DIGITALIZACIÓN DEL SECTOR CACAOCULTOR EN EL MUNICIPIO DE BARAYA - HUILA.

Juan Camilo Salas Diaz
Código: 8992709

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Directora:
Yady Tatiana Solano Correa

Codirector(a)
Mario Milver Patiño Velasco

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 1 DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

La agricultura enfrenta crecientes desafíos en términos de sostenibilidad, competitividad y optimización de recursos naturales. La integración de tecnologías avanzadas como machine learning y el análisis de datos multiespectrales se presenta como una solución prometedora [1]. El cultivo de cacao, con su relevancia histórica y económica en Colombia, destaca como un sector estratégico para aplicar estos avances. Colombia posee aproximadamente dos millones de hectáreas aptas para el cultivo de cacao, lo que lo consolida como uno de los principales productores mundiales.

A nivel nacional, el sector cacaocultor enfrenta una serie de desafíos relacionados con la eficiencia y la sostenibilidad de la producción. Estos desafíos se hacen particularmente evidentes en regiones como el departamento del Huila, que ha mostrado un incremento en la producción de cacao en los últimos años, gracias a mejoras en los rendimientos por hectárea. Sin embargo, el área cosechada ha disminuido significativamente, lo cual manifiesta la necesidad de implementar estrategias que promuevan la sostenibilidad y el aprovechamiento de las potencialidades del sector [1].

En respuesta a estas necesidades, el departamento del Huila ha adoptado propuestas de investigación como el proyecto denominado "Ventajas Comparativas para el Subsector Cacao", ejecutado por el Centro de Investigaciones en Ciencias y Recursos Geo-Agroambientales (CENIGAA), en cooperación con la Gobernación del Huila, la Universidad Surcolombiana, el SENA y el Consejo Superior de Investigaciones de España (CSIC). Este proyecto macro tiene como objetivo proporcionar una solución integral desde el enfoque AGROTECH, abarcando desde la generación de la línea base del sector cacaocultor hasta la comprensión y descripción detallada del mismo, mediante la integración de conceptos científicos y tecnológicos.

El presente proyecto de grado se deriva de la propuesta macro descrita anteriormente, con el fin de brindar una alternativa metodológica que permita comprender las características similares de las unidades productivas de cacao a través de la creación de una base de datos compuesta por datos sociales, agronómicos, física de suelo e imágenes multiespectrales; capturados en el proyecto macro. Junto con la aplicación de modelos de *machine learning,* se busca facilitar el entendimiento, la visualización y la descripción de los grupos de unidades productivas de cacao que comparten características similares. Todo esto contribuirá a una gestión del territorio basada en información precisa, lo que permitirá fortalecer la producción de cacao de una manera sostenible y eficiente.

Para el desarrollo de los objetivos planteados en esta propuesta, el presente documento describe la metodología aplicada que abarca desde el análisis exploratorio de las diferentes fuentes de información, hasta la estructuración y creación de la base de datos. Asimismo, se detalla el procesamiento y la generación de índices de vegetación a partir de imágenes multiespectrales, la aplicación de técnicas de *machine learning* para la identificación de patrones, y finalmente la visualización e interpretación de los resultados obtenidos a través de los modelos implementados.

1 DEFINICIÓN DEL PROBLEMA

1.1 Definición del problema

El cultivo de cacao en Colombia representa una importante oportunidad de desarrollo agrícola, con un potencial de aproximadamente dos millones de hectáreas [2], lo cual posiciona al país como uno de los principales productores de cacao del mundo. Sin embargo, a pesar de las condiciones edafoclimáticas favorables que permiten producir cacao de sabor y fino aroma, el sector enfrenta importantes desafíos, especialmente en el departamento del Huila, donde el área de cultivo ha disminuido considerablemente en los últimos años debido a la falta de integracion tecnologica, asi como tambien los bajos precios, las plagas, enfermedades y los daños a asociados a factores climaticos de la region [3].

En el Huila, el área cosechada de cacao pasó de 9.188,5 hectáreas en 2007 a 6.488 hectáreas en 2016, una disminución del 29,4%. A pesar de esta reducción en la superficie cultivada, la producción aumentó un 6,4% debido al incremento en los rendimientos por hectárea desde 2010 [3]. Esto evidencia un sector resiliente, pero también expone la necesidad de mejorar las prácticas productivas para garantizar su sostenibilidad y competitividad a largo plazo.

Aunque el departamento del Huila es conocido por su destacada y exquisita producción de cacao, esta enfrenta desafíos en la optimización, competitividad y sostenibilidad de sus prácticas productivas [1]. A pesar de su perfil sensorial diferencial en sabor y aroma, la falta de un análisis de datos integrados del sector limita la capacidad de los productores y de los entes gubernamentales para tomar decisiones efectivas. Esta situación subraya la necesidad de un enfoque innovador que incorpore tecnologías avanzadas [2] en el aprovechamiento de la información detallada y específica de cada municipio, así como para realizar análisis de datos que fortalezcan la planificación y el direccionamiento estratégico del sector cacaocultor. Actualmente, no se cuenta con un modelo analítico avanzado que integre y procese eficazmente los datos recopilados en el marco del proyecto de investigación ***"Ventajas Comparativas para el Subsector Cacao"*** cuyo desarrollo metodológico requiere una agrupación de predios con características similares**.** Estos datos incluyen variables socio agronómicas, muestras de suelo e imágenes multiespectrales tomadas por drones, que permiten clasificar y comprender de manera exhaustiva las unidades productivas de cacao. La implementación de un enfoque basado en técnicas de machine learning, tiene el potencial de revelar patrones ocultos y agrupar predios con características similares, lo cual facilitaría la identificación de fortalezas y debilidades específicas en cada unidad productiva y contribuiría a una gestión más informada y eficiente del sector cacaocultor en departamento del Huila.

Según la FAO, estos planteamientos pueden ser abordados desde la óptica de la ciencia de datos, esencialmente para trascender la tradicional gestión agrícola y avanzar hacia una agricultura de precisión que no solo busque aumentar la producción, sino también maximizar la calidad del cacao, haciendo uso efectivo de las ventajas comparativas de Colombia. Al integrar y analizar datos multidimensionales, este trabajo pretende proporcionar una base sólida para políticas y estrategias que promuevan la competitividad y sostenibilidad del sector productivo, abriendo nuevos mercados y oportunidades de valor agregado [4].

1.2 Formulación del problema

Con la integración de técnicas de *machine learning* y el análisis de datos socio-agronómicos y multiespectrales recolectados mediante sensores acoplados a drones, se busca identificar patrones y factores críticos que influyen en la productividad y calidad del cacao. Este enfoque permite desarrollar modelos que faciliten la toma de decisiones agrícolas del sector más informadas que optimicen el uso de los recursos y mejoren la cadena de valor del cacao en el departamento, transformando la producción local en una más competitiva [5], a través de este planteamiento se procede a orientar la lectura de la pregunta de investigación y se genera el siguiente interrogante compuesto:

¿Es posible aplicar técnicas de machine learning para identificar patrones y agrupar unidades productivas de cacao con características similares en el municipio de Baraya-Huila, para facilitar la gestión informada y diseñar estrategias que mejoren la competitividad y sostenibilidad del sector cacaocultor?

1.2.1 Preguntas de Sistematización

- ¿Cómo se puede diseñar de manera estructurada una base de datos que contenga datos sociales, agronómicos y de imágenes multiespectrales que permitan diseñar estrategias que mejoren la competitividad y la sostenibilidad del sector cacaocultor?
- ¿Qué patrones se pueden identificar entre las diferentes unidades productivas de cacao al aplicar técnicas de machine learning y cómo estas influyen en la sostenibilidad del sector?
- ¿Cómo se pueden evaluar los agrupamientos generados mediante modelos de machine learning utilizando métricas específicas de validación de clustering, para determinar la similitud de las características?
- ¿Cómo se pueden interpretar los agrupamientos de características similares generados en relación con la gestión y la toma de decisiones del sector cacaocultor?
- ¿Cómo se pueden visualizar los resultados de los agrupamientos de característica similares para facilitar la interactividad de los patrones y relaciones entre las unidades productivas de cacao?

2 OBJETIVOS DEL PROYECTO

2.1 Objetivo General

Identificar las características similares de las unidades productivas de cacao del municipio de Baraya – Huila, mediante técnicas de machine learning, que faciliten la gestión informada del sector cacaocultor a través de datos socio-agronómicos y de imágenes multiespectrales.

2.2 Objetivos Específicos

1. Crear una base de datos estructurada compuesta por datos sociales, agronómicos e imágenes multiespectrales que facilite la descripción de sector cacaocultor.
2. Desarrollar modelos de machine learning que permita identificar patrones y características similares entre las unidades productivas de cacao que influyen en la sostenibilidad del sector.
3. Evaluar los agrupamientos de características similares generados a partir de la aplicación de modelos de machine learning, mediante métricas de evaluación específicas para clustering.
4. Interpretar descriptivamente los agrupamientos de características similares generados a partir de la aplicación del modelo de machine learning.
5. Desarrollar una estrategia de visualización interactiva que facilite la interpretación de patrones y relaciones entre las unidades productivas de cacao en el municipio de Baraya.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 Marco teórico

A continuación, se presentan los temas que se relacionan con el desarrollo del proyecto, teniendo en cuenta la conceptualización base e implementación de las tecnologías de la Información y la Comunicación (TIC) en la agricultura, de lo que se denomina soluciones AGROTECH [5], un concepto que refiere al manejo de grandes volúmenes de datos agrícolas. Este manejo incluye la recolección, almacenamiento y análisis de datos para generar insights que apoyen la toma de decisiones eficaces en el sector agrícola. Dentro de este contexto, la ciencia de datos emerge como un campo crucial, proporcionando las herramientas y métodos necesarios para procesar y extraer valor de grandes conjuntos de datos [6].

3.1.1 Sistemas agroforestales

Un sistema agroforestal o SAF, es un área donde se combina un cultivo principal con otros cultivos, árboles y, en algunas ocasiones, animales. Las ventajas de un SAF van desde la optimización del suelo pasando por el aumento de los ingresos económicos de la familia productora, hasta la conservación del medio ambiente [7] [8].

- ***Cacao como sistema agroforestal:*** El establecimiento de cultivos semestrales y anuales dentro de áreas de cacao, permiten reducir costos de establecimiento y manejo en los primeros años de vida del cultivo debido a que producen en pocos meses y parte de la producción puede venderse para garantizar el manejo y enfrentar otras demandas del cultivo de cacao [9].

El cacao es un árbol procedente de América de nombre científico Tehobroma Cacao que produce un fruto del mismo nombre que se puede utilizar como ingrediente para alimentos entre los que destaca el chocolate [10]. Su uso se remonta a la época de los mayas, aztecas e incas, y desde entonces se ha usado tanto para fines nutricionales como médicos. Por lo general se dan dos cosechas de cacao al año: una hacia el final de la época lluviosa y el inicio de la seca, y otra al principio del siguiente período de lluvias [11].

3.1.2 Agricultura sostenible y de precisión

La agricultura sostenible satisface las necesidades de las generaciones presentes y futuras, alineadas con los objetivos de desarrollo sostenible que al mismo tiempo garantizan la rentabilidad, la salud ambiental, y la equidad social y económica [12]. Mientras que la agricultura de precisión se basa en técnicas que tienen en cuenta las particularidades en el desarrollo de los cultivos, el estado del suelo o los factores climáticos y se aleja de aplicaciones más tradicionales y homogéneas. Su meta es mejorar la eficiencia de la producción haciendo uso de herramientas tecnológicas como el internet de las cosas, big data, analítica e inteligencia artificial [13].

3.1.3 Drones e imágenes Multiespectrales en la agricultura

Los drones son vehículos aéreos no tripulados que tienes aplicaciones en diferentes sectores. Para la

agricultura son utilizados para transportar cámaras o sensores que capturan imágenes para optimizar el manejo de la variabilidad espacial y temporal del cultivo con sus factores de producción [14].

Las imágenes multiespectrales son capturadas a través de cámaras especiales con la capacidad de capturar una imagen por cada ancho de banda por separado. Estas cámaras generalmente captan información entre 3 y 7 bandas del espectro electromagnético, cada uno de unos 100 nanómetros [nm] de ancho. Entre las bandas espectrales que determinan están la verde (550 nm de longitud de onda, ancho de banda de 40 nm, Green), La banda roja (con un centro de banda de 735 nm y un ancho de banda de 10 nm, correspondiente a la región del **Red Edge**) y la banda del infrarrojo cercano (**NIR**, con longitud de onda de 790 nm y un ancho de banda de 40 nm) son fundamentales para el análisis de la vegetación [14]. En la Figura 1 se ilustran los anchos de banda que componen las imágenes multiespectrales utilizadas, así como las firmas espectrales asociadas al estado de salud vegetal. Por ejemplo, la curva de color negro representa la firma espectral de una vegetación enferma o estresada, mientras que la curva verde corresponde a una vegetación sana.

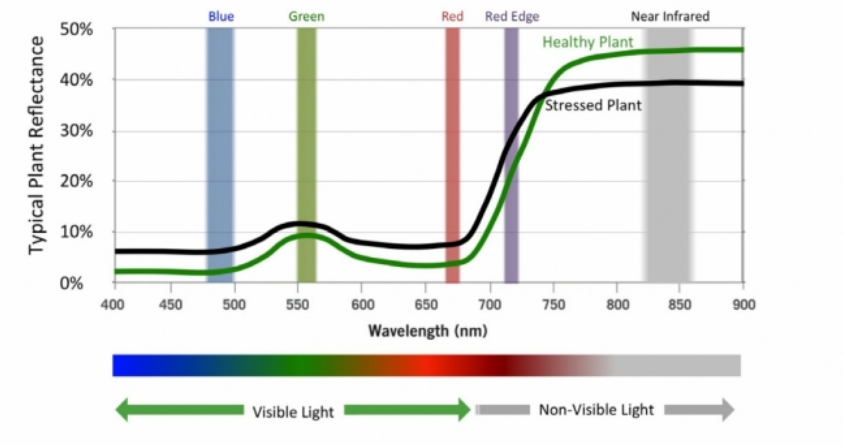


Figura 1. Ancho de banda de imágenes multiespectrales. (adaptado de [15]).

Un índice de vegetación es el resultado de una fórmula algebraica que utiliza una o varias bandas del espectro electromagnético. La relación entre estas bandas tiene detrás un estudio empírico que demuestra la relación directa entre el valor numérico captado por la cámara y la variable de la planta a medir (normalmente biomasa o sanidad vegetal) [15]. Para definir los píxeles de la cobertura se utilizan los índices de vegetación que se clasifican de la siguiente manera: Índices básicos de vegetación, Índices que abordan la reflectancia del suelo e Índices agrícolas [16].

Los índices de vegetación más relevantes son:

- ***NDVI:*** Es el acrónimo en inglés de Normalized Difference Vegetation Index, este índice de vegetación funciona con sensores de infrarrojo cercano (NIR), analizando la respuesta espectral de las plantas en las bandas roja e infrarroja cercana. NDVI es utilizado para monitorear cultivos, detectar déficit hídrico y daños por plagas.

El índice de vegetación NDVI se calcula mediante la fórmula:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \quad (1)$$

Donde, NIR es la reflectancia en el rango del infrarrojo cercano y RED es la reflectancia en la banda roja.

- ***NDRE***: Evalúa el contenido de clorofila en las plantas, así como su absorción de nitrógeno y demanda de fertilizantes, proporcionando un mejor análisis de estados fenológicos más avanzados, ya que consigue obtener datos con mayor profundidad.

$$NDRE = \frac{(NIR - REDEdge)}{(NIR + REDEdge)} \quad (2)$$

Red Edge es la reflectancia al borde rojo.

3.1.4 Píxeles y datos geoespaciales

Un píxel es la unidad fundamental de una imagen, la cual está compuesta por un conjunto de valores numéricos. Estos píxeles poseen propiedades tanto geométricas como radiométricas. En su forma más básica, un píxel se representa mediante un número que describe el brillo y el color en un punto específico de la imagen [17]. Su valor y la tonalidad que muestra están relacionados con su radiancia y la manera en que esta interactúa con la luz y el entorno, lo cual fundamenta su representación geométrica y escalar.

En el caso de las imágenes multiespectrales e hiperespectrales, cada magnitud de un píxel para cada banda corresponde a un perfil espectral particular. Así, cada tipo de material (como suelo, vegetación o superficies diversas) presenta un perfil espectral único, denominado “firma espectral”.

El valor de un píxel es, esencialmente, la medición de la radiación electromagnética asociada a una ubicación específica. Esta propiedad se aprovecha en teledetección para analizar fenómenos en zonas de interés. Por ello, una imagen digital se considera, en última instancia, una medición tanto radiométrica como fotogramétrica [18].

El nivel de detalle de una imagen, o su resolución, es directamente proporcional al tamaño del píxel. Este debe ajustarse de modo que capture la información necesaria y, al mismo tiempo, permita un almacenamiento eficiente. Al emplear píxeles más pequeños, se logra una mayor definición o resolución en la representación de la escena, pero también se generan conjuntos de datos más complejos, lo cual aumenta los requerimientos de almacenamiento y de procesamiento, por ejemplo, en la generación de ortomosaico y en el cálculo de índices de vegetación.

Al trabajar con datos ráster, además de la resolución espacial, es importante considerar otras tres resoluciones:

1. **Resolución espectral**: se relaciona con la capacidad de un sensor para distinguir diferentes longitudes de onda dentro del espectro electromagnético. Cuanto mayor sea la resolución espectral, más estrecha será la banda de longitud de onda que abarca cada canal.

2. **Resolución temporal**: hace referencia a la frecuencia con la que se capturan imágenes de una misma zona de interés, lo que resulta como base para el monitoreo de fenómenos que cambian en el tiempo.

3. **Resolución radiométrica**: describe la capacidad de un sensor multiespectral para diferenciar objetos que reflejan o emiten energía en la misma región del espectro electromagnético. Cuantos más bits posea una imagen, más niveles de intensidad o brillo se podrán representar, facilitando así la detección de diferencias sutiles entre los objetos [28].

3.1.5 Análisis físicos y químicos de suelos

La física del suelo es el estudio de las fases sólida, líquida y gaseosa de los suelos, y sus interacciones. La textura, estructura y densidad aparente del suelo reflejan cómo se combinan las partículas minerales y orgánicas del suelo para formar la matriz del suelo y los espacios porosos [19], los parámetros físicos de suelo evaluados en análisis físicos de suelos son clase textural del suelo, densidad aparente-real y porosidad total.

La química del suelo es una rama de la ciencia del suelo que se ocupa de la composición química, las reacciones y las propiedades químicas en los suelos. Como la capacidad de intercambio catiónico (CIC), pH, potencial redox y la conductividad eléctrica son importantes debido a la influencia en la disponibilidad de nutrientes, el crecimiento de las plantas, el destino de los contaminantes y la actividad biológica [20].

3.1.6 Integración de modelos

Los modelos mixtos que combinan enfoques basados en biofísica y en datos son especialmente relevantes en la agricultura, donde tanto las variables físicas como las observaciones empíricas juegan un papel crítico. La integración de estos modelos permite un análisis más holístico y detallado del comportamiento de los sistemas agrícolas, mejorando la precisión en la modelación y la efectividad de las soluciones propuestas.

El uso de estos modelos no solo se enfoca en la precisión teórica sino también en su aplicabilidad práctica. La visualización de datos, por ejemplo, juega un papel crucial al transformar los resultados del análisis de datos en representaciones gráficas que son fácilmente interpretables por los agricultores y los planificadores. Esto es esencial para que los *insights* generados por técnicas de ciencia de datos sean accesibles y utilitarios.

3.1.7 Agricultura 4.0

La ciencia de datos aplicada a la agricultura extrae conocimiento valioso a partir de grandes volúmenes de información provenientes de diversas fuentes. Este proceso se sustenta en componentes clave como la detección, el monitoreo, la recopilación y transmisión de datos, así como en los sistemas de apoyo a la toma de decisiones. En este contexto, los datos geoespaciales adquieren un papel fundamental dentro de la denominada agricultura 4.0, al permitir una gestión contextualizada del sector basada en variables como el clima, el suelo, los cultivos, las etapas fenológicas y los rendimientos.

Sin embargo, el carácter heterogéneo, volumétrico y dinámico de los datos en la agricultura dificulta su procesamiento y análisis. A medida que los sensores se integren en los entornos productivos, los procesos agrícolas se convierten progresivamente a sistemas guiados por datos (data-driven). Los avances en el Internet de las Cosas (IoT) y la computación en la nube han acelerado esta transformación, consolidando la agricultura 4.0 [21].

A diferencia de la agricultura de precisión que se enfoca principalmente en gestionar la variabilidad espacial dentro del campo, la agricultura 4.0 amplía este enfoque al incorporar información contextual y situacional en tiempo real, guiada por eventos dinámicos como alertas meteorológicas o brotes de enfermedades [22]. En este escenario, se requieren sistemas capaces de brindar asistencia adaptativa y reconfigurable en tiempo real, que apoyen la implementación, el mantenimiento y el uso óptimo de la tecnología, permitiendo acciones ágiles ante condiciones cambiantes.

En ese sentido, la ciencia de datos y la inteligencia artificial (IA) ofrecen herramientas decisivas para maximizar el potencial de esta información. Los distintos campos de la IA permiten monitorear cultivos, detectar plagas y enfermedades, y predecir rendimientos con alta precisión. Gracias a ello, los agricultores pueden optimizar la asignación de recursos, reducir el desperdicio y aumentar la productividad, integrando la IA en procesos de toma de decisiones basados en datos específicos de cada sistema productivo [23].

3.1.8 Machine learning

El aprendizaje automático (machine learning) hace parte de una de las ramas de la inteligencia artificial, mediante la cual es posible desarrollar algoritmos capaces de aprender de los datos para ejecutar tareas específicas como la clasificación o la predicción. Este proceso requiere disponer de un conjunto de datos que permita explorar relaciones, identificar correlaciones y descubrir patrones que reflejen el comportamiento del caso de estudio. En función del tipo de información disponible y del propósito, el aprendizaje automático se clasifica principalmente en supervisado**,** no supervisado y por refuerzo, siendo los dos primeros los de mayor aplicación [24]. A continuación, se describe el enfoque de aprendizaje supervisado y no supervisado, que se centran en entrenar modelos con datos capaces de generar predicciones o clasificaciones.

Clasificación Supervisada: Utiliza datos etiquetados para entrenar algoritmos capaces de categorizar nuevos datos o datos desconocidos. A medida que se introducen datos en el modelo, este ajusta sus ponderaciones hasta lograr una configuración óptima, lo cual se verifica mediante un proceso de validación cruzada. El aprendizaje supervisado ayuda a las resolver una variedad de problemas del mundo real a escala, siempre que se cuente con inputs etiquetados que puedan usarse como variables objetivo. A partir de esta información, el modelo ajusta sus parámetros hasta poder predecir o clasificar correctamente nuevos casos desconocidos.

