regresión como LASSO, Elastic Net y Ridge Regression. Los resultados destacan a Elastic Net como el más eficiente. El modelo LASSO mostró buen rendimiento en Udham Singh Nagar y Nainital, pero pobre en otros distritos. Elastic Net demostró buen rendimiento en todos los distritos, aunque se observó subestimación en la etapa de validación. Por su parte, el modelo Ridge presentó rendimiento variado, siendo bueno en algunos distritos como Udham Singh Nagar, Nainital y Haridwar, pero deficiente en otros.

En el caso colombiano, en el departamento del Cauca, se analizó la viabilidad de la predicción temprana del rendimiento en cultivos de café arábico, basado en el uso de imágenes aéreas multiespectrales. En este caso particular, se emplearon métodos estadísticos como la regresión lineal y la regresión de Descenso de Gradiente Estocástico para modelar la relación entre variables predictoras, como el volumen del árbol, el NDVI, el Índice de Madurez del Café, y el rendimiento del cultivo. La validación del modelo se llevó a cabo mediante un enfoque de validación cruzada, evaluando métricas como R^2. El estudio identificó limitaciones, como el tamaño de la muestra y la necesidad de validar los modelos con más datos. Las conclusiones destacan la viabilidad de la teledetección para prever el rendimiento del café, subrayando la importancia de la segmentación de plantas en el procesamiento de imágenes (Bolaños et al., 2023).

Adicionalmente, en el departamento del Meta, un estudio analizó datos climáticos e índices agroclimáticos para entender el impacto en la productividad de los agroecosistemas de naranja 'Valencia' durante un periodo de tres años (2013-2015). Para ello, se desarrolló un modelo empírico mediante regresión lineal múltiple, identificando que la precipitación fue el factor climático más influyente, explicando el 49% de la variación en la productividad. Aunque el modelo ofrece una herramienta valiosa para estimar la productividad en la región, presenta limitaciones al no generalizar sus resultados a otras áreas y no considerar todos los factores que afectan la productividad (Cleves et al., 2023).

5.3.2 B. Algoritmos de árbol de decisión

5.3.2.1 Árboles de decisión

Representa un de las herramientas más poderosas y populares para el análisis en problemas de clasificación y predicción. Presenta una estructura similar a un diagrama de flujo, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Este algoritmo emplea una estrategia iterativa que divide en submuestras los datos para identificar los puntos de división óptimos dentro de un árbol. La facilidad en la interpretación, flexibilidad y pocos requerimientos en la preparación de los datos son sus principales ventajas. Sin embargo, es un algoritmo con un alto costo computacional, propenso a sobre ajuste y puede generar estimadores con alta varianza(UNIR Revista, 2021).

En el año 2022 en Pahang, Malasia, se emplearon datos históricos de 35 años de rendimiento de frutos frescos de palma de aceite, así como 12 parámetros meteorológicos y tres de humedad del suelo, con el propósito de estimar el rendimiento en el cultivo de palma de aceite. El estudio aplicó diferentes algoritmos de árboles de decisión. Estos modelos fueron comparados con otros métodos de regresión tradicional y el rendimiento general de todos los modelos de regresión basados en árboles es considerablemente mejor en la predicción del rendimiento.
De acuerdo con la revisión de literatura, este modelo se utilizó para la predicción del rendimiento en cultivos de té en Pakistán. Este modelo utilizó datos meteorológicos, de suelo, de cultivo y de gestión agrícola en el periodo de tiempo entre el 2016 al 2019, combinada con datos de simulación, y en términos generales el algoritmo de regresión de aprendizaje automático se comportó mejor en la predicción del rendimiento utilizando menos datos que el modelo de simulación de AquaCrop[1]

En Colombia, específicamente en el departamento de Santander, se llevó a cabo un estudio enfocado en la identificación de factores clave que impactan el rendimiento de cultivos de cacao con el propósito de mejorar la toma de decisiones agrícolas y prever la productividad. Se aplicaron y compararon diversos algoritmos de aprendizaje automático, como máquinas de soporte vectorial (SVM), modelos ensamblados (Random Forest, Gradient Boosting) y el modelo de regresión LASSO. Los predictores considerados incluyeron condiciones climáticas, variedad de cacao, fertilización y exposición al sol en un cultivo experimental en Rionegro, Santander. Los resultados destacan a Gradient Boosting como la mejor alternativa de pronóstico, evidenciando un R2 del 68%,

[1] Modelo de simulación AquaCrop de las Naciones Unidas que simula la respuesta del rendimiento de los cultivos herbáceos al agua y es particularmente adecuado para las condiciones en las que el agua es un factor limitante en la producción de cultivos.

MAE de 13.32 y RMSE de 20.41. La variabilidad del rendimiento se atribuye principalmente a la radiación y temperatura un mes antes de la cosecha, así como a las lluvias acumuladas durante la cosecha. Además, se revela la importancia de la exposición solar en los rendimientos, indicando que la radiación previa a la cosecha es crucial para cultivos a la sombra. ubicaciones, y futuras investigaciones podrían explorar múltiples sitios y considerar variables adicionales como la edad de las plantas y prácticas agrícolas (Lamos et al., 2020).

5.3.3 Métodos kernel

5.3.3.1 Support vector machine

Técnica de aprendizaje supervisado basadas en vectores separados por hiperplanos o límites de decisión. Las máquinas de soporte vectorial son muy flexibles y permiten encontrar la forma óptima de clasificar entre varias clases maximizando el margen de separación entre las clases. Usan varias funciones del kernel y pueden estimar límites de decisión no lineales complejos.

En Pakistán, se realizó un estudio sobre el rendimiento del cultivo del té, utilizando una amplia gama de datos que abarcaban aspectos meteorológicos (temperatura mínima, temperatura máxima, humedad, precipitación y velocidad del viento), condiciones del suelo (tipo de suelo, aplicación de fertilizantes, programación de riego, método de siembra, densidad de siembra y calendario de cultivo), detalles del cultivo (días para la emergencia, cobertura inicial del dosel, días para la madurez, días para la cobertura máxima del dosel y días para la cosecha), así como datos de gestión agrícola y datos de recolección. Se aplicaron diversas técnicas, como la regresión lineal múltiple, el Método de Soporte Vectorial (SVM) y los Bosques Aleatorios, y los resultados destacaron que el algoritmo de regresión lineal múltiple superó al modelo de simulación al lograr una mejor predicción del rendimiento del cultivo, incluso con un conjunto de datos más reducido (all B. D., 2022).
En la región de Menemen Plain en Turquía, se buscaba predecir el rendimiento del cultivo de algodón a partir de datos climáticos, que incluyeron variables como la temperatura, la precipitación y la humedad relativa, y de teledetección, que incluyeron índices de vegetación y otros parámetros relacionados con la vegetación. Esto se logró utilizando una Red Neuronal Artificial (ANN) que pudo predecir con precisión el rendimiento del algodón al explorar diferentes combinaciones de variables explicativas y tamaños de conjunto de datos (T. Yildirim, 2022).

En el Instituto de Investigación en Ingeniería Agrícola, Centro de Investigación Agrícola, Giza, Egipto, se llevó a cabo un estudio en un invernadero controlado para predecir el rendimiento de la lechuga hidropónica. Se emplearon cuatro técnicas diferentes de aprendizaje automático para desarrollar el modelo: Máquina de Vectores de Soporte (SVM), Bosque Aleatorio (RF), Redes Neuronales Artificiales (ANN) y Regresión Lineal Múltiple (MLR). Los datos utilizados en el modelo incluyeron factores ambientales como la temperatura, la humedad y la iluminación. Los resultados demostraron que todos los modelos aplicados arrojaron coeficientes de correlación elevados, superiores a 0,95, y que la desviación estándar (SD) se mantuvo cercana a los valores observados (Ali Mokhtar, 2022).

En Pahang, Malasia, se emplearon datos históricos de 35 años de rendimiento de frutos frescos de palma de aceite, así como datos meteorológicos y de humedad del suelo con el propósito de estimar el rendimiento del aceite de palma. El estudio se basó en dos de los modelos de regresión más destacados, específicamente el modelo Extra Tree y el modelo AdaBoost. Los resultados indicaron que la frecuencia de lluvia, la humedad del suelo en la zona de las raíces y la temperatura podrían ejercer un impacto significativo en el rendimiento del aceite de palma (all K. N., 2022).

5.3.4 Modelos de redes neuronales artificiales

5.3.4.1 Redes neuronales recurrentes

Son una clase de aprendizaje profundo basada en los trabajos de David Rumelhart en 1986. Son conocidas por su capacidad para procesar y obtener información de datos de series temporales (Arana, 2021). Por sus características particulares, existen diferentes tipos de redes recurrentes:

Red simple: Algoritmo que permiten conexiones arbitrarias entre las neuronas y utilizadas de manera predominante para solución de problemas de reconocimiento de la voz o reconocimiento de la escritura a mano (Villanueva, 2020).

Este modelo ha sido utilizado con resultados prometedores en la predicción del rendimiento de cultivos, en particular i). sembrados de trigo realizado en la India, comparado con métodos de regresión pese a subestimación del rendimiento del cultivo en la etapa de validación para todo el distrito (Setiya et al., 2022), ii). Cultivos de algodón en la provincia de Menemen, en Turquía, utilizando variables derivadas de datos climáticos y remotos básicos y fácilmente disponibles

(Yildirim et al., 2022), iii). Cultivos de lechugas hidropónicas de tres sistemas diferentes en Egipto, y con combinación de datos (incluidas de manera iterativa) cómo el número de hojas, consumo de agua, peso seco, longitud del tallo y diámetro del tallo. Este modelo fue comparado con otros Máquinas de soporte vectorial, Random Forest y XGBoost, la elección del modelo se toma en función de la menor cantidad de variables de entrada requeridas (Mokhtar et al., 2022), iv). Cultivos de cebolla y arroz en la india utilizando datos de área de campo, producción, estado, época y cultivo (Rananavare & Chitnis, 2022), v). cultivos de arroz, ragi (cereal de tierra árida), gramo (similar al frijol negro), papa y cebolla, utilizando datos detectados por IoT sobre el medio ambiente, condiciones agrícolas, características de las plantas y necesidades (Apat et al., 2022).

Red LSTM: Modelo capaz de "recordar" un dato relevante en la secuencia y de preservarlo por varios instantes de tiempo, por lo que el modelo puede tener una memoria tanto de corto plazo como también de largo plazo.

En un estudio de pronóstico para la caña de azúcar en Brasil, en el estado de São Paulo, Brasil, utilizando métricas derivadas de la serie de tiempo del índice de vegetación de diferencia normalizada (NDVI) del sensor de espectroradiómetro de imágenes de resolución moderada (MODIS). En este caso, para la selección de las características hicieron uso del modelo de Bosques Aleatorios regularizados (Pham et al., 2022) y se obtuvieron buenos resultados tanto en el ajuste como en la predicción de los datos (Lobato et al., 2017).

5.3.4.2 Redes neuronales convolucionales

Es una arquitectura de red para aprendizaje profundo que aprende directamente de los datos, sin necesidad de extraer características manualmente. Útiles particularmente para encontrar patrones en imágenes a partir del reconocimiento de objetos, caras y escenas, o para analizar otros tipos de datos desestructurados como audio, series temporales y señales.

A nivel funcional utiliza principalmente tres tipos de capas: i). convolucionales que aplican distintos filtros convolucionales a las imágenes agrega término "bais" y utiliza función de activación no lineal; ii). pooling que se encarga de reducir la resolución de la imagen, utilizando generalmente promedios y iii). Densas con las mismas funcionalidades de una red neuronal recurrente.

En lo que respecta a la predicción de rendimiento de cultivos, en Alemania se realizó un análisis de variables de clima, suelo y fenología de cultivos en 271 condados de Alemania desde 1999 hasta

2019 para el trigo de invierno y evaluando el rendimiento del aprendizaje automático y los métodos de aprendizaje profundo. En este caso, la red neuronal convolucional propuesta superó los otros siete modelos comparados[2] (Srivastava et al., 2022).

Adicionalmente, de acuerdo con el estudio realizado para la predicción del rendimiento de cereales cómo cebada, avena y trigo sembrados en primavera, y trigo y centeno sembrados en otoño, con uso de imágenes satelitales correspondiente a 28 mosaicos de Sentinel-2, se evidencia la robustez del modelo ante la presencia de píxeles nublados, lo que sugiere puede aprender el enmascaramiento de las nubes a partir de los datos (Yli-Heikkila et al., 2022).

En Colombia, gracias a la colaboración con miles de pequeños productores, se llevó a cabo un estudio para la estimación de rendimiento de cultivos de café a nivel de árbol. El estudio propone el uso de inteligencia artificial, específicamente el modelo You Only Look Once (YOLO) , para identificar cerezas en imágenes de ramas productivas de café tomadas con teléfonos móviles. La recolección de datos fue realizada por los productores a través de sus propios dispositivos, permitiendo la creación de aplicaciones móviles para predicciones en tiempo real. El análisis estadístico del modelo incluyó métricas como el coeficiente de determinación y el error porcentual absoluto medio. Las conclusiones destacan la aplicabilidad del método para estimar el rendimiento en tiempo real, con posibilidad de escalar a niveles de parcela y explorar nuevas especies. No obstante, el estudio reconoce limitaciones como la necesidad de datos para el entrenamiento del algoritmo, la variabilidad entre regiones y la posible limitación en la implementación a gran escala (Rivera et al., 2023).

5.3.5 Modelos híbridos

Una de las prácticas actuales es la conjugación o modelos híbridos, que conjugan dos o más métodos mencionados anteriormente, por ejemplo, un estudio realizado por Hoa Thi Pham, Joseph Awange y Michael Kuhn, en el que se utilizó un conjunto de datos históricos de rendimiento de 124 cultivos de todos los estados y distritos de India para el periodo de tiempo comprendido entre el año 1995 y 2015. En este caso, se propuso tres modelos híbridos, uno basado en LSTM, dos combinaciones de reducción de dimensionalidad con modelos de boosting PCA-AdaBoost y PCAXGBoost. El modelo de redes neuronales profundas, denominado Stacked Auto Encoder – Crop Yield Predicting Deep

[2] KNN, Random Forest, XGBoost, Regression Tree, Lasso and Ridge Regressions y SVM

Neural Network (LSAE-CYPDNN) obtuvo los mejores resultados, obteniendo un error mínimo óptimo (Pham et al., 2022).

Cómo se puede observar, en los últimos años han surgido diversos estudios y propuestas para la estimación de rendimientos de cultivos con aplicación de herramientas de aprendizaje automático. Sin embargo, la mayoría de estas responden a condiciones muy diferentes a las presentes en la ruralidad de Medellín, entre las que se pueden listar:

1. **Cantidad y disponibilidad de datos para el análisis**: Los estudios citados emplean conjuntos de datos históricos de gran volumen, disponibles en ciertas áreas nacionales o generados a través de la teledetección. Sin embargo, en el caso de Medellín, la disponibilidad de datos es restringida, limitándose principalmente a la información recopilada durante el proceso de caracterización rural llevado a cabo en 2023.
2. **Cobertura geográfica**: En términos generales, los estudios mencionados se realizan en regiones extensas, como países enteros o provincias de gran tamaño. En contraste, en este contexto, se enfoca únicamente en los registros de áreas cultivadas por la agricultura familiar, campesina y comunitaria de los territorios corregimentales de Medellín de acuerdo con los hallazgos de la Caracterización Rural, enmarcada en el Convenio Conexión Medellín Rural – COMER-.
3. **Tamaño de las áreas de análisis**: Los análisis típicamente se centran en áreas extensas destinadas al cultivo y la cosecha de los productos en estudio. No obstante, en este caso, las áreas de interés corresponden a parcelas de pequeña escala, incluso abarcando la producción traspatio o de autoconsumo.
4. **Diversidad en condiciones topográficas y climáticas**: La mayoría de los estudios consultados se llevan a cabo en regiones alejadas del trópico y experimentan estaciones climáticas definidas, lo que resulta en condiciones meteorológicas y climáticas menos heterogéneas para los cultivos en análisis. En contraposición, en los corregimientos de Medellín, las condiciones climáticas pueden variar notoriamente de un área a otra debido a la diversidad topográfica y las variaciones de altitud. A pesar de que, en líneas generales, Medellín presenta un clima tropical de montaña, es evidente que existen marcadas discrepancias entre los corregimientos situados en las regiones más bajas de los valles y aquellos ubicados en altitudes más elevadas. Esta variabilidad climática se atribuye a la

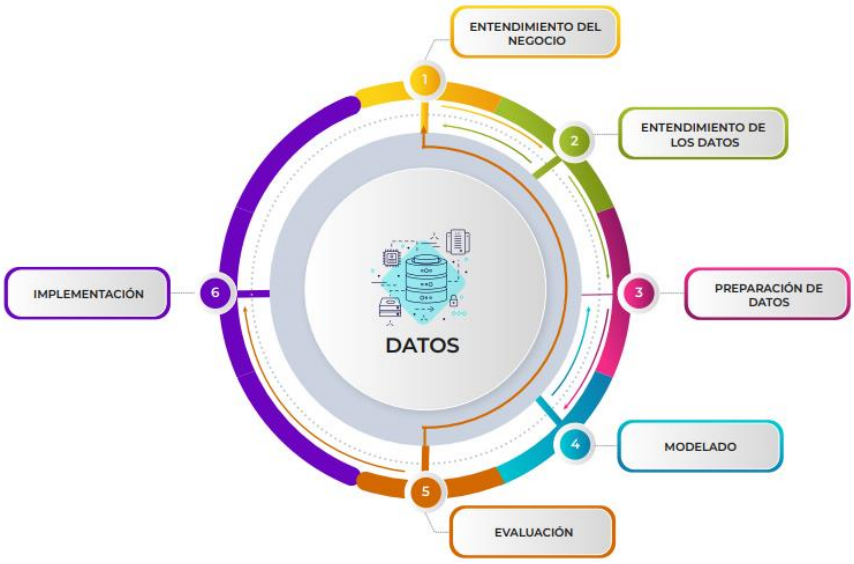
topografía diversa de la región y a las diferencias en altitud, lo que da lugar a microclimas específicos en cada corregimiento y, en última instancia, resulta en condiciones climáticas heterogéneas en todo el territorio.

