DETERMINACIÓN DE LA PRODUCCIÓN DE AGUACATE EN COLOMBIA A PARTIR DE FACTORES METEREOLÓGICOS EMPLEANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Proyecto de Trabajo de Grado

Anyi Marcela Cajas Santacruz

Leandro Potes Urrutia

Director: Msc. Juan Fernando Casanova Olaya
Codirector: PhD. Juan Carlos Corrales Muñoz
Asesor: PhD. Cristhian Nicolás Figueroa Martínez

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Departamento de Telemática

Popayán, diciembre de 2023

DETERMINACIÓN DE LA PRODUCCIÓN DE AGUACATE EN COLOMBIA A PARTIR DE FACTORES METEREOLÓGICOS EMPLEANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Anyi Marcela Cajas Santacruz

Leandro Potes Urrutia

Trabajo de grado presentado a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para optar por el título de: Ingeniero en Electrónica y Telecomunicaciones

Director: Msc. Juan Fernando Casanova Olaya
Codirector: PhD. Juan Carlos Corrales Muñoz
Asesor: PhD. Cristhian Nicolás Figueroa Martínez

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Departamento de Telemática

Popayán, diciembre de 2023

Resumen

El aguacate se ha convertido en uno de los cultivos más importantes de Latinoamérica, desempeñando un papel crucial en la economía y la agricultura de la región. Su versatilidad culinaria, alto contenido nutricional y demanda creciente en los mercados internacionales han impulsado su producción y comercialización en países como Colombia. El cultivo de aguacate se ve afectado directamente por factores meteorológicos cuya alta variabilidad dificulta la predicción de las condiciones futuras del cultivo y por lo tanto la producción del mismo. En este sentido, es necesario proponer herramientas que faciliten una estimación precisa de la producción del cultivo de aguacate. Por ello, en este estudio se implementan una serie de algoritmos de aprendizaje automático para estimar la producción de aguacate a partir de datos meteorológicos y la fenología del cultivo.

Inicialmente, se realizó una revisión sistemática de literatura bajo la metodología de Kitchenham para identificar elementos considerados en otros estudios dirigidos a la estimación de la producción de aguacate y otros tipos de cultivos. También, se implementó la metodología CRISP-DM para realizar el proceso de minería de datos, que permitió el entendimiento y la construcción de conjuntos de datos óptimos para ser modelados con algoritmos de aprendizaje automático, como la regresión de bosques aleatorios, máquinas de vectores de soporte, redes neuronales artificiales y regresiones lineales. Para la evaluación de los modelos se emplearon las métricas RMSE, R^2 y MAE bajo un método de validación cruzada de k iteraciones.

El estudio reveló que la altitud y el comportamiento fenológico del cultivo son factores clave para la estimación de la producción de aguacate. Se encontró que el algoritmo de bosques aleatorios fue la solución óptima, ya que ofreció un equilibrio entre rendimiento y explicabilidad.

Este trabajo contribuye al campo de la agricultura al abordar las brechas en la investigación de la producción de aguacate y al proponer posibles áreas de mejora para estudios futuros, como el uso de fuentes de datos más precisas y la exploración de técnicas de aprendizaje profundo para mejorar la estimación de la producción de aguacate.

Abstract

Avocado has become one of the most important crops in Latin America, playing a crucial role in the region's economy and agriculture. Its culinary versatility, high nutritional content and growing demand in international markets have boosted its production and marketing in countries such as Colombia. Avocado cultivation is directly affected by meteorological factors whose high variability makes it difficult to predict future crop conditions and therefore its production. In this sense, it is necessary to propose tools that facilitate an accurate estimation of avocado crop production. Therefore, in this study a series of machine learning algorithms are implemented to estimate avocado production from meteorological data and crop phenology.

Initially, a systematic review of the literature was carried out using the Kitchenham methodology in order to identify elements considered in other studies aimed at estimating the production of avocado and other types of crops. Also, the CRISP-DM methodology was implemented to carry out the data mining process, which allowed the understanding and construction of optimal data sets to be modeled with machine learning algorithms, such as random forest regression, support vector machines, artificial neural networks and linear regressions. To evaluate the models, the RMSE, R^2 and MAE metrics were used under a k-iteration cross-validation method.

The study revealed that altitude and phenological behavior of the crop are key factors for estimating avocado production. The random forest algorithm was found to be the optimal solution as it offered a balance between performance and explainability.

This work contributes to the field of agriculture by addressing gaps in avocado production research and proposing possible areas of improvement for future studies, such as using more accurate data sources and exploring deep learning techniques to improve the estimation of avocado production.

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

1.1. Planteamiento del problema

La agricultura es una actividad crucial para el desarrollo de los países de Latinoamérica y el Caribe, ya que es una región privilegiada en recursos naturales y tiene el potencial para exportar y proveer alimentos a nivel global [1]. Entre los cultivos que más destacan en los últimos años, se encuentra el de aguacate, un producto que ha ganado reconocimiento mundial por sus cualidades alimenticias y saludables, lo que se ve reflejado en las importaciones del fruto a nivel global que han incrementado un 171.97% durante la década de los 2010 [2]. Colombia es un país que produce grandes cantidades de este producto ya que se registran aproximadamente 54.000 hectáreas que representan el 6% del área sembrada a nivel mundial. Sin embargo, se encontró que el sistema de producción de aguacate no siempre cuenta con el respaldo tecnológico adecuado y se presentan siembras en áreas no apropiadas para esta especie. Estos problemas han llevado a que algunas inversiones terminen en fracaso o en baja sostenibilidad [3].

Existen otro tipo de problemáticas en la cadena productiva que exceden el manejo del cultivo, como los efectos adversos ocasionados por el cambio y la variabilidad climática en los cultivos. En el caso del cultivo de aguacate, esta variabilidad puede afectar el crecimiento y productividad de la planta. Los principales aspectos que inciden sobre la productividad del cultivo de aguacate son la humedad del suelo, la precipitación, los vientos y los cambios en la temperatura. Cuándo se presenta un déficit hídrico severo, la planta puede presentar afectación en la formación de yemas, el desarrollo del órgano floral, la floración y desarrollo del fruto. Así mismo, cuando se presenta este tipo de déficit, el cultivo es más propenso a sufrir el ataque de plagas y enfermedades [4]. La precipitación también es un factor de alta relevancia debido a que las sequías prolongadas provocan la caída de las hojas, afectando la fotosíntesis y el desarrollo de la planta, además, el exceso de precipitación durante la floración y la fructificación puede generar la caída del fruto, estas condiciones causan una reducción en la producción del cultivo [5].

Por otra parte, los vientos afectan la transpiración y la distribución de luz lo que puede impactar sobre el proceso de fotosíntesis. La falta de fotosíntesis limita el desarrollo de frutos grandes y de brotes vigorosos [6], y en caso de vientos de altas velocidades (por encima de 20 km/hora) puede provocar la ruptura de ramas, caída de flores y frutos y quemazón de las hojas y brotes del árbol [7]. Igualmente, las temperaturas bajas afectan el desarrollo del cultivo en la fase de floración, y las temperaturas altas acortan el periodo de apertura de las flores y disminuyen la viabilidad del polen [8]. Otros efectos se observan en el funcionamiento fisiológico adecuado de la fotosíntesis, transpiración y respiración del aguacate impidiendo tener cosechas todos los años [6].

Estos problemas ocasionados por la variabilidad climática generan consecuencias en varios aspectos de los cultivos, incrementando el factor de incertidumbre a todo el proceso e impidiendo al agricultor la toma de medidas de acuerdo con las condiciones climáticas en la zona de su cultivo. De acuerdo con la problemática encontrada sobre la incertidumbre en la producción de aguacate debido a los efectos que tienen las condiciones meteorológicas sobre el cultivo, se plantea la siguiente pregunta de investigación:

¿Cómo estimar la producción de aguacate en Colombia considerando factores meteorológicos?

1.2. Objetivos

1.2.1. Objetivo General

Determinar la producción de aguacate en Colombia a partir de factores meteorológicos empleando técnicas de aprendizaje automático.

1.2.2. Objetivos específicos

- Caracterizar los factores meteorológicos que influyen en la producción de aguacate en Colombia.

- Implementar modelos de aprendizaje automático para la estimación de la producción de aguacate.

- Evaluar la exactitud de los modelos de aprendizaje automático implementados.

1.3. Contenido de la monografía

Este documento está compuesto por 5 capítulos, que se exponen a continuación:

- **1. Introducción:** En este capítulo se presenta el planteamiento del problema y la pregunta de investigación, se definen los objetivos y se expone la estructura del documento realizado.

- **2. Estado actual del conocimiento:** Se presenta el desarrollo de la revisión sistemática de literatura, describiendo las técnicas y métodos empleados por investigaciones similares. También se identifican las brechas encontradas entre los estudios relacionados y la investigación expuesta en este trabajo.

- **3. Conjunto de datos meteorológicos y de producción de aguacate en Colombia:** Presenta la metodología implementada para realizar el proceso de minería de datos, describiendo paso a paso cada fase hasta obtener un conjunto de datos preparado para modelar.

- **4. Modelo para la estimación de producción de aguacate:** En este capítulo se muestra el proceso de modelado empleando algoritmos de aprendizaje automático, se exponen y se analizan los resultados arrojados por los distintos modelos implementados.

- **5. Conclusiones y trabajos futuros:** En este capítulo se exponen las conclusiones generales de toda la investigación y los trabajos futuros propuestos de acuerdo al análisis realizado.

2. Estado actual del conocimiento

2.1. Revisión sistemática de literatura

Para cumplir con los objetivos planteados para este trabajo, es necesario hacer una investigación que permita recolectar la información necesaria para poder plantear el modelo, esto consiste en detalles sobre los cultivos y como las condiciones climáticas pueden afectarlos. También es importante conocer las distintas técnicas (principalmente las de aprendizaje automático) que permiten realizar las estimaciones de la producción o rendimiento de los cultivos, y las formas en que estas mismas pueden ser evaluadas.

Para realizar esta investigación se realiza una revisión sistemática de literatura bajo la metodología desarrollada por Barbara Kitchenham [9], la cual consiste en una serie de pasos que se desarrollan a continuación:

2.1.1. Protocolo de revisión

El primer paso consiste en establecer las preguntas de investigación que se quieren resolver, en este caso las preguntas son realizadas a partir de cada uno de los objetivos específicos planteados para el trabajo:

- ¿Qué factores meteorológicos influyen en la producción de los cultivos?
- ¿Qué técnicas se han utilizado para la estimación de la producción de distintos cultivos?
- ¿Cómo se evalúan estas técnicas o modelos?

A partir de estas preguntas se establecen unas palabras clave (junto con sinónimos de estas) que posteriormente se utilizan para construir la cadena de búsqueda, las palabras claves seleccionadas son las siguientes: cultivo, producción, rendimiento, cosecha, estimación, modelo, simulación, aprendizaje automático, inteligencia artificial, minería de datos, clima, meteorología.

2.1.2. Proceso de búsqueda

Para esta sección se realiza una búsqueda de artículos científicos y libros que contengan información sobre cultivos, los factores meteorológicos que influyen en estos y las técnicas usadas para la estimación de la producción. La búsqueda se realizará en las bases de datos de Science Direct y Web of Science empleando la herramienta de búsqueda avanzada que estas permiten, allí se incluye la cadena de búsqueda establecida a partir de las palabras clave elegidas en el protocolo de revisión. Por el formato de las plataformas, las palabras claves se tradujeron al inglés y se creó una cadena de búsqueda general que permita encontrar la mayor cantidad de artículos relacionados con el tema:

crop AND (forecasting, OR simulation OR model) AND (Production OR Yield OR harvest) AND ("machine learning" OR "artificial intelligence")

Esta cadena de búsqueda se introdujo en las bases de datos seleccionadas, buscándolas en el título, el abstract y las keywords de los artículos. El resultado obtenido consiste en 839 documentos de Web of Science y 367 documentos de Science Direct, para un total de 1206 artículos.

2.1.3. Criterios de inclusión y exclusión

Se deben establecer de manera previa los criterios de inclusión y exclusión, es decir, los parámetros que deben incluir los artículos para que puedan ser aceptados para la revisión. Los criterios establecidos para este caso son los siguientes:

- **Se excluyen aquellos artículos que hayan sido publicados hace más de diez años:** este criterio se establece debido a que el campo de las tecnologías de la información y más específicamente el de la inteligencia artificial ha avanzado mucho en los últimos años, e incluir artículos publicados hace mucho tiempo puede llevar a encontrar información obsoleta y desactualizada.

- **Se excluyen los artículos que no tengan como uno de sus objetivos la estimación de la producción o rendimiento de un cultivo:** se establece este criterio debido a que la estimación de la producción es el objetivo principal de esta revisión por lo cual es sumamente necesario que los estudios encontrados cumplan con esta condición, y a partir de esto, es posible encontrar la información que permita resolver las preguntas de investigación.

- **Se excluyen los artículos que no están disponibles para la lectura:** es posible que al realizar la búsqueda se encuentre la referencia de un artículo que cumpla con los criterios de inclusión, sin embargo, si en la base de datos no está disponible el artículo y no se puede encontrar por medios externos, el artículo será rechazado debido a que no será posible leerlo en su totalidad.

- **Se excluyen todos los artículos duplicados que se encuentren:** este criterio se establece puesto a que se emplea la cadena de búsqueda en dos bases de datos distintas, por lo que es probable que parte de los artículos estén en ambas fuentes, también, se asumen como duplicados las nuevas ediciones de los mismos estudios (solo en caso de ser escritos por los mismos autores).

  Para realizar el filtrado se revisa el título de cada uno de los 1206 artículos resultantes de la búsqueda, empleando los criterios establecidos para definir si cada artículo es aceptado o rechazado. El resultado consiste en 224

artículos, luego se procede a revisar el abstract de esos artículos restantes y se vuelven a aplicar los mismos criterios de inclusión y exclusión para refinar aún más el resultado. De este segundo filtro quedan 162 artículos que van a ser leídos completamente y con los cuales se continúa la revisión sistemática.

2.1.4. Evaluación de la calidad

Para realizar la evaluación de la calidad, se realizan una serie de preguntas de respuesta sí/no/parcialmente, que faciliten establecer qué tan relevante es cada artículo de acuerdo a lo que se busca para la investigación. Luego se le da un puntaje a cada artículo de acuerdo a las respuestas de esas preguntas; para este caso, cada respuesta “Si” tiene el valor de 1 punto, una respuesta “Parcialmente” el valor de 0.5; y una respuesta “No” el valor de 0 puntos. Cada uno de los artículos resultantes del proceso de inclusión y exclusión debe ser leído y las preguntas deben ser respondidas para darle un puntaje final de calidad a dicho artículo.

Las preguntas establecidas son las siguientes:

- ¿El artículo realiza comparaciones entre diferentes técnicas?
- ¿La principal fuente de datos para la estimación, son datos meteorológicos?
- ¿La técnica empleada para la estimación es de aprendizaje automático?
- ¿El artículo no es meramente teórico, sino que realiza un caso de estudio práctico?
- ¿El cultivo estudiado es el aguacate?
- ¿El artículo explica la técnica utilizada?
- ¿El artículo realiza y explica el método de evaluación de la técnica utilizada?
- ¿El artículo incluye los datos utilizados para realizar la estimación?
- ¿En el artículo se hace un estudio del peso/relevancia de las variables?

Como se observa, las preguntas están orientadas a lo que se busca en la investigación, y aquellos artículos que tengan mayor puntaje son aquellos artículos que más probablemente ayuden a responder las preguntas de investigación, y serán los más relevantes para la extracción y el análisis de los datos. Se establece un valor mínimo de 6 puntos para que un artículo se considere de alta relevancia y sea revisado con mayor detalle.

2.1.5. Recolección de datos

La recolección de los datos consiste en adquirir los datos puntuales de cada estudio que permitan posteriormente responder las preguntas de investigación, para esto se debe establecer los datos que se vana extraer antes de iniciar con la lectura de los artículos:

- Nombre del artículo
- Cultivo o cultivos sobre los que se hace la estimación
- País donde se realiza el estudio
- Fecha de recolección de los datos
- Variables utilizadas para la estimación
- Técnica usada para la estimación
- Método de evaluación de resultados
- Notas adicionales (aquí se incluyen las bases de datos, anexos o información adicional que pueda ser útil y que pueda variar en cada artículo)

Luego se organizan estos datos en una tabla para facilitar el análisis de los mismos.

2.1.6. Análisis de datos

Para realizar el análisis de resultados, se optó por abordarlo de dos maneras: en primer lugar, se realiza un análisis cuantitativo sobre los datos obtenidos, mostrando gráficamente cuales son las variables más frecuentes en cada una de las categorías de los datos. Posteriormente se realiza en análisis cualitativo, en el cual se describen los aspectos más importantes obtenidos sobre cada una de las categorías, teniendo en cuenta la información abstraída en la lectura.

2.1.6.1. Análisis cuantitativo de los datos

El análisis cuantitativo de los datos se realiza teniendo en cuenta cada uno de los tipos de datos extraídos de los artículos seleccionados, es decir, se tiene en cuenta el cultivo sobre el que se hace el estudio, el país donde se realiza, las técnicas y variables empleadas para realizar la estimación y los métodos de evaluación de resultados.

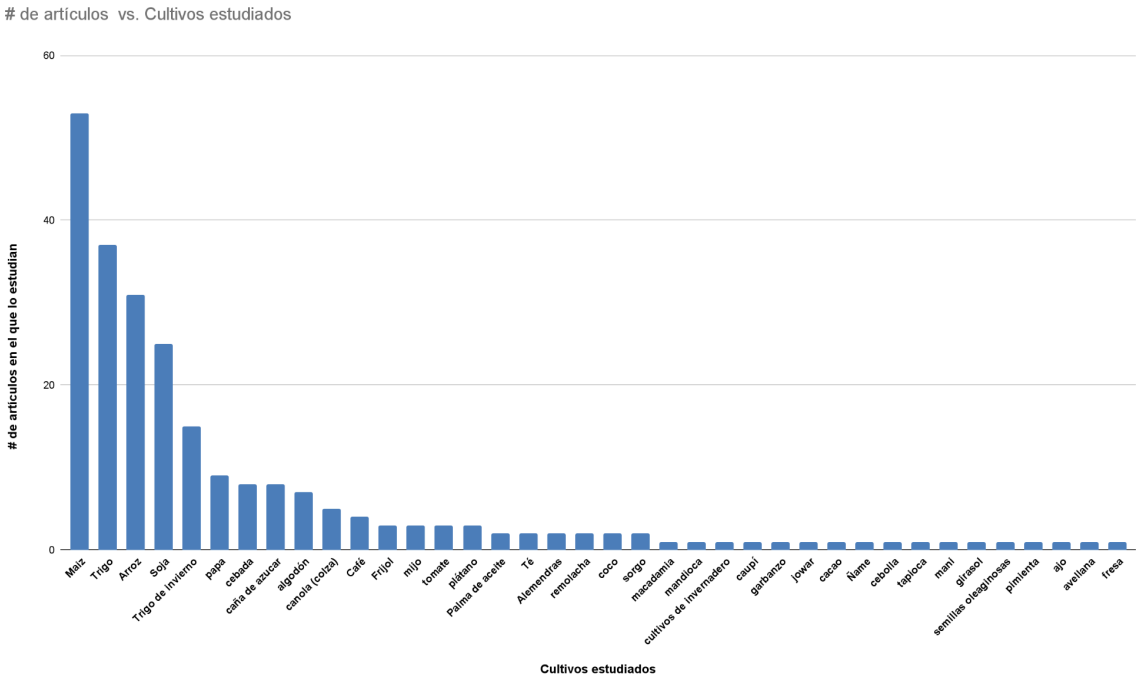

Figura 1. Cultivos estudiados. Fuente propia.

En los artículos encontrados se puede observar que los estudios enfocados en la producción y rendimiento se hacen principalmente sobre granos, siendo el maíz el que destaca sobre el resto, como se evidencia en la **Figura 1**. También, se pueden encontrar otros tipos de cultivos como frutos. El aguacate no es un cultivo sobre el que se hayan hecho estudios de producción y rendimiento dentro de los artículos seleccionados para esta revisión.

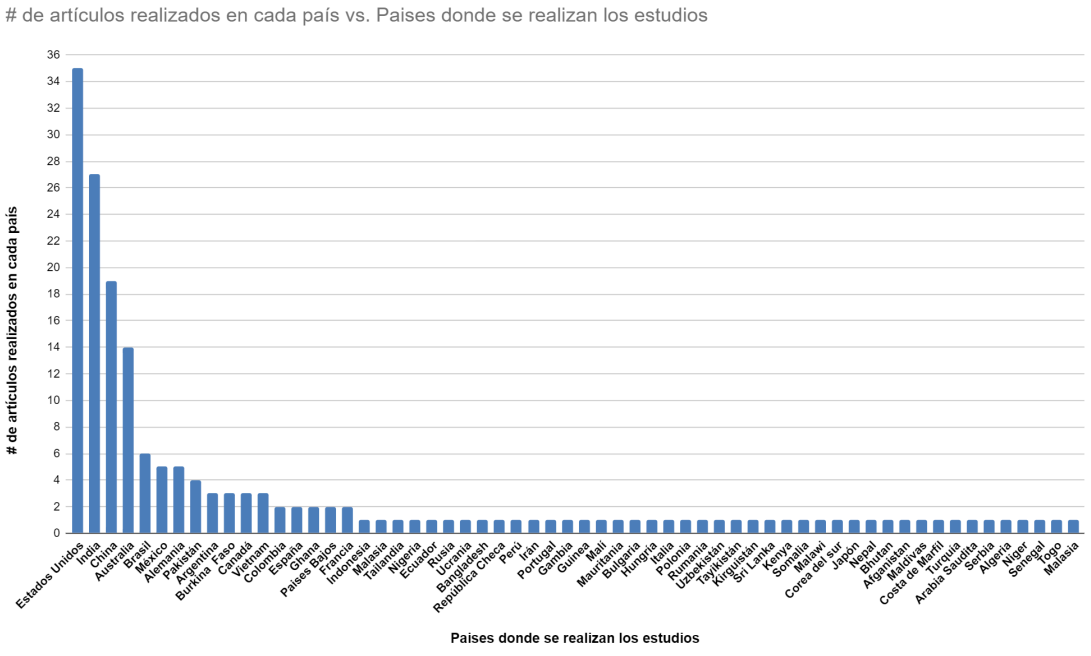

Figura 2. Países donde se realizan los estudios. Fuente propia.