Entre los algoritmos supervisados más utilizados se encuentran la regresión logística, empleada principalmente para problemas de clasificación binaria o multiclase, y los árboles de decisión, que permiten segmentar los datos mediante reglas jerárquicas basadas en atributos. Asimismo, técnicas derivadas como los bosques aleatorios (Random Forest) y los modelos de gradiente (XGBoost, LightGBM) ofrecen altos niveles de precisión al combinar múltiples árboles y ponderar su contribución en el proceso de predicción. Estos modelos, en el ámbito de las soluciones Agrotech resultan

especialmente útiles para predecir rendimientos, estimar la fertilidad del suelo o clasificar tipos de cultivos a partir de características espectrales o edáficas [25].

Clasificación No Supervisada: Se aplican en escenarios en los que los datos carecen de etiquetas predefinidas. Su objetivo es descubrir estructuras subyacentes, agrupamientos naturales o relaciones ocultas entre las observaciones. En el contexto agrícola, este enfoque es clave para identificar zonas de manejo homogéneo, clasificar unidades productivas con condiciones similares o detectar comportamientos anómalos en cultivos. Entre los algoritmos más representativos se destacan K-Means, que agrupa los datos en función de la proximidad entre observaciones utilizando la distancia euclidiana; DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**,** que identifica agrupaciones basadas en densidad y permite detectar puntos atípicos o ruido; y HDBSCAN (Hierarchical DBSCAN)**,** una versión jerárquica y más robusta que no requiere especificar el número de clústeres y es capaz de manejar conjuntos de datos complejos y heterogéneos, como los obtenidos en estudios agrícolas geoespaciales. Otros métodos relevantes incluyen el clustering jerárquico aglomerativo, útil para analizar relaciones de similitud entre parcelas o variables, y los modelos de mezcla gaussiana (GMM), que asumen una distribución probabilística en los datos para definir las agrupaciones [26].

Reducción de dimensionalidad: La reducción de dimensionalidad es una técnica fundamental en la ciencia de datos, especialmente cuando se manejan grandes volúmenes de información, como ocurre en los proyectos Agrotech. En el que se simplifican la representación de los datos al disminuir el número de variables observadas o dimensiones, sin perder la información esencial que describe el caso de estudio. En el contexto agrícola, donde los datos provienen de diversas fuentes como condiciones climáticas, propiedades físico-químicas del suelo, índices espectrales o variables socioeconómicas, la reducción de dimensionalidad resulta indispensable para filtrar el ruido y concentrarse en las variables que realmente influyen en los resultados [27].

Sin embargo, los sistemas agrícolas pueden no presentar relaciones lineales entre sus componentes. La interacción compleja entre factores bióticos, abióticos y de manejo genera comportamientos no lineales y dinámicos, lo que limita la aplicación de métodos clásicos de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA). Este tipo de técnicas asume que la variabilidad de los datos puede describirse mediante combinaciones lineales de las variables originales, lo cual no refleja adecuadamente la estructura intrínseca de los sistemas agrícolas. Por ello, resulta necesario abordar la reducción de dimensionalidad desde una perspectiva más flexible, basada en principios topológicos y estocásticos, capaces de preservar relaciones locales y globales en espacios de alta complejidad [28].

En este sentido, la topología rama de las matemáticas que estudia las propiedades del espacio que permanecen invariantes bajo deformaciones continuas proporciona una conceptualización matemática para comprender cómo se pueden proyectar los datos en espacios de menor dimensión sin perder su forma esencial. La topología parte de la idea de que el espacio no es rígido ni estático, sino elástico y transformable, permitiendo que las estructuras puedan representarse en diferentes escalas o configuraciones sin alterar su conectividad fundamental [29]. Este enfoque resulta particularmente útil en la modelación de datos agrícolas, donde las variables pueden estar relacionadas de manera no lineal o distribuida en múltiples dimensiones.

Técnicas modernas como t-SNE (t-distributed Stochastic Neighbor Embedding) y, especialmente, UMAP

(Uniform Manifold Approximation and Projection), aplican principios topológicos y de geometría diferencial para conservar la estructura de vecindad entre los datos al reducir su dimensionalidad. Estas metodologías no solo preservan las relaciones locales entre puntos similares, sino también la forma global del espacio de datos, permitiendo visualizar y analizar la organización intrínseca de la información en dos o tres dimensiones sin perder la complejidad original. De este modo, los métodos topológicos de reducción de dimensionalidad ofrecen una representación más fiel y robusta de los sistemas agrícolas, facilitando la identificación de patrones, agrupamientos y relaciones que quedarían ocultos bajo técnicas lineales tradicionales.

3.1.9 Métricas de evaluación para clustering

Una vez generadas las representaciones más compactas y coherentes de los datos mediante técnicas de reducción de dimensionalidad, es posible aplicar algoritmos de agrupamiento (clustering) pertenecientes al aprendizaje no supervisado, con el objetivo de revelar estructuras, patrones o comportamientos comunes entre las observaciones. En agricultura, este tipo de algoritmos permite identificar zonas de manejo homogéneo, clasificar unidades productivas con condiciones similares o detectar comportamientos anómalos dentro de los cultivos.

En este contexto los modelos no supervisados no disponen de etiquetas predefinidas ni de una variable objetivo que permita validar directamente los resultados obtenidos. Por esta razón, surge la necesidad de emplear métricas de evaluación que cuantifiquen la calidad, coherencia y estabilidad de los agrupamientos generados. Estas métricas permiten estimar qué tan bien los clústeres formados representan las estructuras del conjunto de datos y qué tan consistentes son frente a variaciones en los parámetros del modelo. En relación con esto, se han desarrollado diferentes medidas internas que evalúan la separación y cohesión de los clústeres, como el Índice de Silhouette, el Índice de Davies–Bouldin, Índice de Calinski–Harabasz y la persistencia del clúster [30].

La persistencia está estrechamente relacionada con el rango de densidad (λ) en el que un clúster se mantiene cohesivo antes de dividirse. Cuanto más amplio sea este rango, mayor será la resistencia del clúster a los cambios estructurales en su forma, y, por consiguiente, mayor su estabilidad. A diferencia de métodos particionales como K-Means, que generan agrupamientos en un solo nivel, los métodos basados en densidades permiten que los clústeres emerjan, evolucionen y se fragmenten jerárquicamente de acuerdo con la densidad del conjunto de datos. De este modo, la persistencia y el rango de densidad proporcionan un soporte teórico para la identificación de las agrupaciones más representativas y confiables dentro de la estructura global del modelo.

La estabilidad de los clústeres puede expresarse matemáticamente mediante la siguiente formulación integral, que describe la acumulación de cohesión del clúster a lo largo de su rango de densidad:

$$Estabilidad = \int_{\rho min}^{\rho max} C(\rho)\, d\rho \tag{3}$$

Donde: ρmin y ρmax son los límites del rango de densidad en el cual el cluster C existe y $C(\rho)$ es el número de puntos que pertenecen al cluster C.

- ***Índice Silhouette:*** Evalúa la calidad de los clusters basándose en la cohesión interna y la separación entre clusters.

- ***Rand Index Ajustado (ARI):*** Evalúa la correspondencia entre los clusters generados y las etiquetas reales (cuando están disponibles).

3.1.10 Visualización geoespacial

La visualización geoespacial es un componente fundamental dentro del flujo analítico de la ciencia de datos aplicada a la agricultura, ya que permite traducir resultados cuantitativos como los clústeres obtenidos mediante técnicas de aprendizaje no supervisado en representaciones espaciales que son interpretadas directamente sobre el territorio productivo. La visualización geográfica asocia cada unidad productiva con su localización exacta y con las variables que la caracterizan (productividad, condiciones de suelo, estado fisiológico del cultivo), facilitando así el análisis territorial de patrones agronómicos y socioeconómicos. En este sentido, los sistemas de información geográfica (SIG) no solo almacenan y analizan datos espaciales, sino que proporcionan herramientas para apoyar la toma de decisiones en campo [30]

Adicionalmente, la visualización geoespacial no solo describe “dónde están” los grupos, sino que ayuda a interpretar “por qué están allí”. Al superponer las clases generadas por algoritmos no supervisados (por ejemplo, HDBSCAN, un método jerárquico basado en densidad que identifica agrupaciones con distintas formas y detecta automáticamente observaciones atípicas) con variables asociadas a manejo agronómico, fertilidad del suelo o respuesta fisiológica del cultivo, es posible inferir relaciones entre desempeño productivo y condiciones biofísicas locales. HDBSCAN se ha utilizado en contextos territoriales precisamente por su capacidad para encontrar clústeres en datos espaciales con densidades variables y formas irregulares, y para marcar unidades anómalas como ruido, lo que facilita priorizar áreas que requieren atención diferencial [31]..

3.2 Antecedentes

Se han adelantado investigaciones científico-tecnológicas a nivel internacional, nacional y regional, enfocadas a estimar y describir el comportamiento de los cultivos a las condiciones que ofrece el ambiente de producción. Este tipo de estudios se realizan igualmente en Colombia, de acuerdo con la plataforma SIEMBRA [32] [33].

A través de la aplicación de vigilancia tecnológica sobre el tema de estudio, se identificaron diversos avances relacionados con el manejo de datos y la identificación de características similares entre grupos mediante técnicas de *machine learning* aplicadas en distintos campos. Uno de los estudios relevantes es “Identifying key features in reactive flows: A tutorial on combining dimensionality reduction, unsupervised clustering, and feature correlation” [34], el cual sirvió como referencia metodológica para el preprocesamiento de datos y la implementación de modelos de aprendizaje no supervisado orientados al análisis de agrupamientos. Este estudio fue tomado como punto de partida para el desarrollo técnico del proyecto de grado, especialmente en aspectos relacionados con las técnicas de clustering, el uso de métodos de compresión estocástica de datos basados en geometría riemanniana y la estrategia general de modelado.

Asimismo, otro aporte clave se encuentra en el estudio “Dimensionality reduction by feature clustering for regression problems” [35], donde se destaca la importancia de una adecuada selección de técnicas de extracción de características —particularmente aquellas relacionadas con la reducción de dimensionalidad— como etapa previa fundamental en problemas de segmentación no supervisada.

Por otro lado, y teniendo en cuenta la propuesta que integra técnicas computacionales aplicadas a la agricultura como la desarrollada por “Guo”, en el que se aborda la problemática de la evolución de las estructuras de plantación de cultivos bajo el impacto del envejecimiento de la población en Qinghai, China. Utilizando datos de teledetección e interpretaciones de imágenes para cartografiar estructuras de plantación desde el año 2000 hasta el 2020, y combinando estos datos con 2000 encuestas a agricultores, el estudio emplea modelos de machine learning y el modelo jerárquico lineal (HLM) para explorar mecanismos de conducción y predecir tendencias futuras de cambios en estas estructuras.

Este estudio [34] introduce un enfoque innovador al integrar modelos de machine learning con análisis HLM para comprender y predecir la estructura de plantación de cultivos en respuesta al envejecimiento de la población. A diferencia de investigaciones previas que se han centrado predominantemente en factores ambientales, el presente trabajo resalta la influencia combinada de las condiciones socioeconómicas de los agricultores junto con las variables ambientales, proporcionando así un modelo más holístico y preciso. Este enfoque permite realizar predicciones detalladas, como se evidencia en el estudio citado, donde se implementó un modelo de *random forest* que alcanzó una precisión del 74%. En contraste, el proyecto de grado aquí propuesto no solo busca validar estos modelos en contextos similares, sino también ampliar el alcance de las técnicas de ciencia de datos aplicadas al sector agrícola. El objetivo es fortalecer las decisiones de política pública y de manejo territorial mediante el uso de información integrada y contextualizada. Este enfoque es adoptado con el fin de generar alternativas concretas para la gestión del territorio, basadas en datos que describen fielmente la realidad del caso de estudio, en este caso, el cultivo de cacao. Además, se destaca la necesidad de incorporar variables demográficas, como el envejecimiento poblacional, como factores clave en la planificación agrícola futura

Desde la parte de suelo también de evidencia aplicaciones relacionadas con la exploración y uso de la espectroscopía visible y cercana al infrarrojo (Vis-NIR) en combinación con algoritmos de aprendizaje automático para predecir rápidamente el contenido de nitrógeno (N), fósforo (P) y potasio (K) en suelos agrícolas de zonas tropicales secas. Utilizando 122 muestras de suelo de la provincia de Aceh, Indonesia, los investigadores aplicaron técnicas de machine learning como k-nearest neighbors (kNN), adaboost y random forest para desarrollar modelos de calibración robustos y precisos para la predicción de nutrientes [35].

Al integrar espectroscopía Vis-NIR y algoritmos de aprendizaje automático para realizar evaluaciones rápidas y no destructivas de nutrientes del suelo, representa un avance significativo sobre los métodos tradicionales de análisis químico. En contraste con el estudio de Guo [34] , que también utilizó técnicas de aprendizaje automático pero enfocadas en la estructura de plantación bajo efectos demográficos, este estudio se centra directamente en la predicción de elementos esenciales para la fertilidad del suelo en un contexto específico de tierras agrícolas secas. Así, mientras Guo et al. Proporcionan un modelo predictivo para ajustes estructurales en la agricultura, Mustaqimah y su equipo platearon soluciones directas para la optimización del manejo de nutrientes, ambos contribuyendo de manera complementaria a la sostenibilidad y eficiencia.

También es importante tener en cuenta el enfoque del uso de las imágenes multiespectrales como recurso para la extracción de firmas multiespectrales, las cuales permiten la predicción de estados fisiológicos de los cultivos en general. De manera similar a la investigación liderada por Hesham Abd en el estudio denominado “Using multispectral imagery to extract a pure spectral canopy signature for predicting peanut maturity” [16], la metodología descrita permite plantear la posibilidad de relacionar esta información con otros aspectos clave del cultivo de cacao, específicamente con la zonificación de metales pesados por cadmio en las unidades productivas del país. Esta información se encuentra alojada en los estudios “The First National Survey of Cadmium in Cacao Farm Soil in Colombia” y “First National Mapping of Cadmium in Cacao Beans in Colombia” [36], [37], los cuales permiten tener un acercamiento al estado actual del sector cacaocultor en este sentido.

Por último, es importante destacar que el valor de estos estudios radica en la convergencia metodológica y conceptual que comparten. Gracias a la transversalidad de los enfoques teóricos utilizados particularmente en lo referente a los modelos de aprendizaje automático y técnicas de análisis multivariado, es posible orientar este conocimiento hacia la visualización interactiva, la gestión territorial y la toma de decisiones informadas en sectores específicos. Un ejemplo de ello se encuentra en el estudio “Deep Learning Techniques for the Exploration of Hyperspectral Imagery Potentials in Food and Agricultural Products” [36], donde se evidencia cómo el uso de imágenes hiperespectrales y técnicas de aprendizaje profundo puede ser aplicado a problemáticas agroalimentarias con un alto nivel de precisión y utilidad práctica.

Asimismo, es relevante señalar que, aunque muchos de estos estudios de referencia están enfocados en áreas de las ciencias agrícolas, ambientales, exactas y naturales, las técnicas de machine learning y clustering también son aplicables en contextos sociales. Tal es el caso del estudio “Método de clustering e inteligencia artificial para clasificar y proyectar delitos violentos en Colombia” [38]**,** el cual propone una metodología de agrupamiento y predicción para clasificar delitos violentos por departamentos, utilizando redes neuronales y datos generados por la Policía Nacional entre 2018 y 2022. Como resultado, se identificaron cuatro clústeres de violencia con características diferenciadas, permitiendo segmentar regiones del país con mayor o menor impacto delictivo.

Este tipo de enfoque abre nuevas posibilidades de investigación en el sector cacaocultor, al permitir integrar datos de naturaleza social como condiciones demográficas, nivel educativo, estructura familiar o prácticas culturales con variables productivas y ambientales. De este modo, se refuerza la importancia de considerar el factor humano como un componente clave dentro de los modelos de gestión agrícola, contribuyendo así a una visión más integral y realista del desarrollo rural en Colombia.

4 ESTRUCTURA METODOLOGICA Y DESARROLLO DEL DOCUMENTO

En la Figura 2 se muestra de forma resumida la metodología usada para el desarrollo de este trabajo aplicado, mostrando las etapas desarrolladas mediante el lenguaje de programación Python en el entorno de trabajo interactivo Google Colab. Inicialmente, se recopilaron y estructuraron los datos provenientes de diferentes fuentes, divididos por capitales: datos sociagronomicos, datos físicos de suelo e imágenes multiespectrales obtenidas mediante drones. En la segunda etapa, se llevó a cabo el preprocesamiento y manipulación de las imágenes multiespectrales, empleando técnicas de análisis raster con Python. A continuación, se construyó e importó una base de datos consolidada en formato CSV, realizando tareas esenciales como limpieza, manejo de datos faltantes y selección o eliminación de variables redundantes. Posteriormente, para reducir la complejidad derivada del alto número de variables disponibles, se aplicó una técnica de reducción de dimensionalidad no lineal (UMAP), facilitando así la visualización y comprensión de las relaciones entre datos. Una vez realizada esta reducción dimensional, se ejecutaron algoritmos de clustering no supervisado, específicamente K-Means y HDBSCAN, con el fin de identificar agrupaciones de unidades productivas con características similares. Seguidamente, se llevó a cabo la evaluación e interpretación de los resultados, mediante métricas de validación internas y una validación por expertos, garantizando coherencia agronómica y aplicabilidad práctica. Finalmente, se diseñó una estrategia de visualización dinámica e interactiva, utilizando herramientas geoespaciales como ArcGIS Dashboard, para facilitar la gestión territorial y la toma de decisiones basada en datos reales. Toda esta secuencia metodológica fue desarrollada a través del lenguaje de programación orientado a objetos Python, aprovechando el entorno de ejecución interactivo y colaborativo proporcionado por Google Colab.

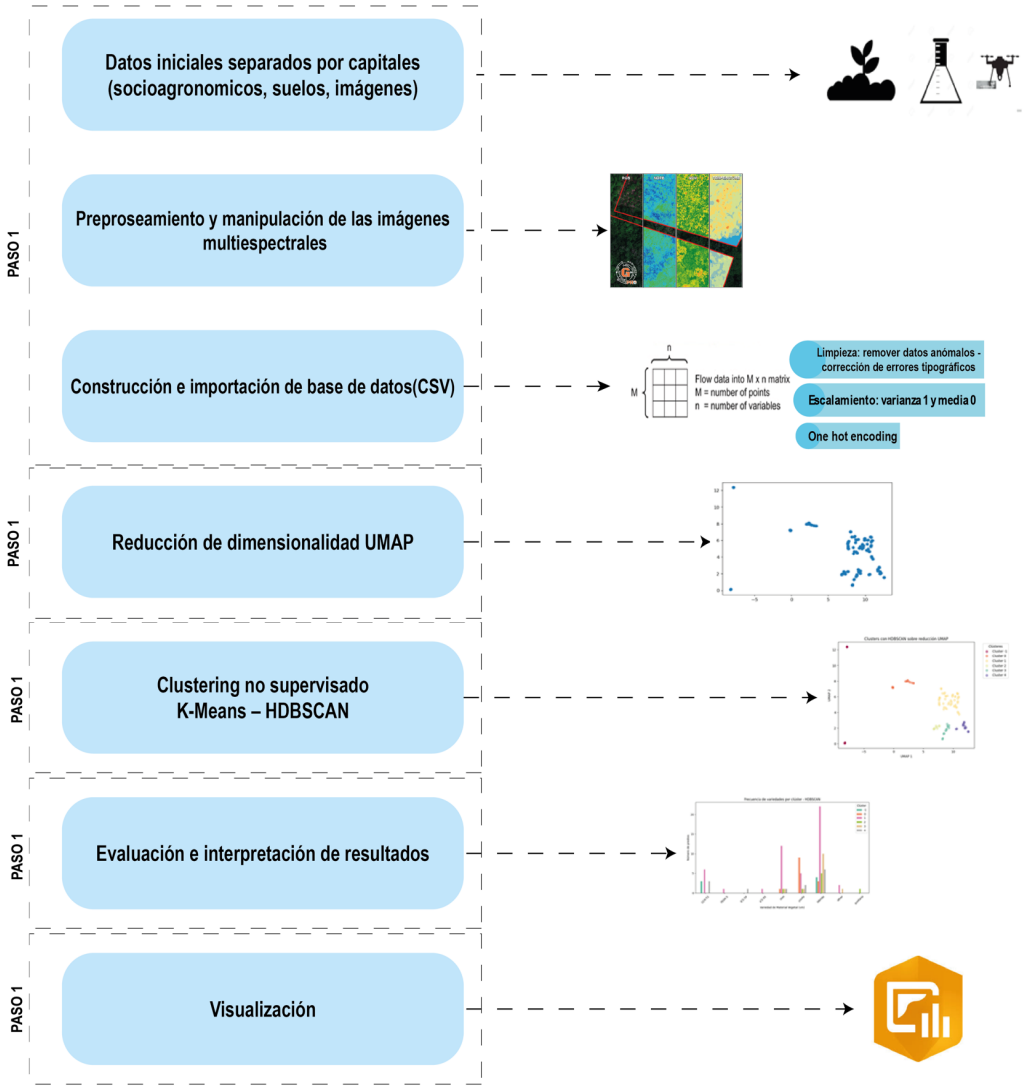


Figura 2. Metodología empleada para la identificación y caracterización de unidades productivas de cacao mediante técnicas de machine learning y análisis geoespacial.

4.1 Área de estudio

La presente propuesta de trabajo de grado se deriva del proyecto de investigación denominado **"Ventajas Comparativas para el Subsector Cacao"**, avalado por el Ministerio de Ciencias. Este proyecto macro busca proporcionar una solución holística desde la visión **AGROTECH** a la comprensión y descripción del sector cacaocultor del departamento del Huila, desde un punto de vista científico soportado en la gestión del territorio basado en información. Esta propuesta de grado se limitó para el municipio de Baraya, uno de los siete municipios beneficiarios del proyecto, debido a que para este municipio se dispone con la totalidad de la información capturada, mientras que para los demás municipios la información aún se encuentra en etapa de recopilación; sin embargo, los resultados obtenidos y las metodologías desarrolladas podrán ser replicados y escalados posteriormente los otros

municipios involucrados en el proyecto macro. Baraya está situado en la parte norte del departamento del Huila, al oriente del valle alto del río Magdalena. Posee una extensión territorial de 737 km² y una temperatura media aproximada de 29 °C. El municipio cuenta con cerca de 10.324 habitantes, de los cuales aproximadamente 3.000 personas dependen económicamente del cultivo de cacao. Actualmente, existen más de 127 hectáreas cultivadas y 103 unidades productivas dedicadas al cacao distribuidas en diversas zonas del municipio.