A pesar de los desafíos inherentes al contexto rural de Medellín, es crucial resaltar las posibilidades y la importancia de este proyecto. La falta de herramientas similares aplicadas en el territorio subraya su relevancia. Al adaptar tecnologías de aprendizaje automático a las condiciones locales, obtenemos resultados más precisos y aplicables a la realidad del territorio. Esto no solo tiene el potencial de mejorar la producción de alimentos y promover la sostenibilidad ambiental al optimizar el uso de recursos, sino que también ofrece una valiosa herramienta para profesionales agrónomos e instituciones. Permite comprender elementos explicativos de la productividad agrícola local y priorizar oportunidades de planes y proyectos que impacten positivamente en el sector a través del análisis de datos. En última instancia, este proyecto tiene el potencial de contribuir al desarrollo sostenible en la zona rural de Medellín.

6 Metodología

Para el desarrollo del trabajo se consideró la metodología CRISP-DM, es quizás la guía de referencia más utilizada en proyectos de analítica en el mundo y particularmente en minería de datos. Esta metodología considera las diferentes fases (ver Figura 2) de un proyecto de este tipo e incluye la descripción de cada una de las tareas requeridas para el alcance de los objetivos de la fase, todo este proceso es considerado como el ciclo de vida del proyecto (IBM, 2016).

Figura 2. Fases empleadas en la metodología CRISP-DM.



Fuente: Adaptado de (IBM, 2016).

El ciclo de vida del proyecto CRISP-DM consta de las siguientes fases fundamentales:

1. **Entendimiento del Negocio:** En esta etapa, se identifican los objetivos del proyecto, los problemas a resolver, las preguntas clave a abordar y los criterios de éxito. También se evalúan los recursos disponibles y las limitaciones, junto con las consideraciones específicas relacionadas con el negocio.
2. **Entendimiento de los Datos:** Esta fase se dedica a comprender y explorar los datos disponibles. Incluye la recopilación de datos, la evaluación de la calidad de estos, la detección de problemas y limitaciones, así como el análisis estadístico y la visualización de los datos para identificar patrones y tendencias iniciales.
3. **Preparación de Datos:** Aquí se seleccionan los datos relevantes para el proyecto y se realiza la limpieza de datos. También se abordan los datos atípicos o faltantes, se aplican transformaciones necesarias, y se prepara un conjunto de datos de entrenamiento y prueba para las fases posteriores.
4. **Modelado:** En esta etapa, se seleccionan los algoritmos adecuados según el marco conceptual y el estado del arte. Se entrenan los modelos y se ajustan los hiperparámetros. Además, se realiza la validación cruzada para mejorar el rendimiento de los modelos.

5. **Evaluación:** Aquí se evalúa el rendimiento del modelo utilizando métricas definidas previamente y se verifica si cumple con los criterios de éxito establecidos. Se realizan pruebas en el conjunto de datos de prueba y se analizan los resultados obtenidos.
6. **Implementación:** En la fase final, se aprovechan los resultados y conocimientos adquiridos en las etapas anteriores y se aplican en un entorno operativo. Es importante destacar que, en este caso, no se propone el despliegue de la solución en producción.

Es importante destacar que, si bien CRISP-DM se distingue por su utilidad en proyectos de este tipo, es crucial reconocer algunas limitaciones. En primer lugar, puede resultar demasiado genérica y no adaptarse completamente a todas las situaciones y necesidades específicas de cada industria o empresa, ya que su flexibilidad se ve limitada debido a la naturaleza secuencial de sus fases. En segundo lugar, el énfasis en modelos estáticos podría no ser la opción óptima para datos evolutivos. Y finalmente, la falta de atención detallada a tipos específicos de datos, como texto o imágenes, plantea desafíos adicionales. Estas consideraciones invitan a revisar la posibilidad de adaptación en el marco específico de cada proyecto, considerando cuidadosamente factores como tipos de datos, objetivos y posibles sesgos (Shearer, 2000).

A continuación, se describen en detalle los procesos específicos llevados a cabo en cada una de estas fases en el contexto del desarrollo de este proyecto[3].

6.1 Entendimiento del negocio

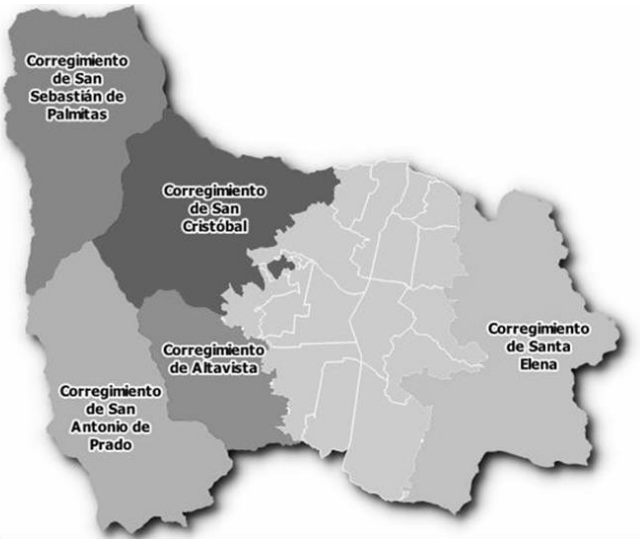
Para diseñar un modelo de predicción del rendimiento de los cultivos predominantes en las áreas rurales cercanas a la ciudad de Medellín mediante técnicas de Machine Learning, es esencial comprender la singularidad geográfica y las dinámicas de este territorio. Medellín está situada en el Valle de Aburrá, en la cordillera Central de los Andes, a una altitud de 1.525 metros sobre el nivel del mar. En la actualidad cuenta con aproximadamente 2.5 millones de habitantes, la mayoría reside en las zonas urbanas, que representan solo el 30% del territorio de la ciudad (Alcaldía de Medellín, 2020).

[3] Para obtener información detallada sobre los recursos utilizados y los procesos llevados a cabo en este proyecto, le invitamos a consultar el siguiente enlace: https://drive.google.com/drive/folders/1qD3MA78s3-7HGo8KcJTjxOWY-S1Bye7X?usp=drive_link.

Las zonas rurales del municipio de Medellín están conformadas por cinco divisiones corregimentales: San Sebastián de Palmitas, San Cristóbal, Altavista, San Antonio de Prado y Santa Elena, que corresponden a aproximadamente el 70% del territorio municipal (Gaviria, 2012). Los primeros cuatro se encuentran en la zona occidental, mientras que el último está en la zona oriental de la ciudad como se muestra a continuación en la Figura 3.

Figura 3. Corregimientos, ubicación, área y división administrativa

Corregimiento	Ubicación	Área	División Administrativa
San Sebastián de Palmitas	Noroccidente	57,54 km² (5.754 ha)	Una cabecera urbana y ocho veredas
San Cristóbal	Noroccidente	49.54 km² (4.954 ha)	Una cabecera urbana y diecisiete veredas
San Antonio de Prado	Extremo suroccidental	60,4 km² (6.040 ha)	Una cabecera urbana y ocho veredas
Altavista	Suroccidente	27.41 km² (2.741 ha)	Una cabecera urbana y ocho veredas
Santa Elena	Oriente	70,46 km² (7.046 ha)	Una cabecera urbana y catorce veredas



Fuente: elaboración propia con base en (Gaviria, 2012)

Los territorios rurales de la ciudad de Medellín enfrentan un desafío crítico en cuanto a la baja producción de alimentos en relación con la creciente demanda urbana. A pesar de que aproximadamente el 70% del territorio de la ciudad es rural, estos espacios apenas contribuyen con el 0.03% de los alimentos necesarios para abastecer a la población urbana. Además, en los últimos años, estos territorios han experimentado una rápida expansión urbana, lo que ha resultado en una disminución del 41% en la producción agropecuaria entre 1971 y 2015, según el Plan Decenal de Seguridad Alimentaria de Medellín 2016-2018.

Uno de los principales retos en la ruralidad de Medellín es el incremento de la producción de alimentos, enfrentando obstáculos socioeconómicos como la pobreza y la falta de recursos, como la carencia de políticas adecuadas para el desarrollo rural, y factores ambientales como el cambio

climático. Esta baja producción impulsa la expansión urbana debido a la migración rural en busca de oportunidades, lo que reduce aún más la tierra disponible para la agricultura. Para afrontar este problema, es vital desarrollar estrategias que fomenten la producción agrícola sostenible y la planificación urbana que preserve la vitalidad rural de la región (Alcaldía de Medellín, 2020)

En este sentido, el presente proyecto se enmarca en el propósito de apoyar la caracterización de la población campesina en las zonas rurales de los corregimientos de Medellín, llevada a cabo a través del convenio COMER- Conexión Medellín Rural entre la Administración Distrital y FAO. Esta caracterización, que incluyó tanto aspectos cualitativos como cuantitativos, su objetivo principal era identificar las condiciones de vida actuales, evaluar la oferta productiva existente, analizar las formas y prácticas de producción, y detectar las limitaciones en el desarrollo de los cultivos, entre otros aspectos relevantes. Los datos recopilados en esta caracterización se consideraron como un insumo fundamental para el presente proyecto, el cual se enfoca en abordar los problemas mencionados anteriormente mediante la aplicación de la ciencia de datos. Específicamente, se analiza la viabilidad de estimar el rendimiento de los cultivos agrícolas, que son predominantemente cultivados en las parcelas de la agricultura campesina familiar y comunitaria, en la ciudad de Medellín. Para lograr esto, se utilizarán modelos de aprendizaje automático previamente probados e implementados, según la literatura, como herramienta para la planificación de las actividades agrícolas. La capacidad de conocer qué cultivos plantar y cuándo hacerlo, en función de las necesidades del mercado, con un enfoque estratégico para incrementar la producción agroalimentaria en los territorios rurales de Medellín.

6.2 Entendimiento de los datos

Aprovechando la participación en el proyecto Conexión Medellín Rural, se utilizarán los siguientes datos como fuentes de información sobre la producción agropecuaria en los territorios rurales de Medellín:

Caracterización rural para la Medellín Futuro: La Caracterización Rural para la Medellín del Futuro fue un proceso de recopilación de información que tuvo lugar en los cinco corregimientos del municipio de Medellín desde el 16 de enero hasta el 10 de mayo de 2023. Este proceso se llevó a cabo en el marco del Convenio firmado entre la administración Distrital de Medellín y la FAO, conocido como Conexión Medellín Rural. A través de un cuestionario semi estructurado se recopiló

información, tanto alfanumérica como espacial, sobre los cultivos presentes en 2.803 parcelas donde se practica la agricultura campesina, familiar y comunitaria en los corregimientos del Distrito.

En este ejercicio, se abarcó la totalidad de las veredas que componen los territorios corregimentales del municipio de Medellín. Durante este barrido, se identificaron un total de 7.239 polígonos dibujados sobre la última ortofoto disponible, correspondiente al año 2021. Cada uno de estos polígonos representa los cultivos presentes en las parcelas caracterizadas en el momento del levantamiento de los datos (ver Figura 4)

Figura 4. Ejemplo información recopilada en el proceso de caracterización rural

Fuente: Convenio Conexión Medellín Rural

Como se mencionó anteriormente, este proceso de recopilación de datos abordó una variedad de elementos más allá de la producción agrícola. Se investigaron aspectos relacionados con la estructura de los hogares, se caracterizó a todos los habitantes presentes, se obtuvo información sobre las características de las viviendas, el acceso a servicios básicos y la seguridad alimentaria de los hogares productores, además de analizar las relaciones territoriales que surgen en la vida cotidiana de los campesinos y campesinas.

En este proyecto se optó por seleccionar las variables que, de acuerdo con la literatura consultada, podrían afectar o influir en el rendimiento de los cultivos. Estas variables se agrupan en dos categorías:

1. **Gestión y Prácticas Agronómicas**: En esta categoría se incluyen datos relacionados con el uso de agroquímicos, modalidades de siembra, conocimiento y asesoramiento técnico, así como el acceso a la infraestructura necesaria para las actividades de cultivo. Estos elementos desempeñan un papel crucial en el rendimiento de los cultivos, ya que permiten la optimización de recursos, la mejora de la calidad del suelo, el control de plagas y enfermedades, la reducción de riesgos climáticos y la promoción de la sostenibilidad a largo plazo en la agricultura.
2. **Producción Actual**: Esta categoría abarca variables relacionadas con los detalles específicos de la producción agrícola en las parcelas, así como su ubicación geográfica. Incluye información sobre el sistema productivo presente en la parcela (agrícola, pecuario o forestal), los productos cultivados, el área ocupada por cada cultivo, la producción por cosecha, y las coordenadas geográficas (longitud y latitud) del centroide de los polígonos de los cultivos.

En la Tabla 1 y en la

Tabla 2, se encuentra una descripción detallada de cada una de las variables asociadas a las categorías.

Tabla 1. Variables inicialmente seleccionadas asociadas a la gestión y prácticas agronómicas en la parcela, a partir de la caracterización rural

Nombre de la variable	Descripción
GlobalID	Corresponde al identificador único de cada registro y el conector con la base de datos de producción
Fecha_Encuesta	Registro de la fecha de realización de encuesta
Código corregimiento	Nombre del territorio corregimental donde se mapeo el dato del productor

Nombre de la variable	Descripción
Código vereda	Nombre de la vereda dónde reside el productor
Código Catastral	Código asociado al predio registrado a nivel catastral y que sirve de para la construcción de la llave que conecta con los datos generados en la plataforma *Geobristol*
Sexo del dueño de la parcela	Identificación del sexo de cada productor caracterizado (Hombre, mujer e intersexual)
¿De dónde provienen principalmente la mano de obra empleada para la producción?	Caracteriza la procedencia de la mano de obra: Familiar, Contratada, Comunitaria (minga), Intercambio de mano de obra
Invernadero	Identifica si el productor tiene invernadero como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.
Poli_Sombra	Identifica si el productor tiene polisombra como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.
Marquesina	Identifica si el productor tiene marquesina como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.
Inf_Bioabono	Identifica si el productor tiene infraestructura para la producción de abonos orgánicos como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.
Tractor	Identifica si el productor tiene tractor como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.
Lombricultivo	Identifica si el productor tiene lombricultivo como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.
¿Utiliza maquinaria agrícola para el arado del suelo?	Identifica cómo práctica sostenible el uso o no de maquinaria agrícola para arar el suelo.
¿Siembra en contra de la pendiente?	Identifica cómo práctica sostenible el sembrar en contra de la pendiente.
¿Tiene cultivos limpios en zonas de alta pendiente que protegen el suelo?	Identifica cómo práctica sostenible la siembra de cultivos en zonas de alta pendiente.
¿Realiza prácticas de conservación y recuperación de suelos en su parcela?	Identifica si el productor realiza alguna práctica de conservación de los suelos en la producción
En su parcela, ¿aplica prácticas de producción sostenible?	Identifica si el productor realiza alguna práctica considerada de producción sostenible (diversificación de cultivos y aplicación de buenas prácticas ambientales, certificación de Buenas Práctica Agrícolas o predio exportador y vinculación de modelos de agricultura sostenible cómo agricultura orgánica, agroecología o agricultura regenerativa)
¿Para la producción agropecuaria, usted hace uso de?	Indaga acerca del uso de los fertilizantes y productos en el manejo agronómico de los cultivos orgánico, químico o mixto
¿Cómo accede al agua para la producción?	Identifica la procedencia del agua utilizado en el riego de los cultivos: acueducto, Río o quebrada, Pozo, Agua lluvia, o sistema de riego
De las siguientes opciones ¿cuáles formas financiación para la producción utiliza?	Identifica la procedencia de los recursos financieros destinados para el financiamiento de la producción: Capital propio, Banco Agrario, Otros Bancos, Comerciantes/intermediarios, Cooperativas, Almacén de insumos, Subsidios, auxilios, Paga diarios (Gota gota)
En el último año ¿Ha recibido capacitación o asistencia técnica?	Identifica si ha sido capacitado a nivel productivo en el último año
Usted como productor, ¿Utiliza o ha utilizado internet para capacitarse y mejorar sus procesos productivos?	Indaga acerca del uso de herramientas tecnológicas y en particular internet para la mejora de procesos productivos
Epsea	Identifica si el productor está inscrito en el Registro único de extensión agropecuaria del Distrito de Medellín.