En la **Figura 2** se puede observar que los países donde más estudios se ha realizado sobre producción y rendimiento de cultivos son: Estados Unidos, India y China; cabe destacar que Australia, a pesar de ser un país con menor población en comparación con los tres países mencionados y que se caracteriza por sus ecosistemas desérticos, tiene una gran cantidad de estudios realizados.

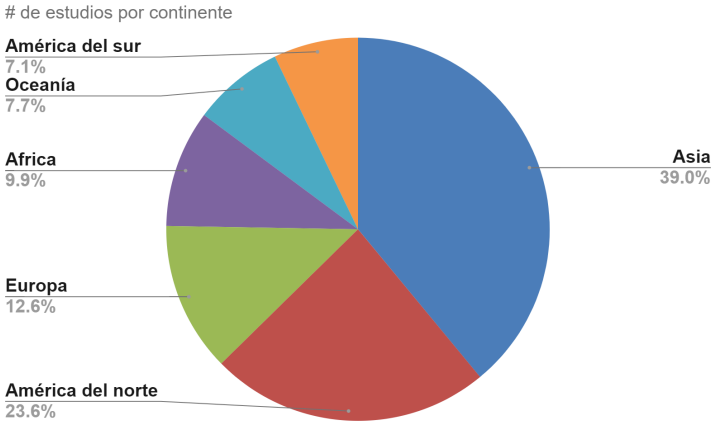

Figura 3. Estudios realizados en cada continente. Fuente propia

También, se realizó un conteo sobre los continentes en donde se realizan los estudios, encontrando que Asia con el 39% de los estudios, es el continente donde más se han realizado estas investigaciones; y Sudamérica quedando atrás con únicamente el 7.1% de los estudios realizados, donde los países que más se destacan son Brasil y Argentina como se muestra en la **Figura 3.** En Colombia han sido realizados 2 estudios sobre la producción y/o rendimiento de cultivos.

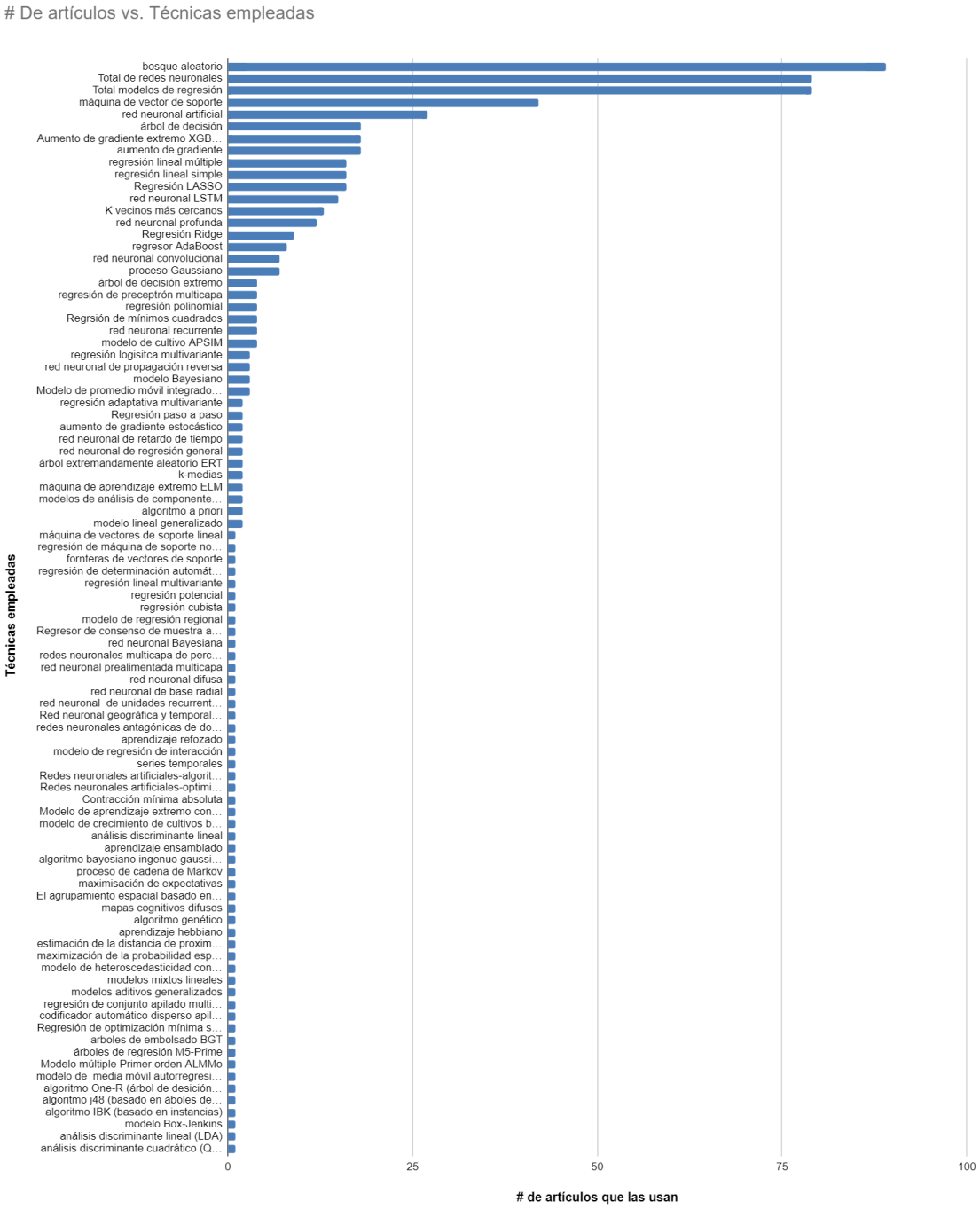

Figura 4. Técnicas empleadas para la estimación. Fuente propia.

La **Figura 4** muestra las técnicas más empleadas para la estimación de producción o rendimiento, se encontró que en su mayoría emplean bosques aleatorios, el cual es utilizado en 89 estudios. Otras técnicas que destacan son las máquinas de vectores de soporte y las redes neuronales. Un aspecto importante a tener en cuenta, es que hay bastantes variaciones de técnicas, sobre todo en redes neuronales o modelos de regresión, por lo cual se añadió manualmente una variable que cuenta el total de estas dos técnicas, encontrando que en ambos casos se emplean en 79 estudios.

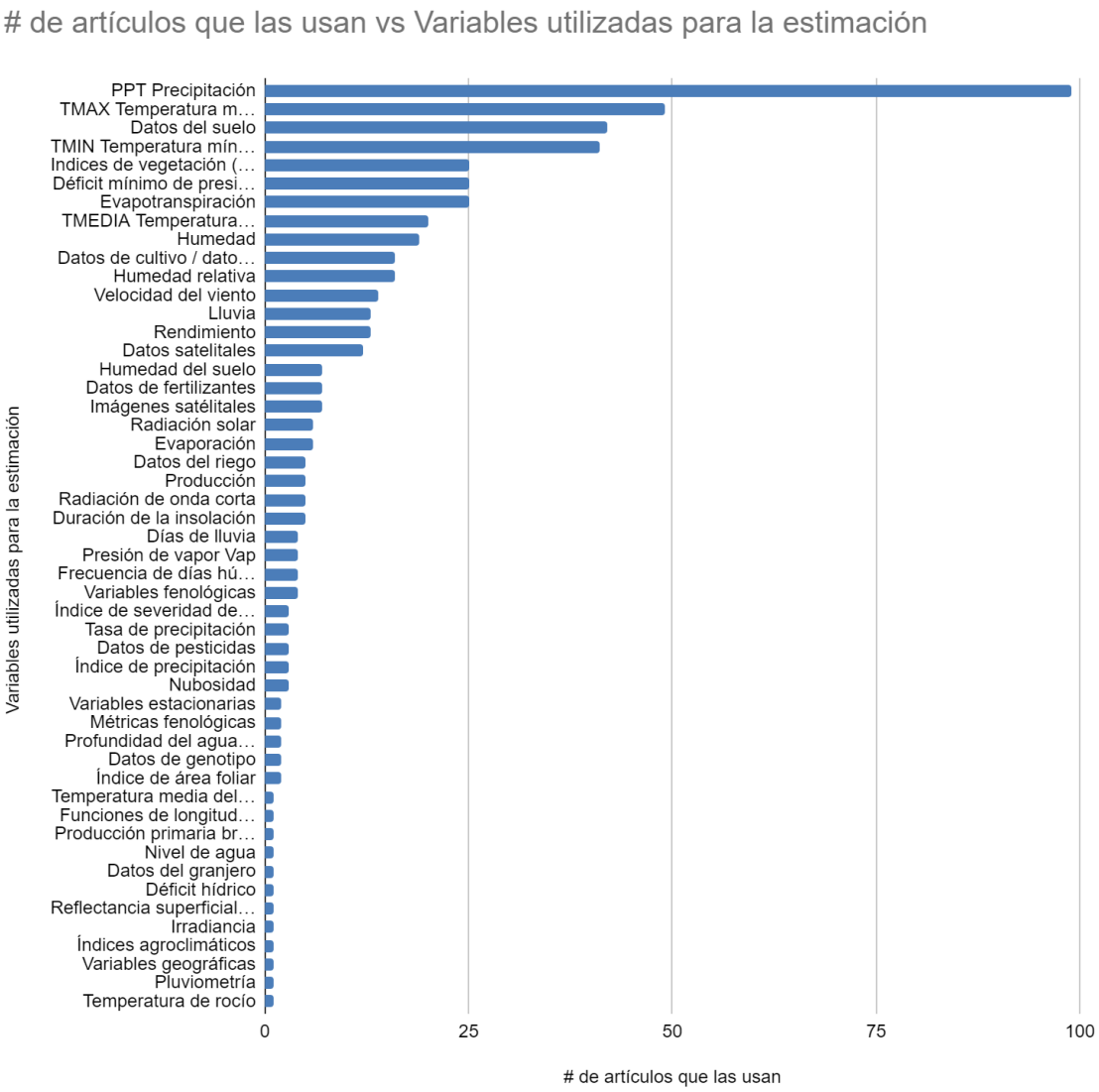

Figura 5. Variables meteorológicas utilizadas para la estimación

En la revisión sistemática se observa que la variable más usada para realizar la estimación de rendimiento en todos los estudios es la precipitación, siendo usada en el doble de estudios en comparación a las otras variables relevantes como: temperatura, datos del suelo, índices de vegetación, déficit mínimo de presión de vapor, evapotranspiración, humedad, datos de cultivo, velocidad del viento, esta información se ve reflejada en la **Figura 5.**

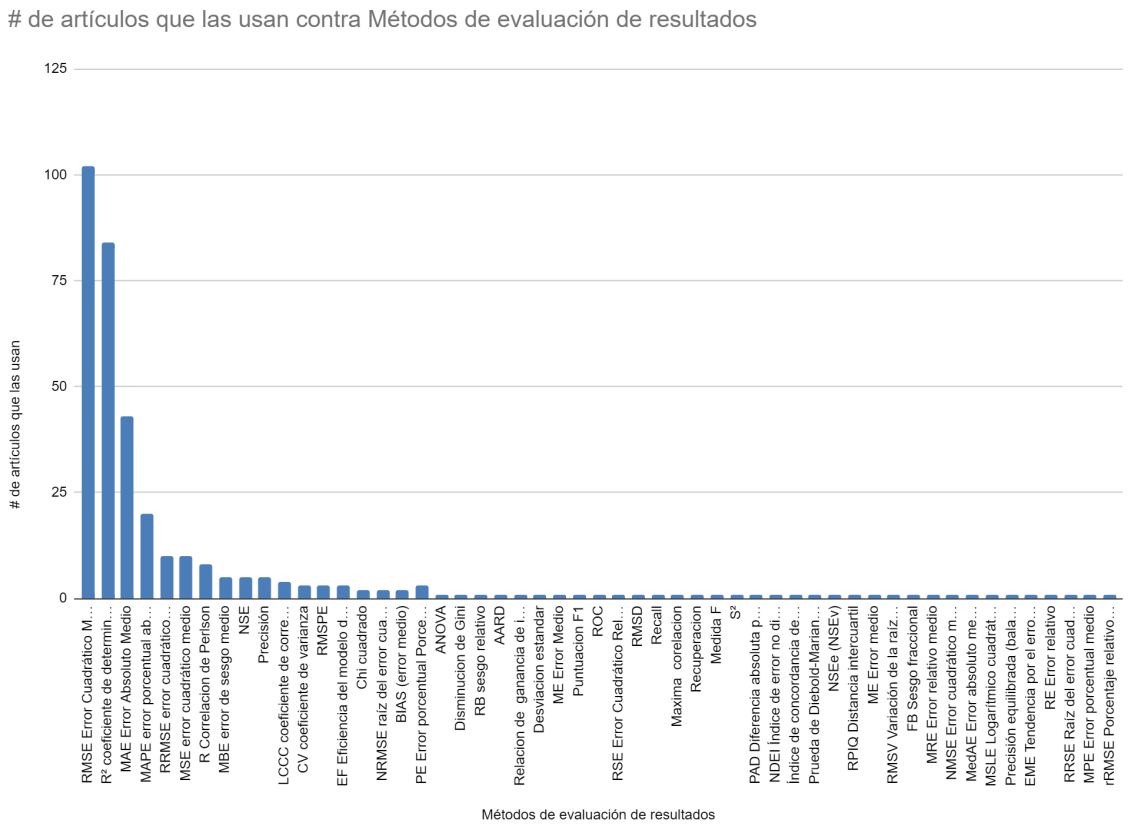

Figura 6. Métodos de evaluación de resultados

En cuanto a los métodos más usados para realizar la evaluación de resultados de las diferentes técnicas, como se muestra en la **Figura 6,** se encuentran el error cuadrático medio RMSE y el coeficiente de determinación R^2, siendo usados en el 66% y 54% de los estudios respectivamente. Otros métodos relevantes son: error absoluto medio MAE y error porcentual absoluto medio MAPE.

2.1.6.2. Análisis cualitativo de los datos

Para realizar el análisis cualitativo se tendrán en cuenta los artículos encontrados en la revisión sistemática que superen el puntaje de corte de 6 puntos establecido en la evaluación de la calidad. También se tendrán en cuenta aquellos artículos cuyo motivo de estudio sean los datos con más relevancia encontrados en el análisis cualitativo de los datos.

Cultivos estudiados y países donde se realizan los estudios

Los cuatro cultivos más utilizados en las investigaciones para estimar la producción son el maíz, trigo, arroz y soya. Los cuatro países más destacados de la investigación son Estados Unidos, India, China y Australia, son los grandes

productores de estos cultivos en el mundo, por lo que es importante tener una comprensión detallada de su producción y la investigación en torno a ellos.

- **Maíz:** en Estados Unidos, el maíz es uno de los cultivos más importantes, en el cinturón de maíz de EU se produce alrededor de 30% de la producción mundial de maíz y juega el papel más crítico en la exportación del cultivo. La investigación en torno a la producción de maíz en Estados Unidos se ha centrado en mejorar la eficiencia de producción, reducir los costos, pronosticar el suministro de alimentos y aumentar la rentabilidad para los agricultores [10].

- **Trigo:** China es uno de los países productores de trigo más grandes del mundo y representa alrededor del 18 % de la producción mundial de trigo, también India es el segundo productor de trigo más grande en el mundo. Por lo cual la investigación en torno a la producción es relevante en ambos países para comprender los factores que la afectan y de esta manera hacer uso de los avances en ciencia y tecnología agrícola, con el objetivo de obtener predicciones precisas y oportunas del rendimiento de los cultivos con datos ambientales para garantizar la seguridad alimentaria [11].

- **Arroz:** en China, el arroz es un cereal básico y su demanda está aumentando sustancialmente con el crecimiento de la población mundial. Predecir con precisión los rendimientos del arroz es de vital importancia para garantizar la seguridad alimentaria en los países como China, donde el arroz representa una quinta parte de la producción agrícola total [12].

- **Soya:** Estados Unidos es el mayor exportador de cereales del mundo, la soya es uno de los cultivos más relevantes con un 34,9% de la producción mundial y es altamente utilizado para estimar la producción agrícola, así como el maíz. Según el USDA, la soya es el segundo cultivo más ampliamente cultivado en el país. La investigación en torno a la producción de soya en Estados Unidos se ha centrado en mejorar la eficiencia, la gestión agrícola, las políticas alimentarias nacionales, los precios mundiales y el comercio internacional de cultivos [13].

Técnicas empleadas para la estimación

En cuanto a las técnicas empleadas, el bosque aleatorio, las redes neuronales artificiales y las máquinas de vectores de soporte destacan sobre el resto.

- **Bosque aleatorio:** en términos generales, se encontró que esta técnica fue empleada en una gran cantidad de estudios debido a que tiene una alta exactitud y precisión, facilidad de uso y utilidad en el análisis de datos. También tiene capacidad para manejar alta dimensionalidad de datos, detección de valores atípicos, robustez contra el sobreajuste y análisis de relevancia de variables que permite seleccionar las mejores y generar

modelos más eficientes (menos ajuste de parámetros, cálculo más rápido y más transparente) [14]. El alto rendimiento del bosque aleatorio es más evidente cuando la respuesta es el resultado de interacciones complejas entre múltiples predictores, como en los sistemas de cultivo [15].

Entre las ventajas que tiene el bosque aleatorio en comparación con otras técnicas es que puede usar múltiples tipos de predictores en un modelo más fácilmente que las regresiones lineales o no lineales múltiples tradicionales [15]. También un bosque aleatorio es capaz de aprender relaciones complejas y no lineales entre el rendimiento y los predictores a partir de datos sin necesidad de modelarlos explícitamente [10].

Una de las desventajas encontradas es que el comportamiento del modelo puede ser menos intuitivo de interpretar que los modelos de regresión tradicionales porque su algoritmo consiste en un conjunto de una gran cantidad de árboles de decisión que pueden no estar completamente descritos mecánicamente [15]. Otra de las desventajas que tiene la técnica de bosque es que su precisión puede disminuir en el caso en que los datos de entrenamiento sean escasos [15].

- **Red neuronal artificial:** se utilizaron distintos tipos de redes neuronales, en algunos estudios se realizaron modelos híbridos que combinan otros tipos de modelos con redes neuronales, mostrando muy buenos resultados.

  Las redes neuronales simples con estimaciones puntuales generalmente requieren una gran cantidad de muestras para el entrenamiento del modelo y están sujetas a sobreajuste en pequeños conjuntos de datos de entrenamiento (el error de entrenamiento es casi igual a 0, pero se muestran algunas fluctuaciones en la validación y los testeos) [16].

  Entre los distintos tipos de redes empleadas destacan las redes neuronales de aprendizaje profundo (Deep Neural Networks) que, al tener múltiples capas ocultas, son más poderosas para revelar la relación no lineal fundamental entre los predictores y los rendimientos. Además, las redes neuronales profundas permiten integrar datos de múltiples fuentes y realizar buenas predicciones a partir de estos [17]. Este tipo de redes son funciones de aproximación universal, lo que significa que pueden aproximarse a casi cualquier función [18], lo cual está en línea con la relación no lineal entre el clima y el rendimiento de los cultivos [19].

  Las desventajas de este tipo de redes consisten en su misma complejidad ya que puede ser muy difícil encontrar la estructura de la red y los parámetros correctos para el entrenamiento, además requieren hardware más avanzado y técnicas de optimización para poder entrenarlas [18].

Los modelos de redes neuronales profundas conducen a predicciones más precisas que otros modelos de aprendizaje automático interpretables, como la regresión lineal. Sin embargo, la naturaleza de caja negra de los modelos de aprendizaje profundo no revela las razones (características) que son importantes para predecir los valores de rendimiento de los cultivos y si estas características están disminuyendo o aumentando los resultados de rendimiento [20].

- **Máquinas de vectores de soporte:** las máquinas de vectores de soporte SVM se caracterizan por su alta flexibilidad para capturar las relaciones no lineales y, además de hacerlo de una manera rápida y robusta [21]. A diferencia de otros modelos de aprendizaje automático que intentan minimizar el error entre los datos observados y simulados, SVM tiene como objetivo encontrar la mejor línea dentro de los valores de umbral [11].

  Las máquinas de vectores de soporte poseen una alta capacidad de generalización lo cual lleva a obtener resultados muy sólidos, además de poder manejar datos de alta dimensionalidad sin necesidad de una selección de características [12].

  Los modelos de regresión de máquinas de vectores de soporte tienen un rendimiento comparable y una mayor precisión de predicción que los modelos lineales debido a que capturan la estructura no lineal de las variables, aunque existe un ligero problema de sobreajuste en los resultados de predicción de algunos modelos de SVM [20]. Las máquinas de vectores de soporte pueden encontrar estas relaciones no lineales entre varias variables ya que estas pueden ser lineales en una proyección de alta dimensión del conjunto de datos donde pueden evaluarse fácilmente mediante métodos lineales [21].

  En términos generales, las SVM muestran un rendimiento más bajo que otros modelos más complejos como el bosque aleatorio o las redes neuronales, sin embargo, en términos de interpretabilidad es superior debido a que con este método es más sencillo entender las relaciones entre las variables de entrada y la salida

Variables utilizadas para la estimación

En el análisis cuantitativo se encontró que las variables meteorológicas más empleadas en los estudios sobre estimación de producción y rendimiento de cultivos son la precipitación, la temperatura (máxima y mínima) y el déficit mínimo de presión de vapor VPD, siendo la precipitación la más empleada con gran diferencia.

- **Precipitación:** además de que la precipitación es la variable que más artículos utilizan, también es la variable que mayor impacto o relevancia tiene sobre la producción y rendimiento de los cultivos en la mayoría de estudios,

esto se puede observar en aquellas investigaciones que realizaron un análisis de la importancia y peso de las variables, ya sea mediante análisis de correlación o mediante algoritmos, como por ejemplo el bosque aleatorio. Está relevancia de la precipitación no sólo se ve en el análisis estadístico, sino que ya es un conocimiento establecido de la agricultura, puesto que se considera ampliamente que la variabilidad de las precipitaciones es la causa directa de la fluctuación del rendimiento de los cultivos en entornos semiáridos de todo el mundo [22].