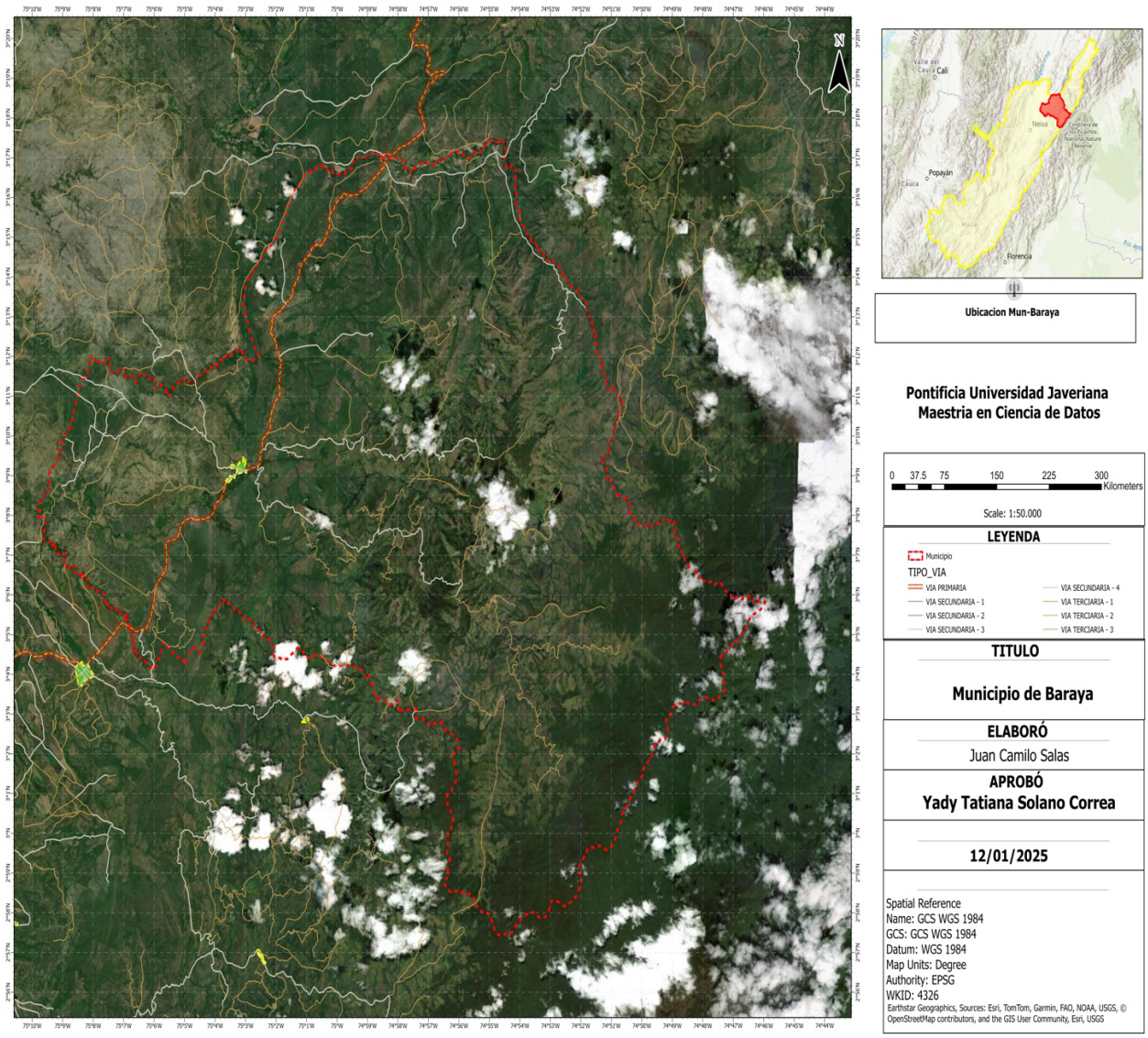


Figura 3. Localización geográfica de Baraya-Huila.

5 CREACIÓN DE BASE DE DATOS ESTRUCTURADA

5.1 Diseño de estructura, requerimientos y estándares de la base de datos

Los datos utilizados en el desarrollo metodológico de este proyecto de grado presentan una naturaleza diversa y provienen de distintas fuentes. Los datos sociales, económicos y agronómicos fueron recolectados a través de encuestas aplicadas a cada productor de cacao del municipio de Baraya, estructuradas en torno a diferentes tipos de capitales: social, natural, financiero y agronómico. Por otro lado, los datos multiespectrales y de suelo se obtuvieron mediante la captura de imágenes a través de plataforma drone y análisis de laboratorio, respectivamente. A continuación, se describen las principales características de cada conjunto de datos:

Capital humano: Características individuales del productor y su núcleo familiar que influyen en la gestión y productividad de la unidad productiva. Este capital incluye variables relacionadas con el nivel educativo, la experiencia en el cultivo, la capacitación recibida, el acceso a asistencia técnica y la disponibilidad de mano de obra en la finca:

Tabla 1. Variables de capital humano.

Variable	**Descripción**
Chum_1	Cuantos años de experiencia tiene en el manejo del cultivo.
Chum_2	Que tanto conocimiento tiene sobre el cultivo de cacao.
Chum_3	Nivel de experiencia de la mano de obra.

Capital físico: Variables relacionadas con el acceso al predio, estado físico de su unidad de vivienda y tenencia del predio.

Tabla 2 Variables de capital físico.

Variable	**Descripción**
Cfis_1	El predio posee tenencia legal (carta vento y/o escritura).
Cfis_2	Calidad de la vivienda (en términos de infraestructura).
Cfis_3	Como es el acceso al predio.

Capital financiero: Comprende variables asociadas a la estabilidad económica del productor, tales como sus fuentes de ingreso, acceso a crédito y tipo de inversión en la producción.

Tabla 3. Variables de capital financiero.

Variable	Descripción
Cfin_1	Cuál es el valor estimado de los ingresos mensuales del núcleo familiar.
Cfin_2	Valor estimado de los ingresos mensuales de la actividad en cacao.
Cfin_3	Tiene acceso a créditos bancarios.

Capital natural: Contiene información sobre las características del entorno natural de la unidad productiva, incluyendo la disponibilidad de recursos hídricos, biodiversidad presente y condiciones agroecológicas.

Tabla 4. Variables de capital natural.

Variable	Descripción
Cnat_1	Posee acceso a fuentes hídrica de buena calidad.
Cnat_2	Cuál es la calidad del suelo para la producción de cacao.
Cnat_3	Como es la calidad de aire.

Capital social: Incluye variables relacionadas con la estructura y dinámica social del productor, como su participación en asociaciones, acceso a redes de apoyo y nivel de cooperación dentro de la comunidad, las principales variables correspondientes a este capital se encuentran codificadas descritas de la siguiente manera:

Tabla 5. Variables de capital Social.

Variable	Descripción
Csoc_1	Pertenece a algún grupo asociativo.
Csoc_2	Como es la relación de confianza con la asociación.
Csoc_3	Como es la relación de confianza con sus vecinos.

Capital agronómico: Se enfoca en aspectos relacionados con el manejo del cultivo, como las prácticas agrícolas utilizadas, la productividad del cultivo, el uso de fertilizantes y la presencia de plagas.

Tabla 6. Variables de capital agronómico.

Variable	Descripción
atp	Área total del predio (ha).
Ctl	Cantidad de lotes de cacao en predio.
Ec	Edad del cultivo.

La encuesta socioagronómica aplicada a cada productor incluye diversas variables agrupadas según los

diferentes tipos de capital evaluados. La descripción completa de cada variable, se encuentra disponible en el diccionario de datos en la sección de Anexo A, también disponible en formato .xlsx en el repositorio correspondiente en GitHub.

Finalmente, la estructura de la encuesta socioagronómica se presenta en formato de tabla de la siguiente manera:

Tabla 7. Estructura de base de datos sociagronomica.

id	viven_en_el_predio_de_caco	chum_2	chum_3
bar1	si	alto_	mano_de_obra_medianamente_calif
bar2	no	alto_	mano_de_obra_medianamente_calif
bar3	no	medio	mano_de_obra_altamente_califica

Análisis de laboratorio de suelos: a través de los resultados de laboratorio de suelo se evalúan las propiedades químicas y físicas, determinando su fertilidad y capacidad para sostener el crecimiento de cultivos de cacao. Este tipo de análisis proporciona información clave sobre la composición del suelo, su estructura, disponibilidad de nutrientes y características que influyen en la retención de agua y la aireación.

DATOS DE LA MUESTRA			
Cultivo:	CACAO	Departamento:	HUILA
Código Muestra:	Bar Fi 1	Municipio:	Baraya
Recep. Muestra:	2022-12-12	Vereda:	Espinalosa
Exp. Resultado:	2023-02-22	Predio:	La Esmeralda

Muestra lab No.	PARÁMETRO		MÉTODO/ TÉCNICA ANALÍTICA	UNIDADES	RESULTADO
	RESULTADOS ANÁLISIS QUÍMICO				
617-22	pH		Relación 1:1 Suelo- Agua / Potenciometría	Unidades de pH	7.13
	Carbono Orgánico		NTC 5403 (Walkley-Black) / Colorimetría	%	1.28
	Materia Orgánica		Cálculo a partir de % CO	%	2.20
	Nitrógeno Total		Cálculo a partir de % MO	%	0.11
	RESULTADOS ANÁLISIS FÍSICO				
	Densidad Aparente		Terrón Parafinado	$g.cm^{-3}$	1.41
	Densidad Real		Picnómetro	$g.cm^{-3}$	2.30
	Porosidad Total		Relación de densidades	%	38.73
	Textura	Arena	Bouyoucos	%	19.30
		Arcilla			26.40
		Limo			54.30
		Clasificación	Franco Limoso - FL		

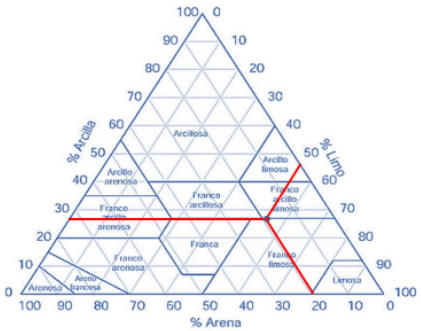


Figura 4. Formato de análisis físico y químico de suelo, fuente propia.

- Parámetros Químicos

- pH: Indica el grado de acidez o alcalinidad del suelo, influyendo en la disponibilidad de nutrientes para las plantas. Un pH de 7 es neutro, valores menores indican suelos ácidos y valores mayores suelos alcalinos.
- Carbono Orgánico (%CO): Mide la cantidad de carbono presente en la materia orgánica del suelo, un indicador de la fertilidad y actividad biológica del suelo.
- Materia Orgánica (%MO): Representa el contenido total de materia orgánica en el suelo, esencial para la retención de humedad y la disponibilidad de nutrientes.
- Nitrógeno Total (%N): Evalúa la cantidad total de nitrógeno en el suelo, un macronutriente fundamental para el crecimiento de las plantas.

- Parámetros Físicos

- Densidad Aparente (g/cm³) [Da]: Mide la compactación del suelo, afectando la infiltración de agua y la penetración de raíces.
- Densidad Real (g/cm³) [Dr]: Representa la densidad de las partículas sólidas del suelo sin considerar el espacio poroso.
- Porosidad Total (%): Indica la cantidad de espacios vacíos en el suelo, lo que influye en la aireación y capacidad de retención de agua.

-Textura del Suelo

La textura del suelo se define según la proporción de tres tipos de partículas:

- Arena (%) [A]: Partículas de mayor tamaño, que mejoran el drenaje, pero reducen la retención de agua.
- Arcilla (%) [AR]: Partículas finas que incrementan la capacidad de retención de agua y nutrientes, pero pueden compactar el suelo.
- Limo (%) [L]: Partículas intermedias que aportan fertilidad y mejoran la estructura del suelo.

Para integrar estos resultados a la base de datos del estudio, los valores obtenidos fueron transformados y organizados en una tabla estructurada como se muestra en la Tabla 8, de manera que se simplificara el preprocesamiento y el análisis posterior junto con otros datos agronómicos, socioeconómicos y multiespectrales de cada unidad productiva.

Tabla 8. Estructura de la base de datos de análisis físico-químicos de suelo.

id	**pH**	**Ntotal**	**A**	**AR**	**L**	**Textura**	**Dr**	**Da**
bar1	7.13	0.1100	19.3	26.4	54.3	FL	2.3	1.40918
bar2	6.33	0.0249	38.67	34.5	26.83	FAr	2.39	1.25
bar3	7.15	0.1341	43.7	32.1	24.2	FAr	2.36	1.27

5.2 Tratamiento de los datos multiespectrales

Las imágenes multiespectrales fueron obtenidas mediante la integración del drone con el sensor Altum de Micasense, el cual tiene la capacidad de capturar múltiples bandas las cuales se detallan en la Tabla 9. Las capturas se realizaron a una altura promedio de 80 metros, con el fin de asegurar una resolución espacial homogénea en todas las unidades productivas, lo que permite realizar comparaciones precisas

y consistentes entre ellas. La salida gráfica generada por estas imágenes se ilustra en la Figura 5, donde se visualiza la composición multibanda utilizada para el análisis agronómico.

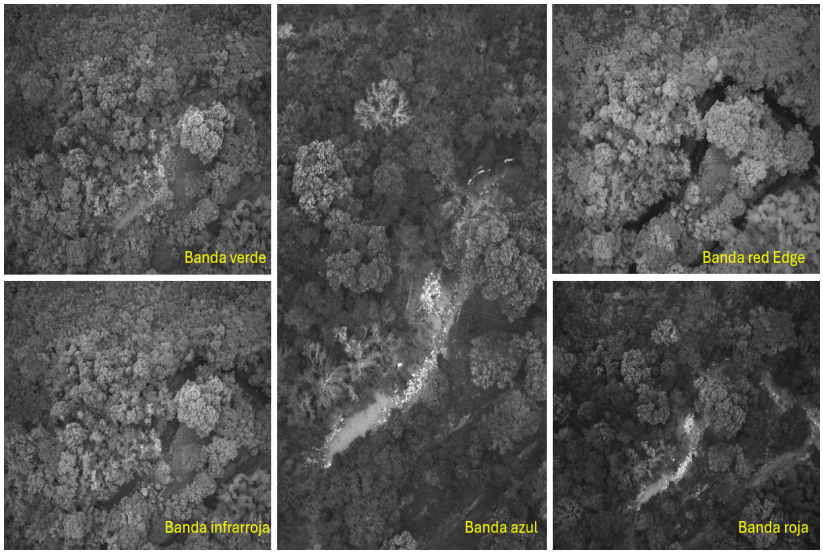


Figura 5. Bandas espectrales provenientes del sensor Altum.

Las bandas espectrales capturadas por el sensor permitieron analizar diversas características fisiológicas y biofísicas del cultivo. La banda azul resulta útil para examinar la estructura celular de las plantas y detectar signos tempranos de estrés. La banda verde está estrechamente relacionada con la concentración de clorofila, por lo que permite evaluar el vigor y la densidad foliar del cultivo. La banda roja refleja la absorción de luz por la clorofila y, por tanto, se asocia con la actividad fotosintética.

La banda Red Edge es particularmente sensible a cambios en la estructura y contenido de clorofila de las hojas, lo que la convierte en un indicador temprano de estrés vegetal antes de que este sea perceptible visualmente. Por su parte, la banda del infrarrojo cercano (NIR) se relaciona con la biomasa y el estado fisiológico del cultivo, siendo fundamental para la estimación de índices espectrales como el NDVI (Normalized Difference Vegetation Index). Finalmente, la banda térmica permite determinar la distribución de temperatura dentro del cultivo; sus valores originales, registrados en micro kelvin, fueron transformados a grados Celsius para facilitar su interpretación y análisis comparativo entre las distintas unidades productivas.

Tabla 9. Centro de bandas multiespectrales.

Banda	**Centro de banda [nm]**
Azul	475
Verde	560
Rojo	668
Red Edge	717
NIR	842
TEMPERATURA	Micro grado Kelvin [μK]

Cada escena multiespectral está compuesta por cinco bandas espectrales y una banda térmica adicional. Cada banda posee un centro de banda determinado en nanómetros mostrado en la Tabla 9. Cada imagen incluye metadatos esenciales, como la coordenada geográfica, altitud y fecha de captura, lo que permite una correcta georreferenciación y alineación de las imágenes durante el procesamiento.

La base de datos de "Imágenes", correspondiente al Proyecto Cacao en Baraya, contiene un total de 8,365 archivos distribuidos en 11 subcarpetas, con un tamaño total de 41.5 GB. Esta colección de archivos corresponde a imágenes multiespectrales capturadas mediante el sensor *Altum* de *Micasense*, acoplado a un *drone*, a una altura promedio de 80 metros, con el fin de garantizar una resolución espacial homogénea entre todas las unidades productivas. Debido al alto volumen de datos, fue necesario un preprocesamiento y estructuración adecuada de la información. Estas imágenes fueron ortorrectificadas y procesadas para la construcción de los ortofotomosaicos utilizando el software especializado *Agisoft Metashape*, el cual permite generar productos geoespaciales de alta precisión a partir de imágenes aéreas.

En la Figura *6* se representa el proceso metodológico diseñado para la organización, análisis y procesamiento de las imágenes multiespectrales, este flujo siguió las recomendaciones sugeridas por el fabricante del sensor multiespectral en relación con el procesamiento de las imágenes y su integración con información auxiliar clave [39], como encuestas socioagronómicas y datos de suelo, para generar insumos que apoyen el análisis de las unidades productivas de cacao del municipio de Baraya-Huila.

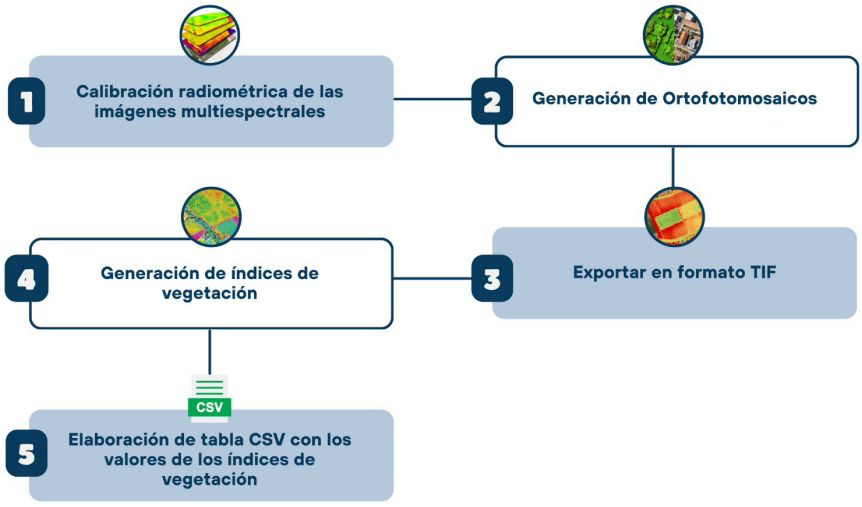


Figura 6. Metodología seguida para el procesamiento de imágenes multiespectrales.

Al usar imágenes multiespectrales, estas deben ser preprocesadas para su respectiva calibración radiométrica. Este paso consiste en ajustar los valores de las imágenes capturadas por el sensor para garantizar que reflejen las condiciones reales de iluminación y reflectancia [40]. Para ello, se utiliza un panel de calibración radiométrica, que permite corregir posibles variaciones en las condiciones de luz durante la captura, como cambios en la nubosidad o en la posición del sol.

Los ortomosaico fueron generados a partir de cinco bandas espectrales, cada una correspondiente a una longitud de onda específica dentro del espectro electromagnético. Entre estas bandas se incluyen el azul *(Blue),* verde *(Green)* y rojo *(Red),* las cuales son fundamentales en la composición de imágenes en color real (RGB). Inicialmente, se optó por trabajar con esta combinación, ya que permite visualizar las ortofotos de una manera más familiar e intuitiva para los usuarios, similar a cómo el ojo humano percibe el entorno.

La representación en *RGB* facilita la interpretación del arreglo forestal y la identificación de elementos dentro de la escena, como caminos, cuerpos de agua y diferentes tipos de cobertura vegetal como se visualiza en la Figura *7*-a. Sin embargo, en aplicaciones más avanzadas como el cálculo de los índices de vegetación, se incorporan otras bandas espectrales como el borde rojo (Red Edge), el infrarrojo cercano *(Near Infrared - NIR)* y la térmica, que aportan información clave sobre el estado fisiológico de los cultivos, permitiendo detectar estrés hídrico, variaciones en la fotosíntesis y diferencias en la temperatura superficial.

Los ortofotomosaicos disponibles cuentan con una resolución espacial de entre 4 y 7 cm por píxel, lo que permite un nivel de detalle más profundo en el análisis de cada unidad productiva. Esta alta resolución es fundamental para la generación de información más precisa, asegurando que cada píxel represente la realidad en campo.

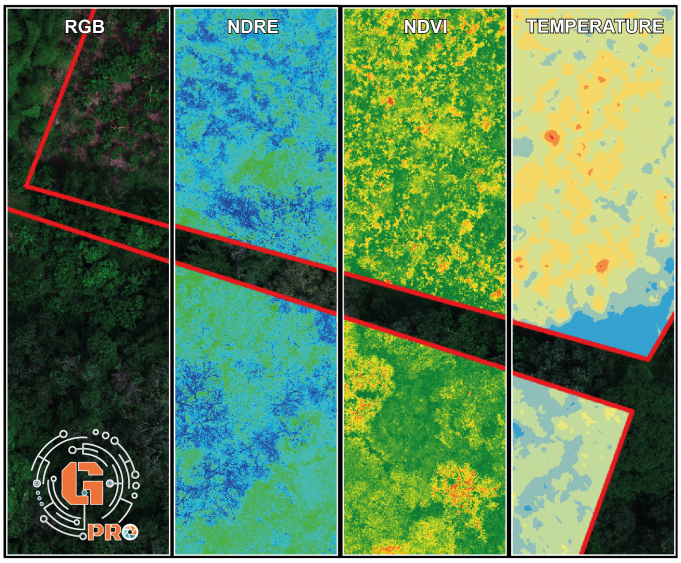


Figura 7. Índices de vegetación de referencia.

Cálculo de índices de vegetación y extracción de datos relevantes: El cálculo de los índices de vegetación se llevó a cabo utilizando el lenguaje de programación orientado a objetos Python, a través de diversas librerías especializadas en el procesamiento de datos geoespaciales, como *Rasterio, Geopandas, NumPy, GDAL*. Estas herramientas permitieron manejar, analizar y manipular imágenes raster de manera iterativa sobre todas las unidades productivas.