Fuente: Convenio Conexión Medellín Rural

Tabla 2. Variables inicialmente seleccionadas asociadas a la producción actual existente en las parcelas, a partir de la caracterización rural

Nombre de la variable	**Descripción**
ParentGlobalID	Corresponde al identificador único de cada registro y el conector con la base de datos de información del productor
Sistema productivo	Identifica el tipo de sistema productivo presente en la parcela: agrícola, pecuario o forestal
Producto producido	Identifica el producto cultivados
Área (ha)	Área sembrada u ocupada por cada uno de los rubros cultivados en la parcela
Volumen (Ton/cosecha)	Producción por cosecha obtenida del área sembrada con el producto cultivado. Variable objetivo
Coordenada X	Longitud de coordenada del centroide de cada uno de los polígonos de los cultivos
Coordenada Y	Latitud de coordenada del centroide de cada uno de los polígonos de los cultivos

Fuente: Convenio Conexión Medellín Rural

El aplicativo de captura de los datos de la Caracterización rural fue diseñado y desarrollado sobre la plataforma *ArcGIS Survey123*, con mapas dispuestos en la plataforma de ArcGIS Online y con acceso de aplicativos móviles desde *FieldMaps*. Esta aplicación permite exportar los datos en archivo plano tipo Excel o csv.

Plataforma de seguimiento y monitoreo de cultivos "Geobristol": *Geobristol* se basa en tecnología de imágenes satelitales y análisis de datos, aprovechando la teledetección para generar información precisa y en tiempo real sobre diversas variables agronómicas. Estas variables incluyen la salud de los cultivos, el estrés hídrico, la fertilidad del suelo y el crecimiento de las plantas, entre otros aspectos relevantes para la agricultura (ver Figura 5)

En el contexto de la caracterización rural, esta plataforma se habilitó mediante una suscripción gratuita por un año para cada uno de los hogares encuestados en el proceso. Esto permitió la vinculación y el monitoreo satelital de los polígonos de los cultivos, generando datos adicionales fundamentales para la predicción de cosechas y la toma de decisiones informadas.

Figura 5. Ejemplo información generada a través de Geobristol

Fuente: Convenio Conexión Medellín Rural

En este sentido, los datos seleccionados para el proyecto, a partir de los generados por la plataforma de seguimiento y monitoreo de cultivos son de dos tipos:

1. **Índice de vegetación (NDVI):** A pesar de que *Geobristol* proporciona información sobre varios índices de vegetación, el Índice de Vegetación de Diferencia Normalizada (NDVI) destaca como uno de los más utilizados en la predicción de cultivos debido a su capacidad para evaluar la salud de los cultivos y detectar problemas tempranos. En la fuente de información utilizada, el NDVI ya está calculado para cada imagen satelital histórica del campo que se analiza y está disponible para su descarga en archivos planos directamente desde el sitio web de la plataforma.

   Para este caso particular se trabaja con la información del índice considerado desde el momento del levantamiento de la información (16 de enero) hasta el 31 de agosto de 2023. Para obtener detalles y descripciones de los campos vinculados desde *Geobristol* al análisis, consulte la siguiente tabla.

Tabla 3. Variables preseleccionadas asociadas al índice de vegetación proporcionados por Geobristol

Nombre de la variable	Descripción
Campo	Identificador de cada uno de los polígonos de cultivos, conformado por el código catastral de la parcela y el cultivo sembrado
Grupo	Corresponde al corregimiento donde está ubicada la parcela y el cultivo
Valor del índice	Índice NDVI -Índice de Vegetación de Diferencia Normalizada-: Evalúa la salud y la densidad de la vegetación en cada cultivo monitoreado a través de la plataforma considerando el valor máximo en el periodo analizado, para cada uno de los polígonos de cada producto en las parcelas objeto de estudio.
Fecha de la imagen	Fecha en la que fue tomada la imagen satelital del cultivo.

Fuente: Convenio Conexión Medellín Rural

2. **Datos climatológicos:** Los datos climáticos son esenciales en la predicción de cultivos debido a su impacto significativo en el crecimiento y rendimiento de las plantas. Contar con información histórica y pronósticos precisos es fundamental para los agricultores, ya que les permite tomar decisiones basadas en datos y optimizar sus operaciones. A través de *Geobristol*, es posible obtener datos meteorológicos con descarga directa en formato plano, provenientes del servicio *World Weather* Online que proporciona una cobertura global a través de una cuadrícula de 9,2 x 9,2 km. Estos datos están centralizados en el municipio de Medellín y el período de tiempo seleccionado para el análisis coincide exactamente con el considerado para el indicador NDVI.

   En la *Tabla 4* se relacionan las variables climáticas seleccionadas inicialmente para el análisis de predicción de cultivos

Tabla 4. Variables preseleccionadas asociadas al clima proporcionados por Geobristol

Nombre de la variable	Descripción
Fecha	Fecha de generación del dato
Máx grado C	Temperatura máxima en grados centígrados
Mín grado C	Temperatura mínima en grados centígrados
Humedad (%)	Porcentaje de humedad del ambiente
Precipitaciones (mm)	Cantidad de agua que cae en forma de lluvia

Fuente: Convenio Conexión Medellín Rural

La totalidad de los datos sujetos a análisis en este proyecto son de acceso público y se manejan de manera anónima, lo que no plantea limitaciones en el proceso. No obstante, es importante destacar algunas consideraciones especiales relacionadas con los datos que podrían influir en los resultados de los modelos propuestos.

1. **Cantidad de datos:** Se dispone de un total de 7.239 registros de cultivos mapeados durante el proceso de caracterización rural, de los cuales 6.303 corresponden a productos agrícolas, que son el enfoque principal de esta investigación. Es importante señalar que esta cantidad de datos podría plantear limitaciones en la implementación de modelos de predicción, especialmente en el caso de las redes neuronales.
2. **Calidad o precisión de los datos:** La variable objetivo del análisis, que consiste en la producción en toneladas por cosecha de cada uno de los cultivos sembrados, se basa en las respuestas proporcionadas por los productores, reflejando sus experiencias previas o cálculos estimados. Esto podría introducir cierto grado de error en las estimaciones.

   Además, no se dispone de información precisa sobre la fecha de siembra de los cultivos objeto de estudio. Solo se conoce que al momento de la encuesta estaban sembrados, pero se desconoce su estado o fase de desarrollo en ese momento.

   En relación con la infraestructura de producción en las parcelas, la encuesta solo aborda la existencia o tenencia de esta infraestructura. No se tiene información sobre si estaba actualmente en uso o a disposición de los cultivos. Por ejemplo, se registra si existe un invernadero en la parcela, pero no se especifica si algún cultivo en particular se encontraba bajo el invernadero en el momento de la encuesta
3. **Heterogeneidad de los cultivos:** En el territorio mapeado, se han identificado un total de 79 cultivos diferentes, cada uno con procesos y temporalidades de producción distintos. Además, dado el pequeño tamaño promedio de las parcelas (0.8 ha en promedio por parcela y 0.1 ha por cultivo) y el enfoque productivo, es común encontrar policultivos o sembrados tipo "chagras[4]", lo que podría afectar la capacidad de los modelos de aprendizaje.
4. **Heterogeneidad del territorio objeto de análisis:** Los corregimientos de Medellín abarcan una amplia variedad de climas, geografías y topografías, lo que se refleja en la diversidad de cultivos presentes en cada uno. Por ejemplo, Santa Elena, situado a 16 kilómetros de la zona urbana de Medellín, tiene una altitud promedio de 2.500 metros sobre el nivel del mar. En contraste, San Sebastián de Palmitas, ubicado al oeste de Medellín a una distancia de 32

[4] Un chagra es un espacio de cultivo que se utiliza para sembrar una variedad de cultivos y plantas en un mismo lugar. En este contexto, una chagra es un jardín o huerto donde se cultivan diferentes tipos de alimentos, hierbas, verduras y plantas medicinales en un espacio relativamente pequeño y generalmente en un ambiente rural. Es común en algunas comunidades indígenas y rurales de América Latina, donde se practica la agricultura de subsistencia y se busca diversificar la producción agrícola para satisfacer las necesidades de la familia o comunidad.

kilómetros, presenta una altitud que oscila entre 1.400 y 2.700 metros. Esto implica que el uso y el potencial del suelo varían significativamente tanto entre los corregimientos como dentro de cada uno de ellos.

6.3 Preparación de los datos

Dado el contexto de limitaciones y elementos particulares previamente enumerados en la fase anterior, la preparación de datos se erige como el fundamento para los análisis posteriores y la creación de modelos predictivos. En esta sección, se presenta una descripción detallada de las actividades llevadas a cabo en el proceso de preparación de datos, desde la adquisición de las fuentes de datos hasta la generación de un conjunto de datos apto para la modelización y el análisis. A continuación, se exponen las etapas clave y se detallan las decisiones tomadas en esta fase crítica del proyecto.

6.3.1 Recopilación de Datos

El acceso a los datos de las dos fuentes utilizadas se llevó a cabo mediante la descarga de archivos planos desde ArcGIS Survey y *Geobristol*, respectivamente. Una vez descargados, estos datos se pusieron a disposición a través de una carpeta en la plataforma web de Google Drive, desde donde se procedió a su lectura y procesamiento usando Python en Google Colab.

6.3.2 Exploración Inicial de Datos

Después de haber recopilado los datos, se procedió a realizar un análisis exploratorio inicial de las cuatro tablas disponibles. En este proceso, se identificaron datos faltantes o con valores igual a cero en la variable respuesta (Volumen_Ton). Además, se detectó información en la tabla de productos relacionada con la producción pecuaria o forestal, la cual no está incluida en el alcance de este proyecto de análisis.

Se realizó un análisis de los productos agrícolas que se encuentran en la tabla, y se identificó una notable dispersión en los datos en lo que respecta a la cantidad de productos, con un total de 79. Además, se observó una discrepancia en las unidades de medida aplicadas a productos como flores, plantas ornamentales, hoja de biao, helecho cuero y pasto King grass.

Al enfocarnos exclusivamente en los rubros agrícolas de interés, se realizó un análisis de la concentración, tanto en términos de registros como de área sembrada, utilizando un enfoque tipo Pareto. Es importante destacar que, al realizar el análisis basado en el área cultivada, se observó que el 80% de la producción estaba concentrada en 10 productos específicos (Café, Plátano, Cebolla de rama, Limón, Aguacate, Cilantro, Banano, Papa, Caña de azúcar y Lechuga). Si consideramos la cantidad de registros, este mismo porcentaje estaba compuesto por 12 rubros agrícolas (Cebolla de rama, Cilantro, Plátano, Café, Lechuga, Frijol, Limón, Maíz, Plantas aromáticas y medicinales, Papa, Col, Tomate, Espinaca, Banano, Yuca y Brócoli).

Se ha observado una considerable dispersión en la variable respuesta, con una presencia significativa de datos atípicos.

6.3.3 Limpieza de Datos

En concordancia con los hallazgos obtenidos durante la fase de análisis exploratorio de los datos, se han llevado a cabo procesos de limpieza que abarcan los siguientes aspectos:

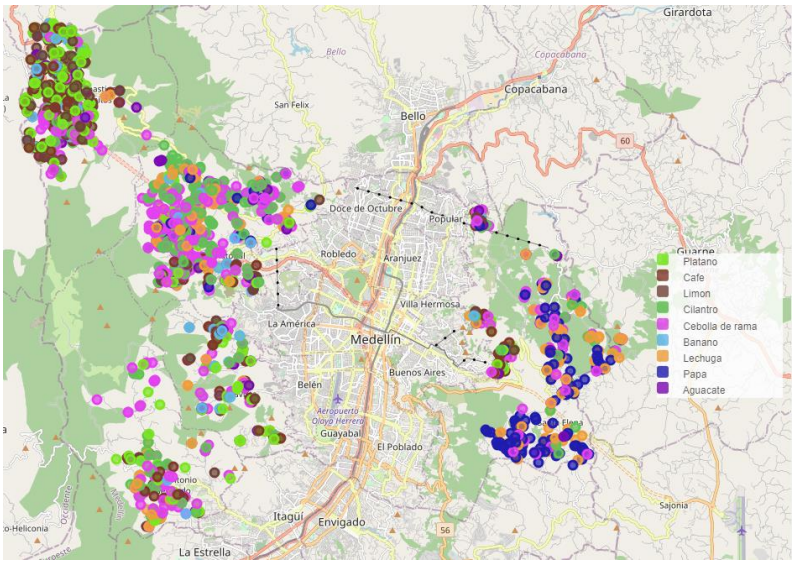
- Revisión de Formatos de Variables: En primer lugar, se efectuó una revisión y ajuste (en casos necesarios) de los formatos de las variables preseleccionadas en cada tabla de datos.
- Filtrado de Datos de Caracterización Rural: Se realizó un proceso de filtrado de datos en la Caracterización Rural con el fin de elegir los cultivos que abarcan el 80% del área utilizada para la producción en la zona rural de Medellín. En la Ilustración 3, se presenta una representación espacial que muestra de manera visual los cultivos específicos seleccionados en este análisis.
- Procesamiento de Datos de *Geobristol*: En cuanto a las tablas provenientes de *Geobristol*, las cuales contienen series de tiempo con registros desde el 10 de enero de 2023 hasta el 31 de agosto de 2023 para el indicador NDVI y variables climatológicas, se generaron medidas resumen para cada campo. Estas medidas serán consideradas en el análisis posterior. Para el NDVI, se eligió el valor máximo como variable predictora en el período de datos disponible. Para las variables de temperatura máxima, temperatura mínima, velocidad del viento, humedad y precipitaciones, se calculó la mediana con el fin de mitigar el impacto de los datos atípicos que se encontraron en los registros.

Las claves de conexión entre las tablas, en el caso de la Caracterización Rural, corresponden al GlobalId o identificador de registro de la parcela. Además, se estableció una relación conjunta entre el código catastral de la parcela y el tipo de cultivo presente en esta, facilitando así la correlación de los datos con las tablas de *Geobristol*.

6.3.4 Selección de Variables

Para el presente proyecto, se decidió trabajar únicamente con los cultivos agrícolas, excluyendo aquellos registros con datos faltantes en el volumen cosechado y los cultivos cuya unidad de medida en producción fuera diferente a toneladas por unidad de área. Adicionalmente, se define trabajar solo con los rubros que representan el 80% de la producción en la ruralidad de Medellín, es decir, los rubros que tenían más productores sembrados en el momento de la encuesta. En total, se han seleccionado nueve rubros agrícolas, que incluyen plátano, café, limón, cilantro, cebolla de rama, banano, lechuga, papa y aguacate, como se muestra a continuación (ver Figura 6.)

Figura 6. Ubicación geográfica de los cultivos seleccionados para el modelo de predicción de rendimientos



Fuente: Convenio Conexión Medellín Rural

También se revisó la relación entre cada una de las variables propuestas como predictoras y la variable respuesta. Esta revisión se realizó de manera diferenciada en los casos donde las variables son categóricas o de tipo numérica, cómo se describe a continuación:

- **Variables categóricas:** Se utilizó la prueba de Kruskal-Wallis para determinar si existen diferencias significativas entre tres o más grupos independientes en una variable numérica continua. La elección de la prueba utilizada obedece a las condiciones particulares de los datos analizados en el presente proyecto, específicamente relacionados con la alta presencia de valores atípicos. Considerando como hipótesis nula (H0) que no hay diferencias significativas entre las medianas de los grupos, mientras que la hipótesis alternativa (H1) sugiere que al menos un grupo tiene una mediana diferente de los demás. La decisión se toma con un nivel de significancia de 0.05.

  Adicionalmente, se realiza un filtrado de las variables con una alta concentración (superior al 80%) en una de sus categorías entendiendo un comportamiento constante que no aporta al modelo.

- **Variables numéricas:** La evaluación de relación de variables predictoras de tipo numérico con la variable objetivo se realizó considerando la correlación de Spearman, aprovechando su capacidad para identificar relaciones monótonas y su robustez ante valores atípicos hacen que sea una opción valiosa para los datos analizados en presente proyecto. Es importante considerar que este coeficiente no detecta relaciones lineales.

A continuación, se presenta el análisis de las variables inicialmente propuestas, la estrategia de análisis y la decisión tomada en cada caso (Ver *Tabla 5)*

Tabla 5. Análisis para la selección de variables a vincular en los modelos.