La precipitación ejerce una influencia en el rendimiento de los cultivos principalmente a través de la alteración de las condiciones de humedad del suelo. De acuerdo a [23], una disminución del 20% en la precipitación podría causar una pérdida de rendimiento de maíz del 8,10%, cuya magnitud es mucho mayor que la ganancia de rendimiento del 5,63 % por un aumento del 20% en la precipitación.

Los efectos de la precipitación varían respecto a los cultivos y a sus etapas fenológicas, algunos cultivos se benefician de la lluvia en etapas como la floración, mientras que a otros los puede perjudicar en otras etapas, por ejemplo, el maíz es más sensible a la sequía durante la etapa reproductiva que la vegetativa, mientras que el exceso de lluvia causaría más pérdidas en el rendimiento del maíz durante la etapa vegetativa que la reproductiva [23]; y en [24] las temperaturas extremas durante la fase de floración y las precipitaciones extremas durante las ventanas de siembra y cosecha influyeron significativamente en el rendimiento de los cultivos.

- **Temperatura:** en una gran cantidad de estudios, no solo se emplea una variable de temperatura, sino que la caracterizan como temperatura máxima, temperatura mínima y temperatura media; esto debido a que los picos de temperatura (principalmente de temperaturas altas) pueden generar importantes efectos adversos sobre la producción y rendimiento de los cultivos.

  En general, las altas temperaturas a menudo conducen a una pérdida de rendimiento al acortar la fase reproductiva, aumentar la senescencia de las hojas y provocar el cierre de los estomas [23]. Además de los efectos directos sobre la fisiología y la fotosíntesis de las plantas, una temperatura alta podría aumentar la demanda de agua y disminuir el suministro de agua del suelo a través de una mayor evapotranspiración, lo que en conjunto conduce a un estrés hídrico elevado para el crecimiento de los cultivos con impactos negativos en el rendimiento [23].

  Al igual que la precipitación, la variación de la temperatura también genera efectos dependiendo la fase fenológica de cada cultivo que terminan afectando la producción y rendimiento del mismo, por ejemplo, durante el período de llenado del grano de maíz, las temperaturas cálidas por encima

del umbral superior provocan una reducción del rendimiento. Las estimaciones del modelo sugieren que, por cada 1°C de aumento en la temperatura, hay una reducción del rendimiento de casi el 10% [24]. También, la alta temperatura (es decir, “estrés por calor”) que ocurre durante la etapa reproductiva del maíz conduce a una reducción en el número de semillas, y la alta temperatura durante la etapa de llenado del grano conduce a un menor peso de la semilla [10].

En el caso de otro grano como el trigo de invierno, es conocido que una temperatura del aire más alta acorta la duración del llenado del grano, reduce la ganancia fotosintética y da como resultado un rendimiento de grano menor [25].

- **Déficit mínimo de presión de vapor VPD:** el déficit de presión de vapor es un indicador de los efectos de las altas demandas atmosféricas de agua, lo que también tiene efectos sobre las plantas, y en consecuencia sobre su producción, por ejemplo, un alto VPD que indicada sequedad atmosférica, aumenta la pérdida de agua de las plantas o el suelo con respecto a la atmósfera. Las plantas responden a un alto VPD cerrando sus estomas para evitar una pérdida de agua más rápida con la consecuencia de una reducción de la tasa de fotosíntesis. Un alto VPD también puede causar un agotamiento más rápido del almacenamiento de humedad del suelo, lo que puede resultar en un mayor déficit de humedad del suelo al final de la temporada [10].

  De acuerdo a [10] el VPD y la temperatura están altamente correlacionadas, por lo cual es difícil diferenciar los efectos que cada variable genera independientemente en la producción del cultivo, sin embargo, se logró concluir que las pérdidas de rendimiento de cultivos se deben principalmente a un alto VPD. Esto nos puede llevar a concluir que el VPD es una variable igual de relevante para el rendimiento y la producción de los cultivos que las dos revisadas en esta sección del análisis cualitativo de la revision sistemática.

Métodos de evaluación de resultados

En cuanto a los métodos para realizar la evaluación de los resultados, el RMSE error cuadrático medio y el R^2 coeficiente de determinación, se destacan sobre el resto ya que permiten evaluar la precisión y la calidad de los modelos predictivos.

- **RMSE error cuadrático medio:** a nivel general este método fue empleado en una gran cantidad de estudios debido a que se utiliza para determinar la precisión de un modelo en la predicción de valores continuos [24]. Cuanto menor sea el valor del RMSE, mejor será el modelo en la predicción de los valores, se destaca por ser una medida de la diferencia entre los valores reales y los valores predichos por el modelo [26]. Usualmente se utiliza en

algoritmos de regresión, como bosque aleatorio y máquinas de vectores de soporte, los cuales fueron los más usados en las investigaciones.

- **R^2 coeficiente de determinación:** este método también fue bastante empleado. El R2 es una medida de la proporción de la variación en los datos, determina qué tan bien se ajustan los datos a la línea de regresión. Varía entre 0 y 1, donde 1 indica una buena correlación entre los datos y el modelo [26]. Es usado de igual manera, en algoritmos de regresión y clasificación, por está razón se usó en un gran porcentaje con los algoritmos más utilizados de las investigaciones.

- **MAE error absoluto medio:** esta métrica de evaluación define la diferencia absoluta entre el valor estimado por el modelo con el valor real, es una medida de errores entre observaciones pareadas que expresan el mismo fenómeno [26].

2.1.7. Artículos Adicionales

Debido a que en la búsqueda avanzada con la cadena de texto realizada en las bases de datos no se encontraron resultados sobre investigaciones realizadas en el cultivo de aguacate, se optó por incluirlas manualmente para tomar como referencia estudios previos sobre este cultivo. Se realizó una búsqueda manual en distintas bases de datos y páginas web de estudios que realizaran una estimación de producción o rendimiento del cultivo de aguacate, encontrando los siguientes resultados:

Automated Avocado Yield Forecasting Using Multi-Modal Imaging [27]

En esta investigación se realiza un modelo para predecir el rendimiento de aguacate en el momento de la cosecha, después de haber hecho un conteo previo del número de aguacates en los primeros meses del mismo año. Para realizar el conteo se realiza un clasificador de aguacates empleando visión de máquina a partir de imágenes térmicas y RGB.
Se utilizó un modelo de predicción Bavendorf (este modelo identifica un pequeño parche en la superficie de un árbol de aguacate y cuenta la cantidad de frutos que se ven dentro de ese parche), para que, a partir del conteo de aguacates con la visión de máquina, se estime el posible rendimiento que pueda generar cada planta (estimar la cantidad y peso de los aguacates que no son visibles en las imágenes debido al follaje de los árboles).

Fruit weight and yield estimation models for five avocado cultivars in Ethiopia [28]

En este estudio se realizó una recolección manual previa de 360 frutos aleatorios (12 frutos de 30 árboles diferentes) y para cada uno se midió la longitud, diámetro, peso

del fruto y carga de frutos (número de frutos por árbol). Luego, se construyeron 16 modelos utilizando ecuaciones de regresión lineal y no lineal basadas en el diámetro y longitud del fruto (variables independientes) para predecir el peso del fruto (variable dependiente).
El desempeño de los modelos se evaluó con métricas estadísticas como el coeficiente de determinación R2, sesgo absoluto promedio MAB, sesgo porcentual PBIAS, raíz del error cuadrático medio RMSE, etc.
La estimación del rendimiento se encuentra al multiplicar el peso estimado del fruto promedio por la carga de frutos. De los 16 modelos, el que arrojó mejores resultados fue una regresión lineal múltiple usando las dos variables independientes.

Potential geography and productivity of "Hass" avocado crops in Colombia estimated by ecological niche modeling [29]

El modelado de nicho ecológico (ENM) es un esfuerzo para estimar los requisitos ambientales de las especies en función de las asociaciones entre las ocurrencias geográficas conocidas y las condiciones ambientales en esos sitios, para permitir la estimación de la distribución potencial de la especie.
En este estudió se empleó el ENM para probar si las áreas de producción de aguacate actuales son las más adecuadas y evaluar las relaciones entre la producción de aguacate Hass y las dimensiones ambientales. Todos los modelos de regresión realizados mostraron relaciones positivas entre la adecuación de las áreas y su producción, lo cual hizo posible que los modelos ENM tuvieran un buen potencial predictivo en términos de la producción.
Los datos recolectados para la creación de este modelo consisten en variables del ambiente como la elevación, la pendiente, índices de suelos y el índice vegetativo NDVI tomado desde imágenes satelitales.

The Effect of Minimum Temperature on Avocado Yields [30]

En este artículo no se realiza una estimación del rendimiento como tal, en su lugar, se busca la relación que existe entre el rendimiento y las bajas temperaturas, con el fin de encontrar los efectos que generan estas temperaturas sobre el cultivo. Para cada combinación de temperatura y periodo del día se obtuvo la correlación, el coeficiente de determinación R2 y la significación estadística. Los resultados muestran una correlación de 0.97 y la posibilidad de explicar los resultados de un alto 93% (R2).

Using WorldView Satellite Imagery to Map Yield in Avocado (Persea americana): A Case Study in Bundaberg, Australia [31]

En este estudio se propone usar imágenes satelitales de alta resolución para realizar una estimación precisa precosecha del rendimiento de aguacate. Se emplearon diferentes índices de vegetación basados en el NDVI (normalised difference vegetation index) que están altamente correlacionados con las variables a estudiar (tamaño del fruto promedio y peso total de frutos). Se realiza un análisis de

componentes principales PCA para determinar cuál de estos índices de vegetación tienen mayor relación con las variables medidas.
Se dividió la zona de estudio en distintos bloques y el rendimiento promedio de cada bloque se calculó sustituyendo el valor de reflectancia de píxel promedio de la imagen satelital de cada bloque en el algoritmo de regresión no lineal de rendimiento del bloque correspondiente.
Para evaluar los modelos con los distintos índices vegetativos, se empleó el coeficiente de determinación R2 y la raíz del error cuadrático medio RMSE, el modelo de regresión con mayor R2 y menor RMSE es el seleccionado como modelo óptimo.

Analítica de datos para el rendimiento en los cultivos de aguacate Hass en Colombia [32]

Esta investigación es la más similar a la que se propone en este documento, puesto que también se utilizan variables meteorológicas para predecir el rendimiento del aguacate en la zona colombiana. En este estudio se emplean como variables independientes la humedad relativa, temperatura, precipitación y radiación solar. Para la predicción se utilizan las técnicas de aprendizaje automático: Regresión logística multi clase y el Bosque aleatorio multiclase. Un aspecto a tener en cuenta de esta investigación es que los modelos empleados no son regresivos, en su lugar son modelos de clasificación que distribuyen el rendimiento en 3 categorías: alto, medio y bajo.
Para determinar la relevancia de las variables con respecto al rendimiento se utiliza la medida chi-cuadrado, y para la evaluación de los modelos de aprendizaje automático se utiliza una matriz de confusión y además se mide la exactitud, precisión y exhaustividad (accuracy, precision y recall) para cada modelo.

2.2. Aportes y Brechas

En la **Tabla 1** se exponen las brechas encontradas en los artículos seleccionados en la revisión sistemática de literatura.

Sección - Trabajo		Brechas
Artículos incluidos en la búsqueda avanzada		- La principal brecha encontrada para estos artículos es que la investigación no es realizada sobre el cultivo de aguacate. - Únicamente dos artículos encontrados fueron realizados en Colombia, lo cual permite un amplio margen de investigación en esta área para la región.
Artículos adicionales (realizados con la	[27]	- La estimación no se realiza empleando técnicas de aprendizaje automático

búsqueda manual)		supervisado. - Las variables meteorológicas no se tienen en cuenta para realizar la estimación
	[28]	- No se consideran las variables meteorológicas para la estimación del rendimiento.
	[29]	- Las variables predictoras se basan en las condiciones del terreno y los índices vegetativos, no se tienen en cuenta las variables meteorológicas.
	[30]	- La estimación que se realiza es sobre la relación entre el rendimiento y cómo las temperaturas bajas influyen en el cultivo, no se realiza una estimación del rendimiento como tal.
	[31]	- Las variables meteorológicas no se tienen en cuenta para realizar la estimación, utilizan índices de vegetación obtenidos por imágenes satelitales.
	[32]	- Se tienen en cuenta las variables meteorológicas para el cálculo de la estimación de producción, pero no se considera el comportamiento fenológico del cultivo sobre la producción. - La estimación del rendimiento se realiza empleando técnicas de aprendizaje automático de clasificación, en lugar de implementar modelos de regresión.

Tabla 1. Brechas encontradas. Fuente Propia

La investigación expuesta en este documento ayuda a fortalecer el área de investigación tecnológica, al ser el aprendizaje automático una de las tecnologías que mejores resultados arroja y que más se está empleando en distintas áreas del conocimiento. También se realiza un aporte al área de la agricultura del país, ya que puede ser una importante investigación para brindar información a los agricultores sobre el estado de sus cultivos y cómo estos pueden evolucionar de acuerdo las condiciones meteorológicas de su región; a partir de esto se pueden tomar las decisiones adecuadas en pro del desarrollo del cultivo.

2.3. Conclusiones de la revisión sistemática de literatura

- Un alto porcentaje de estudios realizados sobre la estimación de la producción y rendimiento de cultivos se han hecho sobre cultivos transitorios, siendo el maíz el

cultivo más estudiado. También se encontró que hay muy pocas investigaciones sobre el rendimiento y producción del cultivo de aguacate.

- Los países con mayor cantidad de estudios realizados son los grandes productores de granos como Estados Unidos, seguido de países asiáticos cómo China e India.

- Dentro de los estudios encontrados se evidenció una escasez de investigaciones que utilizaran técnicas de aprendizaje automático para realizar la estimación de producción o rendimiento en el cultivo de aguacate, asimismo, se observó una carencia de estudios que utilizaran estas técnicas a nivel nacional para estimar la producción de otros cultivos.

- La técnica más empleada para estimar la producción de cultivos es el Bosque Aleatorio, debido a su facilidad de implementación, buenos resultados y su interpretabilidad.

- Los distintos tipos de redes neuronales suelen mostrar mejores resultados, sin embargo, estas requieren una mayor cantidad de datos para ser entrenadas y tienen menor interpretabilidad, mayor complejidad de diseño y requieren más tiempo y potencia de hardware para ser entrenadas.

- Entre las variables empleadas para la estimación destacan la precipitación y la temperatura, debido a que estas variables meteorológicas afectan en gran medida la fenología de los cultivos. Además de manera estadística, los estudios que realizaron análisis de variables, demostraron la alta relevancia de estas variables para la predicción de la producción y rendimiento.

3. Conjunto de datos meteorológicos y de producción de aguacate en Colombia

Para realizar el proceso de minería de datos, desde la selección de los conjuntos de datos, su limpieza y la implementación y evaluación del modelo, se tomó como referencia la metodología CRISP-DM, la cual consiste en una serie de pasos que se muestra en la **Figura 7**:

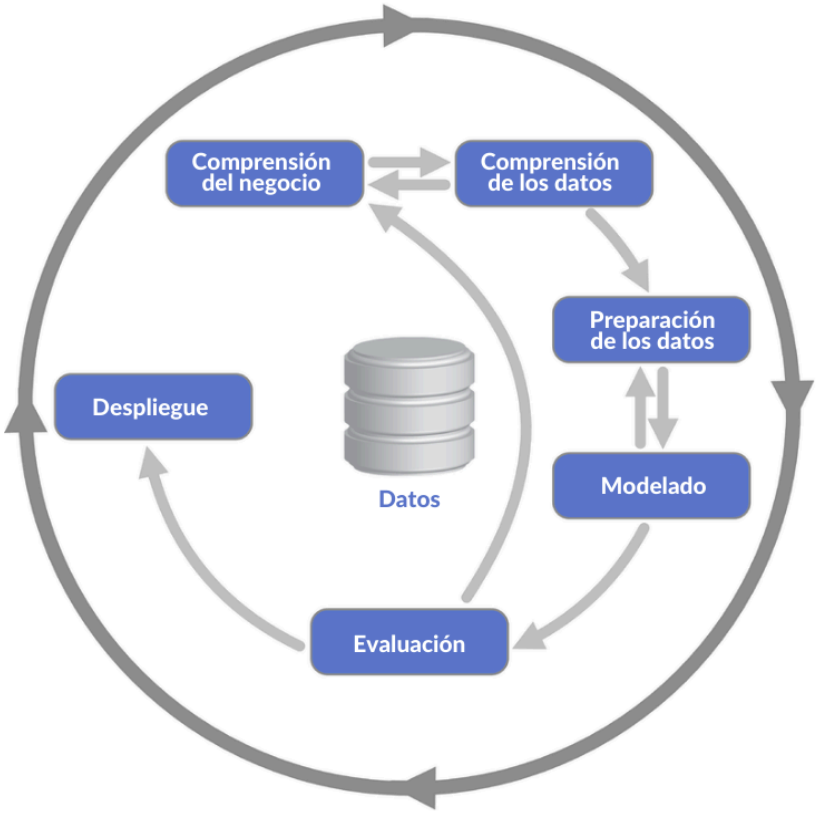

Figura 7. Metodología CRISP-DM. Adaptada de [33]

3.1. Comprensión del negocio

La investigación se centra en predecir la producción de aguacate a partir de un conjunto de datos de variables meteorológicas, por ello es importante entender cómo estas variables influyen en el cultivo. Se realizó una investigación manual, consultando en distintos artículos científicos y páginas web relacionadas, donde se recolectó información sobre las principales variables meteorológicas, su incidencia en el cultivo de aguacate en las diferentes fases de desarrollo de la planta y los

rangos de valores ideales de las variables meteorológicas para mantener un buen estado de la planta.

3.1.1. Fases fenológicas del aguacate

Un estudio fenológico de un cultivo consiste en investigar y analizar cómo los factores meteorológicos afectan a la planta en sus distintos ciclos biológicos mostrados en la **Figura 8**. A continuación, se muestra información relevante que se ha recolectado sobre las fases fenológicas del aguacate relacionadas con la producción (floración y desarrollo del fruto):

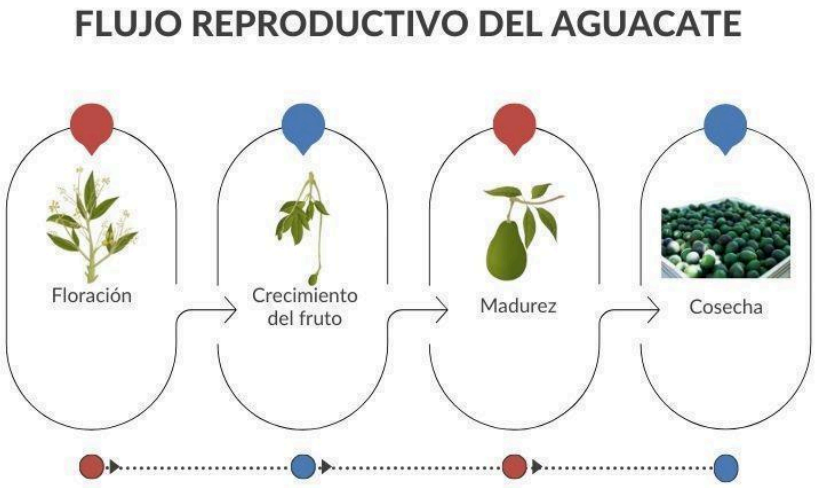

Figura 8. Fases fenológicas del aguacate. Fuente Propia

- **Floración**: los principales factores que influencian la transición a la floración son: fotoperíodo, temperatura y disponibilidad de agua. La temperatura es el factor que bajo condiciones tropicales mayor influencia tiene en la floración. Cuando se presentan períodos con temperatura por debajo de 6°C, se da el estímulo para que el árbol pase del estado vegetativo al estado reproductivo [34]. En el caso de Colombia la floración se da cuando las temperaturas bajan de 20° en el día y 10° en la noche [35].

  El período entre la floración y la maduración fisiológica es característico de cada cultivar. En la variedad Antillana este período dura de 5 a 8 meses, en la Guatemalteca 10 a 15 meses y en los Mexicanos 6 a 8 meses [5].

- **Cuajado de frutos:** el período entre la polinización y la cosecha del fruto depende de la variedad, del cultivo y del clima donde se sitúe el huerto y puede oscilar entre 27 a 60 semanas. Para cultivar fuerte, este tiempo toma de ocho a diez meses; para Hass, en países subtropicales toma de 10 a 12 meses y en Colombia 9 meses [34].

3.1.2. Condiciones meteorológicas adecuadas para el aguacate Hass

La meteorología es un un factor fundamental en el desarrollo de la planta, del fruto y de la producción de la misma puesto que el clima determina las características físicas y químicas de los frutos de aguacate, siendo la temperatura un factor determinante en el tamaño, peso, forma y rugosidad del fruto, así como en el tamaño de la semilla [36]. Por lo tanto, se realizó una investigación sobre la influencia de las principales variables meteorológicas en el cultivo de aguacate.

- **Temperatura:** las temperaturas ideales para el cultivo deben estar entre 18°C y 22°C. En fases previas a la floración las condiciones de temperatura ideales son de 18 y 15 °C en el día y la noche, respectivamente, lo que ocasiona una reducción en el crecimiento de la planta, pero induce la floración [37]. En cambio, durante la floración las temperaturas óptimas deben rondar 20°C en el día y 10°C en la noche [35], esto debido a que durante las fases de floración y cuajado del fruto, la presencia de heladas y bajos promedios de temperatura, son los factores más limitantes [38].

  En la variedad Hass las bajas temperaturas inciden directamente en la duración del periodo de flor a fruto, el cual se alarga a medida que la temperatura disminuye. En zonas frías este periodo dura hasta 10 – 14 meses, mientras que en las zonas cálidas únicamente de 5 a 8 meses [5]. Cabe resaltar que en ciertos lugares del trópico, en los que las bajas temperaturas son breves o irregulares, el estrés hídrico es el factor que más incide [39].

- **Humedad:** la humedad relativa debe estar entre el 75% y el 85%, esto favorece la germinación de polen y el desarrollo del fruto [35].