En la Figura *7*, se representa a manera de ejemplo la salida gráfica el cálculo de los índices de vegetación, fundamentales para el monitoreo y análisis del estado de la cobertura vegetal. Sin embargo, para poder

correlacionar estos cálculos con los datos socioagronómicos y las muestras de suelo, fue necesario aprovechar la naturaleza matricial de los datos raster. Las imágenes multiespectrales están compuestas por matrices de píxeles donde cada banda espectral es una matriz de tamaño *M x N*.

Por medio de la ecuación 1 y 2 se calculan los índices de vegetación como NDVI y NDRE respectivamente, dado que la operación se aplica a cada elemento matricial de toda la imagen. Este cálculo se repite de manera iterativa para cada píxel o elemento matricial de la imagen en la que se almacena un valor de reflectancia de una longitud de onda específica, generando una nueva matriz con los valores de índice de vegetación.

Dado que las imágenes presentan diferentes resoluciones espaciales y, por ende, tamaño matricial, se implementó un proceso de remuestreo para asegurar que todas las matrices tengan un tamaño matricial uniforme. Este *resampling* se efectuó mediante la interpolación de valores de reflectancia, con el fin de ajustar todas las imágenes y garantizar la comparabilidad entre las unidades productivas. La manipulación de los conjuntos matriciales se realizó considerando que cada unidad productiva estaba representada como un polígono georreferenciado en un sistema de coordenadas.

Para extraer el valor promedio de los índices de vegetación dentro de cada unidad productiva, se llevó a cabo un proceso de zonificación espacial, en el cual cada polígono correspondiente a una unidad actúa como una máscara sobre la imagen multiespectral. Esta operación permite delimitar las áreas de análisis y calcular únicamente los valores contenidos dentro de cada polígono.

El valor medio de cada índice se obtuvo aplicando operaciones estadísticas sobre los píxeles pertenecientes a cada unidad productiva, procedimiento que se repitió para todos los índices de vegetación generados. De esta manera, se obtuvieron valores promedio representativos de las condiciones biofísicas de cada unidad

Los datos obtenidos fueron estructurados en formato de tabla, permitiendo asociar cada magnitud calculada a su respectiva unidad productiva. Esta organización de la información permite la integración de los índices de vegetación como el $NDVI(i,j)$, lo que a su vez facilita la generación de mapas de variabilidad espacial y conserva la estructura de la Tabla 10**.**

Tabla 10. Estructura de base de datos – índices de vegetación.

id	**NDRE_Promedio**	**NDVI_Promedio**	**TEMP_Promedio**
bar1	0.414625256	0.851582583	31.79696347
bar2	0.420870698	0.855564547	28.96557507
bar3	0.162613168	0.496476194	33.44620535

5.3 Estandarización y limpieza de los datos

Para garantizar la calidad y coherencia de los datos utilizados en este estudio sobre el municipio de Baraya, se implementó un proceso de estandarización y limpieza de datos. Este proceso consistió en la eliminación de valores inconsistentes, el manejo de datos faltantes y la normalización de variables

categóricas, asegurando así su correcta interpretación y uso en los análisis posteriores.

El conjunto de datos sociagronomica incluía inicialmente 87 variables categóricas de 129, lo que representaba un desafío en términos de redundancia y calidad de los datos. Para garantizar que estas variables fueran consistentes y útiles para el análisis posterior, se realizó un proceso de inspección y estandarización con los siguientes pasos:

- **Detección de redundancias:** Se identificaron variables que contenían información duplicada o que podían ser combinadas en categorías más representativas.
- **Corrección de errores de digitalización:** Se realizó un análisis exhaustivo de errores tipográficos y de codificación que podrían generar inconsistencias al aplicar modelos de machine learning, dado que estas imprecisiones son comunes en la captura manual de datos o de respuestas de preguntas abiertas. Uno de los problemas más frecuentes es la variación en la escritura de una misma respuesta dentro de una misma variable, lo que genera duplicaciones innecesarias y dificulta la estandarización. Por ejemplo, respuestas afirmativas pueden aparecer registradas como ***"Si", "sí", "SI"*** o incluso con espacios adicionales, lo que impide un correcto procesamiento automatizado. Del mismo modo, nombres de cultivos o insumos pueden contener errores ortográficos o diferencias en el uso de mayúsculas y tildes, afectando la coherencia de la base de datos. A continuación, se presentan algunos ejemplos de estos errores en la Tabla 11:

Tabla 11. Detección de respuestas con errores tipográficos.

Variable	**Respuestas**
Dta	['si', nan, 'no', 'SI']
Cfis_2	['media_', 'alta_', 'alta', 'baja_']
af	['si', 'SI', 'no']

Dado el alto número de variables categóricas en la encuesta sociagronomica, corregir manualmente las inconsistencias en cada una se convierte en un proceso dispendioso, por lo que se implementó una estrategia de estandarización automatizada utilizando la librería ***"fuzzywuzzy"***. Se desarrolló una función en Python que recorre cada variable categórica dentro del *DataFrame* de Pandas, analiza los niveles de respuesta y los compara con una lista de términos de referencia predefinidos mediante coincidencia aproximada. Esto permitió detectar y corregir errores tipográficos, agrupar respuestas equivalentes (por ejemplo, unificar "Si", "sí", "SI" y "sí." en "Sí"), eliminar espacios en blanco y caracteres especiales, y reducir la cantidad de categorías únicas, optimizando el procesamiento de datos y mejorando la capacidad de los modelos de machine learning para extraer patrones significativos. Gracias a esta metodología, se logró automatizar la estandarización de las variables categóricas, asegurando mayor coherencia y calidad en los datos sin necesidad de un procesamiento manual.

- **Manejo de valores nulos:** Para garantizar la calidad de los datos y evitar inconsistencias en el análisis, se realizó un tratamiento exhaustivo de los valores nulos en el *dataFrame*. En primer lugar, se identificaron las columnas con valores ausentes y se cuantificó el número de datos faltantes por variable. Se observó que varias columnas presentaban un número significativo de valores nulos,

especialmente aquellas relacionadas con enfermedades, plagas y características socioeconómicas clasificadas como "otras".

Tabla 12. Datos nulos.

Variable	Datos nulos
otra_enfermedad - tenfq	103
otro - cfin_6	103
otra_enfermedad - forest_2	103
otra_enfermedad - forest_1	103

Este patrón de valores nulos se debe a la naturaleza de las preguntas en la encuesta, donde se indagaba si existían otras variedades sembradas distintas a las principales y secundarias ya evaluadas, así como la presencia de enfermedades y plagas adicionales no incluidas en las categorías estándar. En consecuencia, muchas respuestas quedaron en blanco al no aplicar a la situación específica del encuestado, generando así una cantidad considerable de valores nulos en estas variables.

Inicialmente, el *dataframe* sociagronomica tenía un tamaño de 103 filas * 129 columnas, correspondientes a las 103 unidades productivas de cacao que conforman la totalidad del universo de cultivos de este tipo en el municipio. La presencia de valores nulos en ciertas variables se debe exclusivamente a la naturaleza de las preguntas y no a errores en la captura de datos.

En particular, algunas variables presentaban 103 valores nulos en un total de 103 filas, lo que indica que estaban completamente vacías. En estos casos, no era viable aplicar técnicas de imputación, ya que esto introduciría un sesgo inducido en los datos. Además, al no contener información relevante, estas variables no aportaban valor para la aplicación de modelos futuros de machine learning.

Por esta razón, se decidió eliminar dichas variables del análisis, lo que contribuyó a reducir la complejidad del modelo. Finalmente, el *dataframe* correspondiente a la encuesta sociagronomica quedó con un tamaño de 103 filas * 73 columnas.

5.4 Integración y validación de las fuentes de información

Cada conjunto de datos presentaba estructuras diferentes inicialmente, lo que requirió un proceso de estandarización para garantizar su compatibilidad e integración. Para ello, se siguió una serie de pasos fundamentados en principios de ciencia de datos, asegurando una correcta transformación y unificación de la información.

En primer lugar, se llevó a cabo la estandarización de formatos y estructuras, incluyendo la conversión de datos no estructurados, como las imágenes multiespectrales georreferenciadas, en datos estructurados en forma de tablas. Este proceso permitió relacionar espacialmente cada observación con su respectiva unidad productiva, facilitando la integración con otros datos agroclimáticos y socioeconómicos.

Posteriormente, se llevó a cabo el tratamiento de valores nulos e inconsistencias, las cuales provenían principalmente de la encuesta sociagronomica. Para garantizar la calidad y representatividad de los datos, se aplicaron técnicas de limpieza y depuración, asegurando que la información resultante fuera confiable y adecuada para el análisis.

Como resultado de este proceso, se obtuvieron bases de datos estructuradas y sólidas. Un aspecto clave de estas bases de datos es que todas comparten una columna de identificación única (ID), como se muestra en las Tabla 7, 8 y 10. Esta característica permitió realizar un *merge*, integrando las diferentes fuentes de datos en una sola base consolidada con un nuevo tamaño de 103 × 86. La unificación de estas bases de datos no solo facilitó la integración de información clave, sino que también permitió disponer de un conjunto de datos más robusto y coherente para su posterior análisis.

Con el fin de validar la calidad de los datos y detectar posibles comportamientos anómalos, se realizó un análisis exploratorio basado en estadísticas descriptivas de las principales variables del conjunto de datos procesado. Como resultado, la tabla presenta medidas de resumen estadístico como la cantidad de observaciones, la media, la desviación estándar, los valores mínimo y máximo, así como los percentiles 25%, 50% (mediana) y 75%. Este análisis es fundamental para comprender la distribución de los datos y evaluar la presencia de valores extremos o posibles inconsistencias que podrían afectar los resultados de los modelos y análisis posteriores.

Tabla 13. Resumen estadístico de las variables cuantitativas.

	atp	ctl	NDVI-pro	NDRE -pro	rph	TEMP -pro	Ntotal
count	103	103	103	103	103	103	103
media	16.7	16.	0.63	0.22	383.49	28.99	0.09
std	43.68	43.6	0.16	0.14	431.7	3.55	0.05
min	0.5	0.5	0.34	-0.04	0	19.02	0.01
25%	1.5	1.5	0.52	0.11	200	27.29	0.06
50%	3	3	0.6	0.18	300	28.97	0.09
75%	10.5	10.5	0.8	0.38	500	31.12	0.1
max	300	300	0.89	0.49	3000	37.39	0.34

El conjunto de variables numéricas contiene 103 registros, correspondientes a todas unidades productivas de caca0 presentes en el municipio de Baraya. Se incluyen variables de diferentes categorías, como indicadores agronómicos, métricas relacionadas con la calidad del suelo, índices de vegetación calculados a partir de imágenes multiespectrales y características estructurales del cultivo.

Los boxplots presentados en la Figura 8 muestran la distribución de los índices de vegetación "NDVI_Promedio" y "NDRE_Promedio", así como la variable "TEMP_Promedio". Se observa que "NDVI_Promedio" (gráfico b) tiene una mediana cercana a 0.63, con valores que oscilan entre 0.34 y 0.89, reflejando una variabilidad moderada en la cobertura vegetal de las unidades productivas. Por otro lado, "NDRE_Promedio" (gráfico a) presenta una mayor dispersión, con una mediana más baja en comparación con el "NDVI_Promedio" y valores que van desde aproximadamente -0.04 hasta 0.49, lo

que sugiere diferencias significativas en la salud de las plantas evaluadas.

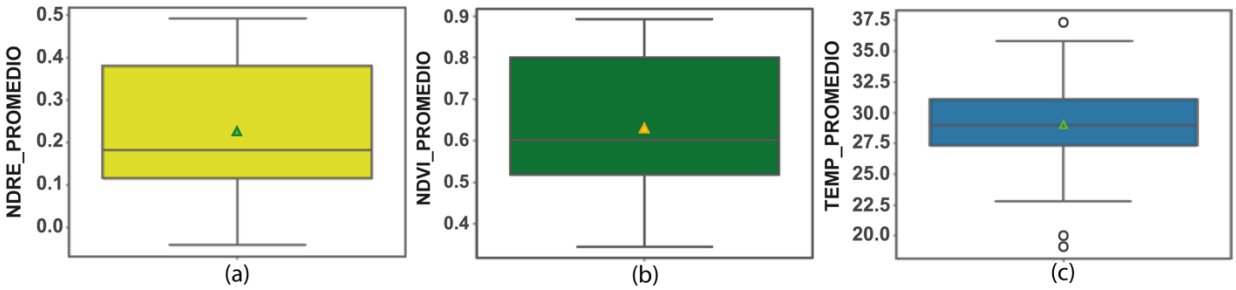


Figura 8. Resumen estadístico de variables de sanidad vegetal derivadas de imágenes multiespectrales.

Es importante destacar la variabilidad observada en algunas variables clave del conjunto de datos. En cuanto a la temperatura promedio ("TEMP_Promedio"), los valores registrados oscilan entre 19.02°C y 37.39°C, con una media de 28.99°C y una desviación estándar de 3.55°C. Esta variabilidad se debe a las condiciones agroclimáticas del área de estudio, reflejando posibles diferencias en altitud o la presencia de microclimas dentro del municipio de Baraya.

Los patrones identificados son explorados en el apartado de visualizaciones interactivas. Asimismo, las variables relacionadas con la extensión del terreno y su productividad evidencian diferencias significativas. La variable *atp*, que representa el área total del predio, presenta una media de 16,74 con una desviación estándar de 43,68, lo que refleja una notable dispersión en el tamaño de las unidades productivas. De forma similar, el rendimiento por hectárea (rph**)** muestra una alta variabilidad, con valores que oscilan entre 0 y 3000, lo que indica diferencias sustanciales en los niveles de producción. Esta heterogeneidad sugiere la coexistencia de predios con productividad nula y otros con rendimientos significativamente superiores, posiblemente asociados a factores como el manejo agronómico, la fertilidad del suelo o las condiciones climáticas locales.

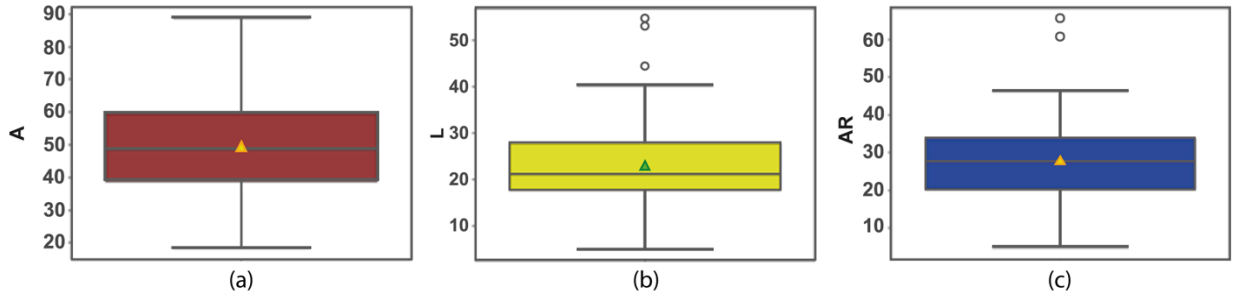


Figura 9. Resumen estadístico de variables de las clases texturales derivadas de los análisis físico-químicos de suelo.

El análisis de la variable "Ntotal" revela en la Tabla 13 una media de 0.09 y una desviación estándar de 0.05, lo que sugiere que la mayoría de las observaciones se encuentran dentro de un rango relativamente estrecho. Esto indica una baja variabilidad en la concentración de nitrógeno total dentro de las unidades productivas evaluadas y puede estar asociados a una mala salud del suelo relacionada con este parámetro que a su vez puede estar asociado con otros factores como social, climático o

agronómico. En contraste, las variables relacionadas con la textura del suelo, como el porcentaje de arcillas ("A"), arenas ("AR") y limos ("L"), presentan distribuciones más dispersas como se evidencia en la figura 11. Sin embargo, sus valores máximos superan más del doble de sus medias, lo que sugiere la posible existencia de casos atípicos o diferencias significativas en la distribución de estos componentes texturales del suelo. Este comportamiento podría estar asociado a variaciones en las condiciones edáficas de las unidades productivas analizadas lo que en términos agronómicos puede ser común si se tiene en cuenta la variedad espacial de los tipos de suelo existentes en el departamento de Huila, lo que a su vez podría influir en el manejo agronómico y en la productividad del cultivo de cacao.

5.5 Creación de la base de datos

Luego de integrar las diferentes fuentes de datos, se obtiene la base de datos final estructurada en formato de tabla, la cual será utilizada en la aplicación de modelos de machine learning para agrupar las unidades productivas que comparten características similares. En este sentido, la base de datos final mantiene un tamaño de 103 filas y 86 columnas, compuestas por variables categóricas y numéricas distribuidas de la siguiente manera:

- **Variables categóricas:** 60 columnas, que incluyen capitales como sistema productivo, prácticas de manejo agrícola y clasificación del suelo según el productor.
- **Variables numéricas:** 26 columnas, entre las cuales destacan índices de vegetación (NDVI, NDRE), parámetros climáticos (temperatura promedio), variables edáficas (pH, materia orgánica, textura del suelo) y características productivas (rendimiento por hectárea, área total del predio, cantidad de lotes).

Esta estructuración de la base de datos proporciona beneficios en las etapas posteriores de análisis y modelado de datos. Algunas de las principales ventajas incluyen:

Tabla 14. Estructura de la base de datos.

Proceso	***Ventaja de la estructura de la base de datos***
Transformación de variables	La separación entre variables categóricas y numéricas permite aplicar estrategias diferenciadas como codificación one-hot para las categóricas y normalización para las numéricas.
Reducción de dimensionalidad	La organización estructurada facilita la aplicación de técnicas como PCA (Análisis de Componentes Principales) o técnicas de reducción de dimensionalidad basada en geometrías no euclidianas para identificar variables más representativas y reducir la carga computacional
Manejo de valores faltantes	La integración clara de variables permite estrategias de imputación según el tipo de dato (media para numéricas, moda para categóricas).
Clustering	La clasificación previa de variables facilita la aplicación de algoritmos como HDBSCAN o K-Means, asegurando que las relaciones entre variables sean consistentes.

6 MODELADO Y ANÁLISIS CON ALGORITMOS DE MACHINE LEARNING

Para garantizar el buen desempeño de un algoritmo de *machine learning*, es fundamental contar con datos de calidad que contengan información útil y factores clave que influyan en su capacidad de aprendizaje. Por ello, fue necesario examinar y preprocesar el conjunto de datos antes de alimentar el modelo, lo que incluye tareas como la transformación y selección de variables.

En el desarrollo actual del trabajo, se ha llevado a cabo el tratamiento de datos faltantes y la corrección de errores tipográficos derivados de la aplicación de las encuestas socioagronómicas. Como se mencionó anteriormente, la base de datos final contiene una gran cantidad de variables categóricas, las cuales deben ser adecuadamente procesadas para su uso en los algoritmos de machine learning.

6.1 Tratamiento de datos categóricos

Este conjunto de datos, que recopila información sobre el estado del sector cacaocultor en el municipio de Baraya, Huila. Es importante destacar las variables categóricas pueden tener características ordinales y nominales.

Las características ordinales son variables categóricas cuyos valores pueden clasificarse u ordenarse según un criterio, mientras que las características nominales representan categorías sin un orden específico [41].

Para el manejo de nuestras variables nominales, se empleó la técnica de codificación one-hot (one-hot encoding). Esta técnica consiste en la creación de una nueva característica ficticia para cada valor único en las diferentes columnas de variables nominales. Por ejemplo, si convertimos la variable *Textura*, que contiene los tipos de suelos presentes en el municipio, y esta tiene 10 niveles de respuesta: ['FL', 'FAr', 'Ar', 'FArA', 'FA', 'F', 'FArL', 'AF', 'A', 'ArA']**,** entonces transformaríamos esta característica en diez nuevas columnas, correspondientes a cada nivel de respuesta. Los valores binarios indicarían la presencia o ausencia de una clase textural específica del suelo. Por ejemplo, 'FL' podría codificarse como: FL = 1, FAr = 0, Ar = 0, FArA = 0, FA = 0, y así sucesivamente para los demás niveles de respuesta.

Para generar estas características ficticias en las diferentes variables categóricas de la base de datos, se utilizó la codificación one-hot a través del método get_dummies de la biblioteca pandas [42]. Al aplicarlo al DataFrame, este método transforma únicamente las columnas categóricas, dejando las demás variables sin modificaciones.

Tabla 15. Transformación de datos.

FL	FAr	Ar	FArA
1	0	0	0
0	1	0	0
0	0	1	0

Al aplicar esta codificación, se tuvo en cuenta que esta metodología puede introducir multicolinealidad

en las variables, lo que podría afectar el rendimiento de los modelos de machine learning. Para reducir la correlación entre las variables, se eliminó una columna de características ficticias generadas por la codificación.

Esta eliminación no afecta la integridad de la información, ya que, por ejemplo, si se suprime la columna correspondiente a la característica 'FL', la información sigue estando presente. Esto se debe a que, si observamos que FAr = 0, Ar = 0, FArA = 0 y FA = 0, podemos inferir que la observación debe pertenecer a la categoría 'FL'.

Para aplicar esta metodología con la función get_dummies, podemos eliminar la primera columna de características ficticias estableciendo el argumento drop_first=True, como se muestra en el código de procesamiento asociado al EDA disponible en el repositorio de GitHub .

6.2 Escalamiento de las características

Escalar las características es un paso fundamental en la aplicación de diversos modelos de *machine learning*, ya que muchas técnicas de aprendizaje automático son sensibles a las diferencias de escala entre las variables. Algunos algoritmos, como los árboles de decisión y los bosques aleatorios, son invariantes a este tipo de transformación porque dividen el espacio de características en función de umbrales y no dependen de la magnitud absoluta de los datos [43].

Sin embargo, en el caso de los algoritmos basados en distancias y densidades, como *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*), la transformación de escala es crucial. Esto se debe a que estos métodos calculan distancias entre puntos y determinan agrupaciones en función de la densidad de datos en el espacio de características [44]. Si las variables tienen escalas muy diferentes, las dimensiones con valores más grandes pueden dominar la métrica de distancia, lo que podría sesgar la detección de agrupaciones y afectar la calidad de los clústers generados.

Con esto, es fundamental comprender los dos enfoques clásicos para expresar diferentes características en escalas equivalentes: normalización y estandarización. Es importante no confundir la normalización con otros usos del término en diferentes disciplinas. Por ejemplo, la normalización aplicada a funciones de onda en mecánica cuántica tiene un significado y una formulación completamente distintos a los de la normalización de características en un proyecto de ciencia de datos.

En el contexto del *machine learning*, la **normalización** se refiere a la transformación de los valores de una variable para ajustarlos dentro de un rango acotado, típicamente [0,1], siguiendo una ecuación del tipo:

$$x_{norm}^{(i)} = \frac{x^{(i)} - }{x_{max} - }\frac{x_{min}}{x_{max}} \tag{5}$$

donde x_{min} y x_{max} representan el valor mínimo y máximo de la característica, respectivamente.