Variable	Decisión	Argumento
Producto producido	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *1.5648903884309134e-84*
Área sembrada (ha)	Continúa como variable predictora	Coeficiente de correlación de Spearman **Valor p:** *8.76956418310759e-118*
Sexo del dueño de la parcela	Se elimina	Análisis experto: Puede ser una variable confusora o generar un sesgo en el análisis, dado que solo el 29% de los registros de cultivos están bajo la responsabilidad de una mujer, pero además

Variable	Decisión	Argumento
		el tamaño promedio de las áreas destinadas es mucho menor que las bajo responsabilidad de hombres. De otro lado, el acceso de las mujeres a recursos, infraestructura y capacitación para la producción es inferior en comparación con los hombres responsables de los cultivos objeto de análisis.
Corregimiento	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *1.0198988131405226e-76*
Tipología de mano de obra utilizada en la producción	Se elimina	Más del 80% concentrado en una de las categorías
Tenencia de invernadero	Se elimina	Más del 80% concentrado en una de las categorías
Tenencia de polisombra	Se elimina	Más del 80% concentrado en una de las categorías
Marquesina	Se elimina	Más del 80% concentrado en una de las categorías
Tenencia de infraestructura para bioabono	Se elimina	Más del 80% concentrado en una de las categorías
Tenencia de tractor	Se elimina	Más del 80% concentrado en una de las categorías
Tenencia de lombricultivo	Se elimina	Más del 80% concentrado en una de las categorías
Utiliza maquinaria agrícola para el arado del suelo	Se elimina	Más del 80% concentrado en una de las categorías
Tenencia de cultivos limpios en zonas de alta pendiente	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *0.0009017813920818858*
Realiza prácticas de conservación y recuperación de suelos	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *1.0240051671402183e-08*
Prácticas de producción sostenible utilizadas	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *8.127163490790547e-17*
Uso de fertilizantes (incluye la aplicación de productos orgánicos, químicos o ambos para la gestión del cultivo)	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *2.698661744873546e-47*
Acceso al agua para la producción	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *1.201936764364714e-50*
Formas de financiación de la producción	Se elimina	Más del 80% concentrado en una de las categorías
Ha recibido capacitación o asistencia técnica para la producción	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *0.00680230016880169*
Uso del internet para capacitarse y mejorar	Se elimina	Prueba de Kruskal-Wallis **Valor p:** *0.30216156145645534*

Variable	Decisión	Argumento
sus procesos productivos		
Está inscrito en el Registro Único de Extensión Agropecuario (RUEA)	Continúa como variable predictora	Prueba de Kruskal-Wallis **Valor p:** *5.454079537550873e-39*
Max Valor del índice NDVI	Continúa como variable predictora	Coeficiente de correlación de Spearman **Valor p:** *2.5795492165781563e-16*
Median Máx grado C	Se elimina	Coeficiente de correlación de Spearman **Valor p:** *0.10776751752513725*
Median Mín grado C	Se elimina	Coeficiente de correlación de Spearman **Valor p:** *0.5023149082666529*
Median Viento (m/s)	Continúa como variable predictora	Coeficiente de correlación de Spearman **Valor p:** *0.004469395052078328*
Median Humedad (%)	Se elimina	Coeficiente de correlación de Spearman **Valor p:** *0.39861118821712127*
Median Precipitaciones (mm)	Se elimina	Coeficiente de correlación de Spearman **Valor p:** *0.6914336953973907*

Fuente: Elaboración propia

6.3.5 Integración de Datos

A partir de las cuatro tablas de datos disponibles, se ha creado una tabla maestra que condensa las variables tras los procesos de selección y limpieza previamente mencionados. En esta tabla consolidada, se han reunido un total de 3.144 registros y se han seleccionado 12 variables, tal como se detalló en el paso anterior.

6.3.6 Transformación de Datos

Se realizaron diversos procesos de transformación de los datos con el objetivo de prepararlos adecuadamente para la generación de modelos. Estos procedimientos se llevaron a cabo de la siguiente manera:

- **Creación de Variables Dummy**: En primer lugar, se crearon variables dummies a partir de las variables categóricas seleccionadas. Este paso permitió representar las categorías de manera numérica, facilitando su inclusión en los modelos de análisis.
- **Estandarización de Variables Categóricas**: Para las variables categóricas, se aplicó un proceso de estandarización utilizando la raíz cuadrada de la probabilidad de ocurrencia de cada dato en cada columna. Esta técnica se empleó con el propósito de preparar los datos

para un posterior análisis factorial, especialmente diseñado para datos mixtos que incluyen tanto variables categóricas como numéricas.

- **Escalamiento de Variables Numéricas**: Todas las variables numéricas, a excepción de la variable objetivo, fueron escaladas utilizando la función *StandardScaler* de la librería sklearn. Este paso aseguró que las variables numéricas tuvieran una distribución común y comparativa, lo que es esencial para muchos modelos de *Machine Learning*.
- **Análisis Factorial para Datos Mixtos**: Se realizó un análisis factorial específico para datos mixtos, teniendo en cuenta la presencia de variables tanto categóricas como numéricas. El propósito principal de este análisis fue reducir la dimensionalidad de la matriz de datos, simplificando así la representación de la información y permitiendo un enfoque más efectivo en la construcción de modelos. El resultado de este análisis corresponde a los datos utilizados en los modelos PCA.

6.3.7 División de Datos

Previo a la implementación de los modelos se propone la división del conjunto de datos a través de validación cruzada. En este sentido y considerando la cantidad reducida de datos y la alta variabilidad de estos se utiliza la función ***KFold*** de la biblioteca Scikit-Learn que permite hacer este procedimiento de manera estratificada. En este caso, se ha configurado para dividir los datos en 5 conjuntos (k=5) y un *random state* igual a 42.

KFold al realizar múltiples divisiones de los datos permite evaluar el modelo en diferentes conjuntos de prueba proporciona una estimación más robusta del rendimiento del modelo y puede ayudar a detectar la variabilidad en las métricas de rendimiento. Adicionalmente, utiliza todos los datos tanto para entrenamiento como para prueba en diferentes iteraciones, lo que maximiza el uso de los datos disponibles, elementos relevantes para el presente proyecto.

Como resultado de la preparación de datos, se ha consolidado un conjunto de datos con 3,144 registros y 12 variables seleccionadas, correspondiente en nueve cultivos agrícolas clave. Este conjunto de datos será la base para los análisis y modelos subsiguientes en el proyecto.

6.4 Modelado

De acuerdo con el objetivo del proyecto y basándonos en la revisión de la literatura y el estado del arte consultados, se han implementado los siguientes modelos, organizados por tipo de algoritmo.

- **Modelos de regresión:** Incluye la regresión lineal múltiple y los métodos de regularización (Lasso, Ridge, y Elastic Net).

  Además, en consideración de los tests robustos y no paramétricos aplicados en las fases previas del proceso, se incorpora la Regresión por Mínimos Cuadrados Robustos. Este método utiliza la mediana en lugar de la media para calcular el error mediano absoluto (MedAE) como métrica de evaluación, lo que puede resultar más apropiado cuando se enfrentan datos con valores atípicos.
- **Algoritmos de Bagging:** Específicamente el Random Forest.
- **Algoritmos de Boosting:** Incluyendo AdaBoosting y XGBoost (Aumento de gradiente extremo
- **Métodos kernel:** Específicamente Support Vector Machine (SVM)

La optimización de hiperparámetros constituye una fase crítica en el proceso de aprendizaje automático, dado que su objetivo principal radica en identificar los valores ideales de los parámetros que influyen en el desempeño de un modelo. En este contexto específico, para llevar a cabo la optimización, se empleó el método de búsqueda en cuadrícula mediante la función GridSearchCV de la biblioteca sklearn (scikit-learn). Este enfoque sistemático permite explorar múltiples combinaciones de hiperparámetros con el fin de determinar cuáles resultan óptimas para los modelos.

El entrenamiento de los modelos supervisados comprendió dos experimentos. En el primer experimento, se utilizó la salida del análisis factorial para datos mixtos como la matriz de variables predictoras. Se aplicó una reducción de dimensionalidad de modo que se lograra capturar el 80% de la varianza de los datos. En el segundo experimento, se implementaron los modelos utilizando los datos originales, con normalización tanto para las variables categóricas como para las numéricas, como se describió en pasos anteriores. En ambos casos, antes de programar los modelos, se realizó una eliminación del 10% de los datos más atípicos, identificados mediante el uso de la distancia de Mahalanobis.

A continuación, (ver *Tabla 6*) se presentan los modelos ajustados, los hiperparámetros óptimos de acuerdo con los experimentos realizados y los resultados en términos de ajuste a los datos objeto de análisis:

Tabla 6. Modelos entrenados para la predicción del rendimiento de cultivos

Modelo propuesto	**Hiperparámetros óptimos (GridSearchCV)**	**Resultado en ajuste**
PCA Regresión Lineal Múltiple	Fit_intercept: True	Coeficiente de Determinación (R^2): *0.25* Coeficiente de Determinación Ajustado (R^2 ajustado): *0.21*
PCA Regresión LASSO	Alpha: 0.01 Fit_intercept: True	Coeficiente de Determinación (R^2): 0.25 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.21
PCA Regresión Ridge	Alpha: 10 Fit_intercept: True	Coeficiente de Determinación (R^2): 0.25 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.21
PCA Elasticnet	Alpha: 0.01 Fit_intercept: True. L1_ratio: 0.1	Coeficiente de Determinación (R^2): 0.25 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.21
PCA Regresión por Mínimos Cuadrados Robustos		No aplica. Se analizará con respecto a la capacidad de predicción.
PCA Random Forest	Max_depth: 20 Min_samples_leaf: 4 Min_samples_split: 2 n_estimators: 200	Coeficiente de Determinación (R^2): 0.72 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.71
PCA ADABoosting	Learning_rate: 0.1 n_estimators: 50	Coeficiente de Determinación (R^2): 0.18 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.14
PCA XGB Boosting	Learning_rate: 0.1, Max_depth: 5, Min_child_weight: 4, n_estimators: 200	Coeficiente de Determinación (R^2): 0.92 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.92
PCA Support Vector Machine	C: 10.0 Epsilon: 0.5 kernel: rbf	Coeficiente de Determinación (R^2): 0.70 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.68
Regresión Lineal Múltiple	Fit_intercept: False	Coeficiente de Determinación (R^2): 0.27 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.20
Regresión LASSO	Alpha: 0.01 Fit_intercept: True	Coeficiente de Determinación (R^2): 0.26 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.21
Regresión Ridge	Alpha: 10 Fit_intercept: False	Coeficiente de Determinación (R^2): 0.27 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.21
Elasticnet	Alpha: 0.1 Fit_intercept: True. L1_ratio: 0.1	Coeficiente de Determinación (R^2): 0.27 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.21
Regresión por Mínimos	No aplica	No aplica. Se analizará con respecto a la capacidad de predicción.

Modelo propuesto	Hiperparámetros óptimos (GridSearchCV)	Resultado en ajuste
Cuadrados Robustos		
Random Forest	Max_depth: 10 Min_samples_leaf: 2 Min_samples_split: 2 n_estimators: 200	Coeficiente de Determinación (R^2): 0.92 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.91
ADABoosting	Learning_rate: 0.1 n_estimators: 50	Coeficiente de Determinación (R^2): 0.72 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.70
XGB Boosting	Learning_rate: 0.1, Max_depth: 5, Min_child_weight: 2, n_estimators: 50	Coeficiente de Determinación (R^2): 0.91 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.90
PCA Support Vector Machine	C: 10.0 Epsilon: 0.1 kernel: rbf	Coeficiente de Determinación (R^2): 0.69 Coeficiente de Determinación Ajustado (R^2 ajustado): 0.67

Fuente: Elaboración propia

Los resultados muestran una variedad de desempeños. Específicamente, en los modelos de regresión (Regresión Lineal Múltiple, Regresión LASSO, Regresión Ridge y Elasticnet) entrenados tanto con los datos originales como con la matriz reducida, se observa un bajo ajuste, con coeficientes de determinación y determinación ajustada cercanos al 20%. Estos resultados son consistentes con los análisis previos en el sentido de que no se identifican relaciones lineales entre los datos analizados.

En contraste, los algoritmos de Boosting, entrenados tanto con los datos reducidos como con los originales, como PCA Random Forest y PCA XGB Boosting, con hiperparámetros específicos, muestran un ajuste mucho más sólido con valores de R^2 y R^2 ajustados de 0.72 y 0.71, así como 0.92 y 0.92, respectivamente. Esto indica que estos modelos, incluso después de la reducción de dimensionalidad, se ajustan notablemente bien a los datos de rendimiento de cultivos.

En línea con lo anterior, los resultados sugieren que los modelos basados en ensamblaje, como Random Forest y XGB Boosting, logran capturar las particularidades de la producción agrícola en los territorios corregimentales y tienden a ofrecer un mejor ajuste a los datos disponibles para el análisis, incluso después de la reducción de dimensionalidad mediante PCA.

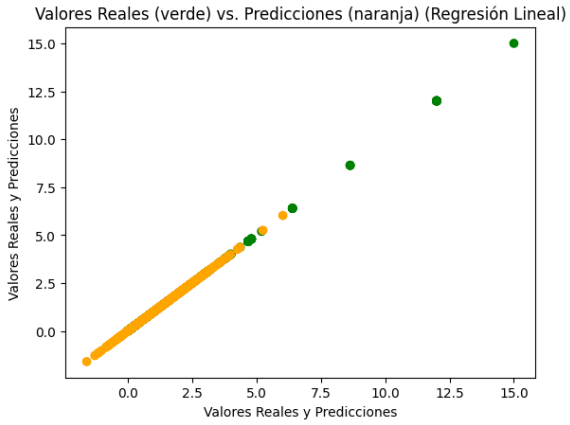
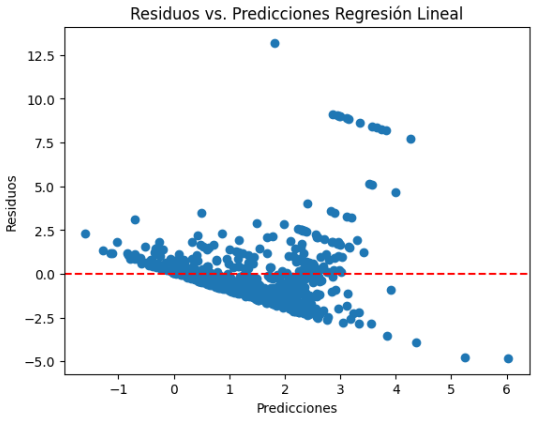
6.5 Evaluación

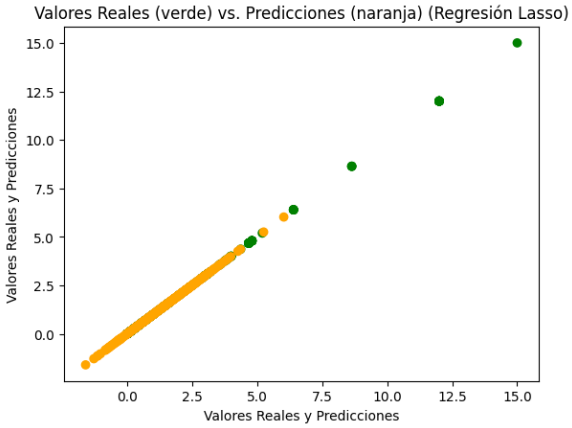
En el análisis de desempeño de los modelos propuestos para la predicción del rendimiento de cultivos, se utilizará como métrica principal el MAE, que representa el Error Absoluto Medio. Esta medida cuantifica la magnitud promedio de los errores entre las predicciones de un modelo y los valores reales. En otras palabras, y bajo el contexto del problema, el MAE indica en promedio cuántas toneladas se equivoca el modelo en sus predicciones. Un MAE más bajo indica un mejor ajuste del modelo a los datos, ya que implica que las predicciones se encuentran más cerca de los valores reales.

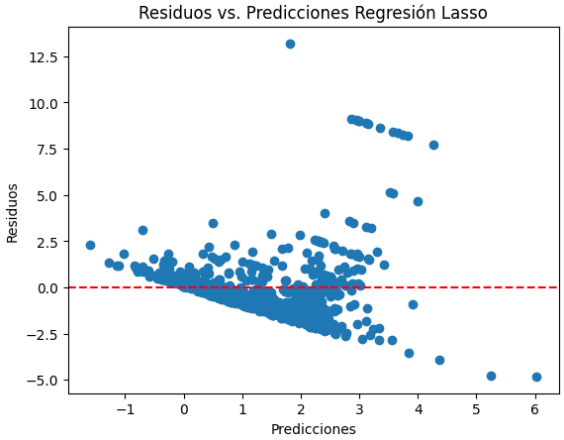
En el caso particular del modelo de Regresión por Mínimos Cuadrados Robustos, su métrica corresponde al Error Mediano Absoluto (MedAE) que es equivalente al MAE pero en vez de utilizar la media, calcula las diferencias con la mediana, lo que lo hace menos sensible a valores extremos.

Adicional a estas métricas, se propone una revisión visual de los gráficos de valores reales vs. valores predichos con el objetivo de analizar cuán precisamente se están ajustando los modelos, así como el gráfico de residuales vs. predicciones del modelo con el objetivo de identificar patrones sistemáticos no capturados por los modelos y la presencia de heterocedasticidad en los errores. En la *Tabla 7*, se presenta cada una de las estrategias de evaluación propuestas para cada uno de los modelos implementados.