- **Elevación**: la altitud ideal para el cultivo de aguacate Hass es entre 1600 y 2100 msnm en las condiciones de trópico. Se debe tener en cuenta que a altitudes elevadas hay baja tasa de fecundación de flores debido a la alta presencia de nubes, provocando menos horas de luz solar, lo que también hace que el periodo de crecimiento sea más lento y la época de floración sea más tardía [35].

- **Precipitación:** la pluviosidad ideal consiste en precipitaciones entre 1200-1800 mm de agua al año, con periodos de sequía menores a 20 días. Para la mejor expresión de desarrollo y producción, el cultivo debe recibir 3 mm de agua al día. Sequías largas provocan la caída de hojas produciendo frutos de menor calibre; y el exceso de precipitación durante la floración puede generar la caída de frutos y afectar la sanidad de la planta, favoreciendo la proliferación de hongos [35].

- **Radiación solar:** los valores de luz solar ideal para cultivar aguacate deben ser entre 1500 y 1800 horas de luz efectiva al año, es decir 4.1 a 4.9 de luz

diaria. Aquellos árboles que están en la sombra (cerca de árboles altos) manifiestan crecimiento excesivo en la longitud del tallo y ramas [35].

- **Velocidad de viento:** no debe haber vientos constantes ni que alcancen velocidades superiores a 20 km/h, ya que tallos y ramas se pueden quebrar, y los frutos se pueden ver afectados. Además, cuando el viento es muy seco durante la floración, reduce el número de flores polinizadas, y por lo tanto también de frutos [35].

3.1.3. Estudios de campo de la fenología del cultivo de aguacate.

La investigación realizada evidenció que la fenología es un factor importante para la producción del cultivo, la cual varía de acuerdo a cada región de estudio, dado que el clima es diferente en cada zona. Por lo tanto, es importante realizar una búsqueda de investigaciones que lleven a cabo un estudio de campo sobre la fenología del aguacate en Colombia, para entender el comportamiento del cultivo en esta región. Se hace un énfasis en las dos fases que más impacto tienen en la producción final del cultivo, la floración y el desarrollo del fruto.

Se encontraron casos de estudios sobre la fenología del aguacate en distintos municipios del país, estos estudios permiten analizar cómo el clima afecta los distintos aspectos del cultivo a lo largo del ciclo productivo del mismo. Por tal motivo, es importante encontrar los patrones y comportamientos cíclicos que permitan entender en términos generales, el comportamiento fenológico del cultivo en la región de estudio.

Cartilla de estados fenológicos-tipo en aguacate Hass para la localidad de Roldanillo, Valle del Cauca [40]

Se realizó un estudio sobre los estados fenológicos del aguacate Hass en Roldanillo, Valle del Cauca. Se encontró que la fase de floración, aunque inició la primera semana de julio, alcanzó su máxima frecuencia (80% de Intensidad Relativa) en la tercera semana de septiembre, que correspondió a la semana 25. Esta etapa de mayor floración, al igual que el estado de yemas en brotación, coincidió con un balance hídrico negativo, pues se presentó baja precipitación y valores altos de evapotranspiración.

Ecofisiología del aguacate cv. Hass en el trópico andino colombiano [41]

Se realizó el estudio para 8 municipios de Antioquia donde se pudo observar que en todos los huertos, la floración de mayor intensidad se observa en el primer trimestre del año. Se observó que existe una influencia climática sobre el comportamiento fenológico del aguacate cv. Hass. Ocho localidades presentaron dos flujos florales por año de distinta intensidad, pero en todos los casos se manifestaron entre enero y

abril y entre julio y septiembre, en las épocas de menor precipitación, lo cual evidenció un comportamiento cíclico en este estado fenológico. Finalmente, se pudo establecer que las precipitaciones influyen en los periodos de crecimiento del árbol, ya que en la época de menor precipitación (enero-marzo) en todas las localidades, los flujos de crecimiento vegetativo disminuyeron.

En este estudio se encontró que la fase de desarrollo del fruto para las distintas localidades del departamento de Antioquia tiene alta variabilidad, empezando desde el segundo hasta el sexto mes después de la floración, y terminando desde el séptimo hasta el décimo segundo mes después de la floración. La duración de la etapa de crecimiento del fruto puede variar entre 4 a 9 meses. Para establecer una generalización de los tiempos de duración de la fase para esta zona, se realiza un promedio de los valores, los cuales se pueden encontrar en la **Tabla 2.**

Fenología del aguacate cv. Hass plantado en diversos ambientes del departamento de Antioquia, Colombia [42]

Este estudio fue realizado en el departamento de Antioquia en donde se encontró que los cuatro municipios estudiados presentaron un flujo floral, de distinta intensidad, pero en todos los casos se manifestaron en el primer semestre del año, mostrando un comportamiento cíclico en este estado fenológico.

En los resultados obtenidos del estudio, se puede observar que la floración tiene su mayor intensidad cerca al mes de febrero, a excepción del estudio realizado en Entrerríos, el cual tuvo la floración máxima en agosto. La cosecha por su parte, fue más variable, en Entrerríos se dió entre febrero y mayo; en Rionegro se dió entre diciembre y marzo; en Jericó fue un periodo de cosecha largo entre diciembre y mayo; y en Támesis se dió un periodo de cosecha más corto y adelantado, en los meses de septiembre a diciembre. En este estudio, los flujos florales tuvieron una duración e intensidad diferente, lo cual implica floraciones sucesivas. Este comportamiento fenológico origina la presencia simultánea de fruto de diferentes edades en el árbol, que es cosechado durante la mayor parte del año en los distintos climas de la región.

El período entre la floración y la cosecha varió entre las localidades, así: en Entrerríos transcurrieron entre 12 y 13 meses, en Rionegro duró de 11 a 12 meses, en Jericó duraron 10 a 11 meses y en Támesis este tiempo fue de 8 a 9 meses.

La precipitación en todas las localidades presentó un régimen bimodal, es decir, se dieron dos periodos de lluvia al año, lo cual tuvo un comportamiento uniforme. Los periodos de mayor precipitación corresponden en todos los casos, a los comprendidos entre marzo a mayo y entre octubre y noviembre; los períodos más secos se presentaron entre enero a marzo y entre junio y septiembre.

Teniendo en cuenta los estudios mencionados, en la **Tabla 2** se resumen los aspectos generales encontrados sobre la fenología del cultivo, en la cual se evidencian los comportamientos cíclicos hallados en las zonas de estudio y que posteriormente permitirán la caracterización del conjunto de datos.

Item	**Descripción**
Floración	En algunos casos hay 1 o 2 floraciones al año, pero la tendencia es que la más importante sea en los primeros meses del año. - Diciembre a abril - Junio a septiembre Se logró concluir que los periodos de sequía inducen a la floración del cultivo debido al estrés hídrico que se produce en la planta.
Crecimiento del fruto	Entre 4 y 10 meses después de la floración, con una duración promedio de 6 meses. Este rango de meses es un promedio de los valores encontrados en los estudios realizados.
Cosecha	En la mayoría de los casos se encuentran dos cosechas al año: - Principal: inicia entre noviembre hasta marzo, con picos en enero y febrero. - Traviesa: se presenta a mitad de año, entre abril y julio. El tiempo entre la floración y la cosecha es variable, pero en general es entre 8 y 13 meses después de la floración.

Tabla 2. Resumen de la información recolectada sobre la fenología del aguacate. Fuente propia

Para concluir la investigación sobre la fenología del aguacate, se determinó que la precipitación es una variable de alta relevancia sobre la floración del cultivo, debido a que los periodos de sequía inducen a esta, lo cual afecta directamente la producción del cultivo. Por tal motivo, es necesario realizar una investigación para conocer estos periodos en las distintas zonas del país. En [43] se encontró un estudio del año 2016 donde se realizó una generalización de los patrones de distribución anual de lluvia en la región colombiana, los cuales se pueden observar en la **Figura 9.**

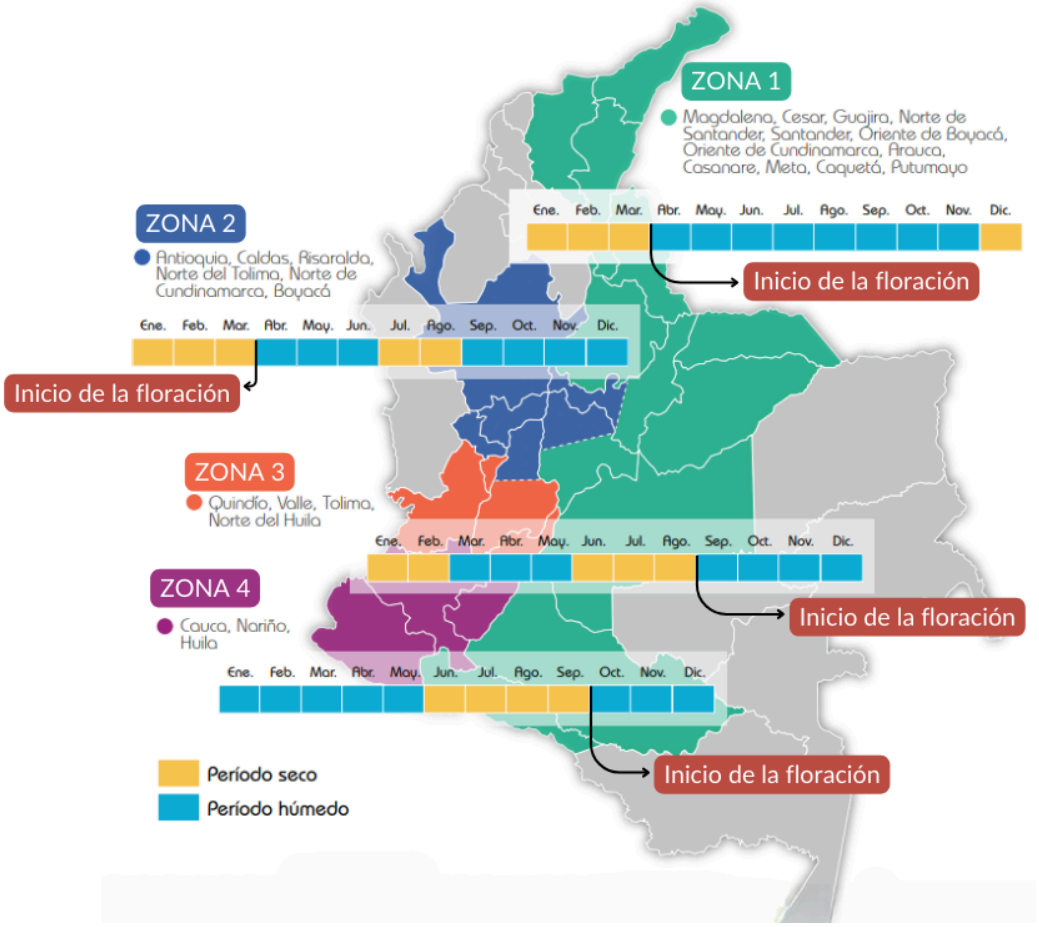

Figura 9. Zonas de distribución de lluvias en Colombia. Adaptado de [43]

Además, se estableció que el desarrollo del cultivo del aguacate se ve afectado por la altitud con respecto al nivel del mar, es decir, a mayor altitud el período entre floración y cosecha va a ser de mayor duración, esto va acorde con lo que se observó en [41] y [44], en donde la altitud también retrasó los tiempos de desarrollo de la planta. De acuerdo con la información encontrada, que fue validada posteriormente con un experto, se establecieron 4 rangos de altitud definidos en la **Tabla 3**, en estos rangos se encuentran las variaciones en el comportamiento del cultivo:

Rango	Altitud [m.s.n.m]
1	0-1500
2	1500-1950
3	1950-2800
4	A partir de 2800

Tabla 3. Rangos de altitudes. Fuente propia

3.2. Comprensión de los datos

Esta fase es importante para entender los datos disponibles con el fin de identificar y evitar problemas de cantidad, calidad y accesibilidad en las fases posteriores, de esta manera se evita repetir procesos o tener que buscar nuevos datos durante la fase de modelado, lo cual genera una pérdida de tiempo y recursos.

3.2.1. Recopilación de datos iniciales

- **Datos de producción de aguacate:** para estos datos se encontraron dos posibles fuentes, los datos de producción de aguacate de la plataforma de datos abiertos de Colombia y los datos de la plataforma Agronet del ministerio de agricultura de Colombia. Se realizó una exploración de los dos conjuntos de datos, encontrando que en ambos casos las variables o columnas tienen los datos necesarios y que estos son datos de producción anuales y a escala municipal. Al realizar una revisión a fondo de los valores numéricos de cada uno de los registros de ambos conjuntos de datos, se encontró que en los dos casos los registros disponibles para cada municipio contenían los mismos datos, es decir ambos conjuntos de datos son iguales, la diferencia radica en que el conjunto de datos abiertos está para los años 2007 a 2018 y los de Agronet están entre el año 2007 y el año 2021. Esto hace que el conjunto de datos de Agronet sea más grande teniendo alrededor de 2000 registros más, por lo cual se optó por emplear este conjunto de datos.

  Para la obtención de estos datos se ingresó a la plataforma en el enlace agronet.gov.co en la sección de estadísticas, luego se seleccionó la opción “Área, Producción, Rendimiento y Participación municipal en el departamento por cultivo”. La plataforma muestra una interfaz gráfica en la que se puede seleccionar el cultivo, el departamento, los municipios y el rango de tiempo (en años) en que se van a mostrar los datos. En este punto se seleccionó el cultivo de aguacate Hass, el departamento, todos los municipios disponibles y el rango máximo de tiempo, es decir de 2007 a 2021 y se descargó el archivo .xlsx. Este proceso se repite para cada departamento de Colombia, por lo que se obtuvieron 32 archivos xlsx que posteriormente se fusionaron para obtener el conjunto de datos a nivel nacional, el conjunto de datos con los archivos ya fusionados se encuentra en el Anexo A de este documento.

- **Datos meteorológicos:** los datos meteorológicos se obtuvieron de dos fuentes distintas con el fin de complementarlas, en primer lugar, se emplearon los datasets meteorológicos del Proyecto POWER del Centro de Investigación Langley de la NASA (LaRC), financiado a través del Programa de Ciencias Aplicadas/Ciencias de la Tierra de la NASA. Se utilizó una API que mediante peticiones HTTP se capturan los datos meteorológicos mensuales entre los años 2007 y 2021 de los municipios de Colombia, luego estos datos se

almacenaron en una hoja de cálculo indicando el municipio, el año y los meses de cada grupo de datos.

La segunda fuente de datos es el IDEAM, en un principio se intentó la recolección de los datos mediante la plataforma DHIME del instituto, sin embargo, se encontró que el proceso era poco eficiente, por lo que se optó por hacer la solicitud de los datos directamente al área de atención al usuario en la página del IDEAM, como resultado se obtuvo un dataset con las variables meteorológicas de las distintas estaciones meteorológicas del país. Se realizó un filtrado de los datos para asegurarse de que quedaran las estaciones con las variables meteorológicas necesarias para el intervalo de tiempo requerido.

3.2.2. Descripción de los datos

- **Datos de producción de aguacate:** el dataset resultante tiene 6280 filas (registros) y 7 columnas (variables), que son las siguientes:

    - **Departamento:** cadena de texto
    - **Municipio:** cadena de texto
    - **Año:** valor numérico entre 2007 y 2021
    - **Área sembrada:** valor numérico expresado en hectáreas
    - **Área cosechada:** valor numérico expresado en hectáreas
    - **Producción:** valor numérico expresado en toneladas
    - **Rendimiento:** valor numérico expresado en toneladas por hectárea

  Estos datos tienen una periodicidad anual, no se encontraron otras fuentes con distintas periodicidades, por lo cual, este registro anual de los datos es el que definió como estándar para el conjunto de datos final. Los otros conjuntos de datos que se recolectaron y se fusionaron con este, fueron adaptados a esta periodicidad.

- **Datos meteorológicos:** los datos meteorológicos deben tener una periodicidad mensual para que los efectos fenológicos del ciclo productivo del cultivo se vean reflejados en los datos, sin embargo, los datos de producción tienen una periodicidad anual, por lo que al hacer la fusión de los conjuntos no coincide la resolución temporal, para solucionar esto, se optó por asignar los registros mensuales meteorológicos a columnas distintas, es decir, 12 columnas (por los 12 meses del año) para cada tipo de dato meteorológico, así, en el registro de un año de cada municipio se verán los 12 meses en columnas separadas.

  Los datos meteorológicos recolectados del proyecto POWER tienen una resolución espacial de 0.5º latitud por 0.625º de longitud, esto causó que algunos municipios cercanos compartieran los mismos datos meteorológicos,

generando posibles problemas en la fase de modelado, por tal motivo se optó por eliminar los municipios con datos repetidos.

Los datos que se recolectan de la API del proyecto POWER de la NASA y del conjunto de datos del IDEAM son las principales variables meteorológicas encontradas en los estudios de la revisión sistemática de literatura, que también corresponden a las variables que mayor influencia tiene sobre el cultivo del aguacate, como se expuso en la fase de comprensión del negocio. Estos datos corresponden a: precipitación, temperatura y humedad. No se cuenta con datos de déficit de presión de vapor.

Las variables meteorológicas recolectadas fueron:

- **Temperatura promedio mensual:** valor numérico en grados Celsius registrado a 2 metros del suelo (T2M)
- **Temperatura máxima promedio mensual:** valor numérico en grados Celsius registrado a 2 metros del suelo (T2M_MAX), este valor es el promedio de los valores de temperatura máxima de cada día dentro del mes.
- **Temperatura Mínima promedio mensual:** valor numérico en grados Celsius registrado a 2 metros del suelo (T2M_MIN), este valor es el promedio de los valores de temperatura mínimos de cada día dentro del mes.
- **Humedad Relativa promedio mensual:** valor numérico expresado como porcentaje, registrado a 2 metros del suelo (RH2M)
- **Precipitación acumulada mensual:** valor numérico acumulado con corrección de sesgo, expresado en milímetros de agua (PRECTOTCORR), esta es la única variable que no se le hace un promedio, en su lugar se calcula todo el acumulado del mes.

Como ya se explicó, cada uno de estos datos se asigna en una columna para cada mes del año, al ser 5 datos climáticos y 12 meses al año, se obtiene un dataset climático con 60 columnas nombrando cada variable con las tres primeras letras del mes y luego el nombre de la variable (ej: ene_T2M_MIN, feb_T2M_MIN, etc.) estas 60 variables se asignan para cada municipio y cada año disponible del conjunto de datos de producción de aguacate, por lo tanto el conjunto de datos meteorológicos consiste en 2581 filas y 60 columnas.

Los conjuntos de datos recolectados de las distintas fuentes seleccionadas, incluyendo el conjunto de datos de datos abiertos de colombia que finalmente fue descartado, se encuentran en el anexo B de este documento.

3.2.3. Exploración de los datos

Para realizar una exploración inicial de los datos, se calcularon las medidas estadísticas que permitan obtener información importante de los datos, ver cómo están distribuidos y detectar valores atípicos. También se realizaron gráficos que permiten entender de forma visual la información que se está generando.
La librería Pandas de python permite con la función describe() obtener información estadística como la cantidad de datos, la media, la desviación estándar y los valores máximos y mínimos de cada una de las variables. A continuación, se muestran las medidas resultantes de 4 de las variables del conjunto de datos, en las que se evidenció un comportamiento inusual:

	Area Sembrada	Area Cosechada	Produccion (ton)	Rendimiento (ha/ton)
count	2581.000000	2581.000000	2581.000000	2581.000000
mean	168.995351	130.304378	1168.535331	8.284510
std	499.665427	426.601955	3325.547889	4.798414
min	0.500000	0.000000	0.000000	0.000000
25%	15.000000	8.000000	52.000000	5.000000
50%	41.000000	28.000000	216.000000	8.000000
75%	126.000000	91.000000	810.000000	10.500000
max	6555.000000	6132.760000	48316.260000	53.280000

Tabla 4. Medidas estadísticas del conjunto de datos de producción. Fuente propia

En la **Tabla 4** se observa que en las 4 variables existe una desviación estándar alta y que los valores máximos registrados están muy por encima del percentil 75, evidenciando una clara existencia de valores atípicos o “outliers”.

Para observar de manera gráfica la distribución de estos valores atípicos, se realizó un diagrama de cajas:

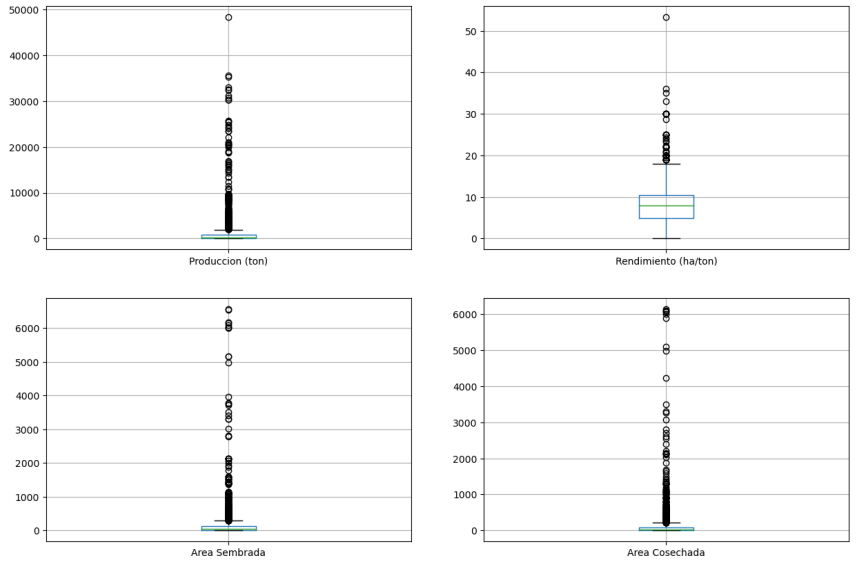

Figura 10. Diagramas de caja de las variables del conjunto de datos de producción. Fuente propia

Los diagramas de caja expuestos en la **Figura 10** permitieron observar que efectivamente existe una considerable cantidad de valores atípicos en las variables analizadas, sin embargo, es necesario revisar en profundidad cada caso para determinar si estos valores corresponden a errores de digitación o son valores atípicos naturales.