El otro enfoque frecuentemente utilizado es la estandarización, la cual es ampliamente empleada en la mayoría de los algoritmos de *machine learning*, especialmente en aquellos basados en optimización y agrupación por similitud de características. En el caso de los algoritmos de clustering basados en distancia, como *k-means* y *DBSCAN*, la estandarización también es fundamental [43]. Dado que estos

métodos utilizan métricas como la distancia euclidiana, si las características tienen escalas diferentes, aquellas con valores más grandes pueden dominar la asignación de *clusters*, afectando la calidad de los grupos formados. Por esta razón, hemos optado por este método para el escalamiento de las características de la base de datos, debido a su compatibilidad con el objetivo de la propuesta [45].

En esta técnica, las columnas de características se transforman para tener una media de 0 y una desviación estándar de 1, obteniendo así los mismos parámetros que una distribución normal estándar. Sin embargo, es importante destacar que esta técnica no altera la forma de la distribución original ni convierte datos no distribuidos normalmente en datos con distribución normal.

Una de las ventajas clave de la estandarización es que, al escalar los datos con media 0 y varianza unitaria, se conserva información útil sobre los valores atípicos (*outliers*), lo cual es fundamental en la estructura de nuestros datos. Por ejemplo, en la variable rendimiento por hectárea (*rph*), podemos encontrar valores muy bajos, como 0 kg/ha, y valores altos, como 3500 kg/ha. Aunque estos últimos pueden considerarse *outliers*, no deben ser tratados como datos anómalos. A diferencia de la normalización, la estandarización es menos sensible a estos valores extremos, lo que la hace más adecuada para nuestro análisis.

La estandarización se expresa matemáticamente de la siguiente manera:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x} \tag{6}$$

Donde, μ_x es la muestra de una columna de características concreta, y σ_x es la desviación estándar correspondiente.
Este procesamiento se ejecutó dentro del Código utilizando la función StandardScaler, perteneciente a la librería sklearn.preprocessing para todas las variables del data-frame.

6.3 Compresión de datos mediante la reducción de la dimensionalidad

La reducción de dimensionalidad es un conjunto de técnicas matemáticas que transforman un conjunto de datos de alta dimensión en un nuevo subespacio vectorial con menor dimensionalidad. De esta manera, los datos reducidos conservan la estructura original y las relaciones entre los puntos, facilitando su visualización, análisis y procesamiento.

Las técnicas de reducción de dimensionalidad se dividen en dos enfoques principales: selección de características y extracción de características [45].

La ***selección de características*** se utiliza cuando se tiene certeza sobre qué variables influyen más en el conjunto de datos. En este caso, la representación de baja dimensión está compuesta por un subconjunto de las variables originales que mejor describen la información, mientras que las demás son descartadas.

Sin embargo, ¿qué ocurre cuando no es posible determinar cuáles son las variables más influyentes? ¿O cuando todas las variables parecen tener una relevancia similar?. En nuestro caso particular, trabajamos con una base de datos que recopila información del sector productivo del cacao en un municipio. Dado

el gran número de variables disponibles, no es evidente cuáles son las más relevantes para aplicar un algoritmo de machine learning que agrupe las unidades productivas según características similares.

Por esta razón, empleamos la metodología de extracción de características, cuyo objetivo es representar el conjunto de datos original en un nuevo conjunto de variables de menor dimensión. Este nuevo conjunto no es más que una combinación de las variables de entrada, permitiendo una mejor representación de la estructura subyacente de los datos.

En el contexto de la reducción de dimensionalidad, la extracción de características se puede entender como un proceso de compresión del conjunto de datos con el objetivo de conservar la mayor cantidad de información relevante. Para ello, una de las técnicas más utilizadas es el Análisis de Componentes Principales (PCA), basado en la factorización matricial.

El objetivo de PCA es expresar los datos originales de alta dimensión como una multiplicación de matrices de menor dimensión. Para lograr esto, se construye la matriz de covarianza, la cual se descompone obteniendo sus valores y vectores propios según la ecuación 7.

$$C_v = lv \tag{7}$$

Donde:

- l son los autovalores (valores propios), que representan la varianza explicada por cada componente.
- v son los auto vectores (vectores propios), que indican las direcciones principales de los datos.
- La solución de obtiene resolviendo:

$$\det(C - lI) = 0 \tag{8}$$

siendo *I* la matriz identidad.

Posteriormente, los valores propios se ordenan en orden decreciente para seleccionar los vectores propios correspondientes a las componentes principales que capturan la mayor parte de la varianza del conjunto de datos. Así, se obtiene una nueva representación en un subespacio de menor dimensión que retiene la mayor cantidad de información posible.

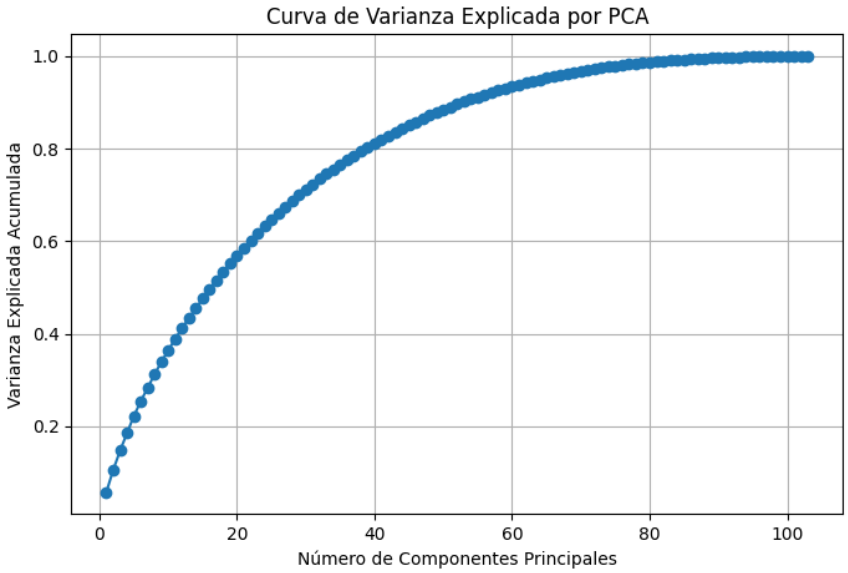


Figura 10. Curva de varianza explicada acumulada por componentes principales.

En este sentido se aplicó PCA en nuestro conjunto de datos considerando que esta técnica es efectiva cuando las variables presentan relaciones lineales, es decir, cuando existe una alta correlación entre ellas. Sin embargo, el hecho de que se requiera un gran número de componentes para capturar el 100% de la varianza (como se muestra en la Figura *10*), sugiere que el *dataset* posee múltiples dimensiones redundantes, en parte debido a la aplicación de One-Hot Encoding, lo que ha introducido colinealidad artificial en los datos.

A pesar de haber eliminado una de las categorías de cada variable codificada para evitar multicolinealidad, el resultado indica que las variables categóricas no siguen una estructura lineal clara, lo que limita la efectividad de *PCA* en este caso en específico. Es importante señalar que la alta dimensionalidad observada no implica que el conjunto de datos original tenga una estructura de naturaleza lineal. De hecho, asumir que la base de datos que describe el sector cacaocultor presenta un comportamiento lineal sería un error metodológico, ya que este tipo de datos esta descrito por factores multidimensionales y no lineales, como condiciones climáticas, manejo agronómico y prácticas socioeconómicas.

Dado que PCA maximiza la varianza bajo una representación lineal, su aplicación en este contexto no es la más adecuada. Por esta razón, utilizamos UMAP (Uniform Manifold Approximation and Projection), una técnica de reducción de dimensionalidad basada en grafos.

UMAP es una técnica de reducción de dimensionalidad no lineal que tiene una fundamentación matemática basada en la geometría de Riemann y en la rama de la matemática que estudia las formas denominada topología algebraica, disciplinas que permiten describir la estructura de los datos más allá de las relaciones lineales. Este enfoque hace que UMAP sea una herramienta práctica y escalable para trabajar con datos del mundo real [46].

Este algoritmo parte de tres consideraciones fundamentales:

1. Los datos de entrada están distribuidos uniformemente sobre una variedad de Riemann.
2. La métrica de Riemann es localmente constante lo que define la distancia entre un espacio curvo (como una superficie).
3. Que en las variedades no hay discontinuidades, lo que permite que el algoritmo construya una representación coherente en espacios de baja dimensión.

Internamente, cuando se utiliza la biblioteca umap-learn en *Python*, desarrollada por Leland McInnes y publicada en 2018, el algoritmo construye primero una representación topológica difusa de los datos en alta dimensión. Es importante recordar que en este punto estamos tratando con espacios geométricos descritos desde el estudio de las formas (topología) y no con enfoques matriciales**,** como lo hace *PCA*, que se basa en la descomposición de la varianza.

Al reducir la dimensionalidad, UMAP busca preservar tanto la estructura global como la local de los datos, minimizando las diferencias espaciales con respecto a la representación original. Matemáticamente, esto se logra mediante la minimización de la entropía cruzada, la cual es clave en el funcionamiento de UMAP.

Es importante recordar que el término **entropía** proviene de la segunda ley de la termodinámica y suele asociarse erróneamente con el nivel de desorden molecular de un sistema. Sin embargo, la interpretación correcta no es el desorden en sí, sino que, cuando un sistema alcanza su máxima entropía, la pérdida de información es muy alta en términos probabilísticos. Por esta razón, *UMAP* busca minimizar la entropía para conservar la mayor cantidad de información del *dataset* original.

Este principio convierte a UMAP en un algoritmo de reducción de dimensionalidad estocástico, que emplea descenso de gradiente estocástico (SGD) para optimizar su función de costo, como se muestra en la ecuación 9. Los detalles matemáticos están disponibles en [45].

$$C = \sum_i \sum_j v_{ij} \log\left(\frac{v_{ij}}{w_{ij}}\right) + (1 - v_{ij}) \log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \quad (9)$$

En la sección anterior se mostró lo pasos seguidos para la transformaciones de variables como la estandarización de las características para garantizar una distancia entre los datos, este paso es fundamental para la representación vectorial de los datos específicos que obedece a que cualquier otra transformación debe aplicarse antes de ejecutar UMAP como se menciona en [47]

Para aplicar este algoritmo a nuestro conjunto de datos preprocesado, ajustamos los hiperparámetros clave de UMAP, como el número de vecinos (n_neighbors), que determina cuántos puntos cercanos se consideran en el espacio de alta dimensión al calcular similitudes. Este valor varía entre 0 y *n* (donde *n* es el número total de observaciones). Valores más altos capturan mejor la estructura global de los datos en la transformación a baja dimensión, mientras que valores más pequeños preservan mejor las estructuras locales.

UMAP permite integrarse con algoritmos de aprendizaje supervisado y no supervisado, pero su mejor performance se alcanza cuando se aplican a algoritmos de aprendizaje no supervisado ya que permite

controlar la densidad de los agrupamientos mediante el hiperparámetros min_dist. Este parámetro define la distancia mínima permitida entre puntos en el espacio de baja dimensión y varía entre 0 y 1. Un valor bajo tiende a agrupar los puntos más estrechamente, generando clústeres más densos, mientras que valores más altos producen una distribución más dispersa.

Para seleccionar la combinación de hiperparámetros que mejor se ajusta a nuestros datos, se realizó una iteración sobre diferentes valores en busca de la configuración óptima que permitiera una mejor agrupación en el nuevo espacio vectorial, reduciendo la dimensionalidad a dos componentes. Como resultado de este proceso, se obtuvieron los siguientes resultados.

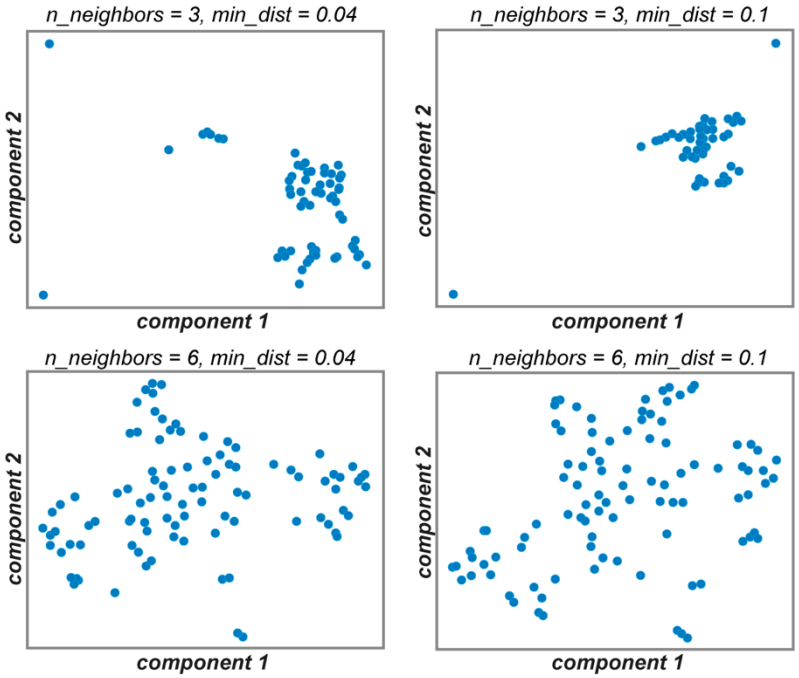


Figura 11. Variaciones de UMAP según n_neighbors y min_dist.

En relación con el objetivo general y el dataset del sector cacaocultor, donde se busca identificar grupos de unidades productivas con características similares en función de variables sociagronomica y espectrales, la ejecución de la reducción de dimensionalidad con UMAP, representada en la Figura *11*, permite evidenciar que con un valor de n_neighbors=3 se detectan estructuras locales más definidas, reflejando posibles diferencias entre las unidades productivas. Además, al ajustar min_dist a un valor de 0.04, los grupos se mantienen más compactos, lo que facilita su identificación en análisis posteriores.

UMAP permite conservar un equilibrio entre la estructura local y global de los datos. Por ejemplo, al iterar con n_neighbors=6, se pierden detalles en la definición de los grupos, lo que sugiere que se están mezclando datos que podrían diferenciarse mejor con un menor número de vecinos. Asimismo, al utilizar min_dist=0.1, los puntos quedaron más dispersos, dificultando la identificación de agrupaciones.

Por esta razón, para el desarrollo posterior seleccionamos la configuración con los hiperparámetros descritos en la Tabla 16, ya que mantiene la estructura local de los datos y permite identificar patrones

específicos en las unidades productivas. Además, en la visualización se forman grupos compactos, lo que facilita la detección de relaciones en el dataset original y evita representaciones dispersas.

Tabla 16. Selección de hiperparámetros.

Hiperparámetros	
n_neighbors	3
min_dist	0.04
random_state	75
n_components	2

Como UMAP es un algoritmo estocástico se emplea un random_state para garantizar que cada vez que se ejecute el algoritmo los resultados sean los mismos durante las ejecuciones.

Este algoritmo se ejecutó en Google colab bajo el entorno de ejecución CPU a través del backend de Google engine en python3, con recursos del sistema asignado de 12.7 GB de memoria RAM y un disco de 10.7 GB. El tiempo estimado de ejecución fue de 4 minutos con la versión de UMAP 0.5.6.

6.4 Selección y definición de los modelos de aprendizaje automático

En la era de la revolución 4.0, caracterizada por la abundancia de datos estructurados y no estructurados, el aprendizaje automático ha evolucionado como un subcampo de la inteligencia artificial. Desde la segunda mitad del siglo XX, esta disciplina ha desarrollado algoritmos capaces de aprender de los datos para hacer pronósticos y tomar decisiones de manera autónoma.

En lugar de depender exclusivamente del análisis manual de grandes volúmenes de datos, el aprendizaje automático ofrece una alternativa más eficiente para extraer conocimiento, permitiendo mejorar progresivamente el rendimiento de modelos predictivos y de clasificación que respaldan la toma de decisiones basada en datos. Su impacto ha trascendido el ámbito de la informática, desempeñando un papel cada vez más relevante en diversas áreas de la vida cotidiana [48], incluida la agricultura.

En relación con el objetivo general de este proyecto de grado, que busca identificar características similares entre las unidades productivas de cacao mediante técnicas de *machine learning*, se debe recordar que este desarrollo se deriva de un proyecto macro de ciencia y tecnología avalado por el Ministerio de las TIC. Este proyecto se enmarca como una iniciativa piloto dentro de la investigación científica, con el propósito de aplicar soluciones basadas en ciencia de datos al sector agrícola, específicamente al subsector cacaocultor.

Considerando que en el proyecto macro es necesario seleccionar unidades productivas representativas para la aplicación de estudios más detallados como los análisis foliares y la caracterización de fruto, surge la necesidad de diseñar una estrategia metodológica que permita optimizar los recursos disponibles y garantizar la pertinencia de la selección.Debido a limitaciones presupuestales, estos estudios avanzados no pueden aplicarse a la totalidad de las unidades productivas, por lo que se requiere una metodología que permita identificar y agrupar predios con características similares.

Como punto de partida, se cuenta con información base de todas las unidades productivas, incluyendo

encuestas socioagronómicas, análisis de suelos y estudios multiespectrales. A partir de estos datos, se desarrolló un modelo de *machine learning* que agrupa predios con características similares, con el objetivo de seleccionar unidades representativas para estudios más profundos. De esta manera, los resultados obtenidos son extrapolables a otros predios dentro de cada grupo identificado.

Esta estrategia metodológica siguió el enfoque del método *Multigrid*, que consiste en partir de una grilla con la máxima resolución de información disponible y reducir progresivamente la dimensionalidad de los datos, generando representaciones más compactas que conservan la estructura del universo inicial. De esta manera, se logró una representación que permite seleccionar unidades productivas de manera eficiente, asegurando que los análisis detallados sean extrapolables a otros predios con características similares.

Para aplicar un método de aprendizaje supervisado, es fundamental definir con claridad cuál es la variable objetivo (*target*), es decir, el criterio con el que se construirá la clasificación. Esto nos lleva a cuestionarnos: *¿cuál o cuáles son las variables que determinan las características de estos predios?* Aunque esta pregunta parece sencilla, en el caso específico del cultivo de cacao en el departamento del Huila, su respuesta es compleja.

A diferencia de otros sectores, en la agricultura específicamente en cacao del departamento no existe una variable única y clara que permita determinar con certeza qué define un predio con características especiales. Como menciona *Bugbee en* el estudio titulado *"LEDs for photons, physiology and food"* [49], en la agricultura existen ocho factores clave para una producción precisa y eficiente: suelo, agua, energía, CO_2, viento, velocidad del viento, nutrientes y oxígeno. Sin embargo, si a esto se añaden variables relacionadas con la distribución espacial, el manejo agronómico y factores socioeconómicos, la complejidad del análisis aumenta significativamente.

Por otro lado, el aprendizaje por refuerzo no es contemplado en este estudio debido a su naturaleza y funcionamiento, que no se ajustan a los objetivos del proyecto. En este contexto, el enfoque más adecuado es el *aprendizaje no supervisado*, ya que no requiere etiquetas ni variables objetivo-específicas, permitiendo descubrir patrones ocultos en los datos. Este enfoque se adapta tanto a la naturaleza de la información disponible como al objetivo de la investigación, proporcionando una base metodológica sólida para la agrupación de unidades productivas de cacao en el Huila.

6.5 Clustering

De acuerdo con lo anterior, el desarrollo de esta propuesta se basa en el uso de algoritmos de aprendizaje no supervisado. Para ello, se realizó una revisión exhaustiva de los algoritmos de agrupamiento más adecuados según la naturaleza de nuestros datos, considerando que las técnicas de clustering fueron diseñadas inicialmente para identificar y agrupar puntos con características similares. Los algoritmos de clustering generalmente están desarrollados para trabajar en conjuntos de datos unidimensionales o bidimensionales, por esta razón se aplicaron después de la ejecución de algoritmos de reducción de dimensionalidad como *UMAP* para nuestro caso particular y como se desarrolla en [50].

Los algoritmos de clustering**,** al igual que otras técnicas de aprendizaje automático, se dividen en tres categorías principales: las basadas en particiones, jerárquicas y basadas en densidad. Los métodos de partición, como K-Means es uno de los más empleados en esta clase, requieren definir previamente el

número de clústeres y ajustar iterativamente sus límites para minimizar la métrica de error, el cual utilizamos como línea base en nuestro desarrollo. Por otro lado, los algoritmos jerárquicos construyen un dendograma, una estructura en forma de árbol que permite agrupar o dividir los datos de manera secuencial, determinando el número óptimo de clústeres mediante una métrica de rendimiento. Finalmente, los algoritmos basados en densidad, como DBSCAN, identifican regiones de alta y baja densidad, lo que les permite descubrir grupos con características similares y detectar valores atípicos en zonas dispersas. En nuestro caso, seleccionamos HDBSCAN, una versión avanzada que combina el enfoque jerárquico con clustering basado en densidad, mejorando la flexibilidad de DBSCAN al introducir una estructura jerárquica en el proceso de agrupamiento. Esta elección se debió a su robustez, modernidad y perfomance con datos espaciales, además de su gran desempeño al combinarlo con UMAP, lo que optimiza la representación de datos en espacios de menor dimensión como se evidencia en trabajo desarrollado por I. de Zarza denominado "UMAP for geospatial data visualization" [51]. En consecuencia, implementamos K-Means y HDBSCAN, comparando sus resultados, los cuales se describen en detalle en las secciones siguientes.

6.5.1 K-Means

K-means es uno de los algoritmos basados en particiones más utilizados debido a la facilidad de implementar y a que computacionalmente es muy eficiente en comparación con otros algoritmos de clustering. El funcionamiento de este algoritmo consiste en definir inicialmente unos centroides en la representación espacial de los datos, determinados mediante un valor denominado *k*. Dicho valor (*k*) representa el número de grupos que deben ser especificados previamente. Luego, todos los puntos son asignados al centroide más cercano. La posición de cada centroide se recalcula iterativamente mediante la minimización de la suma de cuadrados de las distancias dentro de cada grupo. Matemáticamente, esta expresión tiene la siguiente forma:

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - m_j \right\|^2 \quad (10)$$

Dónde son $x_i \in C_j$ son puntos y el grupo interior C_j y $m_j$son el promedio de puntos en el cluster C_j.

Debe tenerse en cuenta que K-means presenta ciertas limitaciones, siendo una de las más importantes la necesidad de elegir previamente el número de grupos (*k*). Una selección inadecuada de *k* puede generar resultados poco eficientes o imprecisos en el proceso de clustering. Además, este algoritmo tiende a mostrar un mejor desempeño al identificar clústeres con formas esféricas.