Tabla 7. Evaluación de los modelos para la predicción del rendimiento de cultivos

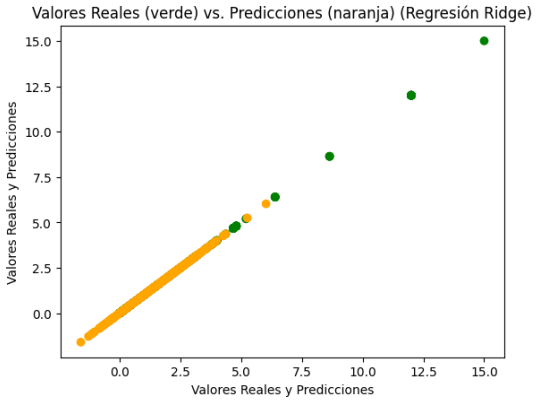
PCA Regresión Lineal Múltiple	
Error Absoluto Medio (MAE)	1.19
Análisis gráfico (reales vs predichos y residuales)	 
PCA Regresión LASSO	
Error Absoluto Medio (MAE)	1.19
Análisis gráfico (reales vs predichos y residuales)	

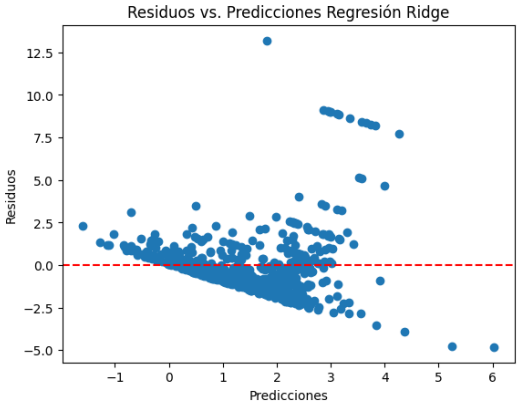




PCA Regresión Ridge	
Error Absoluto Medio (MAE)	1.19

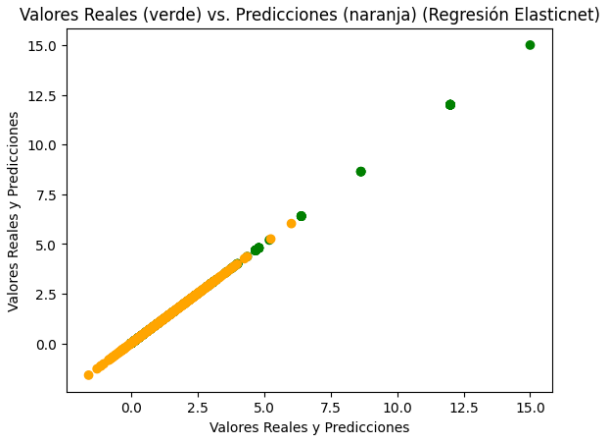
Análisis gráfico (reales vs predichos y residuales)

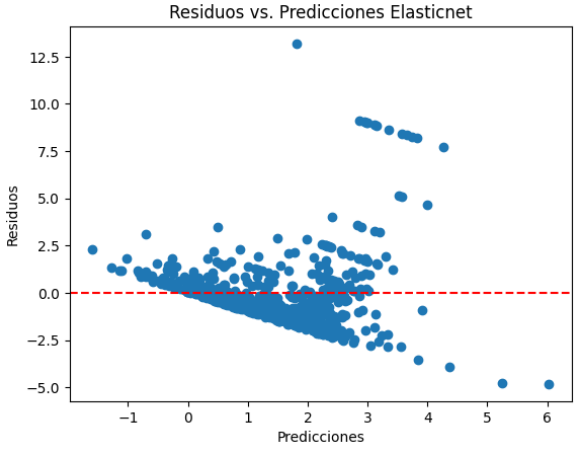




PCA Elasticnet	
Error Absoluto Medio (MAE)	1.19

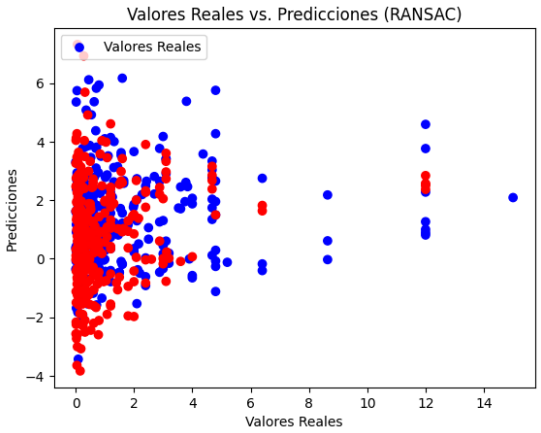
Análisis gráfico (reales vs predichos y residuales)

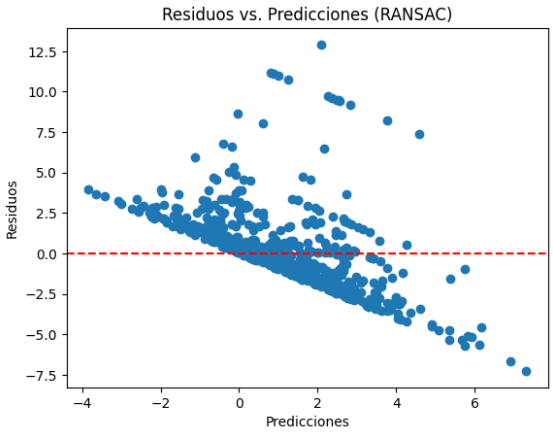




PCA Regresión por Mínimos Cuadrados Robustos	
Error Mediano Absoluto (MedAE)	1.37

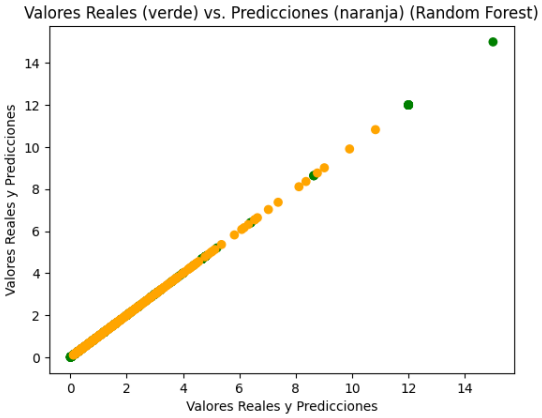
Análisis gráfico (reales vs predichos y residuales)

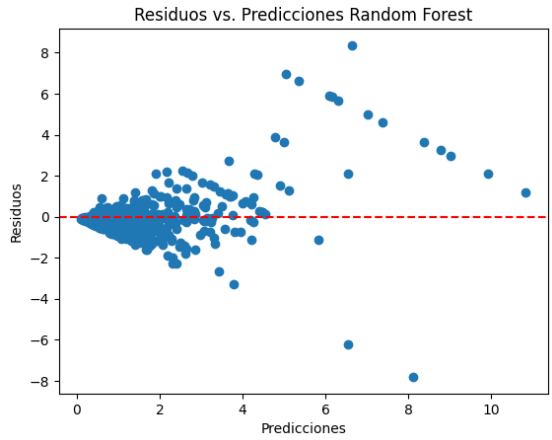




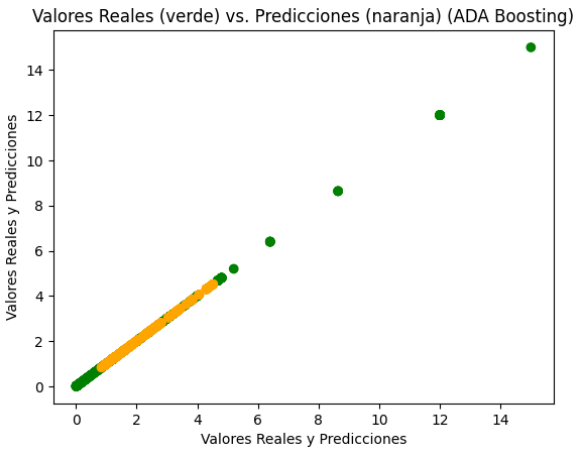
PCA Random Forest	
Error Absoluto Medio (MAE)	0.63

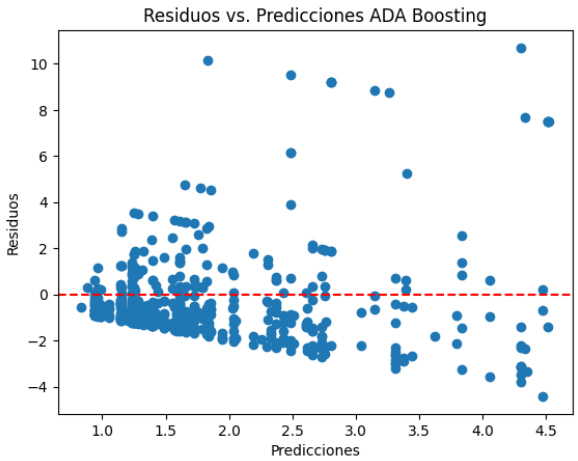
Análisis gráfico (reales vs predichos y residuales)



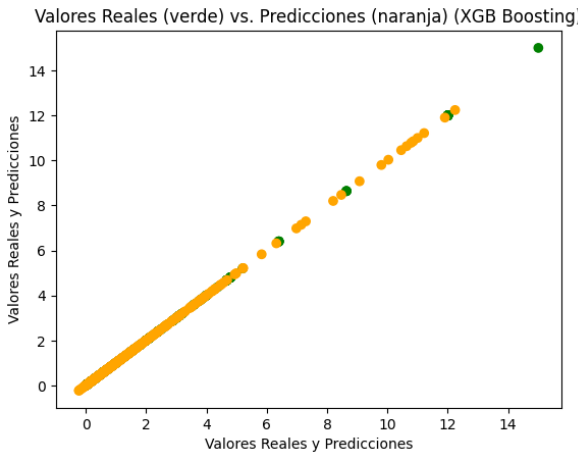


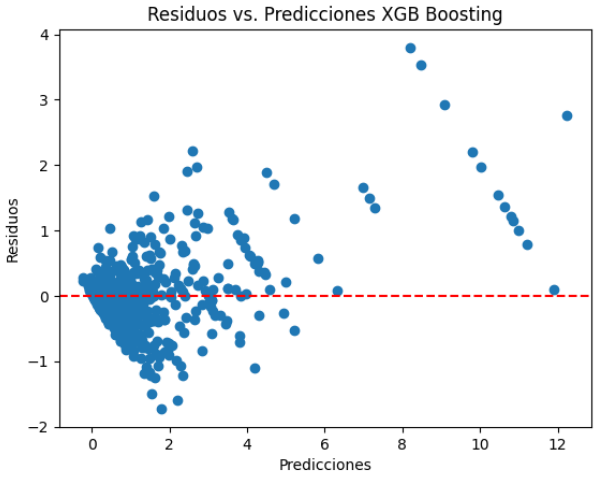
PCA ADABoosting	
Error Absoluto Medio (MAE)	0.63



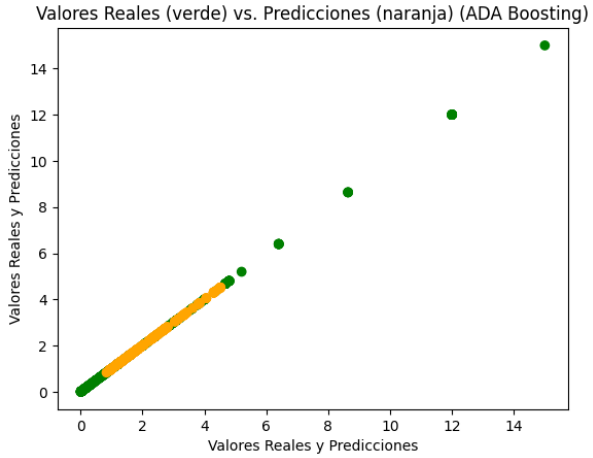


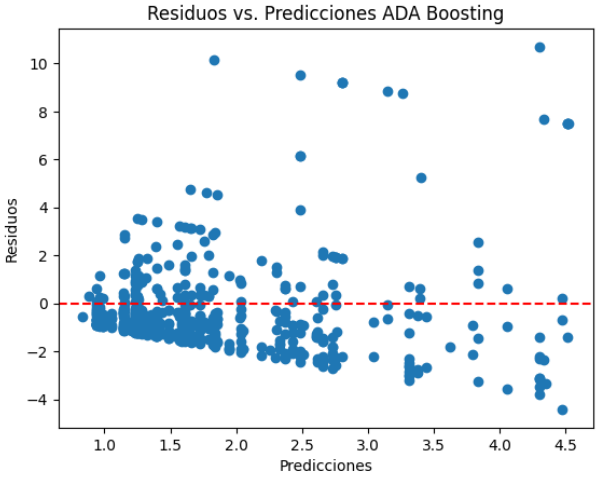
PCA XGB Boosting	
Error Absoluto Medio (MAE)	0.39



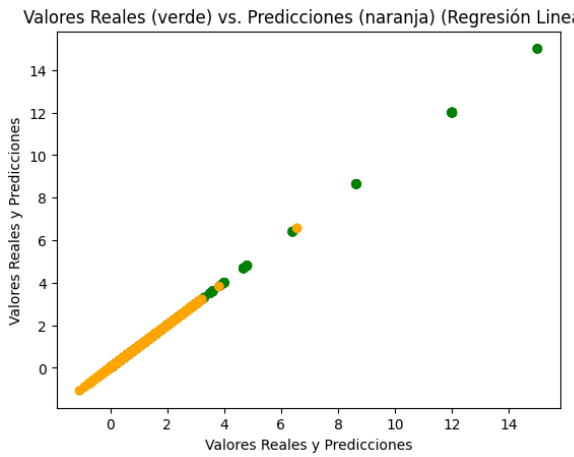


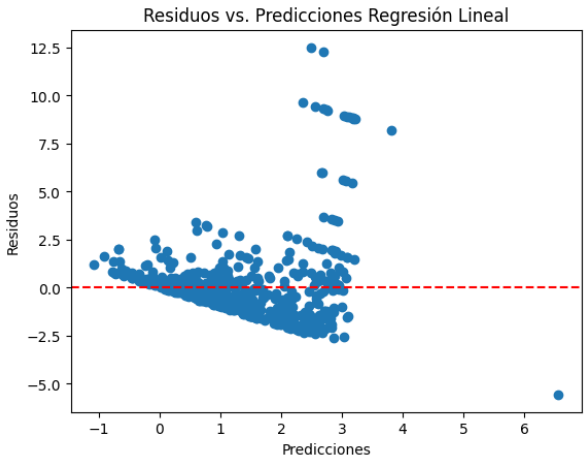
PCA Support Vector Machine	
Error Absoluto Medio (MAE)	0.64



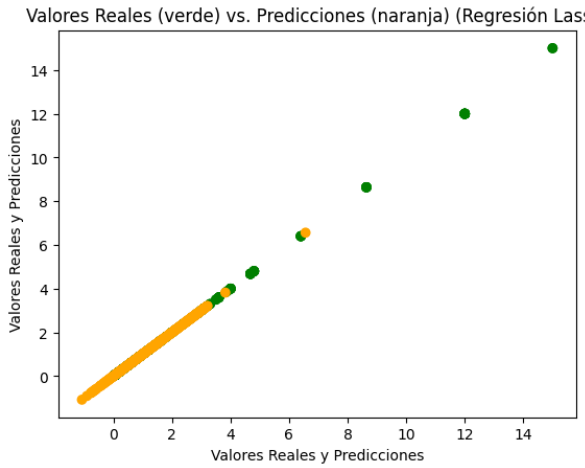


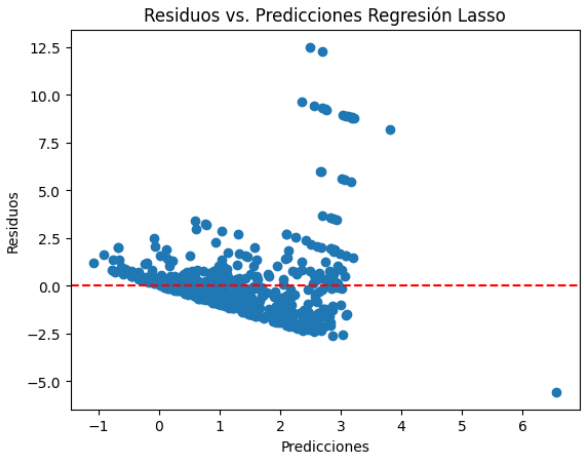
Regresión Lineal Múltiple	
Error Absoluto Medio (MAE)	1.12



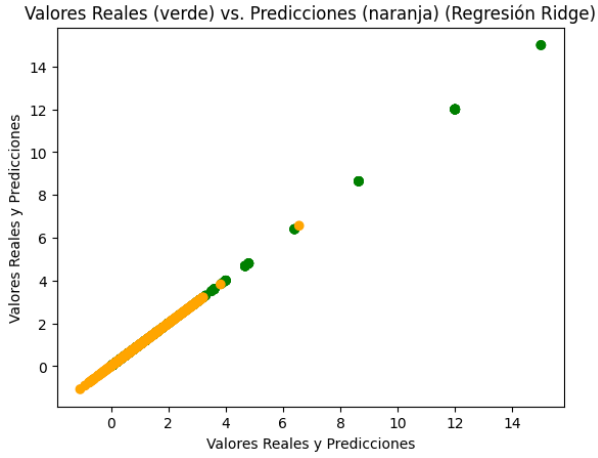


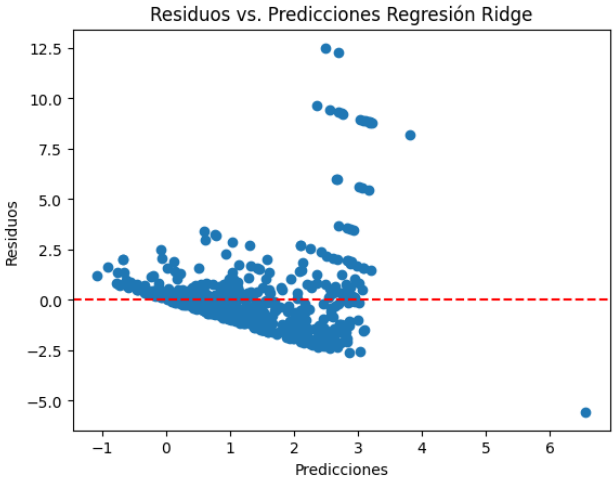
Regresión LASSO	
Error Absoluto Medio (MAE)	1.11