En cuanto a las variables meteorológicas, se encontró que tienen un comportamiento más estable con desviaciones estándar de valores bajos. Sin embargo, la precipitación presentó valores de desviación estándar más altos que el resto de variables, esto se evidencia en la diferencia entre el valor máximo registrado de cada mes y el percentil 75, como se muestra en la **Tabla 5**:

	ene_PRECTOTCORR	feb_PRECTOTCORR	mar_PRECTOTCORR	abr_PRECTOTCORR	may_PRECTOTCORR	jun_PRECTOTCORR
count	2544.000000	2547.000000	2551.000000	2539.000000	2539.000000	2534.000000
mean	38.524380	39.571456	66.434015	100.441208	107.681984	78.215559
std	58.196263	53.103224	78.564846	110.188653	120.052154	100.304128
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.812903	2.384298	4.961290	7.073333	7.100000	4.230833
50%	10.550000	10.550000	31.640000	58.010000	64.680000	31.640000
75%	58.010000	63.280000	110.740000	168.750000	179.300000	127.572500
max	585.350000	348.050000	611.720000	543.160000	769.920000	717.190000
	jul_PRECTOTCORR	**ago_PRECTOTCORR**	**sep_PRECTOTCORR**	**oct_PRECTOTCORR**	**nov_PRECTOTCORR**	**dic_PRECTOTCORR**
count	2534.000000	2532.000000	2539.000000	2541.000000	2541.000000	2534.000000
mean	72.620630	68.245470	71.046766	101.438942	102.691215	62.834996
std	99.573068	92.454855	87.736255	99.462454	106.463331	81.288908
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.162903	3.694994	4.398333	7.865385	8.173333	4.477868
50%	21.090000	26.370000	36.910000	79.100000	68.550000	26.370000
75%	116.020000	105.470000	116.020000	174.020000	174.020000	105.455000
max	922.850000	759.380000	711.910000	632.810000	648.630000	645.540000

Tabla 5. Medidas estadísticas del conjunto de datos climáticos. Fuente propia

Este comportamiento se puede dar debido a que la variable precipitación, a diferencia de las otras variables meteorológicas, es acumulada y no un promedio. Esto puede causar que haya algunos meses con mucha más pluviosidad que otros y genere estos valores que pueden parecer atípicos.

Finalmente se realizó un análisis de correlación lineal entre las variables independientes del conjunto de datos y las variables dependientes (producción y rendimiento).

	Produccion (ton)	Rendimiento (ha/ton)
zona distribucion lluvias	-0.034227	-0.031756
Altitud [m.s.n.m]	-0.016151	-0.010765
zona distribucion altitud	-0.000094	-0.014090
Area Sembrada	0.903440	0.056168
Area Cosechada	0.909714	0.043506
Produccion (ton)	1.000000	0.195579
Rendimiento (ha/ton)	0.195579	1.000000
Año	0.063946	0.040852
ene_T2M	0.066694	0.038735
ene_T2M_MAX	0.042763	0.035692
ene_T2M_MIN	0.042404	0.018874
ene_RH2M	-0.005487	-0.051733
ene_PRECTOTCORR	-0.017836	-0.052345
feb_T2M	0.062493	0.042597
feb_T2M_MAX	0.042912	0.039865
feb_T2M_MIN	0.045894	0.036902
feb_RH2M	-0.003900	-0.048741
feb_PRECTOTCORR	0.013785	0.006778
mar_T2M	0.063325	0.038738
mar_T2M_MAX	0.038493	0.032903
mar_T2M_MIN	0.051527	0.037622
mar_RH2M	-0.005198	-0.027532
mar_PRECTOTCORR	-0.011237	0.005113
abr_T2M	0.064512	0.026455
abr_T2M_MAX	0.035503	0.020631
abr_T2M_MIN	0.057835	0.024953
abr_RH2M	-0.014630	-0.005223
abr_PRECTOTCORR	-0.027257	0.024241
may_T2M	0.066627	0.020745
may_T2M_MAX	0.034176	0.007960
may_T2M_MIN	0.061092	0.019843
may_RH2M	-0.020570	0.006142
may_PRECTOTCORR	-0.020926	0.047981
jun_T2M	0.073327	0.016936
jun_T2M_MAX	0.044725	0.007676
jun_T2M_MIN	0.057195	0.017760
jun_RH2M	-0.016431	0.020535
jun_PRECTOTCORR	-0.016176	0.028259
jul_T2M	0.074798	0.011428
jul_T2M_MAX	0.045846	-0.003273
jul_T2M_MIN	0.059604	0.016979
jul_RH2M	-0.030075	0.033500
jul_PRECTOTCORR	-0.033094	0.047684
ago_T2M	0.072179	0.017752
ago_T2M_MAX	0.041567	0.001796
ago_T2M_MIN	0.057152	0.019738
ago_RH2M	-0.021004	0.026562
ago_PRECTOTCORR	-0.018485	0.039311
sep_T2M	0.062841	0.023097
sep_T2M_MAX	0.028269	0.010145
sep_T2M_MIN	0.049275	0.027693
sep_RH2M	-0.006153	0.011637
sep_PRECTOTCORR	-0.009862	0.044138
oct_T2M	0.058809	0.028929
oct_T2M_MAX	0.028204	0.014884
oct_T2M_MIN	0.051005	0.029178
oct_RH2M	-0.004673	-0.010231
oct_PRECTOTCORR	-0.018277	0.040811
nov_T2M	0.057360	0.027557
nov_T2M_MAX	0.033621	0.016856
nov_T2M_MIN	0.049903	0.030258
nov_RH2M	-0.008189	0.006348
nov_PRECTOTCORR	-0.028087	0.038823
dic_T2M	0.057041	0.031811
dic_T2M_MAX	0.036717	0.021881
dic_T2M_MIN	0.043227	0.023309
dic_RH2M	-0.009865	-0.018774
dic_PRECTOTCORR	-0.013160	0.020280

Figura 11. Correlación lineal entre variables dependientes e independientes. Fuente propia

Los resultados expuestos en la **Figura 11** muestran que hay una baja o nula correlación lineal entre las variables meteorológicas y la producción y el rendimiento del cultivo, sin embargo, el uso de modelos multivariables o no lineales si podrían encontrar relaciones entre las variables y ser óptimos para la estimación.

3.2.4. Verificación de calidad de datos

De acuerdo a la exploración de los datos realizada en el paso anterior, se encontraron aspectos de los datos que comprometen la calidad de los mismos. En primer lugar, algunas variables tienen claros valores atípicos, lo que puede afectar el modelo que se construya, la solución para estos valores es analizarlos y ver si corresponden a errores en la digitación de los datos, o son valores extremos. Otro aspecto a tener en cuenta, es la presencia de datos que tienen valores nulos o campos faltantes.

3.3. Preparación de los datos

La fase de preparación de los datos consiste en hacer los debidos procesos para que el conjunto de datos esté listo para ser modelado.

3.3.1. Selección de datos

La selección de los datos en una primera instancia se realizó simultáneamente en la fase de recolección de los mismos debido a que el resultado de la revisión sistemática de literatura permitió establecer las variables meteorológicas necesarias para llevar a cabo la investigación. Esto permitió que los datos se obtuvieran desde la API del proyecto POWER de la NASA y al IDEAM únicamente las variables climáticas que se van a usar para el intervalo de tiempo y las zonas geográficas establecidas. En cuanto a los datos de producción, se precedió de una manera similar ya que se seleccionaron únicamente las columnas de la base de datos de la plataforma AGRONET que iban a ser relevantes para el estudio.

Una vez realizada la selección inicial de los datos, se pasó a una segunda fase donde se revisaron los conjuntos de datos resultantes y se hizo una segunda selección de acuerdo a las filas y columnas de estos:

Selección de registros (filas): se realizó una revisión de los departamentos y se excluyeron aquellos que tenían restricciones técnicas al no contar con suficientes datos de producción en el rango de tiempo a trabajar (2007 a 2021). Los departamentos que se descartaron fueron: Amazonas, Chocó, Guainía, Guaviare, San Andrés y Providencia, Vaupés.

Adicionalmente, se debe tener en cuenta que el conjunto de datos meteorológicos tiene menos registros que el conjunto de datos de producción, como se explicó en la fase de descripción de los datos. Por tal motivo es necesario descartar todos los registros del conjunto de datos de producción, para los cuales no existen registros en el conjunto de datos meteorológicos, reduciendo así las dimensiones del conjunto de datos de producción de aguacate.

Selección de atributos (columnas): para evaluar el comportamiento fenológico del cultivo se planteó la selección de algunas de las columnas de acuerdo a las zonas de distribución de lluvias y de altitud establecidas en la fase de la comprensión del negocio de la metodología, la explicación de la selección de estas columnas se expone en el desarrollo de la fase de construcción de nuevos datos.

3.3.2. Limpieza de datos

La limpieza de los datos consiste en realizar los procesos debido dentro del conjunto de datos, para que los comportamientos indeseados y problemas encontrados no afecten la fase de modelado.

En primer lugar se encontró un problema en la periodicidad de los datos meteorológicos: los datos que se necesitan son de periodicidad mensual, debido a que de esta manera se pueden ver reflejados los efectos fenológicos del cultivo descritos en la fase de la comprensión del negocio de la metodología, pero los datos del IDEAM fueron entregados con una periodicidad diaria. Por tal motivo fue necesario realizar un promedio de los datos diarios entregados, para lo cual se utilizó un script en python. Una vez calculados los promedios, se organizaron los datos para que tengan el mismo formato y columnas que los datos del proyecto POWER de la NASA.

Otro aspecto a tener en cuenta en la limpieza de los datos es encontrar y tratar los valores nulos y valores atípicos dentro del conjunto de datos. Para los valores nulos se debe tomar la decisión de omitir las filas que contienen estos valores o rellenarlas teniendo en cuenta filas similares utilizando algoritmos y técnicas específicas de imputación.

En cuanto a los datos de producción del cultivo, se encontraron municipios con registros de años faltantes, es decir, no existían valores registrados para algunos años dentro del rango de tiempo, se tomó la decisión de no incluir estos casos dentro del conjunto de datos. También cabe resaltar, que dentro de los datos se encontraron valores de producción de 0 para los cuales se hace una revisión posterior.

Por otro lado, los datos meteorológicos obtenidos del proyecto POWER de la NASA se encuentran completos. Sin embargo, en el conjunto de datos meteorológicos del IDEAM se encontraron datos faltantes, pero representan un bajo porcentaje del total de registros, como se observa en la **Figura 12.**

```
Data columns (total 60 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ene_T2M          2510 non-null   float64
 1   ene_T2M_MAX      2516 non-null   float64
 2   ene_T2M_MIN      2516 non-null   float64
 3   ene_RH2M         2463 non-null   float64
 4   ene_PRECTOTCORR  2543 non-null   float64
 5   feb_T2M          2508 non-null   float64
 6   feb_T2M_MAX      2513 non-null   float64
 7   feb_T2M_MIN      2513 non-null   float64
 8   feb_RH2M         2465 non-null   float64
 9   feb_PRECTOTCORR  2546 non-null   float64
 10  mar_T2M          2517 non-null   float64
 11  mar_T2M_MAX      2522 non-null   float64
 12  mar_T2M_MIN      2522 non-null   float64
 13  mar_RH2M         2476 non-null   float64
 14  mar_PRECTOTCORR  2550 non-null   float64
 15  abr_T2M          2499 non-null   float64
 16  abr_T2M_MAX      2506 non-null   float64
 17  abr_T2M_MIN      2506 non-null   float64
 18  abr_RH2M         2455 non-null   float64
 19  abr_PRECTOTCORR  2538 non-null   float64
 20  may_T2M          2499 non-null   float64
 21  may_T2M_MAX      2505 non-null   float64
 22  may_T2M_MIN      2505 non-null   float64
 23  may_RH2M         2458 non-null   float64
 24  may_PRECTOTCORR  2538 non-null   float64
 25  jun_T2M          2501 non-null   float64
 26  jun_T2M_MAX      2505 non-null   float64
 27  jun_T2M_MIN      2505 non-null   float64
 28  jun_RH2M         2459 non-null   float64
 29  jun_PRECTOTCORR  2533 non-null   float64
 30  jul_T2M          2504 non-null   float64
 31  jul_T2M_MAX      2512 non-null   float64
 32  jul_T2M_MIN      2512 non-null   float64
 33  jul_RH2M         2464 non-null   float64
 34  jul_PRECTOTCORR  2533 non-null   float64
 35  ago_T2M          2491 non-null   float64
 36  ago_T2M_MAX      2499 non-null   float64
 37  ago_T2M_MIN      2499 non-null   float64
 38  ago_RH2M         2460 non-null   float64
 39  ago_PRECTOTCORR  2531 non-null   float64
 40  sep_T2M          2498 non-null   float64
 41  sep_T2M_MAX      2502 non-null   float64
 42  sep_T2M_MIN      2502 non-null   float64
 43  sep_RH2M         2464 non-null   float64
 44  sep_PRECTOTCORR  2538 non-null   float64
 45  oct_T2M          2495 non-null   float64
 46  oct_T2M_MAX      2498 non-null   float64
 47  oct_T2M_MIN      2498 non-null   float64
 48  oct_RH2M         2461 non-null   float64
 49  oct_PRECTOTCORR  2540 non-null   float64
 50  nov_T2M          2495 non-null   float64
 51  nov_T2M_MAX      2500 non-null   float64
 52  nov_T2M_MIN      2500 non-null   float64
 53  nov_RH2M         2444 non-null   float64
 54  nov_PRECTOTCORR  2540 non-null   float64
 55  dic_T2M          2490 non-null   float64
 56  dic_T2M_MAX      2496 non-null   float64
 57  dic_T2M_MIN      2496 non-null   float64
 58  dic_RH2M         2435 non-null   float64
 59  dic_PRECTOTCORR  2533 non-null   float64
```

Figura 12. Valores no nulos del conjunto de datos meteorológicos. Fuente propia.

3.3.2.1. Imputación de datos faltantes

Para realizar la imputación de los datos faltantes, se realizó una investigación sobre las posibles técnicas de imputación de datos y su implementación en Python. En la investigación se encontró que para determinar una técnica de imputación óptima, es necesario en primer lugar, determinar el tipo de datos faltantes, estos pueden pertenecer a 3 posibles clasificaciones: datos faltantes completamente al azar MCAR , datos faltantes al azar MAR o datos faltantes no al azar MNAR [45].

Los datos MCAR se dan cuando la pérdida de los datos no depende de las otras variables independientes en el conjunto de datos. En el caso de MCAR, cualquier observación tiene igual probabilidad de perderse, esto significa que los datos se recolectaron al azar, y no dependen de ninguna otra variable en el conjunto de datos, en MAR las observaciones faltantes están condicionadas por variables dentro del conjunto de datos y en el caso de MNAR ocurre cuando los datos faltantes no son ni MCAR ni MAR, es decir, que la razón de la pérdida de los datos, se debe directamente a un comportamiento que no se encuentra dentro de los datos [45].

Para el caso de esta investigación, se encuentra que la ausencia de datos es completamente al azar, es decir no se encuentra una razón dentro de las variables para definir esos datos faltantes, esto debido a que en este caso, los datos faltantes se deben a información que se pierde dentro de las estaciones meteorológicas y/o que no ha sido recolectada correctamente. Teniendo en cuenta el tipo de datos

faltantes, al ser datos numéricos de variables meteorológicas, se seleccionan tres técnicas de imputación que se adaptan a la situación: imputación por cálculo de la media aritmética, Imputación mediante K-vecinos más cercanos e Imputación por Interpolación. El desempeño de los modelos que emplean cada una de estas técnicas también es evaluado en fases futuras de la metodología.

Cabe resaltar que la imputación de los datos se realizó teniendo en cuenta las variables meteorológicas, para que los valores de producción y rendimiento del cultivo no interfieran en la generación de los nuevos datos y no se genere un sesgo en el entrenamiento del modelo predictivo de producción y rendimiento. Además, se tuvo en cuenta dos configuraciones del conjunto de datos: 1) Se implementan las técnicas teniendo en cuenta el conjunto de datos completo, y 2) Se realiza la implementación agrupando los datos de acuerdo al municipio, esto debido a que los datos de municipios con condiciones climáticas muy diferentes pueden interferir en el cálculo de otros municipios y dar como resultado nuevos datos incoherentes.

- **Imputación por cálculo de la media aritmética:** para está técnica se realizó un cálculo del promedio de los valores de cada columna de manera independiente y se imputa este nuevo valor sobre los campos faltantes de la columna. El cálculo se realizó utilizando la función “mean” de la librería pandas de Python.

  Para el caso del conjunto de datos completos se encontró que, al imputar los nuevos datos, el valor promedio calculado se reemplazó sobre campos faltantes de municipios que tenían valores muy distintos al calculado, obteniendo así un conjunto de datos resultante con muchos valores incoherentes, por ello se descartó el uso de este conjunto de datos completo y se continuó con la implementación del cálculo de la media sobre el conjunto de datos agrupado de acuerdo a los municipios.

  Finalmente se calculó la media para cada grupo y posteriormente se realizó la imputación de los datos. Se obtuvo un conjunto de datos imputados con valores diferentes para cada municipio. En algunos municipios no se contaba con valores para ningún registro, lo que impidió el cálculo de la media, en estos casos se procedió a remover estos municipios. El resultado del proceso es un dataset de 2553 registros sin valores nulos.

- **Imputación mediante K-vecinos más cercanos:** está técnica consiste en encontrar mediante el cálculo de la distancia euclidiana, los registros más similares al que tiene los datos faltantes e imputar el valor promedio de estos registros. Para implementar esta técnica se empleó la función KNNImputer de la librería Scikit-Learn de Python, buscando los 5 registros más cercanos para cada dato faltante. El resultado es un dataset de 2579 registros sin valores nulos.

Para esta técnica de imputación se encontró que utilizar el conjunto de datos agrupado de acuerdo al municipio iba a limitar la búsqueda de los registros más cercanos, ya que es posible obtener datos importantes de otros municipios que tengan condiciones meteorológicas similares. Por ello se optó por utilizar únicamente el dataset completo para la implementación de esta técnica.

- **Imputación por interpolación:** la interpolación consiste en la estimación de nuevos puntos intermedios teniendo en cuenta los datos ya existentes, la librería Pandas de Python contiene una función que permite interpolar y hallar los nuevos valores para un conjunto de datos. Al igual que con la media, el cálculo de los valores interpolados se hizo para cada columna del conjunto de datos de manera independiente y se obtuvo un valor que es imputado sobre los campos faltantes, por eso se tomó la misma decisión de implementar esta técnica de imputación únicamente sobre el conjunto de datos agrupado por municipio. El cálculo de la interpolación no se puede realizar sobre los municipios que no tengan ningún dato de la variable a imputar, por lo que estos municipios fueron removidos del conjunto de datos. El resultado final consiste en un conjunto de datos de 2499 registros.

3.3.2.2. Eliminación de valores atípicos

Para eliminar los valores atípicos fue necesario establecer los límites o umbrales de cada una de las variables. Lo primero que se realizó fue revisar las variables que durante la exploración de los datos mostraron altos valores de desviación estándar y grandes diferencias entre los valores medios y los valores máximos registrados. Las 4 variables que se analizaron fueron área sembrada, área cosechada, producción y rendimiento. Posteriormente, se hizo una comparación con otras fuentes de datos, como artículos, para corroborar si los datos corresponden a valores atípicos.

- **Área sembrada:** se encontraron valores máximos dentro de los datos entre 6000 y 6500 hectáreas, correspondientes al municipio de Fresno, Tolima para los años entre 2014 y 2021. Para corroborar estos valores se realizó una búsqueda en otras fuentes. En el portal de noticias el nuevo día [46] se menciona que de acuerdo a los datos de AsoFrutos, Fresno tiene más de 6500 hectáreas de aguacate sembradas, que alcanzan una producción de hasta 30000 toneladas anuales. El conjunto de datos de Datos Abiertos coincide en que el municipio con mayor área sembrada es Fresno con 6535 hectáreas sembradas para el año 2016. Con esto se puede concluir que los valores máximos de área sembrada no corresponden a valores atípicos.

- **Área cosechada:** los valores de área cosechada corresponden a los de área sembrada en el conjunto de datos, por lo que se asume que estos valores no son atípicos.

- **Producción:** En los datos de producción se encuentran valores máximos de hasta 48316 toneladas, este corresponde al municipio de Fresno, Tolima para el año 2021. Para corroborar estos datos se consulta el informe del ministerio de agricultura “Cadena productiva del aguacate” [47] para el año 2021, donde se muestra un histórico departamental de la producción de aguacate en Colombia de los últimos años. En el informe se evidencia que la producción máxima registrada en todo el departamento de Tolima en el año 2021 fue de 91479 toneladas, por lo que el valor de producción para el municipio de Fresno si puede estar dentro de un rango factible considerando que el área sembrada en este municipio es la mayor en todo el país. Se realiza una revisión con este mismo sistema para los demás municipios que muestran este tipo de comportamiento (los datos de producción del municipio son cercanos a los de todo el departamento para el mismo periodo de tiempo)

A pesar de que los datos encontrados fueron reales, los valores elevados podrían afectar el desarrollo del modelo, y como se puede observar en la **Figura 13**, la distribución de la variable producción muestra valores muy elevados en registros poco representativos.

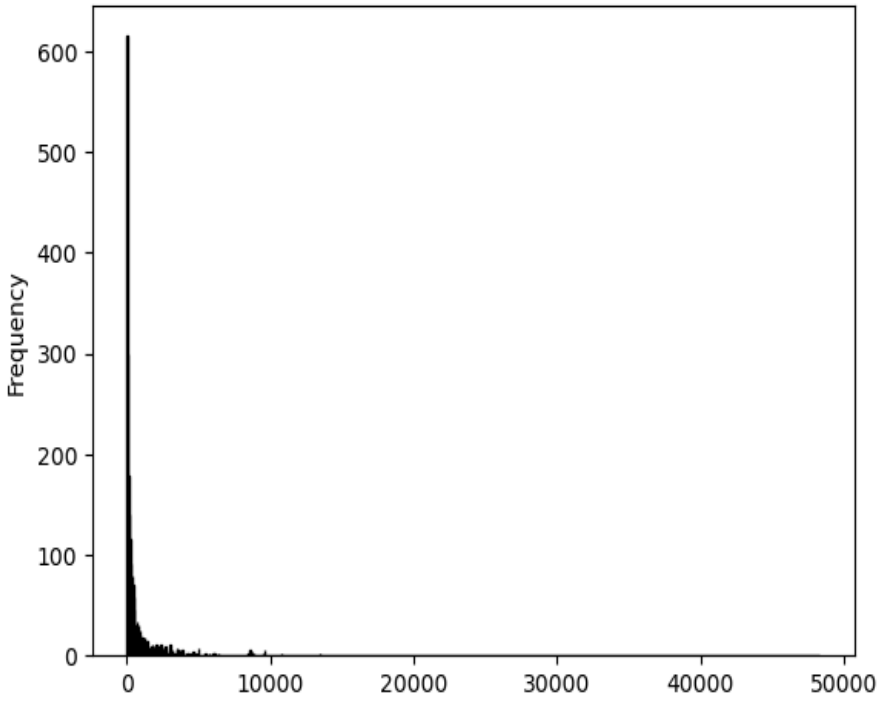

Figura 13. Distribución de la variable de producción. Fuente propia.