Uno de los principales desafíos a resolver del aprendizaje no supervisado es que no se conocen las etiquetas de las clases reales que permitan una agrupación adecuada de unidades productivas de cacao con respecto a una variable específica, esto puede resumirse de manera general e indistinta a la aplicación, como el hecho de que inicialmente no existe una variable privilegiada con respecto a las demás. Por lo tanto, para evaluar la calidad del clustering, se emplearon métricas propias del algoritmo como la suma de los errores cuadráticos (SSE, por sus siglas en inglés), que mide la cohesión interna de los clústeres mediante un enfoque iterativo para minimizar esta suma. La implementación del algoritmo *K-means* se realizó utilizando la librería *Scikit-learn* en *Python*, que permite acceder directamente al valor del SSE de cada clúster mediante el atributo denominado inertia_, evitando así la necesidad de

calcularlo explícitamente.

Con base en el valor de SSE obtenido, se utilizó una herramienta gráfica conocida como el método del codo para estimar el número óptimo de clústeres (*k*). En la Figura *12*, se presenta el gráfico del método del codo aplicado al conjunto de datos analizado, mostrando la relación entre el número un óptimo de clústeres. Podemos decir que, cuando k aumenta, la distorsión del clúster disminuye. Esto gracias a que los data points estarán más cerca del centroide al que han sido asignados. La idea central del este método es poder identificar el valor adecuado de k en el que la distorsión empieza a aumentar más rápidamente, lo que se verá claramente si representamos la distorsión para diferentes valores de k.

La Figura 12 muestra que a partir del número de k entre 3 a 6, la reducción en la inercia (SSE) es menos pronunciada. Esto indicaría que el valor óptimo de *k* estaría en estos valores, ya que es allí donde se produce el "codo" o punto de inflexión más notorio. A partir de este punto, añadir más clústeres no aportaría una mejora sustancial en la explicación de la variabilidad de los datos, lo que indica que la segmentación óptima de los datos para k-means estaría en torno de 3 a 6 grupos.

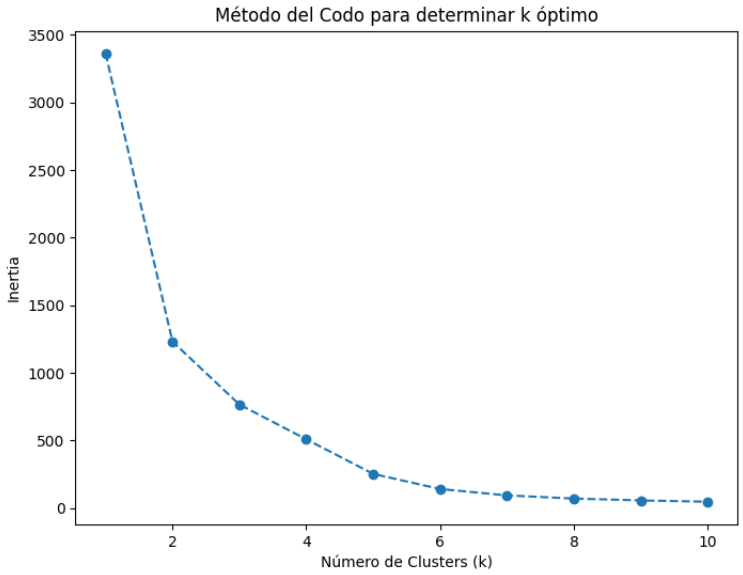


Figura 12. Gráfica del Método del Codo para determinar el número óptimo de clústeres.

Para determinar de forma más precisa el agrupamiento óptimo, se llevaron a cabo diversas ejecuciones del algoritmo k-means, ajustando tanto los valores de *k* como las métricas de distancia utilizadas, como se muestra en la Tabla *17*. Este procedimiento permitió visualizar y comparar los resultados de cada iteración, estableciendo una referencia clara para la selección final del número de clústeres. Además, se evaluó la cohesión interna de los grupos mediante métricas como la inercia y la puntuación del coeficiente de silueta, lo que proporcionó una validación cuantitativa de la calidad del agrupamiento obtenido. De esta manera, fue posible identificar el valor de *k* que mejor representaba la estructura subyacente de los datos, asegurando un balance entre la simplicidad del modelo y la representatividad de las agrupaciones formadas.

Tabla 17. Parametros K-means.

Iteración	# K	metrics
a	4	euclidean
b	6	euclidean
c	5	cosine
d	7	manhattan
e	4	cosine

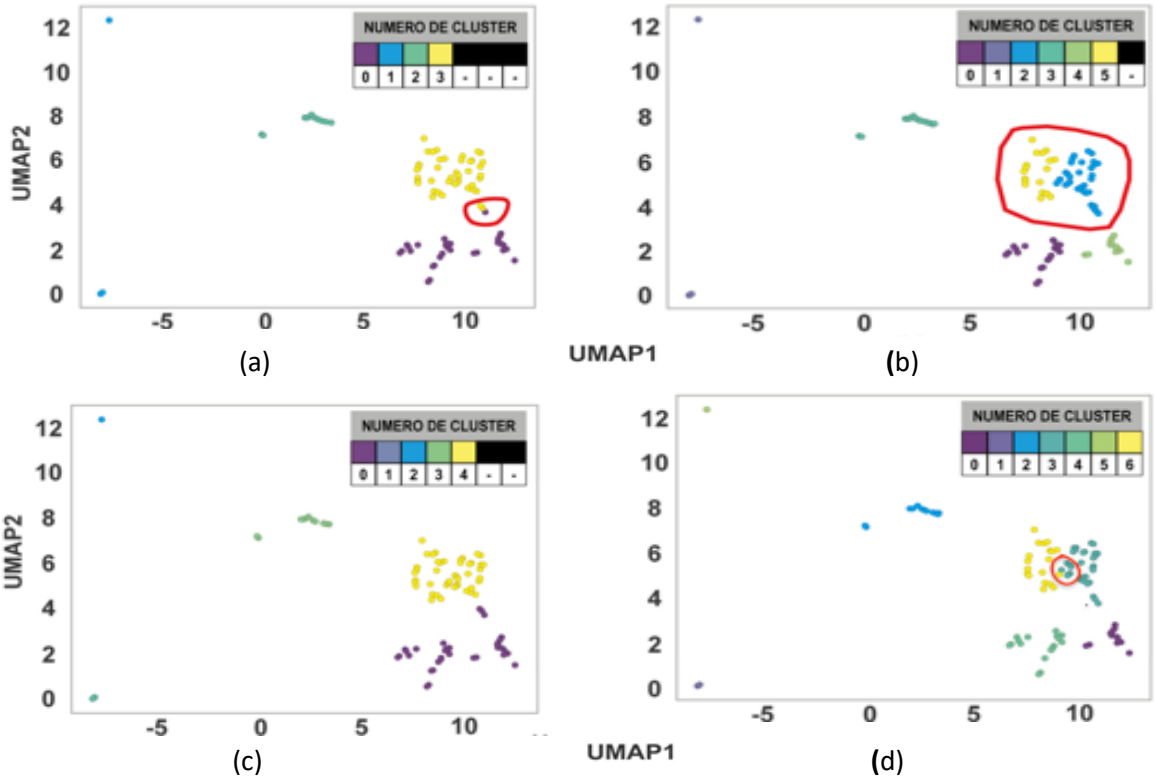


Figura 13. Comparación de resultados de agrupamiento con diferentes parámetros de k-means.

Los resultados de las distintas ejecuciones se presentan en la Figura 13. De manera general, se observa que k-means no identifica valores atípicos, debido a que el propio algoritmo tiene la limitación de identificar estos valores y etiquetarlos dentro de un único clúster. Además, al utilizar un valor de k menor que 3, el algoritmo tiende a generalizar en exceso y a formar grupos amplios, sin tener en cuenta la distribución espacial real de los datos.

En la Figura 13-a, donde se emplea k = 4 y una métrica de distancia euclidiana relacionado con la Tabla 17, se forman cuatro grupos diferenciados; sin embargo, se evidencia una superposición que es resaltada en rojo entre los clústeres 3 y 0. Al incrementar el número de clústeres a k = 6 manteniendo la misma métrica de distancia, como puede verse en la Figura 13-b, desaparece esa superposición presente en la ejecución anterior. No obstante, el clúster 3 de la Figura 13-a, que era el más grande, ahora se divide en dos grupos separados, al igual que el clúster 0.

En la Figura 13-c, con k = 5 y métrica de distancia coseno, el resultado general se asemeja bastante a la configuración de la Figura 13-a con k = 4 y distancia euclidiana. La diferencia principal radica en que los puntos que se encuentran más alejados pasan a etiquetarse como clústeres independientes, producto de la forma en que la métrica coseno calcula la similitud. Por su parte, en la Figura 13-d se observa un comportamiento muy parecido al de la Figura 13-c: dichos puntos alejados se agrupan de manera separada y, adicionalmente, se presenta una nueva superposición entre los clústeres 6 y 3, similar a la de la Figura 13-c.

En términos generales, la métrica euclidiana tiende a brindar un mejor desempeño con datos que describen escenarios de la vida real al aplicar k-means, especialmente cuando se han estandarizado las características para asegurar escalas uniformes. Las visualizaciones adicionales están disponibles en la sección de Anexo B, se aprecian patrones de agrupamiento similares que confirman estas descripciones.

Teniendo en cuenta estas consideraciones, podemos inferir que el mejor desempeño de k-means se obtiene con la configuración de *k = 6* y la métrica de distancia euclidiana. Esto se debe a que, de manera visual, se distinguen con claridad seis grupos, cada uno con un núcleo compacto, lo que indica que el algoritmo segmenta adecuadamente los datos. Además, se observa muy poca superposición entre clústeres cercanos, comportamiento que no se evidencia bajo otras configuraciones. Asimismo, esta elección ofrece un equilibrio entre la cantidad de grupos y la variabilidad de los datos: si se incrementa el valor de k por encima de 6 se dividen los datos más de la cuenta haciendo que el comportamiento tienda a generando clústeres poco justificables; por el contrario, con valores de *k* menores (por ejemplo, 3 o 4), se agruparían demasiados puntos en un mismo clúster haciendo que se generalicen los grupos.

El Código completo relacionado con el clustering de k-means e iteraciones está disponibles en el repositorio de GitHub .

6.5.2 HDBSCAN

El algoritmo K-means fue ejecutado paralelamente con HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), un algoritmo de agrupamiento espacial basado en densidades jerárquicas capaz de identificar clústeres significativos en presencia de valores atípicos (outliers). En comparación con K-means, HDBSCAN no necesita determinar el número de clústeres (*k*) de manera preliminar, sino que determina de forma automática el número óptimo basándose en la estructura inherente de los datos, generando así agrupamientos más robustos frente a la presencia de datos atípicos.

La ejecución del algoritmo *HDBSCAN* se realizó en *Python*, utilizando la librería hdbscan versión 0.8.40. En esta implementación, se configuraron parámetros fundamentales como: min_samples, que define el número mínimo de puntos que se necesitan para considerar una región como núcleo de densidad;

min_cluster_size, que establece el tamaño mínimo aceptable para formar un clúster válido; *metric*, que determina la métrica utilizada para calcular la distancia entre puntos (por ejemplo, Manhattan o Euclidiana); y *allow_single_cluster*, parámetro booleano que especifica si se permite o no la generación de un único clúster [52].

En busca de un agrupamiento óptimo, se llevaron a cabo diferentes iteraciones con el fin de evaluar la robustez y el rendimiento con diferentes configuraciones de parámetros. En cada iteración se variaron principalmente los parámetros descritos los cuales son mostrados en la Tabla 18

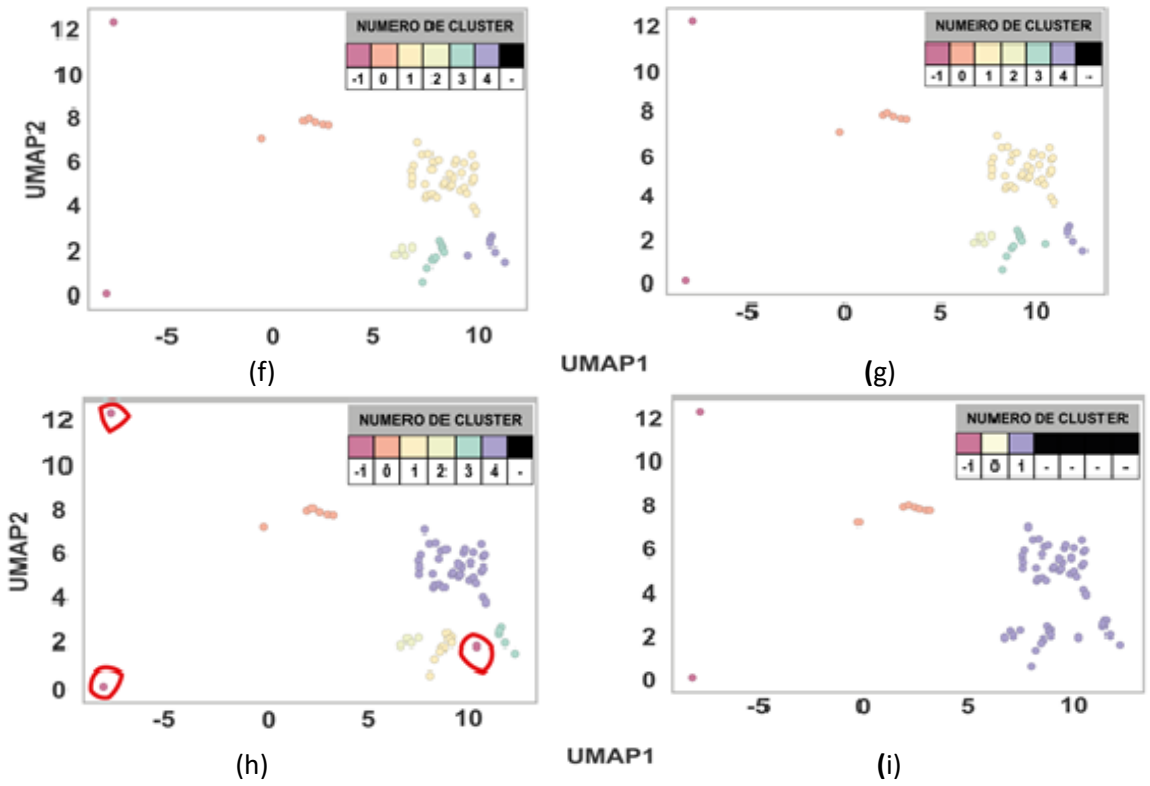


Figura 14. Comparación de resultados de agrupamiento con diferentes parámetros de HDBSCAM.

Tabla 18. Parametros HDBSCAN.

#iteración	min_samples	min_cluster_size	metric	#_de_cluster	Silhouette_Score
f	2	6	Euclidean	6	0.58
g	2	8	Manhattan	6	0.56
h	5	6	Euclidean	6	0.59
i	5	13	Euclidean	3	0.66

j	5	13	Manhattan	4	0.55
k	5	8	Euclidean	6	0.59

En la Figura 14 se muestran cuatro iteraciones principales de HDBSCAN con diferentes configuraciones de parámetros y distintos resultados de silhouette score. En todas estas ejecuciones, HDBSCAN logró identificar valores atípicos (outliers) que fueron etiquetados como parte del clúster -1. Al comparar los resultados de las Figura 14-f y 15-g, se observa un comportamiento muy similar: en ambos casos se identifican seis clústeres bien definidos, con *silhouette scores* de 0.58 y 0.56 respectivamente, lo cual indica una estabilidad en la formación de los grupos.

En la Tabla 18 se describen los parámetros utilizados para obtener estos resultados. Es importante señalar que en ambos ejemplos se utilizó el mismo valor de min_samples, pero se variaron tanto el tamaño mínimo de clúster (min_cluster_size) como la métrica de distancia. Estas diferencias en la métrica tienen implicaciones directas en la forma en que se agrupan los puntos. Es decir, en la Figura *14*-f, el clúster etiquetado como 4 incluye un grupo de puntos intermedio entre los clústeres 4 y 3, mientras que en la Figura *14*-g, dichos puntos aparecen integrados dentro del clúster etiquetado como 3.

Esto se debe a que la métrica Euclidiana utilizada en la Figura *14*-f, calcula la distancia como una línea recta entre dos puntos como se muestra en la ecuación 11, siendo más adecuada cuando las variables se encuentran en la misma escala o han sido estandarizadas previamente como en el desarrollo metodológico planteado en esta propuesta entre otros. Por otro lado, la métrica Manhattan utilizada en la Figura *14*-g, se basa en la suma de las diferencias absolutas, resultando menos sensible a valores atípicos y, por lo general, ofrece un mejor rendimiento con datos de mayor dimensionalidad, ya que mide la distancia acumulada en cada dimensión.

$$eculidean_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{11}$$

Por otro lado en la Figura 14-h y 15-i se presentan otras dos iteraciones adicionales del algoritmo HDBSCAN con distintas configuraciones de parámetros. Al comparar estos resultados, se observan diferencias importantes respecto a las figuras anteriores. En particular, la Figura 14-h muestra la presencia de varios puntos clasificados como valores atípicos lo que sugiere que la configuración utilizada en esta iteración fue más sensible a la dispersión o aislamiento de ciertos puntos en el espacio reducido por UMAP, Sin embargo, a pesar de que estos puntos se encuentran relativamente cerca de los clústeres **3** y 4, terminan etiquetados como outliers.

Este resultado contradice la conceptualización teórica de los modelos basados en densidad, distancia y métrica de distancias aquí empleadas, ya que no debería considerarse a dichos puntos como atípicos cuando su posición, dentro del espacio comprimido por la reducción de dimensionalidad estocástica *UMAP*, los ubica de manera clara cerca de otros grupos. Aunque en esta configuración el coeficiente de silueta resulta elevado, se decidió descartar esta representación por no coincidir adecuadamente con los supuestos del modelo.

En la Figura 14-i, se observa una reducción drástica en la cantidad de clústeres identificados con dos grandes agrupaciones junto a presencia de valores atípicos. Esto indica que un ajuste inadecuado de parámetros como min_samples, min_cluster_size o la métrica de distancia puede conducir a un exceso de *generalización*, agrupando puntos que antes se consideraban en clústeres separados y perdiendo información relevante sobre la estructura de los datos. Este comportamiento se repite en las iteraciones j y k las cuales no fueron incluidas, pero están disponibles en Anexo C; el uso de valores elevados para min_samples y min_cluster_size también reduce el número de clústeres a tan solo dos. Esa situación impide identificar las características similares de cada grupo, pues prácticamente todos los puntos quedan asignados a un mismo clúster.

Estas diferencias resaltan la importancia de realizar una selección adecuada y fundamentada de los parámetros en HDBSCAN, considerando cuidadosamente el equilibrio entre sensibilidad al ruido, capacidad de identificación de valores atípicos, y la estabilidad general del agrupamiento.

En relación de lo anterior y teniendo en cuenta las comparaciones entre las distintas configuraciones, se seleccionó como resultado la clusterización con HDBSCAN correspondiente a la Figura 14-f, Esta elección se debe a que, con dichos valores, se logra un equilibrio apropiado entre la identificación de grupos claramente diferenciados, una identificación de valores atípicos y un coeficiente de silueta estable, condiciones relevantes para el desarrollo de la propuesta. Al contar con clústeres coherentes y bien delimitados, se facilita la posterior integración de estos resultados en el proceso de segmentación y análisis.

7 EVALUACIÓN DE LOS MODELOS

Al segmentar el conjunto de datos que describe el sector cacaocultor del municipio de Baraya – Huila, se llevó a cabo un perfilamiento detallado de las unidades productivas, con el objetivo de descubrir patrones ocultos e identificar tanto características comunes como diferencias entre los productores. Para garantizar la validez del análisis de segmentación aplicado sobre el conjunto de datos transformado, se utilizaron métricas de evaluación orientadas a medir la calidad de los agrupamientos generados por los algoritmos de clustering empleados, los cuales corresponden a técnicas de aprendizaje no supervisado.

Considerando que se trabaja con datos multidimensionales, estos fueron proyectados a un espacio bidimensional utilizando técnicas de reducción de dimensionalidad. Esta transformación permitió validar visualmente la distribución de los grupos generados, es decir, observar qué tan bien los algoritmos lograron separar y cohesionar los datos en torno a su estructura local y global. La visualización en espacios de baja dimensionalidad es una herramienta clave para la interpretación cualitativa de los resultados, especialmente cuando se representan mediante las Figura 13 y la Figura 14. Sin embargo, es importante resaltar que, a medida que se incrementa la dimensionalidad (más de tres dimensiones), este análisis visual pierde eficacia y se hace más complejo.

La calidad de la segmentación generada se evaluó bajo el enfoque de validez de *clustering*, el cual permite determinar qué tan apropiada es una agrupación en función de criterios objetivos [54]. En este caso se adoptó específicamente el criterio de validez interna, el cual es el más comúnmente utilizado en contextos de aprendizaje no supervisado. Este enfoque se basa únicamente en la información interna del conjunto de datos, considerando aspectos como la cohesión dentro de los clústeres y la separación entre ellos, Es importante señalar que otros métodos de validación fueron abordados en el marco teórico correspondiente.

5.1. Validación K-MEANS

Para evaluar la calidad del agrupamiento generado por *K-means* en la iteración correspondiente a la Figura 13-b, configurada con seis clústeres y previamente identificada como la de mejor desempeño en comparación con las demás tal como se describió anteriormente, se aplicaron métricas de evaluación interna. Entre ellas, se utilizó el índice de *Silhouette,* que mide qué tan similar es un punto con los elementos de su propio clúster en relación con los puntos de otros clústeres.

- Esta métrica toma valores en un rango de -1 a 1, donde: Un valor cercano a 1 indica que el punto está bien agrupado.
- Un valor cercano a 0 sugiere que el punto se encuentra en el límite entre dos clústeres.
- Valores negativos indican una posible asignación errónea del punto al clúster.

Para esta configuración, se obtuvo un valor promedio de coeficiente de *Silhouette* de 0.57, lo que indica que los grupos están bien definidos, con buena cohesión interna y una adecuada separación entre grupos. La Figura 15 muestra que los clústeres presentan valores positivos y relativamente consistentes,

lo cual sugiere que los puntos se encuentran bien asignados. En particular, los clústeres etiquetados como 4 y 5 presentan los valores más altos del coeficiente de silueta, reflejando una segmentación más clara y efectiva. Por el contrario, el clúster etiquetado como 1 evidencia cierta dispersión en sus coeficientes de silueta, lo que podría deberse a la presencia de valores atípicos. Aunque estos puntos se encuentran ampliamente separados espacialmente, el algoritmo *K-Means* los agrupa en uno solo, esto debido a la limitación de identificar valores atípicos. Para los demás clústeres se observa una estructura coherente, lo que valida la partición en seis grupos como una de las más adecuadas entre las diferentes configuraciones evaluadas con variaciones en los hiperparámetros de *K-Means*, como se evidencia en la Figura 13-b.