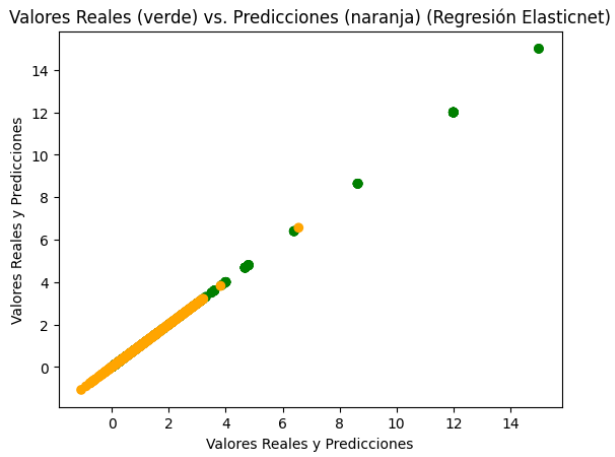


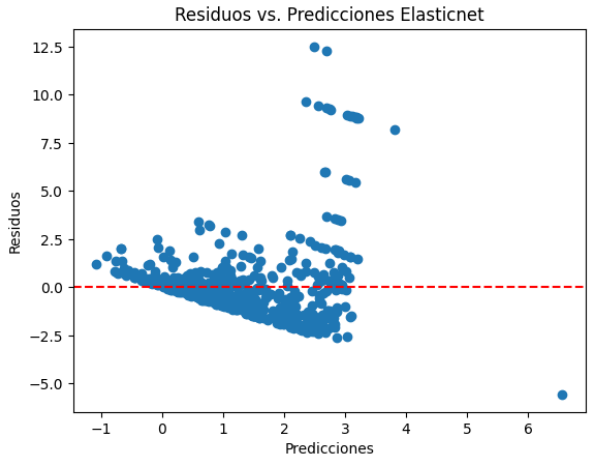
Regresión Ridge	
Error Absoluto Medio (MAE)	1.12



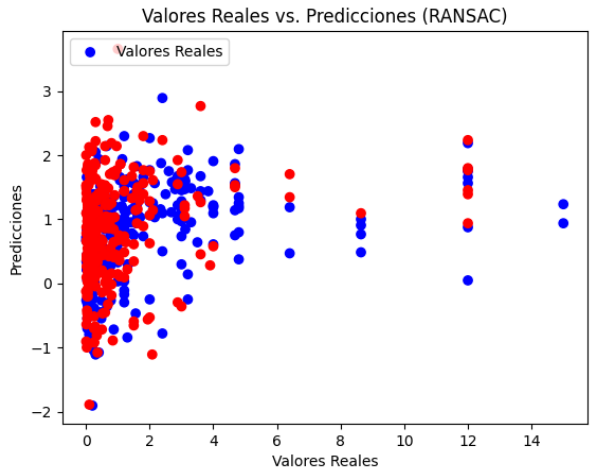


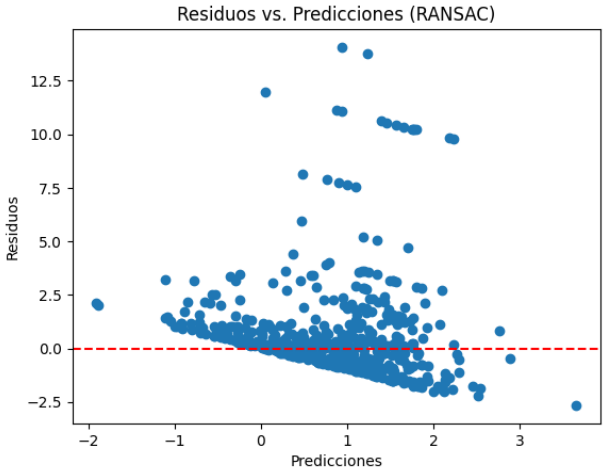
Elasticnet	
Error Absoluto Medio (MAE)	1.11



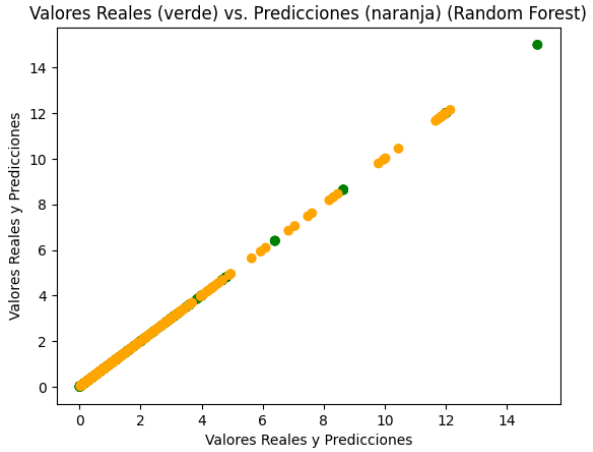


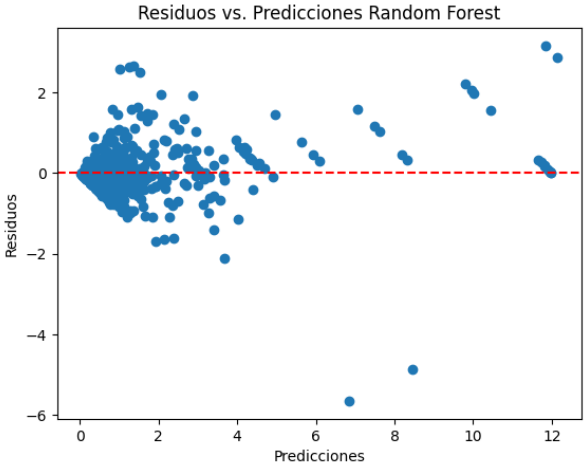
Regresión por Mínimos Cuadrados Robustos	
Error Mediano Absoluto (MedAE)	0.69



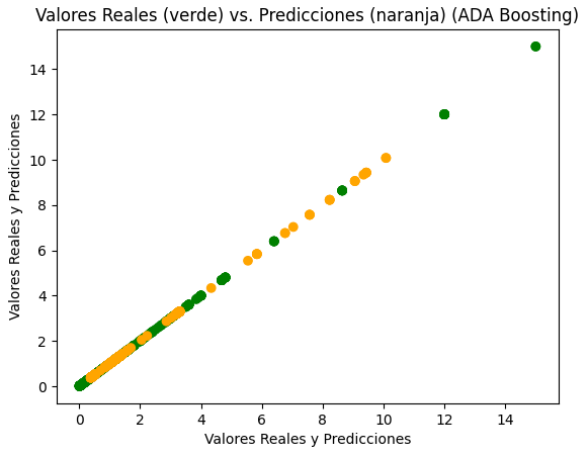


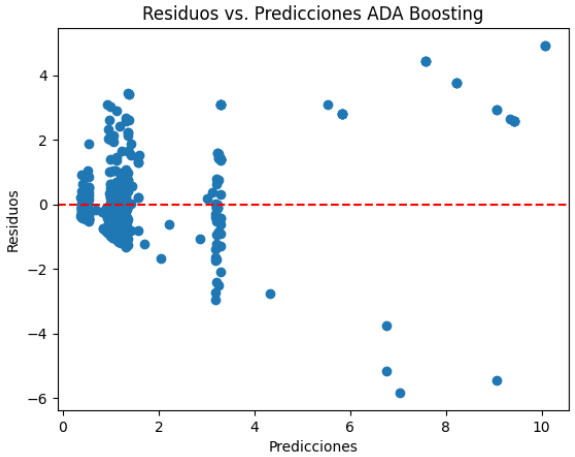
Random Forest	
Error Absoluto Medio (MAE)	0.40



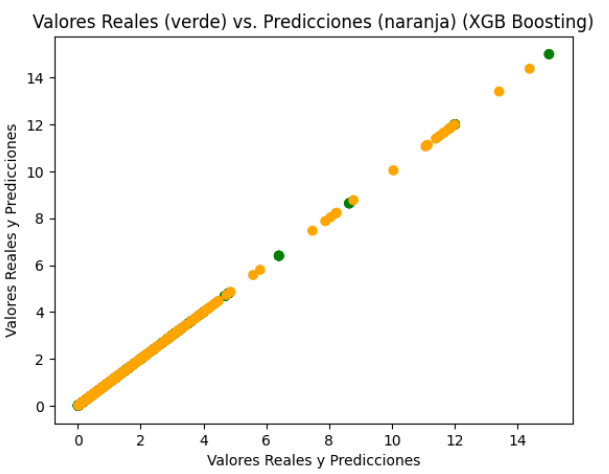


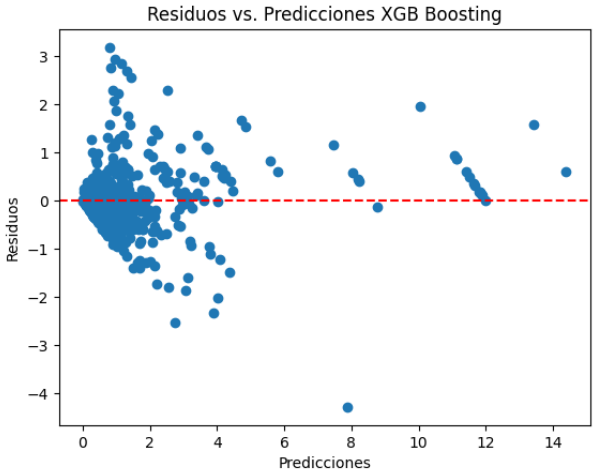
ADABoosting	
Error Absoluto Medio (MAE)	0.85



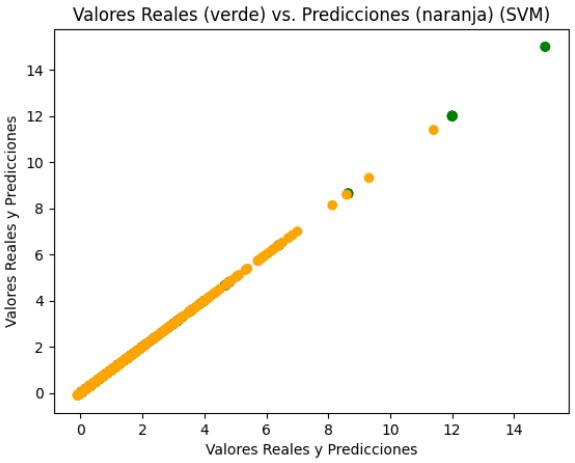


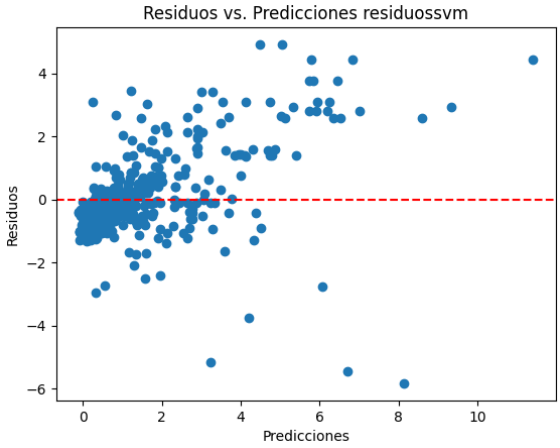
XGB Boosting	
Error Absoluto Medio (MAE)	0.46





Support Vector Machine	
Error Absoluto Medio (MAE)	0.53





Fuente: Elaboración propia

Los resultados obtenidos muestran un rendimiento variado. En el caso de los modelos de regresión, como mencionamos anteriormente, se observa un ajuste muy deficiente a los datos, lo que se traduce automáticamente en errores de predicción elevados, tanto para los modelos entrenados con los datos originales como para aquellos a los que se les aplicó reducción de dimensionalidad. En estos casos, el error supera una tonelada, lo cual, dadas las condiciones de producción del territorio, se considera excesivo.

Cuando exploramos el modelo de Regresión por Mínimos Cuadrados Robustos, se observa que es el modelo con las mayores diferencias en los dos experimentos desarrollados. Notablemente, cuando se entrena con los datos reducidos, el error se duplica, lo que sugiere que al transformar los datos se pierde información y relaciones relevantes.

Por otro lado, los algoritmos de Boosting (PCA Random Forest, PCA ADABoosting y PCA XGB Boosting) presentan en general un error en la predicción del rendimiento mucho más bajo, oscilando entre 400 y 600 kilogramos aproximadamente, y muestran resultados ligeramente mejores cuando se entrenan con los datos originales. Esto sugiere que estos algoritmos de ensamblaje, incluso después de la reducción de dimensionalidad, son muy efectivos para predecir el rendimiento de los cultivos.

En resumen, los modelos basados en ensamblaje, como Random Forest y XGB Boosting, destacan por su precisión, mientras que los modelos de regresión tradicionales tienen dificultades para captar correctamente la información contenida en los datos. Además, es importante señalar que el análisis gráfico revela patrones sistemáticos, presencia de datos atípicos y heterocedasticidad en general, lo que sugiere que los modelos tienen dificultades para ajustarse a los datos.

7 Discusión

La predicción del rendimiento de los cultivos en el trópico representa un desafío complejo debido a las diversas condiciones climáticas y geográficas, así como a la variabilidad en la calidad del suelo. Los ciclos de lluvia impredecibles, eventos climáticos extremos y la topografía ejercen una influencia significativa en la producción agrícola. Estos elementos subrayan la importancia de la construcción de modelos de aprendizaje automático, que constituye el enfoque central de este proyecto en la zona rural de Medellín.

En el contexto de Medellín, un área con una amplia diversidad de cultivos, prácticas agrícolas tradicionales y un promedio de edad de productores de 58 años, los resultados de este proyecto cobran una importancia aún mayor. El tamaño promedio de las parcelas en la región es de 0.8 hectáreas, y aproximadamente el 36% de la producción se destina al autoconsumo, clasificándose como una producción traspatio o de subsistencia. Esto enfatiza la necesidad de una alta precisión en las predicciones, ya que estas pueden tener un impacto directo en la seguridad alimentaria de la comunidad local.

Los resultados obtenidos revelaron una considerable variabilidad en el desempeño de los modelos. A excepción de los modelos lineales, los algoritmos de regresión de aprendizaje automático supervisado demostraron una destacada capacidad para aprender patrones complejos a partir de una

diversidad de datos, que incluyen información sobre las prácticas de producción, el índice de vegetación y datos meteorológicos, en especial los modelos de ensamble entrenados Boosting (PCA Random Forest, PCA ADABoosting y PCA XGB Boosting).

Para maximizar el potencial de este proyecto, es fundamental considerar la recopilación continua de datos. Una herramienta de seguimiento y monitoreo constante, como *Geobristol*, junto con el acompañamiento técnico a los agricultores, puede ayudar a obtener datos constantes y mapear los cambios en los cultivos de la ruralidad. Además, la medición más precisa de la producción por unidad de área sembrada puede ser un elemento crucial que, vinculado a los modelos propuestos, mejore aún más su rendimiento.

A pesar de las limitaciones identificadas, los modelos entrenados superaron las expectativas y son de gran valor para el sector agrícola de Medellín. Proporcionan una comprensión más profunda de los desafíos específicos que enfrentan los agricultores al estimar su producción. Esto puede conducir a la identificación de áreas de mejora en la recopilación de datos y la exploración de enfoques de modelado alternativos que podrían resultar más adecuados para esta región.

En síntesis, este proyecto se presenta como una herramienta de gran potencial para mejorar la toma de decisiones en la producción agrícola en la ruralidad de Medellín. Tanto profesionales del sector agrícola como instituciones pueden beneficiarse enormemente de la aplicación de la ciencia de datos en este contexto, incluyendo la transición hacia la planificación escalonada de cultivos y la programación de cosechas, elementos cruciales para optimizar el uso de las áreas disponibles, maximizar los rendimientos y mitigar los riesgos inherentes a la agricultura.

8 Conclusiones y recomendaciones

El proyecto, ha revelado conclusiones significativas que enfatizan la relevancia de implementar modelos de aprendizaje automático en la agricultura, especialmente en regiones caracterizadas por condiciones climáticas variables, eventos climáticos extremos y topografía compleja, como la zona rural de Medellín.

En primer lugar, se destaca la importancia crucial de los modelos de aprendizaje automático, especialmente los modelos de Boosting, como PCA Random Forest y PCA XGB Boosting, en la predicción precisa del rendimiento de los cultivos en esta región. Estos modelos han demostrado un ajuste sólido a los datos disponibles, respaldando su utilidad en la agricultura tropical.

La buena precisión en las predicciones del rendimiento de los cultivos se erige como un factor crítico, dado el énfasis en el autoconsumo y la producción traspatio en la comunidad rural de Medellín. La seguridad alimentaria local depende en gran medida de estas predicciones precisas.