Los valores elevados son casos puntuales, es decir, son pocos registros del conjunto de datos que muestran este comportamiento, por lo que realizar modelos aparte, o agrupar estos registros no arrojaría resultados relevantes. Por lo tanto, se optó por eliminar estos registros teniendo en cuenta el rango intercuartil (IQR).

$$IQR = Q3 - Q1$$

Donde, el primer cuartil Q1 hace referencia al valor numérico donde se encuentran el 25% de los datos y el tercer cuartil el 75% [49]. Para la eliminación de los valores atípicos, se tiene en cuenta el umbral superior considerando la siguiente ecuación:

$$Umbral\ superior = Q3 + 1.5 * IQR$$ [49]

Todo valor que esté por encima del umbral superior, fue eliminado del conjunto de datos. Además, se optó por remover los registros con valores de producción cero, debido a que también podrían afectar el modelo.

- **Rendimiento**: el valor máximo de rendimiento encontrado en el conjunto de datos es de 53.28 ton/ha que puede corresponder a un valor atípico, ya que es superior al segundo mayor valor (36 ton/ha). Al revisar la fuente de datos de producción de Datos Abiertos, se encontró que el máximo valor de rendimiento registrado es de 36 ton/ha, lo que dista del valor máximo encontrado, por ello se opta por considerar el registro como un valor atípico y removerlo del conjunto de datos.

  Los siguientes valores del conjunto de datos rondan las 30 ton/ha, estos valores si coinciden con los valores máximos encontrados en el conjunto de datos de Datos Abiertos, sin embargo para este rango de valores se va a realizar una revisión más profunda debido a que en [44], el rendimiento máximo registrado a nivel departamental es de 14 ton/ha, sin embargo cabe la posibilidad de que al ser un promedio de los valores de todos lo municipios, estos valores altos no se vean reflejados, por lo cual se hace una investigación más a fondo, buscando en artículos e investigaciones los rendimientos que puede llegar a tener el cultivo de aguacate en zonas más limitadas.

  En [47] se realiza un estudio para la maximización del rendimiento de aguacate en zonas subtropicales, aquí se menciona que “los promedios respectivos para los mejores productores son 12-15 y 20-25 ton/ha. El rendimiento objetivo de superar las 30 ton/ha todavía parece realista, pero probablemente inalcanzable a gran escala durante un período de años con el germoplasma y la tecnología actual”, además de esto, se afirma que “se ha demostrado que se pueden lograr rendimientos sostenidos de más de 20 ton/ha tanto para 'Hass' como para 'Fuerte'.” Para el caso de los rendimientos en área colombiana, se encontró que en [49] se realizan modelos que “estiman rendimientos potenciales de 29, 13 y 7.5 ton/ha para niveles de tecnología alta, media y baja, respectivamente.” También concluyen: “los modelos indicaron que el uso de tecnologías apropiadas en la producción de aguacate en Colombia en sitios apropiados puede aumentar sustancialmente los rendimientos a cantidades cercanas al potencial máximo informado de 32,5 ton/ha”. Otro ejemplo se encuentra en [50] en donde se obtienen

rendimientos de 6 y 8 ton/ha, pero se menciona que en el estudio realizado por Salazar-García en [51] se logra obtener un rendimiento de hasta 28.2 ton/ha.

De acuerdo a estas investigaciones, se encuentra que los valores de rendimiento que rondan las 30 ton/ha si pueden ser naturales y no corresponder a errores en los datos. Por lo tanto, después de haber removido el valor atípico de 53.28 ton/ha no se establece un umbral superior, dado que los otros datos de mayor valor son cercanos a las 30 ton/ha.

- **Datos meteorológicos:** se encontró un valor atípico en la temperatura mínima del mes de enero como se observa en la **Tabla 6**, este es un valor negativo, que para ser un promedio mensual y teniendo en cuenta las condiciones climáticas tropicales de Colombia, es altamente probable que corresponda a un error.

	ene_T2M_MIN
count	2517.000000
mean	16.543015
std	4.879684
min	-0.020000
25%	13.210000
50%	16.690000
75%	20.620000
max	26.477419

Tabla 6. Medidas estadísticas de la variable Temperatura mínima. Fuente propia

Se verificó en la fuente de los datos (el proyecto POWER de la NASA) y se encontró que este dato de -0.02°C corresponde al valor mínimo diario registrado (posiblemente en la noche de un municipio con bajas temperaturas) que en el momento de calcular el promedio mensual genera el fallo y lo establece como el promedio de la temperatura mínima mensual, esto corresponde a un error y es removido del conjunto de datos.

Para revisar los posibles valores atípicos encontrados en las variables de precipitación del conjunto de datos del proyecto POWER de la NASA, se realizó una verificación con fuentes de datos alternas, en este caso, los datos de la plataforma DHIME del IDEAM. Se encontró que los municipios del conjunto de datos que tienen estos valores máximos, son también municipios con alta pluviosidad de acuerdo a los datos del DHIME, por lo tanto, los valores no son considerados como atípicos. Cabe resaltar que la fuente de datos de la plataforma DHIME, difiere de los datos solicitados previamente al IDEAM que fueron incluidos dentro del conjunto de datos de este estudio.

3.3.3. Agregación de atributos

De acuerdo a la investigación realizada se agregaron 3 nuevas columnas al conjunto de datos considerando las zonas de distribución de lluvia y los rangos de altitud, así como se explica a continuación:

Inicialmente, se establece una variable que identifique la zona de distribución de lluvias de cada registro. La nueva columna llamada “zona distribución lluvias” tiene un valor entero entre 1 y 4. Esta categoría se asigna a cada departamento del conjunto de datos. Seguidamente, se establecen dos columnas de acuerdo a los rangos de altitud definidos en la **Tabla 3**, por lo tanto, se añadió una columna Altitud[m.s.n.m] al conjunto de datos, que contiene la elevación con respecto al nivel del mar de cada uno de los municipios registrados y se añade una columna llamada “Zona de distribución de altitud” que tiene un valor entero entre 0 y 3 a partir del rango de menor valor, la cual define el rango de altitud en el que ha sido asignado cada municipio de acuerdo a su elevación.

3.3.3.1. Selección de datos a partir de las zonas de distribución de lluvias y altitud

Una vez establecidas las columnas que contienen la información sobre la distribución de las lluvias y la altitud, se procede a seleccionar las columnas de variables meteorológicas que se van a emplear para que el conjunto de datos refleje el comportamiento de la fenología del cultivo.

De acuerdo a la investigación realizada, se encontró que la floración se produce después de un periodo de estrés hídrico, por lo tanto, se tomó como mes de inicio de la floración el primer mes después del periodo seco más relevante de cada zona de distribución de lluvia. Los periodos secos y el mes de inicio de la floración para cada zona de distribución se observa en la **Tabla 7:**

Zona distribución de lluvias	Periodo seco principal	Mes de inicio de la floración
1	Diciembre a marzo	Abril
2	Enero a marzo	Abril
3	Junio a agosto	Septiembre
4	Junio a septiembre	Octubre

Tabla 7. Inicio de la floración para las Zonas de distribución de lluvias. Fuente propia

Posteriormente, se establecieron siete pruebas, las cuales permiten caracterizar la influencia de la fenología en la producción del cultivo en los rangos de tiempo y fases reproductivas del cultivo definidos en la fase de comprensión del negocio. Las 4 primeras pruebas fueron definidas con la asesoría de un experto, ya que en estas

se hace énfasis en el comportamiento fenológico del cultivo. Las pruebas 5 y 6 son variaciones de las primeras pruebas con el objetivo de establecer una solución alterna a un problema de dimensionalidad de los datos que se explica detalladamente en el apartado de cada una de las pruebas, por último se estableció la prueba 7 la cual sirve como referencia se usa para comparar la influencia de la fenología en la estimación.

- **Prueba 1: Promedio de los datos de acuerdo al año productivo del aguacate:**

Para este prueba, se tuvieron en cuenta todos los meses dentro del año productivo del cultivo, es decir desde la floración hasta la cosecha. Se obtiene para cada municipio, el mes en el que se produce la floración de acuerdo a la zona de distribución de lluvias y la duración del periodo de acuerdo con los rangos de distribución de altitud establecidos en la **Tabla 3**, los meses seleccionados se muestran en la **Tabla 8.**

Rangos distribución de altitud [m.s.n.m]	Duración entre floración y cosecha	Zona de distribución de lluvias	Meses seleccionados
0-1500	8 meses	1	Abril a noviembre
		2	Abril a noviembre
		3	Septiembre a abril del siguiente año
		4	Octubre a mayo del siguiente año
1500-2000	9 meses	1	Abril a diciembre
		2	Abril a diciembre
		3	Septiembre a mayo del siguiente año
		4	Octubre a junio del siguiente año
2000-2800	11 meses	1	Abril a febrero del siguiente año
		2	Abril a febrero del siguiente año
		3	Septiembre a julio del siguiente año
		4	Octubre a agosto del siguiente año

A partir de 2800	13 meses	4	Octubre a octubre del siguiente año

Tabla 8. Selección de Meses para la Prueba 1. Fuente propia.

- **Prueba 2: Promedio de los datos de acuerdo al período seco:**

Para esta prueba se configuró el dataset para tomar los meses en los que se presenta un estrés hídrico, teniendo en cuenta las zonas de distribución de lluvias. Por lo tanto, se seleccionó para cada municipio dentro de cada zona de distribución de lluvias los meses correspondientes al periodo seco, como se muestra en la **Tabla 9**.

Zona de distribución de lluvias	Meses seleccionados
1	Diciembre del año previo a marzo
2	Enero a marzo
3	Junio a agosto
4	Junio a septiembre

Tabla 9. Selección de meses para la Prueba 2. Fuente propia.

- **Prueba 3: Promedio de los datos de los 6 meses previos a la floración**

En esta prueba se evalúa la afectación de las variables meteorológicas 6 meses previos a la floración, en este caso solo se considera la zona de distribución de lluvias para establecer el mes de la floración y así seleccionar los meses respectivos para cada municipio, estos meses son los correspondientes a cada zona, como se muestra en la **Tabla 10**.

Zona de distribución de lluvias	Meses seleccionados
1	Octubre del año previo a marzo
2	Octubre del año previo a marzo
3	Marzo a agosto
4	Abril a septiembre

Tabla 10. Selección de meses para la Prueba 3. Fuente propia.

- **Prueba 4: Promedio de los datos de acuerdo al periodo del crecimiento del fruto**

En esta prueba se evaluó la incidencia que tiene la meteorología y la altitud en la etapa del crecimiento del fruto. De acuerdo con la información obtenida de los estudios sobre fenología del aguacate realizados en Colombia, se encontró que la duración de la fase de crecimiento del fruto puede variar entre los 4 y 10 meses, a partir de esto, se establecen los tiempos y los meses para cada uno de los rangos de altitud y zonas de distribución de lluvias, como se muestra en la **Tabla 11.**

Rangos distribución de altitud [m.s.n.m]	**Fase de crecimiento del fruto**	**Zona de distribución de lluvias**	**Meses seleccionados**
0-1500	4 meses	1	Abril a julio
		2	Abril a julio
		3	Septiembre a diciembre
		4	Octubre a enero del siguiente año
1500-2000	6 meses	1	Abril a septiembre
		2	Abril a septiembre
		3	Septiembre a febrero del siguiente año
		4	Octubre a marzo del siguiente año
2000-2800	8 meses	1	Abril a noviembre
		2	Abril a noviembre
		3	Septiembre a abril del siguiente año
		4	Octubre a mayo del siguiente año
A partir de 2800	10 meses	4	Octubre a julio del siguiente año

Tabla 11. Selección de meses para la Prueba 4. Fuente propia

En cada prueba, fue necesario seleccionar meses de años anteriores o posteriores, por lo que, fue necesario crear nuevas columnas temporales que almacenan los datos de estos meses en cada registro del conjunto de datos.

Una vez seleccionados los meses para cada municipio en los 4 pruebas, incluyendo las columnas temporales con los años previos y posteriores, se realizó el cálculo del promedio de las variables meteorológicas de los meses correspondientes a cada

caso, para obtener los conjuntos de datos únicamente con las cinco variables meteorológicas para cada registro.

- **Prueba 5: Mediana y Desviación estándar de los datos del año productivo**

Debido a que en los pruebas anteriores se realizó el cálculo del promedio, se evalúa cómo varían los resultados empleando otras medidas estadísticas, por ello, en este prueba, se emplearon los meses del año fenológico (Prueba 1) pero en lugar de calcular el promedio de las variables meteorológicas, se calculó su mediana y su desviación estándar.

Estos valores se calcularon para cada una de las variables meteorológicas y se almacenaron en nuevas columnas, obteniendo un total de 10 columnas de datos.

La selección de los meses para este prueba se observa en la **Tabla 8.**

- **Prueba 6: Datos meteorológicos mensuales 6 meses previos a la floración**

Otra aproximación de modelado que se realizó es pasar al modelo directamente los datos de los meses seleccionados y no el promedio. El problema de esta aproximación es que en la mayoría de las pruebas la cantidad de meses seleccionados varía de acuerdo a la duración del periodo seco y a la altitud, lo que crea el inconveniente de que los registros pueden tener diferente dimensionalidad, generando problemas en el modelado. Por tal motivo, se decidió evaluar este método sobre la prueba 3, es decir, el que evalúa los 6 meses previos a la floración, en donde los registros van a tener la misma cantidad de meses y se evita el problema de la dimensionalidad.

Los meses seleccionados para evaluar esta prueba se observan en la **Tabla 10**.

- **Prueba 7: Datos del año calendario (Modelo de referencia)**

En esta prueba se tomó el conjunto de datos de referencia para evaluar los resultados con respecto a las pruebas donde se realiza la selección de meses de acuerdo a la fenología, es decir se tomaron las variables meteorológicas de todos los meses del año calendario.

3.3.3.2. Generación de nuevos conjuntos de datos

Finalmente, se implementan las siete pruebas establecidas dentro de los tres conjuntos de datos meteorológicos generados mediante las técnicas de imputación seleccionadas, obteniendo como resultado veintiún conjuntos de datos

meteorológicos que permiten evidenciar el comportamiento fenológico del cultivo en sus distintas fases y rangos de tiempo.

Estos veintiún conjuntos de datos meteorológicos son los que se integran con los datos de producción y rendimiento del cultivo para ser modelados como se muestra en la siguiente fase de la metodología.

3.3.4. Integración de datos

La integración de los datos consiste en fusionar los datos de producción y rendimiento de aguacate con los doce conjuntos de datos meteorológicos. Debido a que la recolección de los datos meteorológicos se hizo teniendo en cuenta los rangos de tiempo y las locaciones geográficas disponibles en el conjunto de datos de producción, ambos conjuntos tienen las mismas dimensiones, lo cual facilitó su integración. Los conjuntos de datos se integraron en una hoja de cálculo, teniendo en cuenta que las variables comunes son “Año” y “Municipio”.

3.3.5. Formato de datos

En el formato de datos se realizó una modificación sintáctica a los conjuntos de datos. Se revisaron los nombres de las columnas para identificar que no tuvieran el carácter “ñ” o tildes que podían ocasionar errores.

Otro aspecto a tener en cuenta corresponde a los signos de puntuación de los decimales y los miles, por ejemplo, en el conjunto de datos de producción los miles están identificados con una coma ”,” y en el resto de datos no hay identificador, por lo que se estableció como criterio común que no haya identificador para los miles y el punto “.” para los decimales. Esto se implementó dentro del código de python en el momento de realizar la carga del conjunto de datos fusionado.

Finalmente se estableció el formato de los conjunto de datos, que se guardan como archivos .csv debido a la simplicidad y compatibilidad de este tipo de formato con las distintas herramientas software utilizadas.

Una vez finalizadas las fases de comprensión del negocio, comprensión de los datos y preparación de los datos, se obtuvieron como resultado los distintos conjuntos de datos ya listos para ser modelados y evaluados en las siguientes fases de la metodología.

3.3.6. Conclusión de la fase de preparación de los datos

El proceso realizado en las tres primeras fases de la metodología permitió completar la caracterización de los factores meteorológicos que inciden en la producción de aguacate, esto debido a que se utilizó la información de la fase de comprensión del

negocio, donde se encontró que las variables meteorológicas seleccionadas (precipitación, temperatura, humedad) y la elevación influyen sobre la fenología del cultivo, y por ende en la producción de aguacate. Además, el análisis realizado en la fase de comprensión de datos permitió entender y encontrar comportamientos en los datos para construir el conjunto de datos necesario. Finalmente, a través de las pruebas establecidas, se definió dentro del conjunto de datos construido, cómo las variables meteorológicas seleccionadas afectan la producción de aguacate.

Los conjuntos de datos resultantes del proceso de limpieza y agregación de datos se encuentran disponibles en el Anexo B de este documento, además los códigos de python, donde se realizaron todos los procesos para llevar a cabo la preparación de los datos se encuentran en el Anexo C.

4. Modelo para la estimación de producción de aguacate

Con la fase de preparación de los datos finalizada, se obtuvieron los doce conjuntos de datos listos para ser modelados y poder estimar la producción de aguacate. Para la fase de modelado se desarrollaron los siguientes pasos:

4.1. Selección de los algoritmos de modelado

La selección de algoritmos se realizó a partir de la revisión sistemática de literatura (**Figura 4**), considerando las técnicas de aprendizaje supervisado más utilizadas y que mejores resultados mostraron en las investigaciones similares. De acuerdo con lo anterior, las técnicas seleccionadas son: Regresión de Bosques Aleatorios, Regresión de Máquinas de Soporte, Red neuronal artificial y Regresión multivariable.

Las cuatro técnicas seleccionadas se implementaron mediante la librería Scikit-Learn de python, utilizando los siguientes algoritmos de la librería:

- Sklearn.ensemble.RandomForestRegressor()
- Sklearn.svm.SVR()
- Sklearn.neural_network.MLPRegressor()
- Sklearn.linear_model.LinearRegression()

El código fuente, donde se encuentra la implementación de los modelos, se pueden encontrar en el Anexo C de este documento.

4.2. Generación de un plan de prueba

Para realizar las pruebas se hizo uso de la validación cruzada de K iteraciones. Este método puede ser óptimo debido a que tiene como ventaja que todos los datos son utilizados para entrenar y validar el modelo, lo que genera resultados más robustos y más representativos, garantizando que los casos de entrenamiento y validación no se repitan, pero tiene como desventaja un alto coste computacional debido a las múltiples validaciones que se deben realizar, sin embargo, esto no es un inconveniente en los casos en que el conjunto de datos es pequeño [52]. La implementación de esta técnica de validación se realiza mediante la función respectiva de la librería Scikit-Learn sklearn.model_selection.KFold().

Otro aspecto a tener en cuenta es la selección de los hiper parámetros de los modelos, para esto, se empleó la técnica Grid Search que encuentra la mejor combinación de hiper parámetros que dan resultados óptimos para el modelo [53], al igual que la validación cruzada, esta técnica requiere de alto costo computacional que no es inconveniente en este caso debido al tamaño del conjunto de datos. Se

empleó la función sklearn.model_selection.GridSearchCV() de Scikit-Learn para ejecutar esta selección de hiper parámetros.

Una representación visual de cómo funciona la validación cruzada de k iteraciones, se puede observar en la **Figura 14.**

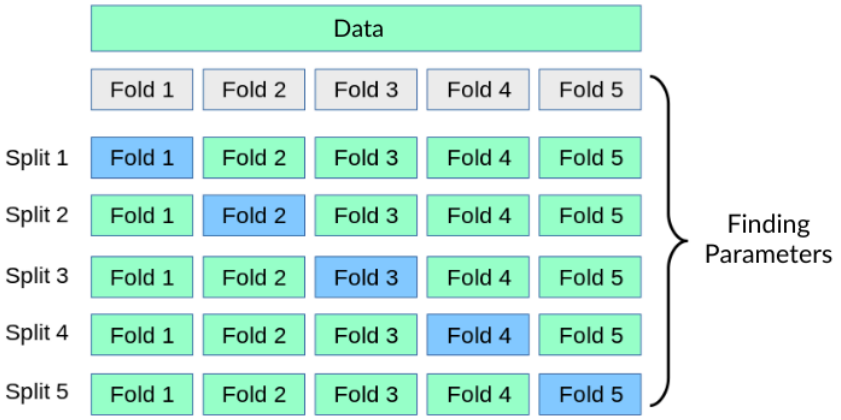

Figura 14. Ilustración de validación cruzada k-fold. Adaptada de [54]

4.3. Determinación de las métricas de evaluación que se calcularán para evaluar los modelos

Las métricas de evaluación de los modelos se eligieron de acuerdo a los resultados de la revisión sistemática de literatura (Figura 6), en donde se encontró que la raíz del error cuadrático medio RMSE, el coeficiente de determinación R^2, y el error medio absoluto MAE son las métricas más empleadas para evaluar estimaciones de rendimiento y producción de cultivos. Las funciones de Scikit-Learn para implementar estas métricas son las siguientes:

- Sklearn.metrics.mean_absolute_error()
- Sklearn.metrics.r2_score()
- Sklearn.metrics.mean_squared_error()

Además de estás métricas, también se tuvo en cuenta la interpretabilidad y explicabilidad de las técnicas seleccionadas, de acuerdo a [55] la explicabilidad es la medida en que la mecánica interna de un sistema de aprendizaje automático se puede explicar en términos humanos y la interpretabilidad de un resultado de un modelo es la descripción de cómo se ha producido la salida arrojada por el modelo, de acuerdo a esto se infiere que estas dos características son deseables en un modelo, debido a que permiten un mayor entendimiento del modelo y de los resultados obtenidos. Esto se complementa con la información obtenida en la revisión sistemática de literatura donde se encontró que algunas de las técnicas seleccionadas tiene una buena interpretabilidad y explicabilidad y otras de las técnicas carecen de estas características.