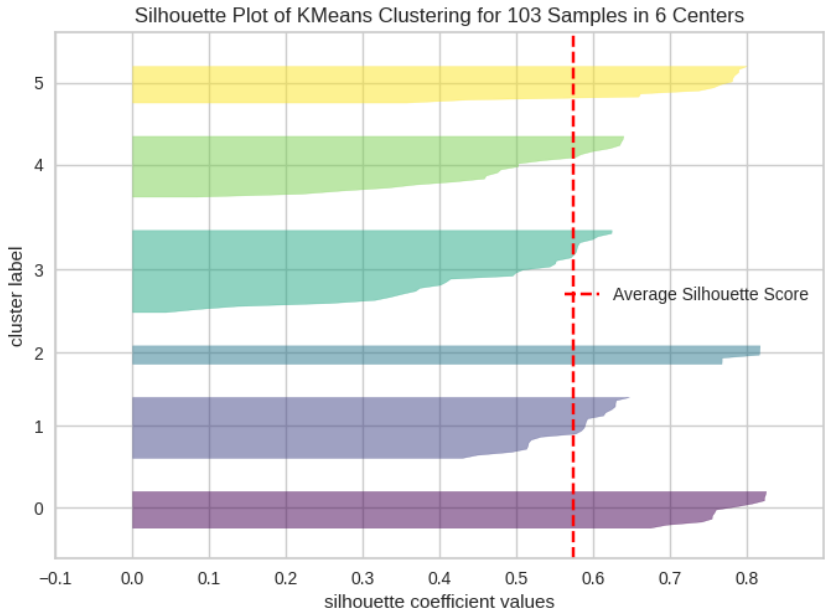


Figura 15. Gráfico de Silueta para Evaluar la Calidad del Agrupamiento con k-means (k=6).

Asimismo, la Figura 15 muestra que la mayoría de los clústeres presentan anchos consistentes y valores del coeficiente de silueta positivos, lo que indica una adecuada asignación de los puntos dentro de sus respectivos grupos. La ausencia de valores negativos sugiere que no se evidencian asignaciones anómalas. Sin embargo, se observa que los clústeres etiquetados como 0 y 1 presentan una mayor variabilidad en sus coeficientes.

Para el cálculo del índice de *Calinski-Harabasz* (CH Score) donde se evalúa la calidad del agrupamiento calculando la relación entre la dispersión entre clústeres y la dispersión dentro de cada uno de ellos. Un valor más alto indica una mejor separación entre los grupos y una mayor cohesión interna. En este análisis, el valor obtenido fue de 166.15, lo cual sugiere que la estructura de clústeres generada por *K-means* presenta una buena definición entre grupos y un agrupamiento compacto dentro de cada segmento.

Por otro lado, en el cálculo del índice de *Davies-Bouldin* (DB Score) el cual mide el promedio de las similitudes entre cada clúster y el clúster más similar, considerando tanto la dispersión interna como la

distancia entre clústeres. A diferencia del índice de *Silhouette* y el de *Calinski-Harabasz*, en este caso valores más bajos indican una mejor calidad de la segmentación. El valor obtenido en este estudio fue de 0.66, lo cual también respalda la calidad aceptable del agrupamiento, reflejando una adecuada separación entre los clústeres y una coherencia interna razonable.

La Tabla 19. Resume las métricas empleadas para *K-means* y sus respectivos puntajes:

Tabla 19. Métricas de evaluación k-mean.

Métrica	Score
# cluster	6
Silhuette	0.57
Calinski-Harabasz	155
Davies-Bouldin	0.66

En resumen, el valor promedio de silueta cercano a **0.6** respalda la validez del número de clústeres seleccionado (**k=6**) como una estructura razonablemente buena para K-means.

5.2. Evaluación HDBSCAN

En la Tabla 18 se presentan las distintas iteraciones del modelo *HDBSCAN* junto con sus respectivos hiperparámetros y los coeficientes de silueta obtenidos para cada ejecución. También, se llevó a cabo una validación visual de cada resultado, seleccionando finalmente la configuración representada en la Figura 14-g por mostrar un mejor perfomance que los demás. Esta elección no solo está fundamentada en la métrica de silueta, sino también en la interpretación estructural que permite la naturaleza jerárquica del algoritmo HDBSCAN.

Debido a esta característica jerárquica, *HDBSCAN* proporciona información adicional para la comprensión y validación de los datos. A través del análisis de la jerarquía condensada de los clústeres, como se muestra en la Figura 16. En esta figura, el dendograma representa la estructura de agrupamiento donde el eje vertical corresponde a los valores de λ *(lambda)*, una métrica de densidad definida como la inversa de la distancia del núcleo que se calcula internamente por medio de la Ecuación 12. Estas agrupaciones que son seleccionadas y formadas automáticamente por el algoritmo son aquellas que muestran la mayor persistencia, es decir, las que permanecen estables durante un mayor intervalo de valores de λ. Estas agrupaciones son resaltadas mediante formas ovaladas sobre sus respectivas ramas, mientras que las subdivisiones descartadas no.

$$l(x_i) = \frac{1}{d_c(x_i)} \tag{12}$$

Los valores mayores de λ implican descartar de forma progresiva los puntos menos densos de cada clúster. Esto puede llevar a que un clúster se divida en subconjuntos más pequeños si se debilita su cohesión interna. En este contexto, cuanto mayor sea la extensión vertical de una rama en el dendograma, mayor será la estabilidad del clúster, lo que equivale a una alta resistencia frente a su disolución. Por tanto, esta persistencia representada en la Figura 16 es un análogo gráfico de la

estabilidad cuantitativa de los clústeres, y respalda la robustez de las agrupaciones identificadas en la configuración óptima aplicada.

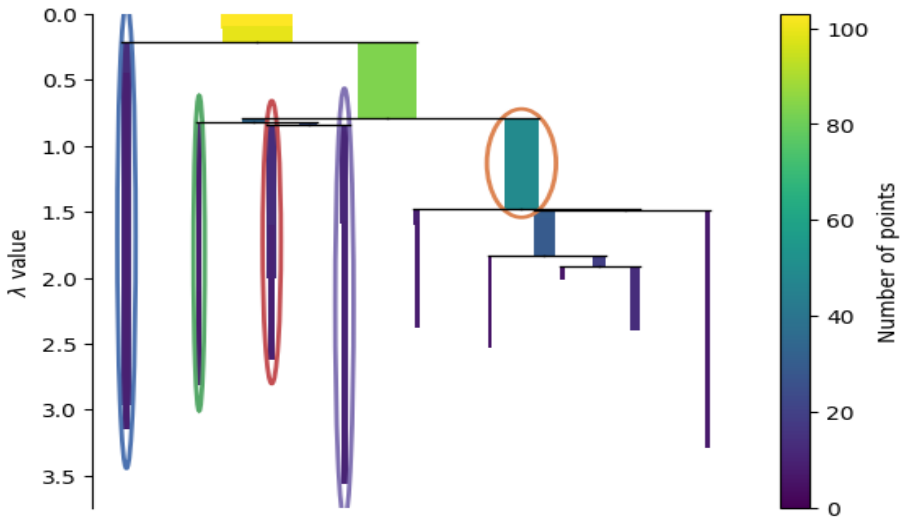
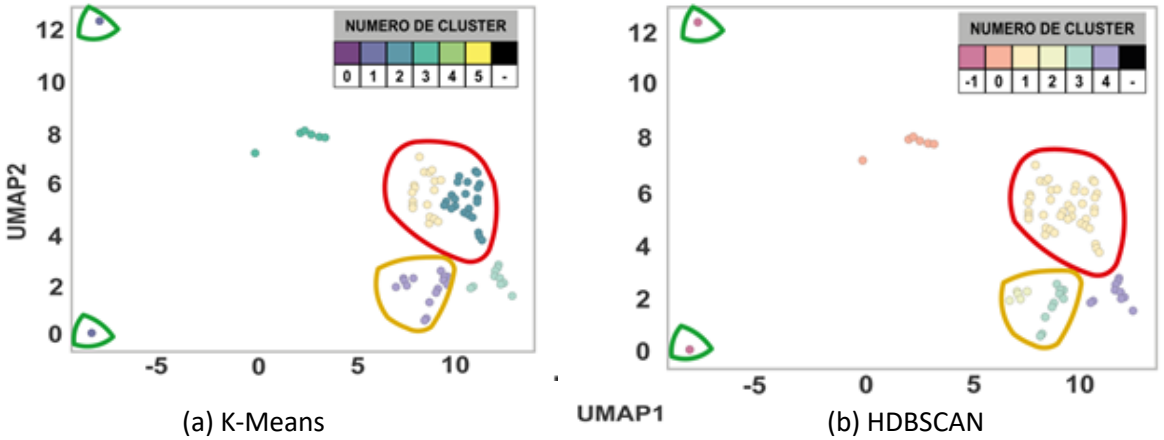


Figura 16. Dendograma de Clustering Jerárquico basado en *HDBSCAN*.

Al analizar los valores cuantitativos de persistencia de los clústeres identificados por *HDBSCAN*, calculados a través de cluster_*persistence*_ de la librería sklearn.metrics en python. Para este caso, se obtuvieron que para los cinco clústeres generados incluyendo el etiquetado con -1, que corresponde a los valores atípicos; los niveles de persistencia respectivos son de 0.586, 0.192, 0.520, 0.417 y 0.563. Estos valores indican la estabilidad relativa de cada agrupación a lo largo del proceso jerárquico de densidad. En particular, los clústeres con persistencias superiores a 0.5 (clústeres 0, 2 y 4) se consideran altamente estables, ya que mantienen su cohesión interna durante un amplio rango de valores de λ. Por otro lado, el clúster con menor persistencia (0.192) presenta una menor robustez, lo cual podría reflejar una estructura menos definida o una mayor sensibilidad a cambios en los parámetros del modelo. Esta métrica de persistencia, en conjunto con la visualización jerárquica, muestra coherencia espacial y estabilidad matemática a lo largo de la jerarquía de densidad.

Al analizar los resultados de validez obtenidos por los algoritmos de clustering, los cuales se aplicaron sobre la misma reducción dimensional obtenida mediante el método estocástico UMAP, de modo que la base de datos transformada es la misma para ambos casos. No obstante, los resultados varían según los hiperparámetros empleados para cada algoritmo. Para HDBSCAN, se debe especificar el tamaño mínimo de grupo que será considerado como un clúster, así como la métrica de distancia, y el algoritmo determina automáticamente el número de clústeres resultantes. En el caso de K-means, es necesario establecer de antemano la cantidad de clústeres a generar. Para este estudio, dicho parámetro se definió combinando el criterio del método del codo con los resultados obtenidos a partir de diferentes iteraciones, en las cuales se variaron los hiperparámetros del modelo. Adicionalmente, se tuvo en cuenta el número de clústeres identificados por el algoritmo HDBSCAN como referencia complementaria, con el fin de validar y optimizar la elección del valor de *k*.



(a) K-Means (b) HDBSCAN

Figura 17. Comparación de resultados *k-means* y *HDBSCAN*.

Al comparar los resultados obtenidos con ambos algoritmos de clustering en sus configuraciones de mejor desempeño, se observan algunas diferencias en la forma en que estructuran los clústeres. Para *K-means* (Figura *17*-a), aunque se observa la formación de seis grupos, uno menos que con *HDBSCAN*, el clúster más grande se divide en dos subgrupos etiquetados como clústeres 5 y 2. En comparación, *HDBSCAN* (Figura 17-b) mantiene estos puntos densos agrupados en un único clúster etiquetado como clúster 1. Otra diferencia relevante está relacionada con el grupo etiquetado como clúster 0 en K-means, compuesto por dos subgrupos espacialmente separados. La distancia entre estos subgrupos es incluso mayor que la existente entre los clústeres 5 y 2; sin embargo, *K-means* los agrupa en un solo clúster. *HDBSCAN*, por su parte, identifica correctamente esta separación y asigna dos clústeres distintos. Además, *K-means* presenta limitaciones al identificar puntos atípicos, agrupándolos todos en un mismo clúster sin importar su distancia espacial, este comportamiento es debido a que el algoritmo no cuenta con un método intrínseco para etiquetar valores atípicos. En cambio, *HDBSCAN* logra clasificar adecuadamente estos puntos atípicos mediante el uso del parámetro *min_samples*, el cual establece la cantidad mínima de puntos requeridos para conformar el núcleo de un clúster. A pesar de estas diferencias, ambos algoritmos mantienen una estructura similar en otros grupos, indicando coherencia en los resultados obtenidos.

En la Tabla 20 se presenta el número de puntos que componen cada clúster en ambos algoritmos, lo cual permite observar las similitudes y diferencias en la distribución de las observaciones y validar la consistencia estructural entre *K-Means* y *HDBSCAN*.

Tabla 20. Comparación del número de observaciones por clúster entre *K-means* y *HDBSCAN*.

Etiquetas de Clúster	K-means	HDBSCAN
-1	NA	7
0	21	13
1	7	49
2	28	8
3	13	13
4	13	13
5	21	NA

En *HDBSCAN*, los puntos etiquetados como -1 corresponden a observaciones consideradas como ruido o valores atípicos. En comparación, *K-means* no cuenta con la capacidad intrínseca para identificar este tipo de datos, por lo que tiende a agruparlos forzosamente dentro de uno de los clústeres predefinidos; en este caso, fueron asignados al clúster **1**. Tal como se ha mencionado previamente, una de las ventajas más significativas de *HDBSCAN* radica en su capacidad para excluir puntos atípicos mediante el parámetro *min_samples*, que establece el umbral mínimo de densidad requerido para que un conjunto de observaciones sea considerado un clúster válido. Por su parte, K-means, al ser un algoritmo particional, impone una clasificación obligatoria para todos los datos, incluso cuando algunas observaciones se encuentran alejadas espacialmente del resto.

A pesar de estas diferencias, ambos algoritmos tienden a agrupar estos puntos atípicos bajo un mismo grupo, lo que sugiere cierto grado de coherencia en la detección de regiones de baja densidad o con mayor dispersión dentro del conjunto de datos.

Adicionalmente, al comparar los resultados de segmentación generados por ambos algoritmos, se evidencian comportamientos similares en la formación de ciertos grupos. Por ejemplo, el clúster 3 de *K-means* y el clúster 0 de *HDBSCAN* agrupan 13 observaciones que son equivalentes en la representación espacial evidenciado en la Figura 17. Este patrón indica la presencia de una estructura consiste para estos grupos ya que los datos de ambos métodos logran capturar de forma coherente. Del mismo modo, el clúster 4 presenta una correspondencia directa en ambos algoritmos, manteniendo una segmentación estable a pesar de sus enfoques distintos.

No obstante, el manejo de los valores atípicos sigue siendo un punto diferenciador clave. Mientras que *HDBSCAN* los identifica y los etiqueta claramente como ruido, *K-means* los incorpora en alguno de sus clústeres, lo que puede llevar a interpretaciones erróneas. Aunque visualmente este comportamiento puede parecer similar, en realidad no lo es. *K-means* basa sus agrupamientos en una métrica de distancia y en un número fijo de clústeres (k), por lo que incluir puntos espacialmente dispersos dentro de un mismo grupo puede reflejar un mal desempeño en presencia de ruido. Esta limitación es inherente al propio algoritmo, ya que la elección del valor de *k* afecta directamente la estructura de los clústeres resultantes.

Por el contrario, *HDBSCAN*, al estar basado en una jerarquía de densidades, es capaz de identificar automáticamente regiones de baja densidad como ruido. Esta capacidad le otorga una ventaja considerable en contextos en la que los datos presentan irregularidades o dispersión significativa.

En conjunto, estos resultados refuerzan la hipótesis de que, si bien ambos algoritmos muestran consistencia en ciertos grupos, *HDBSCAN* ofrece una segmentación más robusta y realista en escenarios con ruido o datos atípicos. La coherencia observada en varios clústeres comunes también sugiere la presencia de una estructura subyacente sólida en los datos.

Como se mencionó previamente, ambos algoritmos presentan comportamientos equivalentes en la formación de ciertos grupos, sin embargo, también se evidencian diferencias significativas en la manera en que se forman otros clústeres. Como sucede para el clúster etiquetado como 1 resuelto por *HDBSCAN*, tal como se observa en la Figura 17 y la Tabla 19, en el que se agrupan 49 observaciones. Este conjunto de observaciones es resulto por K-means formando dos clústeres separados, a pesar de que los grupos resultantes están espacialmente cercanos. Esta división no parece responder a un

comportamiento lógico en términos de espacialidad, lo que sugiere una segmentación provocada por la necesidad que tiene K-means de asignar todos los puntos a un número fijo de grupos definidos previamente, sin considerar la densidad local de los datos. En contraste, *HDBSCAN*, al basarse en una estrategia de agrupamiento por densidad, reconoce la continuidad espacial entre los puntos y mantiene el grupo unido bajo una misma etiqueta.

De manera contradictoria, K-means agrupa 27 observaciones dentro de un único clúster etiquetado como 0, a pesar de que estas presentan una separación espacial considerable que, HDBSCAN, resuelve dividiéndolos en dos agrupaciones distintas, etiquetadas como clústeres 2 y 3. Este comportamiento refuerza la idea de que HDBSCAN es más sensible a la estructura espacial y a las variaciones locales de densidad, lo que le permite generar una segmentación más precisa, especialmente en conjuntos de datos heterogéneos o con distribución irregular.

En conjunto, estos patrones refuerzan la hipótesis de que, si bien ambos algoritmos presentan coincidencias en ciertas agrupaciones, HDBSCAN ofrece una segmentación más robusta y realista en escenarios con presencia de ruido, datos atípicos o estructuras espaciales complejas.

La comparación entre K-means y HDBSCAN evidencia que, si bien ambos algoritmos pueden coincidir en la identificación de ciertos grupos, sus diferencias metodológicas influyen significativamente en la calidad de la segmentación. Mientras que K-means tiende a forzar la asignación de todos los puntos a un número fijo de clústeres, HDBSCAN adapta la agrupación en función de la densidad local, permitiendo una detección más precisa de valores atípicos y estructuras espaciales heterogéneas. Esto hace que HDBSCAN sea más robusto en contextos con ruido o alta dispersión, como ocurre en el análisis de unidades productivas del sector cacaocultor.

En el siguiente repositorio de GitHub asociado al proyecto se encuentra el notebook en el formato .ipynb con el procesamiento completo desde la reducción de dimensionalidad la ejecución de los algoritmos de clustering junto con sus respectivas evaluaciones descritas anteriormente y las visualizaciones descriptivas de las exploraciones estadísticas de cada clustering formados.

8 INTERPRETACION DE RESULTADOS

8.1 Descripción y comparación de grupos productivos

En esta fase se llevó a cabo un análisis detallado de los resultados obtenidos mediante la aplicación de los algoritmos de clustering K-means y HDBSCAN sobre el conjunto de datos que describen al sector cacaocultor del municipio de Baraya. Ambos modelos permitieron identificar seis agrupaciones, de las cuales tres presentan similitudes entre sí. No obstante, se evidenciaron diferencias en la forma en que ciertos clústeres fueron conformados por cada algoritmo, tanto en lo referente a su distribución espacial como en la cantidad y composición de unidades productivas asignadas a cada grupo.

El análisis está enfocado inicialmente en los clústeres que presentan diferencias entre los dos modelos, ya que estos casos permiten establecer comparaciones directas para determinar cuál algoritmo ofrece un desempeño más adecuado en el contexto del sector cacaocultor analizado, seguido de los grupos que son equivalentes entre ambos modelos.

En función de la naturaleza de los datos los cuales fueron capturados y estructurados para su análisis por capitales, como se describió previamente se procede a realizar una comparación entre los grupos obtenidos a partir de las técnicas de clusterización aplicadas. Para ello, se asignó a cada unidad productiva una etiqueta correspondiente al clúster al que pertenece, regresando así del espacio reducido por UMAP al espacio original de las variables.

En la Tabla 21 se presenta el análisis de cinco de las variables más representativas del capital agronómico: pH, rendimiento por hectárea (rph), edad del cultivo (ec), frecuencia de dotación de agua (fdta) y nitrógeno total (Ntotal). En esta tabla se reportan los valores medios de cada variable por clúster, lo cual permite comparar y entender las características particulares de cada grupo de unidades productivas.

Como se explicó previamente, la aplicación de los algoritmos K-means y HDBSCAN produjo tanto clústeres equivalentes como clústeres distintos. Dentro de las diferencias encontradas, se destacan dos casos específicos:

- En el Caso 1, el clúster etiquetado como 1 por HDBSCAN fue subdividido por K-means en dos clústeres: KMeans_2 y KMeans_5.
- En el Caso 2, sucede lo contrario: K-means agrupó ciertas unidades bajo el clúster 0, mientras que HDBSCAN las dividió en dos clústeres distintos, etiquetados como 2 y 3.

Comparar estos casos permite identificar cómo varía la composición interna de los grupos según el algoritmo utilizado. Estas diferencias constituyen la base para la evaluación por parte de expertos, quienes, mediante criterios técnicos y experiencia en el sector agrícola, podrán valorar cuál de los modelos de agrupamiento representa de manera más coherente la realidad productiva del territorio.

Tabla 21. Caso 1: HDBSCAN_1 vs KMeans_2 y KMeans_5.

Cluster	Count	rph	fdta	ec	pH	Ntotal
HDBSCAN_1	49	410.59	11.1	20.51	6.96	0.09
KMeans_2	28	486.39	10.64	23.29	6.95	0.08
KMeans_5	21	309.52	11.71	16.81	6.97	0.09

En la Tabla 21 se presenta la comparación entre el clúster HDBSCAN_1 y los clústeres KMeans_2 y KMeans_5, resultado de la segmentación del mismo conjunto de unidades productivas. En este caso, HDBSCAN agrupa 49 unidades productivas dentro de un único clúster, interpretando que dichas observaciones conforman un grupo homogéneo bajo el criterio de densidad. Sin embargo, K-means subdivide este mismo grupo en dos clústeres: KMeans_2, con 28 unidades productivas, y KMeans_5, con 21, lo cual sugiere la existencia de subgrupos que según Kmeans tienen diferencias significativas que HDBSCAN no separa.

El análisis de las variables permite evidenciar estas diferencias. El clúster KMeans_2 presenta un rendimiento promedio de 486.39 kg/ha, el más alto entre los tres grupos, acompañado de una frecuencia de riego más frecuente (fdta = 10.64 días) y una edad del cultivo mayor (23.29 años). Estos factores en la que a mayor frecuencia de dotación de agua y cultivos más maduros podrían explicar su mejor desempeño productivo, ya que están directamente relacionados con el desarrollo fisiológico del cultivo.