A pesar de que el proyecto aún no se considerará como una herramienta definitiva para la toma de decisiones relacionadas con la cantidad de producción esperada en los cultivos locales, representa un punto de partida valioso. Los resultados obtenidos hasta el momento han proporcionado una comprensión más profunda de los desafíos específicos que enfrentan los agricultores al estimar su producción.

En términos de recomendaciones, se motiva a la mejora continua de los modelos de aprendizaje automático, especialmente los de Boosting, incorporando más datos históricos y la exploración de hiperparámetros específicos para optimizar su desempeño. Además, se enfatiza la importancia de la recopilación periódica de datos de alta calidad a través de herramientas de seguimiento y monitoreo constante, como *Geobristol*, respaldada por el acompañamiento técnico a los agricultores.

Finalmente, se propone como trabajos futuros la exploración de enfoques alternativos de modelado que puedan adaptarse mejor a las particularidades de la producción agrícola en los corregimientos de Medellín, incluyendo la consideración de variables adicionales, la generación de modelos diferenciales por tipo de cultivo o la aplicación de técnicas de reducción de dimensionalidad más avanzadas.

Estas conclusiones y recomendaciones subrayan la importancia de la aplicación de modelos de aprendizaje automático en la agricultura tropical y señalan el camino hacia una mayor precisión en la predicción del rendimiento de los cultivos, lo que beneficiará directamente a la seguridad alimentaria y la nutrición en el municipio.

9 Referencias Bibliográficas

Alcaldía de Medellín. (2020, June 17). *Distrito Rural Campesino*.

Amat, J. (2020). *Regularización Ridge, Lasso y Elastic Net con Python*. https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html

Apat, S. K., Mishra, J., Raju, S. kotagiri, & Padhy, N. (2022). The robust and efficient Machine learning model for smart farming decisions and allied intelligent agriculture decisions. *Journal of Integrated Science and Technology*, *10*(2). https://www.pubs.iscience.in/journal/index.php/jist/article/view/1463/814

Arana, C. (2021). *Redes neuronales recurrentes: análisis de los modelos especializados en datos secuenciales*. https://ucema.edu.ar/publicaciones/download/documentos/797.pdf

Bolaños, J., Corrales, J. C., & Campo, L. V. (2023). Feasibility of Early Yield Prediction per Coffee Tree Based on Multispectral Aerial Imagery: Case of Arabica Coffee Crops in Cauca-Colombia. *Remote Sensing*, *15*, 282. https://doi.org/10.3390/rs15010282

CEO, & Alcaldía de Medellín. (2016). *Caracterización de los productores agropecuarios rurales de Medellín*.

Cleves, J., Ramírez, L., & Díaz, E. (2023). Proposed empirical model for estimating 'Valencia' orange (Citrus sinensis L. Osbeck) productivity in the Colombian low tropics. *Revista Colombiana de Ciencias Hortícolas*, *3*. http://www.scielo.org.co/scielo.php?pid=S2011-21732021000300007&script=sci_arttext

FAO. (2018). *El futuro de la alimentación y la agricultura: Vías alternativas hacia el 2050*. https://www.fao.org/3/CA1553ES/ca1553es.pdf

FAO. (2022). *Estrategia de la FAO para la ciencia y la innovación*. https://www.fao.org/3/cc2273es/cc2273es.pdf

FAO. (2023). *El estado mundial de la agricultura y la alimentación 2023*. https://www.fao.org/documents/card/es/c/cc7724es

Gaviria, J. E. (2012). Medellín: Territorio y ruralidad. Una mirada a las cinco zonas corregimentales. *Escenarios*, *14*(2), 145–164.

IBM. (2016). *Guía de CRISP-DM de IBM SPSS Modeler*. https://www.ibm.com/docs/es/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf

Jhajharia, K., & Mathur, P. (2022). A comprehensive review on machine learning in agriculture domain. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *11*(2), 753–763. https://ijai.iaescore.com/index.php/IJAI/article/view/21357/13391

Lamos, H., Puentes, D., & Zarate, D. (2020). Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia. *Revista Facultad de Ingeniería*, *29*(54). http://www.scielo.org.co/scielo.php?pid=S0121-11292020000100018&script=sci_arttext

Lobato, J., Favilla, N., & Dalla, J. (2017). Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. *International Journal of Remote Sensing*, *38*(16). https://doi.org/10.1080/01431161.2017.1325531

Mokhtar, A., El-Ssawy, W., He, H., Al-Anasari, N., Sammen, S. S., Gyasi-Agyei, Y., & Abuarab, M. (2022). Using Machine Learning Models to Predict Hydroponically Grown Lettuce Yield. *Front Plant Sci*, *3*(13). https://doi.org/10.3389/fpls.2022.706042

Pham, H. T., Awange, J., & Kuhn, M. (2022). Evaluation of Three Feature Dimension Reduction Techniques for Machine Learning-Based Crop Yield Prediction Models. *Sensors: Advances in Time Series Analysis*, *22*(17), 6609. https://doi.org/10.3390/s22176609

Rananavare, L. B., & Chitnis, S. (2022). Crop Yield Prediction Using Temporal Data. *EEE*

International Conference on Electronics, Computing and Communication Technologies (CONECCT), 1–5. doi: 10.1109/CONECCT55679.2022.9865791

Rivera, J., Bunn, C., Rahn, E., Liitle-Savage, D., Schmidt, P., & Ryo, M. (2023). Co-Developing a Deep Learning-Based Crop Yield Estimation Method in Collaboration with Thousands of Smallholder Coffee Producers. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4457375

Sciforce. (2022). *Aprendizaje automático en agricultura: aplicaciones y técnicas*. https://sitiobigdata.com/2019/12/24/aprendizaje-automatico-en-agricultura/#

Setiya, P., Satpathi, A., Nain, A. S., & Das, B. (2022). Comparison of weather-based wheat yield forecasting models for different districts of Uttarakhand using statistical and machine learning Techniques. *Journal of Agrometeorology*, *24*(3), 255–261. https://journal.agrimetassociation.org/index.php/jam/issue/view/61

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4). https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Singh, A., Panda, R. K., & Dwivedi, B. S. (2023). Prediction of wheat yield using historical and weather data: A comparative study of machine learning models. *Computers and Electronics in Agriculture*, *203*.

Srivastava, A. K., Safae, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., & Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*, *12*, 3215. https://doi.org/10.1038/s41598-022-06249-w

UNIR Revista. (2021). *Árboles de decisión: en qué consisten y aplicación en Big Data*. Ingeniería y Tecnología. https://www.unir.net/ingenieria/revista/arboles-de-decision/

Villanueva, J. (2020). *Redes neuronales desde cero*. Machine Learning. https://www.iartificial.net/redes-neuronales-desde-cero-i-introduccion/

Yildirim, T., Moriasi, D. N., Starks, P. J., & Chakraborty, D. (2022). Using Artificial Neural Network (ANN) for Short-Range Prediction of Cotton Yield in Data-Scarce Regions. *Agronomy*, *12*(828), 19. https://rrpress.utsa.edu/bitstream/handle/20.500.12588/845/agronomy-12-00828-v2.pdf?sequence=1&isAllowed=y

Yli-Heikkila, M., Wittke, S., Luotamo, M., Puttonen, E., Sulkava, M., Pellikka, P., Heiskanen, J., & Klami, A. (2022). Scalable Crop Yield Prediction with Sentinel-2 Time Series and Temporal Convolutional Network. *Advances in Deep Learning Techniques for the Analysis of Remote Sensing Time Series*, *14*(17), 4193. https://doi.org/10.3390/rs14174193

10 Anexos

Anexo I. Relación de variables utilizadas en el proyecto

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
GlobalID	Corresponde al identificador único de cada registro y el conector con la base de datos de producción	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Código único por registro	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Fecha_Encuesta	Registro de la fecha de realización de encuesta	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Tiempo	Serie de tiempo desde el 16/01/2023 al 10/05/2023							
Código corregimiento	Nombre del territorio corregimental donde se mapeo el dato del productor	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Altavista San Antonio de Prado San Sebastián de Palmitas Santa Elena San Cristóbal	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Código vereda	Nombre de la vereda dónde reside el productor	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Corresponde al código geográfico asociado a cada una de las 52 veredas del Municipio de Medellín	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Código Catastral	Código asociado al predio registrado a nivel catastral y que sirve de para la construcción de la llave que conecta con los datos generados en la plataforma *Geobristol*	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Código correspondiente al identificador catastral de los predios donde se encuentra ubicadas las parcelas caracterizadas	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
Sexo del dueño de la parcela	Identificación del sexo de cada productor caracterizado	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Hombre Mujer Intersexual	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿De dónde provienen principalmente la mano de obra empleada para la producción?	Caracteriza la procedencia de la mano de obra: Familiar, Contratada, Comunitaria (minga), Intercambio de mano de obra	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Familiar Contratada Comunitaria (minga) Intercambio de mano de obra	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Invernadero	Identifica si el productor tiene invernadero como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Poli_Sombra	Identifica si el productor tiene polisombra como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Marquesina	Identifica si el productor tiene marquesina como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Inf_Bioabono	Identifica si el productor tiene infraestructura para la producción de abonos orgánicos como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
Tractor	Identifica si el productor tiene tractor como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Lombricultivo	Identifica si el productor tiene lombricultivo como infraestructura en su parcela para la producción. La tenencia de la infraestructura no implica uso.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿Utiliza maquinaria agrícola para el arado del suelo?	Identifica cómo práctica sostenible el uso o no de maquinaria agrícola para arar el suelo.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿Siembra en contra de la pendiente?	Identifica cómo práctica sostenible el sembrar en contra de la pendiente.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿Tiene cultivos limpios en zonas de alta pendiente que protegen el suelo?	Identifica cómo práctica sostenible la siembra de cultivos en zonas de alta pendiente.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿Realiza prácticas de conservación y recuperación de suelos en su parcela?	Identifica si el productor realiza alguna práctica de conservación de los suelos en la producción	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
En su parcela, ¿aplica prácticas de producción sostenible?	Identifica si el productor realiza alguna práctica considerada de producción sostenible	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Diversificación de cultivos y aplicación de buenas prácticas ambientales Certificación de Buenas Práctica Agrícolas o predio exportador Vinculación de modelos de agricultura sostenible cómo agricultura orgánica, agroecología o agricultura regenerativa)	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿Para la producción agropecuaria, usted hace uso de?	Indaga acerca del uso de los fertilizantes y productos en el manejo agronómico de los cultivos orgánico, químico o mixto	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Orgánico Químico Mixto	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
¿Cómo accede al agua para la producción?	Identifica la procedencia del agua utilizado en el riego de los cultivos	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Acueducto Río o quebrada Pozo Agua lluvia Sistema de riego	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
De las siguientes opciones ¿cuáles formas financiación para la producción utiliza?	Identifica la procedencia de los recursos financieros destinados para el financiamiento de la producción:	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Capital propio Banco Agrario Otros Bancos Comerciantes/intermediarios Cooperativas Almacén de insumos Subsidios, auxilios Paga diarios (Gota gota)	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
En el último año ¿Ha recibido capacitación o asistencia técnica?	Identifica si ha sido capacitado a nivel productivo en el último año	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
Usted como productor, ¿Utiliza o ha utilizado internet para capacitarse y mejorar sus procesos productivos?	Indaga acerca del uso de herramientas tecnológicas y en particular internet para la mejora de procesos productivos	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Epsea	Identifica si el productor está inscrito en el Registro único de extensión agropecuaria del Distrito de Medellín.	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Si No	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
ParentGlobalID	Corresponde al identificador único de cada registro y el conector con la base de datos de información del productor	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Identificador único de cada parcela	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Sistema productivo	Identifica el tipo de sistema productivo presente en la parcela: agrícola, pecuario o forestal	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Agrícola, pecuario o forestal	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Producto producido	Identifica el producto cultivados	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Plátano Café Limón Cilantro Cebolla de rama Banano Lechuga Papa Aguacate	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
Área (ha)	Área sembrada u ocupada por cada uno de los rubros cultivados en la parcela	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Numérica	No aplica	0.12	0.30	0.00	0.01	0.04	0.11	6.89
Volumen (Ton/cosecha)	Producción por cosecha obtenida del área sembrada con el producto cultivado. Variable objetivo	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Numérica	No aplica	1.38	2.75	0.00	0.16	0.60	1.60	44.32
Coordenada X	Longitud de coordenada del centroide de cada uno de los polígonos de los cultivos	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Coordenada del centroide del polígono de cada cultivo	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Coordenada Y	Latitud de coordenada del centroide de cada uno de los polígonos de los cultivos	Proceso de recolección de información primaria a partir de la Caracterización Rural (Convenio Conexión Medellín Rural)	Categórica	Coordenada del centroide del polígono de cada cultivo	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Campo	Identificador de cada uno de los polígonos de cultivos, conformado por el código catastral de la parcela y el cultivo sembrado	Índice de vegetación proporcionados por Geobristol: plataforma de seguimiento y monitoreo de cultivos. (Convenio Conexión Medellín Rural)	Categórica	Identificador único de cada cultivo	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica
Grupo	Corresponde al corregimiento donde está ubicada la parcela y el cultivo	Índice de vegetación proporcionados por Geobristol: plataforma de seguimiento y monitoreo de cultivos. (Convenio Conexión Medellín Rural)	Categórica	Altavista San Antonio de Prado San Sebastián de Palmitas Santa Elena San Cristóbal	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica	No aplica

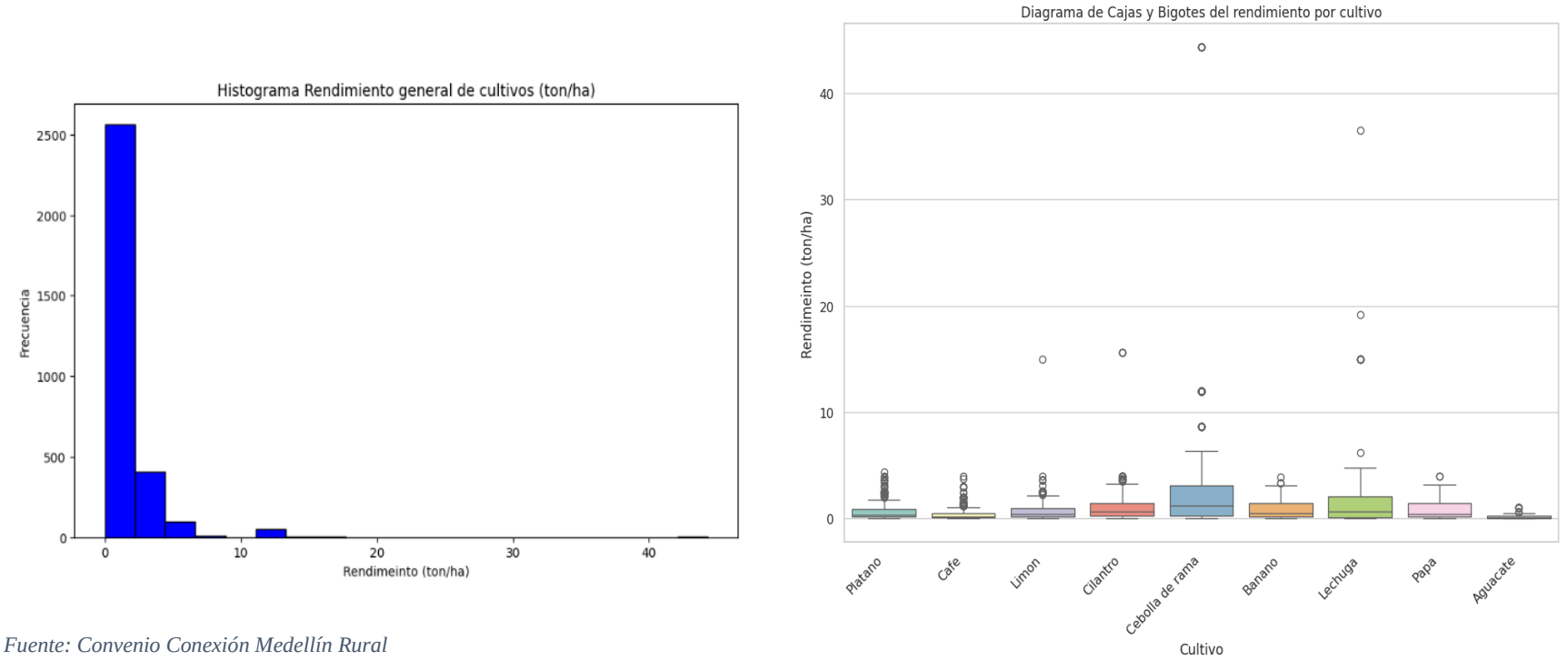
					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
Valor del índice	Índice NDVI -Índice de Vegetación de Diferencia Normalizada-: Evalúa la salud y la densidad de la vegetación en cada cultivo monitoreado a través de la plataforma considerando el valor máximo en el periodo analizado, para cada uno de los polígonos por producto en las parcelas objeto de estudio.	Índice de vegetación proporcionados por Geobristol: plataforma de seguimiento y monitoreo de cultivos. (Convenio Conexión Medellín Rural)	Numérica	No aplica	0.72	0.12	0.24	0.65	0.74	0.81	1.00
Fecha de la imagen	Fecha en la que fue tomada la imagen satelital del cultivo.	Índice de vegetación proporcionados por Geobristol: plataforma de seguimiento y monitoreo de cultivos. (Convenio Conexión Medellín Rural)	Tiempo	Serie de tiempo desde el 16/01/2023 al 31/08/2023							
Fecha	Fecha de generación del dato	Datos descargados a través de Geobristol, provenientes del servicio World Weather Online (Convenio Conexión Medellín Rural)	Tiempo								
Máx grado C	Temperatura máxima en grados centígrados	Datos descargados a través de Geobristol, provenientes del servicio World Weather Online (Convenio Conexión Medellín Rural)	Numérica	No aplica	23.19	2.81	16.00	22.00	23.00	25.00	31.00
Mín grado C	Temperatura mínima en grados centígrados	Datos descargados a través de Geobristol, provenientes del servicio World Weather Online (Convenio Conexión Medellín Rural)	Numérica	No aplica	12.74	1.01	10.00	12.00	13.00	13.00	16.00

					Medidas de tendencia central (numéricas)						
Nombre de la variable	**Descripción**	**Fuente**	**Tipo variable**	**Opciones respuesta (categóricas)**	**Media**	**Desviación Estándar**	**Mínimo**	**0,25**	**0,5**	**0,75**	**Máximo**
Humedad (%)	Porcentaje de humedad del ambiente	Datos descargados a través de Geobristol, provenientes del servicio World Weather Online (Convenio Conexión Medellín Rural)	Numérica	No aplica	79.79	7.47	59.50	74.50	80.50	85.00	97.50
Precipitaciones (mm)	Cantidad de agua que cae en forma de lluvia	Datos descargados a través de Geobristol, provenientes del servicio World Weather Online (Convenio Conexión Medellín Rural)	Numérica	No aplica	19.58	18.73	0.00	2.65	15.20	31.10	86.10

Anexo II. Gráficos: análisis descriptivo de los datos

Al análisis preliminar, realizado antes de proponer el modelo para predecir el rendimiento de cultivos, sirve como piedra angular para la toma de decisiones fundamentadas y fortalece la solidez y validez de los análisis posteriores. En este contexto, se presentan gráficos de respaldo que facilitan la comprensión de los datos empleados en el proyecto, destacando las características esenciales, las distribuciones y las tendencias clave que respaldan la creación y evaluación de los modelos predictivos. Este análisis proporciona una panorámica integral del conjunto de datos, facilitando así la comprensión y validación de los resultados obtenidos en la predicción del rendimiento de los cultivos.