4.4. Construcción de los modelos

El proceso que se llevó a cabo para la construcción de los modelos consistió en modelar los veintiún conjuntos de datos resultantes de la combinación de las siete pruebas establecidas de acuerdo a la fenología y las tres técnicas de imputación de datos. Estos veintiún conjuntos de datos se modelaron con las cuatro técnicas de aprendizaje automático elegidas, obteniendo un total de ochenta y cuatro modelos que fueron evaluados con las tres métricas de evaluación seleccionadas, empleando la validación cruzada de k iteraciones.

De acuerdo a la documentación oficial de la librería Scikit-Learn, el número de iteraciones de la validación cruzada (el valor K) es por defecto cinco, es decir, el conjunto de datos se divide en cinco partes y cada parte será el conjunto de validación en una de las cinco iteraciones del proceso. Se optó por mantener el número de K como cinco en lugar de otros valores más altos que suelen ser utilizados, por ejemplo diez, debido a que el conjunto de datos en este caso es bastante reducido y dividir el conjunto de datos en más partes puede significar que los subconjuntos de validación sean muy pequeños y no tengan la suficiente cantidad de registros para arrojar resultados robustos.

Otro aspecto importante en el proceso de modelado es la selección de hiper parámetros, ya que cada modelo cuenta con una serie de variables que se deben configurar en busca del mejor desempeño. Considerando la documentación de Scikit-Learn se tomaron rangos de posibles valores que puede tener cada hiper parámetro, para posteriormente establecer esos valores utilizando la técnica Grid Search y finalmente obtener la mejor combinación de estos para cada modelo.

4.5. Evaluación de los resultados

Se encontró que a nivel de RMSE y MAE el modelo con mejor rendimiento fue el que empleó la red neuronal artificial (Perceptrón multicapa) sobre la prueba 1, con la imputación de datos calculados con el promedio. Los 20 modelos con mejores métricas de evaluación se pueden observar en la **Tabla 12**. En el Anexo D de este documento se puede encontrar la tabla con los resultados de los 84 modelos implementados.

Algoritmo	Prueba	Técnica de imputación	RMSE	R^2	MAE
MLP	1	Promedio	246.613	0.633	153.66
MLP	1	KNN	248.435	0.64	155.571
MLP	1	Interpolación	250.293	0.64	156.466
MLP	4	KNN	250.606	0.638	156.668
MLP	4	Promedio	252.1	0.637	156.657
MLP	5	Promedio	253.027	0.638	158.331
MLP	5	KNN	255.945	0.634	161.029
MLP	4	Interpolación	256.966	0.634	159.61

MLP	7	KNN	257.403	0.629	161.965
MLP	7	Promedio	257.431	0.625	164.995
MLP	5	Interpolación	260.767	0.637	162.748
MLP	7	Interpolación	262.287	0.632	168.111
MLP	2	Promedio	269.263	0.649	167.102
MLP	2	Interpolación	273.071	0.648	169.343
MLP	2	KNN	273.806	0.646	169.328
MLP	3	Promedio	274.036	0.643	168.06
MLP	6	Promedio	277.22	0.635	172.417
MLP	3	KNN	278.347	0.644	171.452
MLP	3	Interpolación	278.403	0.649	172.866
MLP	6	KNN	281.241	0.636	174.263

Tabla 12. Mejores resultados de acuerdo al RMSE. Fuente propia

Como se puede observar en la **Tabla 12**, los mejores resultados corresponden a la red neuronal en distintas combinaciones de pruebas y técnicas de imputación, además, se encuentra una tendencia en que los mejores resultados corresponden a las pruebas 1,4 y 5. Para poder evidenciar y comparar los resultados entre las distintas técnicas de imputación, algoritmos y pruebas de una manera más clara, se realizó un cálculo de los promedios de los RMSE, MAE y R^2 de cada modelo, obteniendo las siguientes tablas y gráficas:

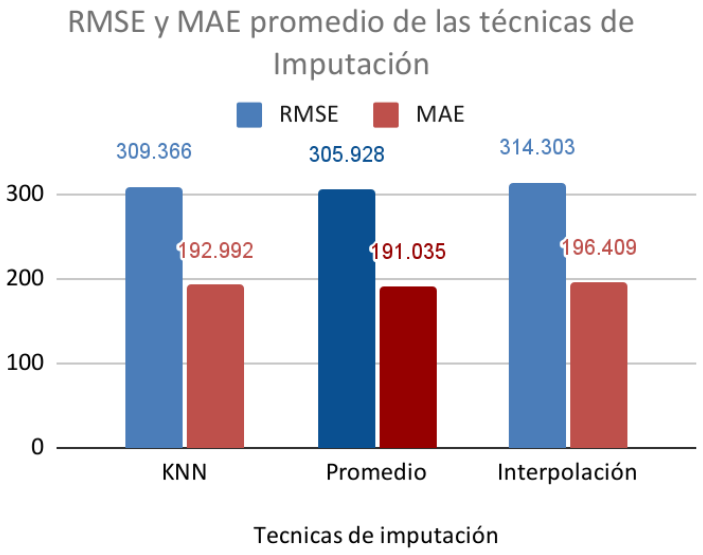

Figura 15. RMSE y MAE promedio de los resultados agrupados de acuerdo a las técnicas de imputación. Fuente propia.

Como se muestra en la **Figura 15** las técnicas de imputación no mostraron grandes diferencias, la imputación mediante el cálculo del promedio obtuvo los mejores resultados tanto de RMSE como MAE, además, se encontró que el promedio fue la mejor técnica de imputación para cada algoritmo de manera independiente como se puede observar en la **Figura 16**.

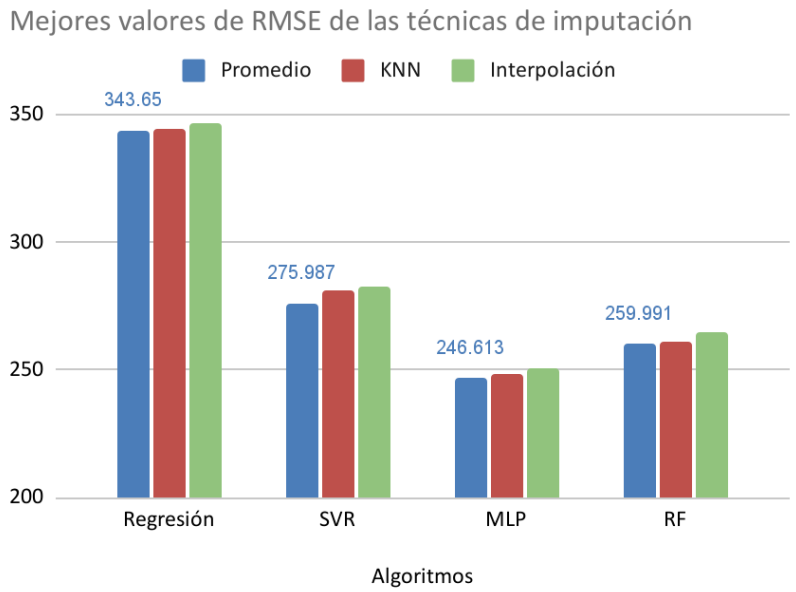

Figura 16. Mejores valores de RMSE de las técnicas de imputación para cada algoritmo. Fuente propia.

Sin embargo, al observar la Tabla 12, se evidencia que los mejores resultados tienen una mayor influencia de los algoritmos y las pruebas en comparación con las técnicas de imputación, concluyendo que la elección de la técnica de imputación no afecta de manera significativa en el rendimiento de los modelos.

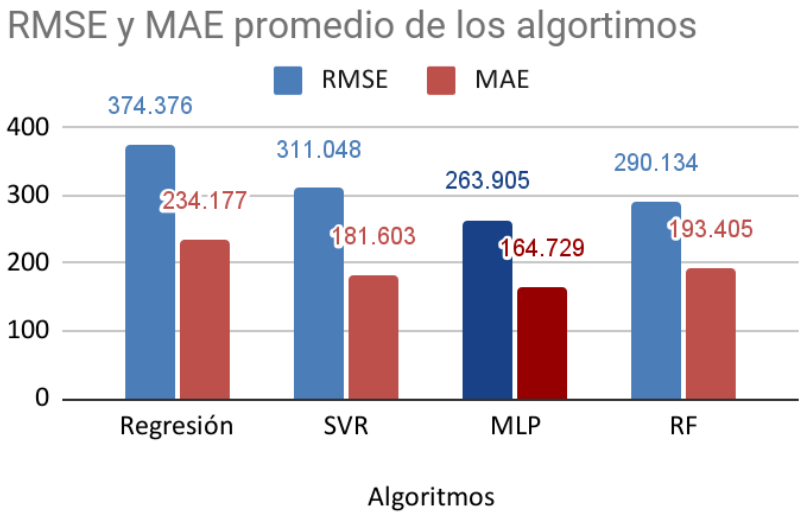

Figura 17. RMSE y MAE promedio de los resultados agrupados de acuerdo a los algoritmos de aprendizaje automático. Fuente propia.

En cuanto a los algoritmos se encuentra que los resultados del Perceptrón multicapa, tienen un menor RMSE y MAE que los otros algoritmos implementados, además en la **Tabla 12.** se observa que los diez valores más bajos de RMSE corresponden todos a la red neuronal, concluyendo que este algoritmo es el que mostró mejor desempeño, cabe resaltar que a pesar de que hay una técnica de

aprendizaje automático con los mejores resultados, los valores de RMSE y MAE del bosque aleatorio y la regresión de vectores de soporte son cercanos a los de la red neuronal como se observa en la **Figura 17**, sin embargo se puede observar que a pesar de que el MAE promedio de SVR es menor, tiene un alto RMSE, lo que significa que se presentaron errores de mayor magnitud en está técnica, mostrando un desempeño poco eficiente, de acuerdo a esto, se establece que la técnica que mostró los segundos mejores resultados, fue el bosque aleatorio.

De acuerdo a lo anterior, se realizó un análisis a profundidad de los resultados de las dos técnicas que mostraron un mejor rendimiento, se calculó el promedio de los valores de RMSE de los dos algoritmos para cada una de las siete pruebas realizadas, la gráfica resultante se puede observar en la **Figura 18.**

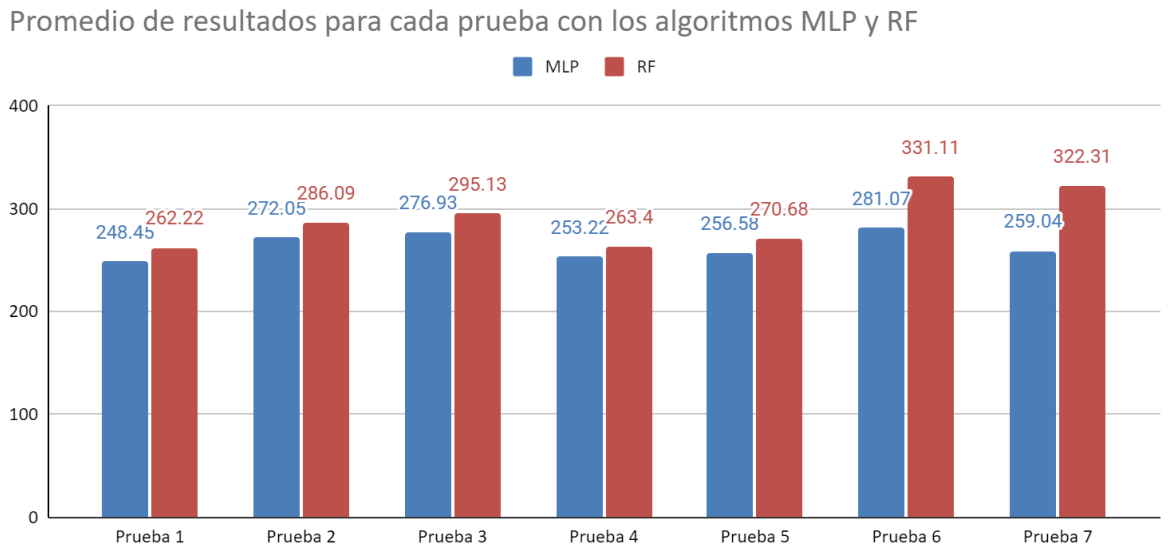

Figura 18. Promedio de RMSE para cada prueba con los mejores algoritmos MLP y RF. Fuente propia.

La **Figura 18** evidencia que el desempeño de la red neuronal, fue superior al del bosque aleatorio en todas las pruebas, sin embargo, cabe destacar que la prueba 7, es decir, la prueba de referencia que no refleja el comportamiento fenológico, es aquella que muestra una mayor diferencia entre los dos modelos. Las pruebas que tienen en cuenta la fenología del cultivo, mostraron una menor diferencia entre los desempeños de las dos técnicas.

Para finalizar con los resultados de las técnicas de aprendizaje automático, se realizó una comparación de las métricas de evaluación establecidas entre los dos algoritmos que mostraron un mejor desempeño, esta comparación se puede observar en la **Tabla 13.**

ITEM	ALGORITMOS	
	MLP	RF

RMSE del modelo con mejores resultados	246.61	259.91
MAE del modelo con mejores resultados	153.66	169.57
R^2 del modelo con mejores resultados	0.63	0.59
Interpretabilidad y Explicabilidad	No es posible demostrar y explicar los resultados obtenidos	Permite realizar un análisis de relevancia de variables

Tabla 13. Comparación entre la técnica de bosque aleatorio y red neuronal. Fuente propia

Además si comparamos el mejor modelo de cada técnica, se evidencia que la red neuronal vuelve a mostrar un mejor en las tres métricas seleccionadas, sin embargo, teniendo en cuenta la interpretabilidad y explicabilidad de estas técnicas encontramos que el algoritmo de bosques aleatorio presenta la ventaja de contar con una funcionalidad que permite identificar las variables predictoras más importantes para la estimación, por otro lado, como se encontró en la revisión sistemática de literatura la red neuronal no permite entender cómo las variables afectan las estimaciones debido a su naturaleza de caja negra, esto se traduce en una baja interpretabilidad y explicabilidad de esta técnica.

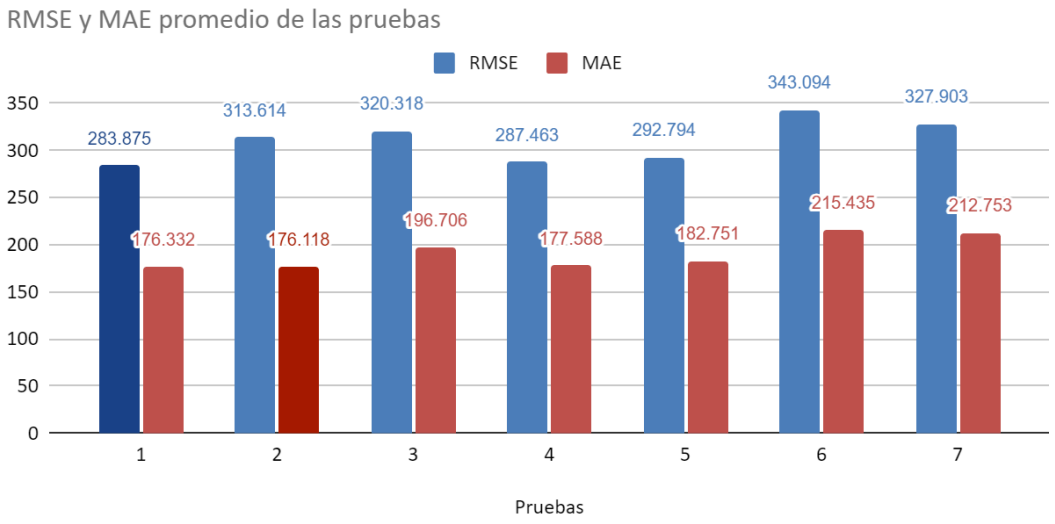

Figura 19. RMSE y MAE promedio de los resultados agrupados de acuerdo a las pruebas. Fuente propia.

De acuerdo a la **Figura 19**, el mejor promedio de RMSE lo tuvo la prueba 1, seguido de las pruebas 4 y 5, cabe resaltar que la prueba 2 presenta el mejor MAE pero un RMSE alto, esto significa que los modelos con esta prueba presentaron errores puntuales de mayor magnitud, debido a que el RMSE es más sensible a grandes errores que el MAE. Por tal motivo, se considera que la prueba número 2 a pesar de su bajo valor de MAE, no muestra un desempeño destacable.

Además, las pruebas 1, 4 y 5 mostraron los mejores resultados de RMSE para cada técnica de aprendizaje automático de manera independiente, como se puede observar en la **Figura 20.**

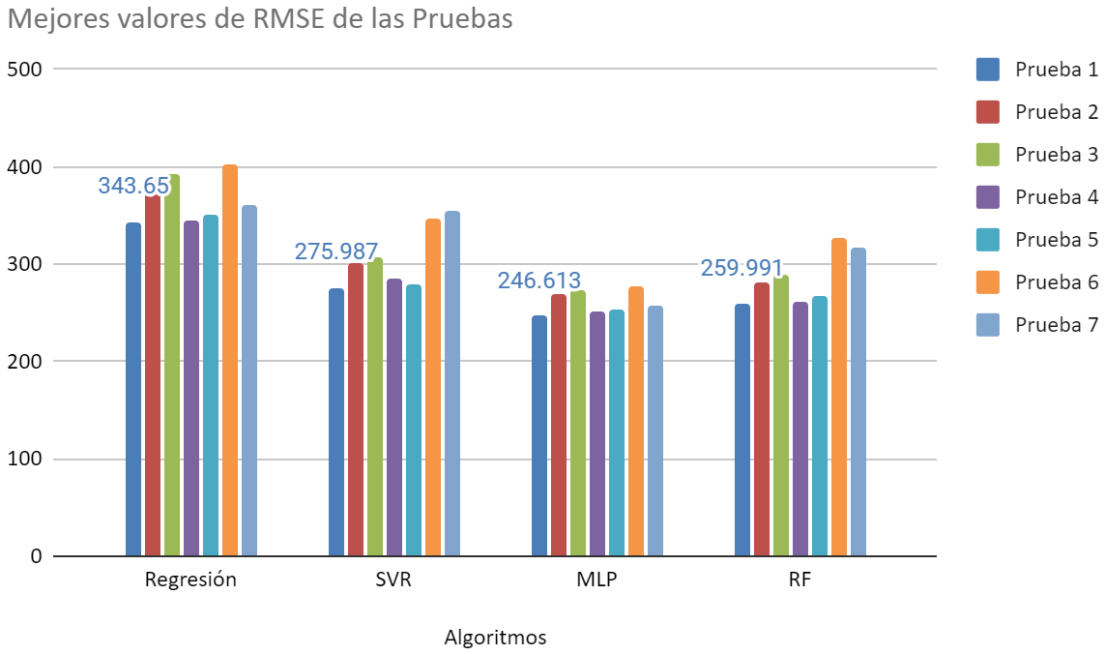

Figura 20. Mejores valores de RMSE de las Pruebas para cada algoritmo. Fuente propia.

Debido a lo ilustrado anteriormente y como se muestra la **Tabla 12**, los tres mejores resultados corresponden a la prueba 1, se puede concluir que la elección de la prueba muestra una relevancia significativa en el desempeño de los modelos. Además, se evidencia que el análisis del comportamiento fenológico del cultivo tuvo un impacto positivo en la estimación de la producción, lo que es evidente al analizar la **Figura 19** donde los resultados de la prueba 7 (el cual no tiene en cuenta la fenología), muestra los segundos valores más altos de RMSE y MAE.

Otro aspecto que resalta al analizar las pruebas es que en los resultados de la estimación de la producción en las pruebas 1, 4 y 5, donde se consideraron los meses después de la floración, se obtuvieron menores valores de RMSE y MAE en comparación a las pruebas 2, 3 y 6, que tomaron los meses previos a la floración, de esta manera, se puede concluir que seleccionar los meses posteriores a la floración muestra un mejor desempeño que la selección de meses previos.

Por último, se encontró que la distribución de altitud mostró una influencia significativa en el desempeño de los modelos de estimación ya que justamente los modelos con las tres pruebas de menor RMSE (pruebas 1, 4 y 5) son los que tienen en cuenta la altitud de los municipios.

Con respecto al desempeño de las técnicas de aprendizaje automático se encontró que los menores valores de RMSE y MAE fueron obtenidos por la red neuronal artificial, que funcionó correctamente con un conjunto de datos de tamaño reducido, lo que no es un comportamiento característico de este tipo de algoritmos. Esto puede darse debido a que las redes neuronales son capaces de encontrar patrones complejos y relaciones no lineales dentro de los datos, existiendo la posibilidad de obtener un modelo óptimo a pesar del tamaño del conjunto de datos. Sin embargo, debido a la naturaleza de caja negra de las redes neuronales, no es posible demostrar y explicar los resultados obtenidos. Debido a esto, se tomó la decisión de elegir el algoritmo de regresión de bosques aleatorios como solución óptima para un caso práctico de estimación de producción de aguacate, puesto a que el bosque aleatorio presentó los segundos valores más bajos de RMSE y MAE, y se prioriza la mejor explicabilidad e interpretabilidad que proporciona esta técnica.

5. Conclusiones y trabajos futuros

En este capítulo se presentan las conclusiones encontradas a lo largo del desarrollo de este trabajo de investigación, también se plantean los posibles trabajos futuros que se pueden derivar a partir de este estudio teniendo en cuenta las limitaciones encontradas.

5.1. Conclusiones

En este trabajo se logró realizar una estimación de la producción del cultivo de aguacate a partir de factores meteorológicos, llevando a cabo un proceso investigativo que permitió en primer lugar, entender el estado actual del conocimiento en el área para establecer las bases y brechas para desarrollar esta investigación. En segundo lugar, se desarrolló un proceso de minería de datos con el fin de crear una serie de conjuntos de datos meteorológicos que caracterizan el comportamiento fenológico del cultivo, para posteriormente implementar los modelos de aprendizaje automático que dieron como resultado la estimación de la producción de aguacate.