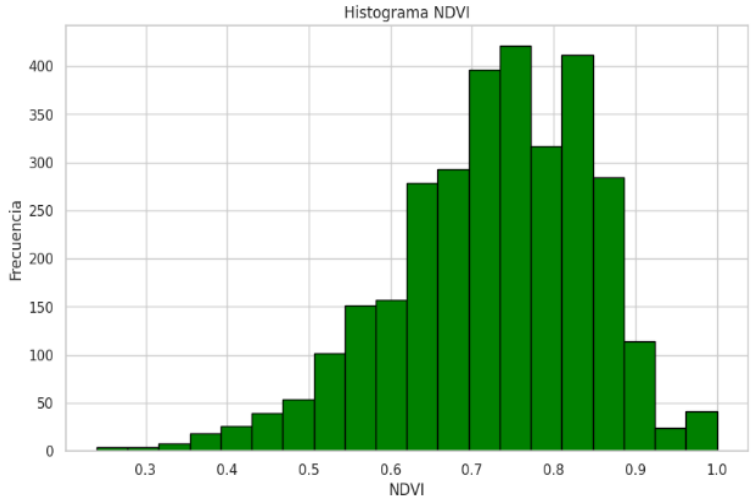
Figura A7. Histograma del rendimiento general y por cultivos

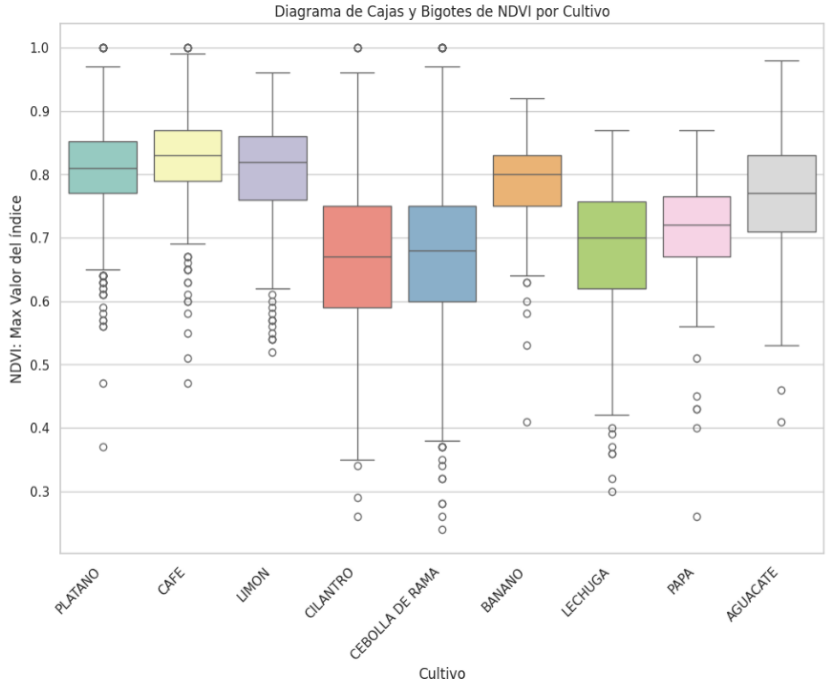


Fuente: Convenio Conexión Medellín Rural

Al examinar el gráfico de producción agrícola, se observa que el plátano lidera con la mayor producción media, seguido por la papa y el limón. No obstante, la cebolla de rama destaca por su notoria variabilidad en la producción, seguida de cerca por el cilantro y el aguacate. Resulta llamativo identificar valores atípicos en cilantro, cebolla de rama y lechuga, indicando posibles condiciones excepcionales que podrían influir en su rendimiento. Además, la distribución asimétrica sugiere que la mayoría de los cultivos tienen una concentración de datos por debajo de la mediana, revelando patrones de producción que podrían derivarse de prácticas agrícolas consistentes, pero también señalando posibles desafíos específicos en algunos cultivos.

Figura A8. Histograma del rendimiento general y por cultivos



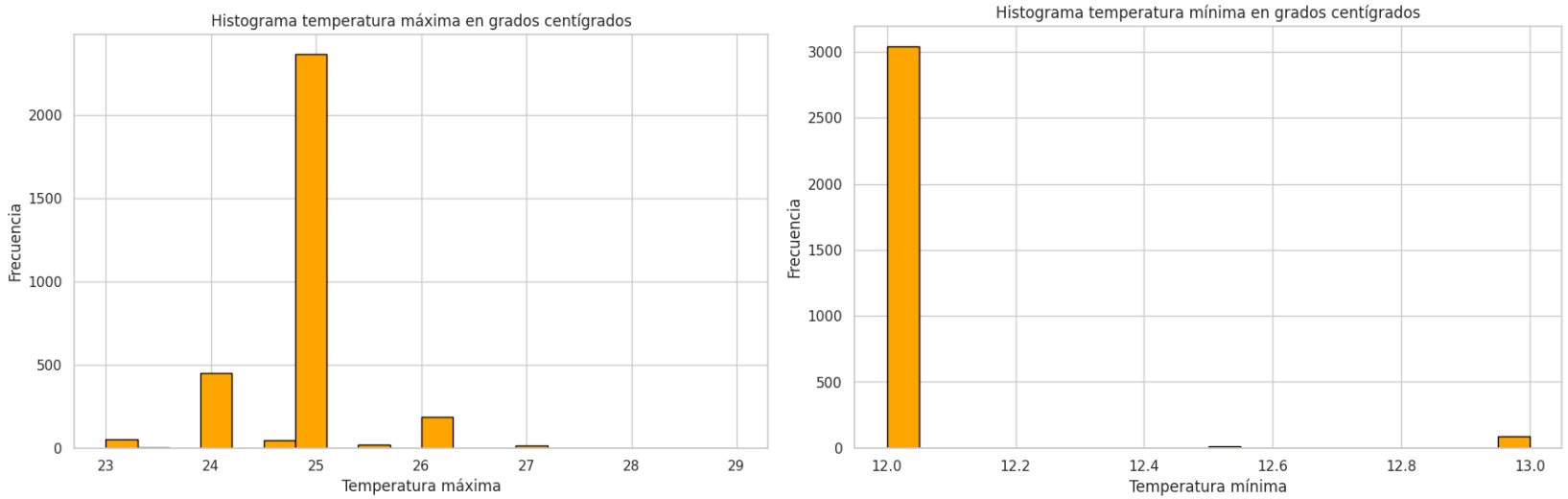


Fuente: Convenio Conexión Medellín Rural

Con respecto al NDVI, se destaca el plátano con el mayor valor mediano, seguido de la papaya y el limón, indicando un vigor vegetativo significativo en estos cultivos. No obstante, se observa que la cebolla de rama presenta la mayor variabilidad en el NDVI, seguida por el cilantro y el aguacate. La detección de valores atípicos, especialmente en cilantro, cebolla de rama y lechuga, sugiere condiciones

particulares que podrían afectar su índice de vegetación. Además, la distribución asimétrica del NDVI en la mayoría de los cultivos, con la concentración de datos por debajo de la mediana, revela patrones distintivos en el vigor vegetativo, proporcionando valiosa información para entender las condiciones ambientales y optimizar la gestión agrícola.

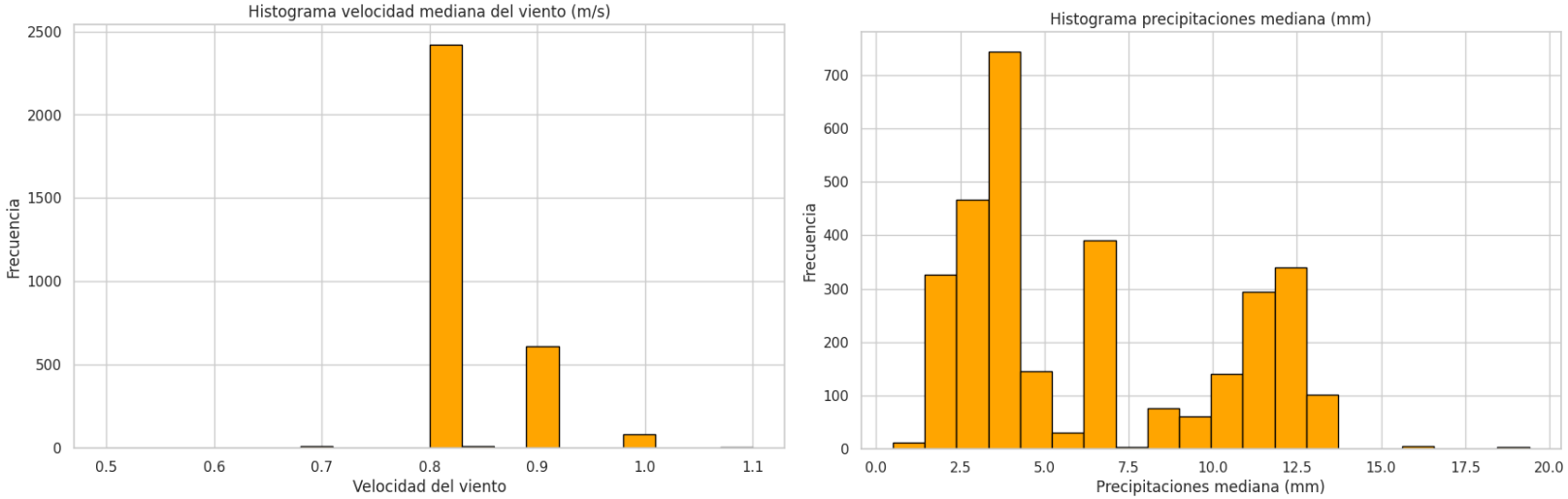
Figura A9. Histograma general variables climáticas: temperatura máxima y mínima



Fuente: Convenio Conexión Medellín Rural

Los histogramas revelan patrones particulares en las variables climáticas analizadas. La temperatura máxima presenta una distribución unimodal, indicando que la mayoría de las temperaturas máximas se concentran entre 25 y 30 grados C°. Sin embargo, se identifican algunos valores atípicos por encima de 35 grados °, sugiriendo eventos climáticos excepcionales. Por otro lado, la temperatura mínima, revelando una distribución bimodal. Este patrón indica la presencia de dos grupos distintos de valores, con la mayoría oscilando entre 15 y 20 grados Celsius. Notablemente, se observan algunos valores atípicos por debajo de 10 grados °, que pueden indicar condiciones climáticas inusuales.

Figura A10. Histograma general variables climáticas: velocidad mediana del viento y precipitaciones medianas



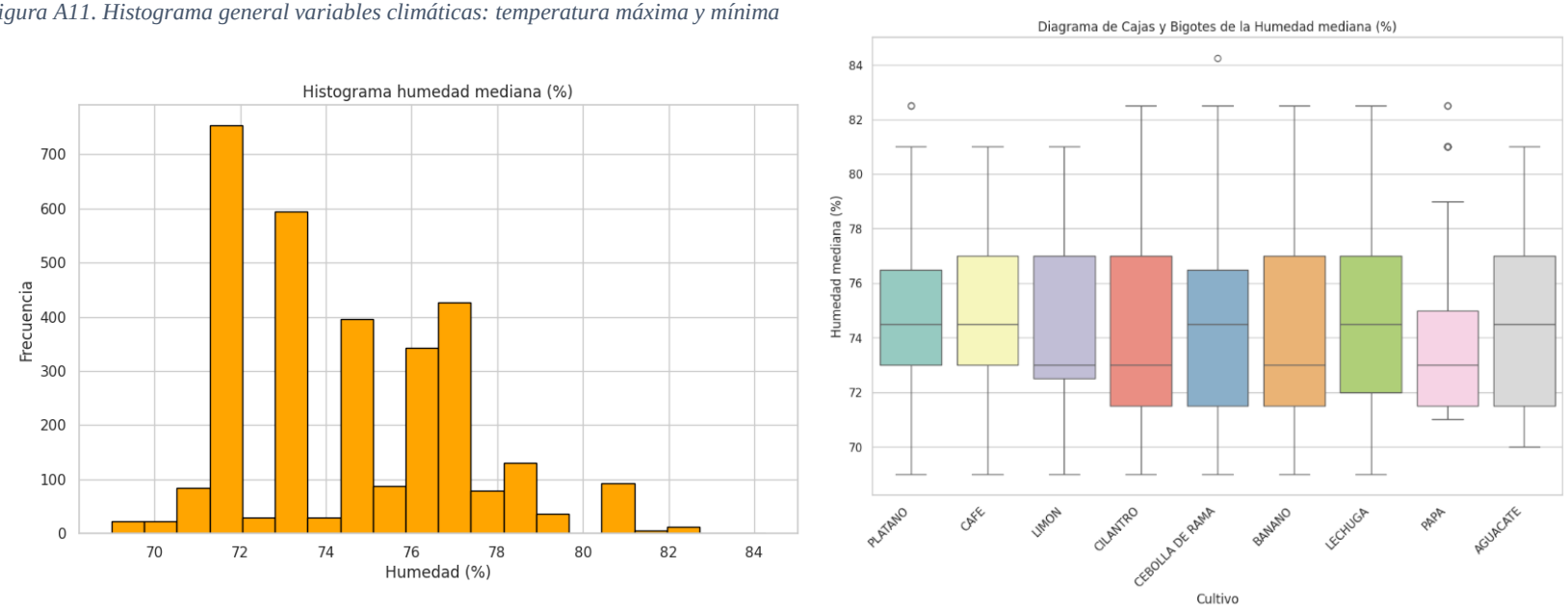
Fuente: Convenio Conexión Medellín Rural

Con respecto a la velocidad del viento, se evidencia una distribución unimodal con la mayoría de las velocidades entre 2 y 4 metros por segundo. No obstante, se detectan valores atípicos por encima de 6 metros por segundo, indicando momentos de intensidad ventosa significativa.

La precipitación, presenta una distribución unimodal con la mayoría de los valores agrupados entre 0 y 5 milímetros. Sin embargo, se destacan algunos valores atípicos por encima de 10 milímetros, indicando episodios de precipitación excepcionalmente intensa.

Las temperaturas máxima y mínima y la velocidad del viento presentan distribuciones unimodales y bimodales, respectivamente, evidenciando la complejidad de las condiciones climáticas. La presencia de valores atípicos en cada histograma subraya la importancia de considerar eventos climáticos excepcionales en el análisis y la planificación.

Figura A11. Histograma general variables climáticas: temperatura máxima y mínima



Fuente: Convenio Conexión Medellín Rural

Analizando el porcentaje de humedad y el rendimiento de cultivos, la gráfica sugiere una relación positiva entre ambas variables En términos generales, se observa que a medida que la humedad del suelo aumenta, el rendimiento de los cultivos tiende a ser más elevado. No obstante, es crucial señalar que existen excepciones a esta tendencia, evidenciadas en los casos del cilantro y la cebolla de rama, los cuales presentan rendimientos divergentes a pesar de niveles similares de humedad en el suelo.