Para llevar a cabo la revisión sistemática de literatura, se tomó como referencia la metodología kitchenham, que permitió conocer las investigaciones similares y el estado actual del conocimiento, encontrando las técnicas y métricas más empleadas para ejecutar este tipo de soluciones, que posteriormente se usaron como una guía para la implementación de este estudio.

Al realizar la implementación usando como referencia la metodología CRISP-DM para la ejecución del proceso de minería de datos se logró cumplir con los objetivos planteados para el proyecto, ya que en primer lugar se construyó una serie de conjuntos de datos que caracterizan el comportamiento fenológico de los cultivos, que además son aptos para realizar un proceso de estimación de la producción. En segundo lugar, se implementaron una serie de modelos de aprendizaje automático que estiman la producción de aguacate a partir de los conjuntos de datos meteorológicos construidos. En tercer lugar, se evaluó el desempeño de los modelos de aprendizaje automático utilizando las métricas de evaluación definidas.

El desarrollo de la revisión sistemática de literatura llevada a cabo y la implementación del proceso de minería de datos permitieron llegar a las siguientes conclusiones:

- La revisión sistemática de literatura permitió evidenciar la escasez de este tipo de investigaciones en un ámbito nacional y principalmente enfocadas hacia el cultivo de aguacate, ya que a pesar de que existen numerosos estudios que realizan estimaciones de la producción y rendimiento de cultivos, en su mayoría son realizados sobre cultivos transitorios en países asiáticos y en los Estados Unidos. Esto genera el espacio para la ejecución de este tipo de investigaciones en un contexto local, dando una mayor relevancia a los aportes que se hacen en este estudio.

- En el estado actual del conocimiento no se encontraron estudios donde se realizara un análisis fenológico del cultivo para mostrar la influencia que tiene el clima en las distintas fases del desarrollo de la planta, afectando directamente la estimación de la producción, lo cual añade un valor adicional al estudio expuesto en este trabajo. Además, se obtuvieron resultados de RMSE y MAE más bajos en los modelos con los conjuntos de datos que consideraban la fenología del cultivo, en comparación a aquellos que no, evidenciando que a nivel de los datos, hubo una importante influencia del análisis fenológico realizado.

- El algoritmo de bosques aleatorios se escogió como el óptimo para ejecutar la solución planteada en este estudio, lo cual va acorde a lo encontrado en en la revisión sistemática de literatura, donde no solo fue la técnica de aprendizaje automático más usada, sino que también fue la que mejores resultados mostró en la mayoría de los estudios.

5.2. Trabajos futuros

A partir de la investigación propuesta en este documento, se identificaron nuevos enfoques en los cuales se puede profundizar. De acuerdo a esto, se proponen los siguientes trabajos futuros:

- Una de las limitaciones encontradas durante el desarrollo de este estudio, fue la periodicidad anual de los datos de producción del cultivo, lo que impidió caracterizar en los datos el comportamiento del cultivo mes a mes, perdiendo información detallada sobre la fenología del cultivo. Por lo tanto, se propone realizar un estudio empleando datos más completos y con una temporalidad más exacta, recurriendo si es posible a fuentes de datos del sector privado y de esta manera, realizar un análisis fenológico más detallado teniendo en cuenta estos aspectos que afectan el cultivo. Esto puede llevar a una mejora considerable en el desempeño de los modelos y en los resultados arrojados. También se recomienda tener en cuenta datos que no sean únicamente meteorológicos y que tengan una directa afectación sobre la planta, pueden ser datos de los suelos o de manejo del cultivo, esto también dependería de la disponibilidad de las fuentes de datos.

- Para escoger los algoritmos de aprendizaje automático implementados en este estudio se tuvo en cuenta la revisión sistemática de literatura y la cantidad y el tipo de datos disponibles para una mejor adaptación de los datos a los modelos. Teniendo en cuenta lo anterior, se recomienda realizar un estudio que implemente técnicas de mayor complejidad o variaciones de las técnicas escogidas que puedan adaptarse mejor a otras fuentes de datos. Además, en caso de tener conjuntos de datos con grandes cantidades de registros, se puede explorar con técnicas de aprendizaje profundo, ya que las redes neuronales de una sola capa oculta implementadas en este estudio

mostraron los menores valores de RMSE y MAE, por lo que profundizar en este tipo de técnicas puede llevar a una considerable mejora en los resultados obtenidos.

- En este estudio se tuvo en cuenta las fases del flujo reproductivo del cultivo de aguacate donde se evidenció la influencia de este comportamiento en la producción. En primer lugar, se propone profundizar en esta fase seleccionando distintos intervalos de tiempo, como por ejemplo periodos más largos donde se pueda considerar el comportamiento del cultivo a mayor escala y evaluar la evolución de la producción del cultivo durante varios años. Además, si se cuenta con datos del cultivo con una periodicidad más exacta, es posible analizar más a detalle cada una de las fases reproductivas del cultivo, no sólo considerando la fase de la floración como referencia.

  En segundo lugar, se propone implementar pruebas adicionales teniendo en cuenta otras fases del desarrollo de la planta no consideradas en este estudio, por ejemplo la del flujo vegetativo, esto con el objetivo de identificar la influencia del clima durante estas fases, lo que podría incidir posteriormente en la producción.

6. Referencias

[1] Marín, D. (2019). *La agricultura en Latinoamérica y el Caribe y las claves para su futuro* [online]. Disponible en: https://www.efe.com/efe/america/economia/la-agricultura-en-latinoamerica-y-el-caribe-las-claves-para-su-futuro/20000011-4103838

[2] Peinado, X. (2020). ¿Existe un 'boom' del aguacate en Europa?. Forbes México. [online]. Disponible en: https://www.forbes.com.mx/economia-existe-boom-aguacate-europa/

[3] J. Guillermo Ramirez-Gil et al. (2018), "Potential geography and productivity of "Hass'' avocado crops in Colombia estimated by ecological niche modeling," Sci. Hortic. (Amsterdam)., vol. 237, pp. 287–295.

[4] F. E. Martínez et al. (2015). "Aptitud agroclimática e identificación de nichos productivos de bajo riesgo de déficit hídrico para aguacate en fresno, Colombia."

[5] J. A. A. Bartoli. (2008) *Manual técnico del cultivo del aguacate hass. Fundación Hondureña de Investigación Agrícola*. [Online]. Disponible en: https://www.avocadosource.com/books/AlfonsoJose2008.pdf

[6] Del Pedregal, J.S. (2018). Funcionamiento fisiológico del aguacate en condiciones de cambio climático. INIA. [online]. Disponible en: https://www.icia.es/icia/download/Aguacate/04.pdf

[7] Tamayo A., Cordoba O., Londoño .M. (2008), "Tecnología Para El Cultivo Del Aguacate". Centro de investigación CORPOICA. Rionegro, Antioquia, Colombia. pp 45-48.

[8] A. Álvarez-Bravo et al (2017), "Escenarios de cómo el cambio climático modificará las zonas productoras de aguacate 'Hass' en Michoacán," Rev. Mex. ciencias agrícolas, vol. 8, no. SPE19, pp. 4035–4048.

[9] B. Kitchenham et al. (2009) "Systematic literature reviews in software engineering – A systematic literature review," Inf. Softw. Technol., vol. 51, no. 1, pp. 7–15.

[10] Xu, T., Guan, K., Peng, B., Wei, S., & Zhao, L. (2021). Machine Learning-Based Modeling of Spatio-Temporally Varying Responses of Rainfed Corn Yield to Climate, Soil, and Management in the U.S. Corn Belt. *FRONTIERS IN ARTIFICIAL INTELLIGENCE*, *4*. https://doi.org/10.3389/frai.2021.647999

[11] Li, L., Wang, B., Feng, P., Li Liu, D., He, Q., Zhang, Y., Wang, Y., Li, S., Lu, X., Yue, C., Li, Y., He, J., Feng, H., Yang, G., & Yu, Q. (2022). Developing machine learning models with multi-source environmental data to predict wheat yield in China. *Computers and Electronics in Agriculture*, *194*, 106790. https://doi.org/https://doi.org/10.1016/j.compag.2022.106790

[12] Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Bryant, C. R., & Senthilnath, J. (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. *ECOLOGICAL INDICATORS*, *120*. https://doi.org/10.1016/j.ecolind.2020.106935

[13] Kim, N., Ha, K.-J., Park, N.-W., Cho, J., Hong, S., & Lee, Y.-W. (2019). A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006-2015. *ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION*, *8*(5). https://doi.org/10.3390/ijgi8050240

[14] dos Santos Luciano, A. C., Picoli, M. C. A., Duft, D. G., Rocha, J. V., Leal, M. R. L. V., & le Maire, G. (2021). Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. *Computers and Electronics in Agriculture*, *184*, 106063. https://doi.org/https://doi.org/10.1016/j.compag.2021.106063

[15] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, *11*(6). https://doi.org/10.1371/journal.pone.0156571

[16] Ma, Y., Zhang, Z., Kang, Y., & Ozdogan, M. (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *REMOTE SENSING OF ENVIRONMENT*, *259*. https://doi.org/10.1016/j.rse.2021.112408

[17] Sun, J., Lai, Z., Di, L., Sun, Z., & Tao Jianbin and Shen, Y. (2020). Multilevel Deep Learning Network for County-Level Corn Yield Estimation in the US Corn Belt. *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE*, *13*, 5048–5060. https://doi.org/10.1109/JSTARS.2020.3019046

[18] Hague, F. F., Abdelgawad, A., Yanambaka, V. P., & Yelamarthi, K. (2020). Crop Yield Prediction Using Deep Neural Network. *2020 IEEE 6TH WORLD FORUM ON INTERNET OF THINGS (WF-IOT)*.

[19] Hague, F. F., Abdelgawad, A., Yanambaka, V. P., & Yelamarthi, K. (2020). Crop Yield Prediction Using Deep Neural Network. *2020 IEEE 6TH WORLD FORUM ON INTERNET OF THINGS (WF-IOT)*.

[20] Srivastava, A. K., Safaei, N., Khaki, S., Lopez Gina and 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*(1). https://doi.org/10.1038/s41598-022-06249-w

[21] Lischeid, G., Webber, H., Sommer, M., & Nendel Claas and Ewert, F. (2022). Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. *AGRICULTURAL AND FOREST METEOROLOGY*, *312*. https://doi.org/10.1016/j.agrformet.2021.108698

[22] Feng, P., Wang, B., Liu, D. L., Xing, H., Ji Fei and Macadam, I., Ruan, H., & Yu, Q. (2018). Impacts of rainfall extremes on wheat yield in semi-arid cropping systems in eastern Australia. *CLIMATIC CHANGE*, *147*(3–4), 555–569. https://doi.org/10.1007/s10584-018-2170-x

[23] Yin, X., Leng, G., & Yu, L. (2022). Disentangling the separate and confounding effects of temperature and precipitation on global maize yield using machine learning, statistical and process crop models. *ENVIRONMENTAL RESEARCH LETTERS*, *17*(4). https://doi.org/10.1088/1748-9326/ac5716

[24] Maitah, M., Malec, K., Ge, Y., Gebeltova, Z., Smutka, L., Blazek, V., Pankova, L., & Maitah Kamil and Mach, J. (2021). Assessment and Prediction of Maize Production Considering Climate Change by Extreme Learning Machine in Czechia. *AGRONOMY-BASEL*, *11*(11). https://doi.org/10.3390/agronomy11112344

[25] Murakami, K., Shimoda, S., Kominami, Y., Nemoto, M., & Inoue, S. (2021). Prediction of municipality-level winter wheat yield based on meteorological data using machine learning in Hokkaido, Japan. *PLOS ONE*, *16*(10). https://doi.org/10.1371/journal.pone.0258677

[26] Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid Deep Learning-based Models for Crop Yield Prediction. *APPLIED ARTIFICIAL INTELLIGENCE*, *36*(1). https://doi.org/10.1080/08839514.2022.2031823

[27] Koob Endowment, R. D., Student, I., Woodson, M., & Ross, C. (2017). *Automated Avocado Yield Forecasting Using Multi-Modal Imaging II.* http://digitalcommons.calpoly.edu

[28] Mokria, M., Gebrekirstos, A., Said, H., Hadgu, K., Hagazi, N., Dubale, W., & Bräuning, A. (2022). Fruit weight and yield estimation models for five avocado cultivars in Ethiopia. *Environmental Research Communications*, *4*(7). https://doi.org/10.1088/2515-7620/ac81a4

[29] Ramírez-Gil, J. G., Morales, J. G., & Peterson, A. T. (2018). Potential geography and productivity of “Hass” avocado crops in Colombia estimated by ecological niche modeling. *Scientia Horticulturae*, *237*, 287–295. https://doi.org/10.1016/j.scienta.2018.04.021

[30] Zamet, D. N. (1990). The Effect of Minimum Temperature on Avocado Yields. In *California Avocado Society* (Vol. 74).

[31] Robson, A., Rahman, M. M., & Muir, J. (2017). Using worldview satellite imagery to map yield in avocado (Persea americana): A case study in Bundaberg, Australia. *Remote Sensing*, *9*(12). https://doi.org/10.3390/rs9121223

[32] Lozano Vásquez, C. A., Esteban, J., & Cabrera, S. (2019). *Analítica de datos para el rendimiento en los cultivos de aguacate Hass en Colombia*.

[33] R. Wirth and J. Hipp. (2000) “CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39,” Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min., no. 24959, pp. 29–39, [Online]. Disponible en: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining.

[34] C. Díaz & J. Bernal. (2008) Tecnología para el cultivo del aguacate. [online]. Disponible en: http://hdl.handle.net/20.500.12324/13459.

[35] I. E. Paz, V. F. Teran, V. Narvaez. (2010) *Manejo técnico para la producción de aguacate (Persea americana miller) variedad Hass en la meseta de Popayán.* Volumen 1. Popayán: Sello Editorial UNICAUCA,

[36] Salazar, S., Medina, R. E., & Álvarez, A. (2016). Evaluación inicial de algunos aspectos de calidad del fruto de aguacate ‘Hass’ producido en tres regiones de México. Revista Mexicana de Ciencias Agrícolas, 7(2), 277- 289. Retrieved from

http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-09342016000200277

[37] Reyes-Alemán, Juan Carlos, Mejía-Carranza, Jaime, Monteagudo-Rodríguez, Omar Ricardo, Valdez-Pérez, María Eugenia, González-Diaz, Justino Gerardo, & Espíndola-Barquera, María de la Cruz. (2021). Fenología del aguacate 'Hass' en el Estado de México, México. Revista Chapingo. Serie horticultura, 27(2), 113-134. Epub 13 de diciembre de 2021.https://doi.org/10.5154/r.rchsh.2020.09.020

[38] Campos, O. (2012). Zonificación agroecológica del aguacate (Persea americana Mill. var. 'Hass') en la cuenca del río Duero. Instituto politécnico nacional. Jiquilpan, Michoacán, México. 36 pp.

[39] Gazit, S. y Degani, C. (2002). Reproductive Biology. En A. W. Wiley; B. Schaffer and B. N. Wolstenholme (Eds.), Avocado: Botany, Production and Uses (pp. 101-134). Cambridge, MA: CABI Disponible en. https://books.google.com.co/books/about/The_Avocado.html?id=CxmvpAYkL54C&redir_esc=y

[40] Vásquez Amariles, H, Saavedra, R, Marín Beitia, E, Guerrero Cobos, J y Sánchez, C. (2022). Cartilla de estados fenológicos-tipo en aguacate Hass para la localidad de Roldanillo, Valle del Cauca. Universidad Nacional de Colombia.

[41] C. A. Diaz, J. A. Bernal & Á. d. Tamayo. (2020) Ecofisiología del aguacate cv. Hass en el trópico andino colombiano. [online]. Disponible en: http://hdl.handle.net/20.500.12324/36875.

[42] J. Bernal, L. Vazquez, J. Cartagena. (2017) *Fenología del aguacate cv. hass plantado en diversos ambientes del departamento de antioquia, colombia.* Memorias del V Congreso Latinoamericano del Aguacate. Ciudad Guzmán, Jalisco, México. [Online]. Disponible en: https://www.avocadosource.com/Journals/Memorias_VCLA/2017/Memorias_VCLA_2017_PG_292.pdf

[43] A. Jaramillo. (2016) *Épocas recomendadas para la siembra de café en Colombia.* Avances técnicos Cenicafé Cenicafé, Manizales, .

[44] Bárcenas O., A.E.; N.A. Martínez, P.S. Aguirre, y C.P. Castro. (2002). Fenología del aguacate (Persea americana Mill.) var. Hass en cuatro diferentes altitudes del municipio de Uruapan, Michoacán. Revista Divulga 5:23-30.

[45] P. Mandeville, "Tema 24: Observaciones perdidas," CIENCIA-UANL, ISSN 1405-9177, Vol. 13, No. 3, 2010, pags. 313-324, 01 2010.

[46] C. Jimenez. (2022) *Aguacate, el 'oro verde' de Fresno.* El nuevo día. [online]. Disponible en: https://www.elnuevodia.com.co/nuevodia/tolima/476762-aguacate-el-oro-verde-de-fresno

[47] Ministerio de agricultura y desarrollo rural, Junio de 2021. (2021). Cadena productiva del aguacate.

[48] Castellanos, Juan. (2012) “Método de detección temprana de outliers”, Pontificia Universidad Javeriana. Disponible en: https://repository.javeriana.edu.co/bitstream/handle/10554/10347/MorenoCastellanosJuanGabriel2012.pdf

[49] Ramírez-Gil, J. G., Morales, J. G., & Peterson, A. T. (2018). Potential geography and productivity of “Hass” avocado crops in Colombia estimated by ecological niche modeling. *Scientia Horticulturae*, *237*, 287–295. https://doi.org/10.1016/j.scienta.2018.04.021

[50] R. García, (2021) et al. Rendimiento, calidad y comportamiento poscosecha de frutos de aguacate ‘Hass’ de huertos con diferente fertilización. Rev. Mex. Cienc. Agríc [online]. vol.12, pp.205-218. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-09342021000200205&lng=es&nrm=iso.

[51] S. Salazar, L. E. Cossio & I. Gonzalez. (2009) *La fertilización de sitio específico mejoró la productividad del aguacate 'Hass' en huertos sin riego*. Agric. Téc. Méx [online], vol.35, n.4, pp.439-448. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0568-25172009000400009&lng=es&nrm=iso

[52] Pérez-Planells, Ll, Delegido, J, Rivera-Caicedo, J, Verrelst, J. (2015) “Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos”, REVISTA DE TELEDETECCIÓN. http://dx.doi.org/10.4995/raet.2015.4153.

[53] Adnan M, Alarood AAS, Uddin MI, ur Rehman I. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. PeerJ Computer Science 8:e803 https://doi.org/10.7717/peerj-cs.803

[54] Cross-validation: evaluating estimator performance. Scikit-Learn. [online]. Disponible en: https://scikit-learn.org/stable/modules/cross_validation.html

[55] Management Solutions (2023) “Explainable artificial intelligence (XAI) Desafíos en la interpretabilidad de los modelos”. [online], pp.23. Disponible en: https://www.managementsolutions.com/sites/default/files/minisite/static/22959b0f-b3da-47c8-9d5c-80ec3216552b/iax/pdf/explainable-artificial-intelligence-sp.pdf

7. Anexos

7.1. Anexo A: Resultados de la revisión sistemática de literatura

Se muestra un hoja de cálculo que contiene la recolección de los datos de los artículos de la revisión sistemática de literatura y el resultado del proceso de evaluación de la calidad, donde está la calificación asignada a cada uno de los artículos leídos. Además, se encuentra el análisis cuantitativo de los datos, con los conteos de los datos recolectados y las gráficas construidas a partir de ello.

Disponible en:

https://drive.google.com/drive/folders/1756zxSOFudoUECEERx5NyQaCm-t7VXMr?usp=drive_link

7.2. Anexo B: Conjuntos de datos recolectados y generados

Contiene tres carpetas que organizan los distintos conjuntos de datos recolectados y generados de la siguiente manera:

- Una carpeta contiene los datos recolectados, los datos meteorológicos del IDEAM y del proyecto POWER de la NASA, y los datos de producción de Agronet, también se añade el conjunto de datos de producción de datos abiertos, que finalmente fue descartado.
- Una carpeta con los datos fusionados, aquí se encuentran los conjuntos de datos meteorológicos en sus distintas fases de filtrado, primero con datos repetidos, luego con los datos repetidos eliminados y reemplazados con los del IDEAM, pero con campos faltantes, y finalmente los tres conjuntos de datos con datos imputados en los campos faltantes, con las tres técnicas seleccionadas.
- Una carpeta con los conjuntos de datos seleccionados por la fenología generados a partir de las siete pruebas establecidas, aquí se encuentran los veintiún conjuntos de datos que pasaron a ser modelados, están los datos de las siete pruebas para cada una de las tres técnicas de imputación seleccionadas

Disponible en:

https://drive.google.com/drive/folders/15c9YRfUJEEaTQNmFPvFJj1Yi0-BB0eeR?usp=drive_link

7.3. Anexo C Código fuente de la preparación de los datos y la implementación de los modelos

Este anexo contiene los códigos en python y notebooks de Jupyter que se utilizaron para la recolección y organización de los datos, los del análisis exploratorio de los datos y la limpieza de los mismos.

Luego, se muestra el código con el proceso de caracterización de los datos meteorológicos de acuerdo a las 7 pruebas establecidas, y finalmente se encuentra el código con la implementación de los 4 modelos de aprendizaje automático.

Disponible en:

7.4. Anexo D Resultados de las estimaciones de los modelos

Contiene los resultados obtenidos por los modelos en una hoja de cálculo, aquí se muestran los resultados individuales de los 84 modelos generados con las tres métricas de evaluación seleccionadas.

En una segunda página de la hoja de cálculo se muestran las agrupaciones realizadas con el fin de crear las gráficas de resultados mostradas en la sección de evaluación de resultados del documento. Por último, en una hoja se hace la comparación en detalle de las dos técnicas que mostraron mejores resultados, en este caso el bosque aleatorio RF y la red neuronal MLP.

Disponible en:

https://drive.google.com/drive/folders/1mIU21Qez9xpUbXXgiYbCYv0gjWkBaDKV?usp=drive_link