El clúster KMeans_5 muestra un rendimiento significativamente menor (309.52 kg/ha), una frecuencia de riego más larga (11.71 días) y una edad del cultivo más joven (16.81 años), lo cual puede asociarse con un menor desarrollo productivo del cultivo. Por su parte, las variables de pH y nitrógeno total (Ntotal) presentan valores similares entre los tres grupos, sin diferencias significativas.

En la Tabla 22 se muestran los resultados para el caso 2, el cual tiene un comportamiento inverso.

Tabla 22. Caso 2: KMeans_0 vs Hdbscan_2 y Hdbscan_3.

Cluster	Count	rph	fdta	ec	pH	Ntotal
KMeans_0	21	410.5	11.0	21.3	6.95	0.08
Hdbscan_2	8	430.2	12.1	20.5	6.9	0.09
Hdbscan_3	13	395.1	10.8	19.8	7.0	0.08

En comparación, el Caso 2**,** K-means agrupa 21 unidades productivas en un solo clúster etiquetado como cluster 0, mientras que HDBSCAN logra diferenciar dos subgrupos (cluster 2 y cluster 3) que presentan variaciones productivas, especialmente en el rendimiento por hectárea. Este comportamiento muestra que para este caso HDBSCAN puede detectar estructuras internas basadas en la densidad de los datos como se relaciona en la Figura 17.

Las diferencias observadas entre los algoritmos K-means y HDBSCAN se explican principalmente por la lógica interna y los supuestos que cada uno utiliza para formar los clústeres. K-means parte de la suposición de que todos los clústeres son esféricos, de tamaño similar y con una densidad homogénea, lo que limita su capacidad para adaptarse a estructuras más complejas o con formas irregulares. En cambio, HDBSCAN puede detectar clústeres de forma arbitraria y con distintas densidades, lo que le permite adaptarse mejor a la heterogeneidad real de los datos.

Un aspecto clave en la diferencia de resultados radica en cómo cada algoritmo maneja los puntos atípicos o dispersos. K-means obliga a que cada punto sea asignado a algún clúster, sin importar su distancia respecto al resto, lo que puede generar agrupamientos poco representativos. Este comportamiento se evidencia, por ejemplo, en el clúster 0 de K-means, que agrupa puntos con amplia dispersión espacial. Por el contrario, HDBSCAN puede dejar fuera puntos que no cumplan con un umbral mínimo de densidad, etiquetándolos como ruido o valores atípicos, bajo la etiqueta -1. Esta capacidad le otorga una mayor flexibilidad para reflejar estructuras más realistas, especialmente en entornos donde no todos los datos pertenecen necesariamente a un grupo bien definido.

Además, HDBSCAN está soportado por un enfoque jerárquico basado en la conectividad entre puntos cercanos y la persistencia de los grupos en distintas escalas de densidad, lo cual permite capturar tanto la estructura local como global del conjunto de datos.

Con el objetivo de contrastar estos comportamientos con los clústeres en los que ambos algoritmos coincidieron, en la Tabla 23, se presentan los valores medios de las variables del capital agronómico para los grupos equivalentes. En este análisis se consideran únicamente los clústeres generados por HDBSCAN, incluyendo el grupo -1, correspondiente a los valores atípicos. Si bien este grupo puede parecer equivalente al clúster 0 de K-means en cuanto a composición, conceptualmente representan situaciones distintas: mientras que en K-means los puntos son forzados a pertenecer a un grupo, en HDBSCAN su clasificación como ruido surge directamente de la incapacidad del algoritmo para encontrar una densidad mínima coherente con algún clúster existente. Esta diferencia teórica tiene importantes implicaciones metodológicas en la interpretación de los resultados.

Tabla 23. Cluster equivalentes.

Cluster	Count	rph	fdta	ec	pH
Hdbscan_0	13	557.69	10.0	39.92	6.46
Hdbscan_-1	7	314.29	12.86	50.86	7.29
Hdbscan_4	13	354.62	8.77	28.08	6.50

La equivalencia de estos grupos entre ambos algoritmos indica que existe una estructura robusta y fácilmente detectable, independientemente del algoritmo empleado. En particular, el grupo etiquetado como HDBSCAN_0 agrupa a los cultivos más productivos del municipio de Baraya, con un rendimiento promedio de 557.69 kg/ha, una edad media de 39.92 años y una frecuencia de riego de diez días, considerada óptima. Otra variable destacable es el pH, cuyo valor promedio de 6.46 se encuentra dentro

del rango ideal para el desarrollo del cacao (entre cinco y siete), siendo especialmente favorable cuando se aproxima a valores de seis.

Por otro lado, el grupo HDBSCAN_-1, clasificado como valores atípicos por el algoritmo, reúne las unidades productivas con las condiciones más adversas para el cultivo. Estos cultivos son los más longevos (edad promedio de 50.86 años) y presentan el menor rendimiento (314.29 kg/ha), además de un pH elevado (7.29), menos adecuado para el desarrollo óptimo del cacao. Asimismo, se caracterizan por tener las frecuencias de riego más espaciadas (12.86 días), lo cual implica un menor acceso al agua.

Finalmente, el grupo HDBSCAN_4 está compuesto por cultivos más jóvenes, con una edad media de 28.08 años y riego más frecuente (8.77 días). Sin embargo, su rendimiento (354.62 kg/ha) es inferior al del grupo HDBSCAN_0.

En la Figura 18 se presenta la distribución de la variable material vegetal (vm) según los clústeres generados por el algoritmo HDBSCAN. En el que se revelan patrones diferenciados en la composición del material de cada grupo. El clúster 1 es el grupo más grande y diverso, al agrupar la mayor cantidad de predios con una fuerte presencia de las variedades híbrido (22 predios) y clon (12 predios), además de otros materiales como criollo, CCN-51, FEAR-5, ICS-95 y otros.

Por otro lado, el clúster 0 muestra una concentración alta en la variedad criollo (9 predios), lo que refleja un perfil de productores con materiales que también son denominados nativos con condiciones de sabor y aroma diferenciadoras. En contraste, el clúster -1, correspondiente a valores atípicos o ruido según la lógica de HDBSCAN, incluye predios con híbrido (4) y CCN-51 (3), lo cual sugiere que, si bien estos cultivos son comunes, se encuentran dispersos o con baja densidad espacial, lo que impide su integración en agrupamientos definidos

Asimismo, los clústeres 2, 3 y 4, muestran patrones particulares: el clúster 3 tiene una alta proporción de híbrido (10 predios); el clúster 2 incluye una única unidad con trinitario, no presente en otros clústeres, diferenciada; mientras que el clúster 4 contiene un conjunto más equilibrado entre CCN-51, clon y criollo.

La descripción y ficha técnica de cada variedad del material vegetal está disponible en el Anexo E.

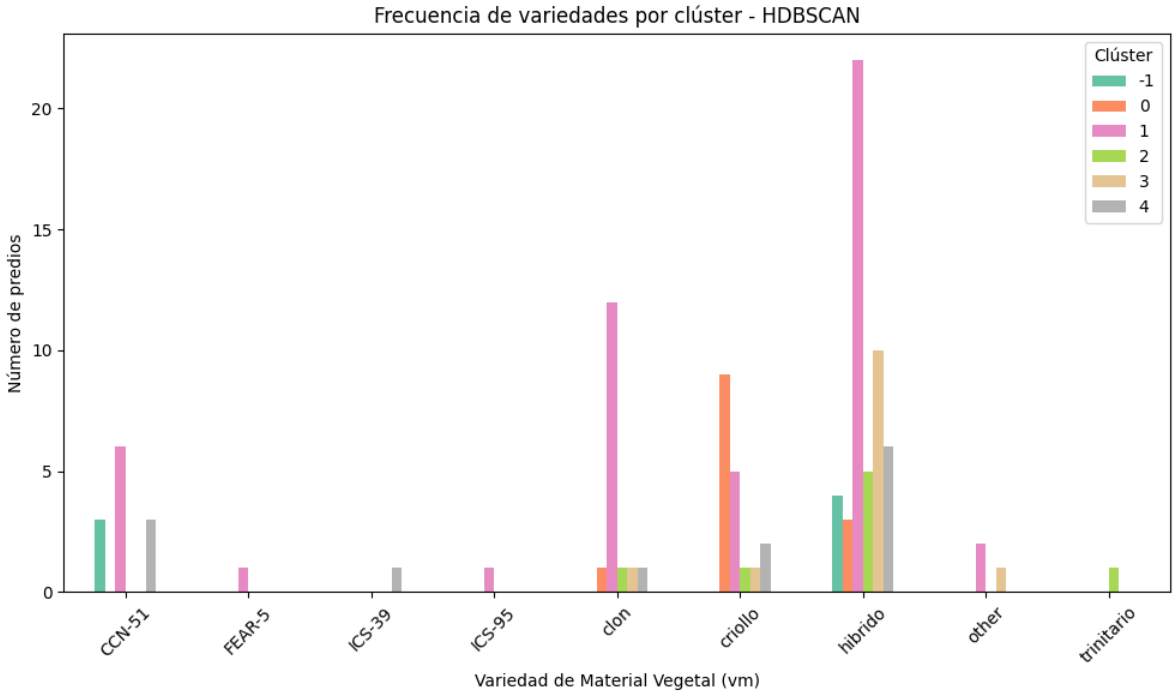


Figura 18. Frecuencia de variedades de materiales vegetales de cacao por grupo de clúster.

Al analizar los grupos segmentados por la clase textural del componente suelo, esta variable es fundamental para el crecimiento y desarrollo del cultivo. En la Figura *19* se presentan la distribución de clases texturales, este análisis revela que el clúster 1 concentra la mayor diversidad y frecuencia de clases texturales, destacándose la presencia predominante de suelos franco arcillo arenosos (FArA) con 19 predios, seguido por franco arcillosos (FAr) y francos arenosos (FA). Esta diversidad indica condiciones edáficas favorables para el cultivo de cacao, ya que estas texturas combinan buen drenaje con capacidad de retención de humedad. Por su parte, el clúster 4 presenta una concentración destacada de suelos arcillosos (Ar), lo cual puede implicar suelos más pesados y con mayor retención de agua, pero también con posibles limitaciones de aireación, este factor explica desde el punto de vista físico del suelo el bajo rendimiento de este grupo, ya que el desarrollo de la agricultura en este tipo de suelo es desafiante y altamente demandante nutricionalmente. El clúster 2, aunque de menor tamaño, muestra un perfil mixto entre texturas arcillosas y franco arcillosas. Cabe resaltar que los predios etiquetados como valores atípicos (clúster -1) también presentan texturas favorables como FArA, pero su exclusión del resto de los grupos podría estar relacionada con condiciones adicionales adversas como problemas de fertilidad o manejo agronómico. Al relacionar esta distribución textural con el rendimiento (rph) y la edad del cultivo (ec), se evidencia que los grupos con mejores texturas como FArA y Far tienden a coincidir con los clústeres de mayor productividad, reforzando la importancia de la textura del suelo como un componente clave en el desempeño agronómico del cultivo de cacao.

Frecuencia de variedades de materiales vegetales de cacao por grupo de clúster.

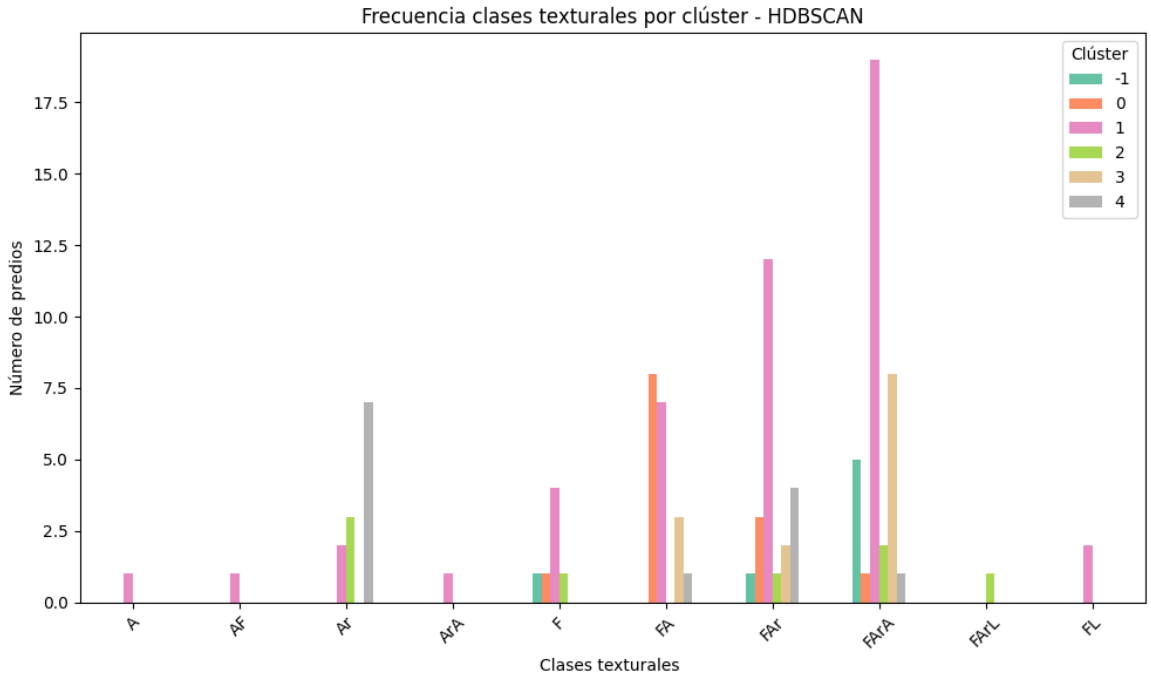


Figura 19. Frecuencia de clases texturales por grupo de clúster.

Desde la perspectiva del componente económico, los resultados asociados a los ingresos provenientes de la actividad cacaocultora, medidos en salarios mínimos legales vigentes (SMLV), permiten identificar patrones relevantes entre los grupos de productores. El clúster 1 presenta una concentración significativa de ingresos bajos: 43 de las 49 unidades productivas se encuentran en el rango de menos de un SMLV. Solo cuatro predios se ubican entre uno y dos SMLV, y únicamente dos superan los cinco SMLV. Esta distribución evidencia una baja rentabilidad económica generalizada dentro del grupo.

En cuanto a los clústeres 2, 3 y 4, todos los predios se agrupan únicamente en el rango de menos de un SMLV, lo que indica una situación crítica en términos de ingresos. Al relacionar este comportamiento con el rendimiento agrícola (rph), se observa que los clústeres 3 y 4 presentan los menores niveles de productividad, lo cual está directamente vinculado con los bajos ingresos generados por estas unidades productivas.

Por el contrario, el clúster 0 el grupo más productivo, con un rendimiento medio superior a los 550 kg/ha, tiene una mayor heterogeneidad en los niveles de ingreso. Nueve predios se ubican en el rango de uno a dos SMLV, y aunque también existen casos con ingresos inferiores a un salario mínimo, se identifica una mayor proporción relativa de ingresos más altos en comparación con los demás grupos. Esto sugiere que una mejor condición agronómica (como textura del suelo, pH adecuado y prácticas de riego eficientes) se traduce en una mayor rentabilidad para los productores de este grupo.

Finalmente, el clúster -1, identificado como conjunto de valores atípicos, está conformado por siete unidades productivas cuyos ingresos se encuentran exclusivamente por debajo de un SMLV. Este grupo también coincide con condiciones agronómicas desfavorables y edades avanzadas de los cultivos, lo cual representa una situación de alta vulnerabilidad.

En conjunto, estos resultados reflejan que el sector cacaocultor del municipio de Baraya enfrenta limitaciones económicas. Exceptuando el clúster 0, la mayoría de los productores obtienen ingresos bajos, lo que compromete su sostenibilidad. Además, esta situación puede estar vinculada con factores sociales más profundos como el envejecimiento de la población rural, la falta de relevo generacional, rezago tecnológico y la ausencia de incentivos para la innovación o acceso a mercados. En la Tabla 24 resume la frecuencia de ingresos económicos por grupo.

Tabla 24. Distribución de ingresos económicos por cluster.

Clúster	1 a 2 SMLV	Más de 5 SMLV	Menos de 1 SMLV
-1	0	0	7
0	9	0	4
1	4	2	43
2	0	0	8
3	0	0	13
4	0	0	13

8.2 Validación con expertos

En complemento con la evaluación realizada a través de métricas internas de validación de cada algoritmo, se diseñó un instrumento de evaluación cuantitativo basado en la metodología conocida como evaluación por el ojo experto. Esta metodología consiste en la valoración de los resultados por parte de un panel de especialistas, quienes, a partir de su experiencia técnica, evaluarán la coherencia de los agrupamientos generados.

El panel estuvo conformado por tres expertos del área agrícola: el PhD. Armando Torrente, PhD. en Ciencias Agrícolas y editor de la Revista Colombiana de la Sociedad del Suelo; el MSc. Jorge Chavarro, estudiante del Doctorado en Sistemas Complejos de la Universidad Politécnica de Madrid; y el Ing. Wilmer Valenzuela, ingeniero agrónomo con amplia experiencia en el manejo técnico de cultivos de cacao. El perfil detallado de cada miembro del panel se encuentra disponible en el Anexo D.

La evaluación se realizó a través de un conjunto de cinco preguntas orientadas a comparar el comportamiento de cada clúster en relación con variables relevantes. El objetivo de esta etapa es establecer cuál de los modelos refleja la realidad del territorio y de las unidades productivas. Esta validación experta constituye un componente fundamental del proceso de análisis, ya que complementa la interpretación técnica de los datos con la perspectiva contextual y práctica del sector agrícola, permitiendo no solo evaluar la validez interna de los agrupamientos, sino también su aplicabilidad en escenarios reales de planificación y toma de decisiones.

Se formularon cinco preguntas orientadas a la evaluación de los resultados de agrupamiento generados por los algoritmos K-means y HDBSCAN. Estas preguntas fueron diseñadas para capturar la percepción de los expertos sobre la coherencia, utilidad y aplicabilidad agronómica de cada modelo. La recolección y análisis de las respuestas se realizó mediante el aplicativo ***ArcGIS Survey123***

A continuación, se presenta la estructura con las respuestas respectivas del cuestionario aplicado al

panel de expertos:

a- ¿Cuál de los dos modelos (K-means o HDBSCAN) presenta una segmentación más clara y comprensible desde el punto de vista productivo?
b- ¿Cuál algoritmo generó agrupaciones más coherentes desde el punto de vista agronómico (variedad, edad del cultivo, suelo, riego)?

c- ¿Cuál algoritmo manejó mejor los casos atípicos o predios con condiciones extremas o muy particulares?
d- ¿Cuál segmentación considera más útil para priorizar acciones técnicas diferenciadas?

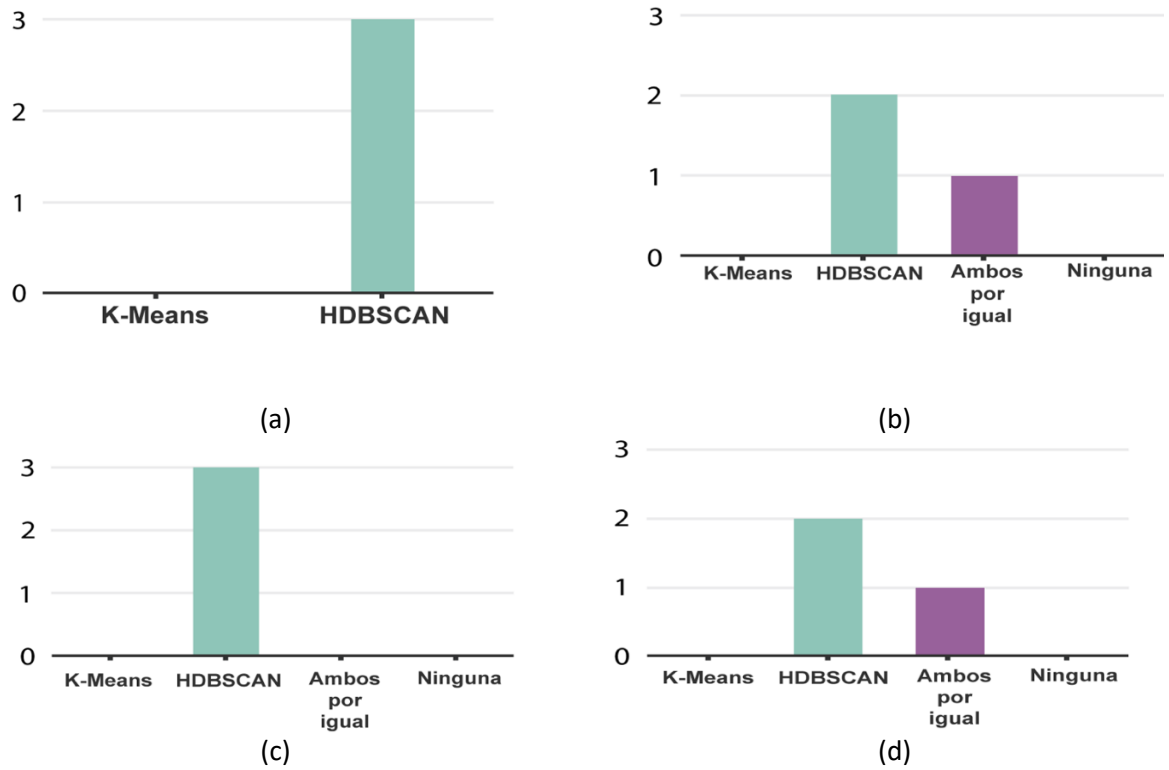


Figura 20. Resumen de respuesta de expertos.

a) Claridad de la segmentación productiva

- En la Figura 20 (a), los expertos coinciden en que HDBSCAN ofrece una segmentación más clara y comprensible desde el punto de vista productivo. Esto refuerza la idea de que la capacidad de este algoritmo para detectar estructuras de densidad no homogénea facilita una interpretación más coherente con la realidad del sector cacaocultor.

b) Coherencia agronómica de los clústeres

- En la Figura 20 (b), HDBSCAN también es el algoritmo mejor valorado, aunque un experto consideró que ambos algoritmos generaron agrupaciones coherentes. Esto sugiere que, aunque

hay consenso sobre la ventaja de HDBSCAN, K-means no fue completamente descartado, especialmente en grupos con estructuras más claras.

c) Manejo de valores atípicos

- La Figura 20 (c) muestra un respaldo total a HDBSCAN en el manejo de valores atípicos, una característica intrínseca de su diseño que permite identificar y etiquetar como ruido aquellos puntos que no pertenecen a ningún grupo denso. Esto evidencia una fortaleza técnica clara frente a K-means, el cual forzosamente asigna todos los puntos a un clúster, incluso si no encajan con el patrón.

d) Utilidad para decisiones técnicas diferenciadas

- Finalmente, en la Figura 20 (d), HDBSCAN se destaca como el más útil para priorizar intervenciones diferenciadas, aunque nuevamente un experto reconoció valor en ambas segmentaciones. Esto indica que, desde una perspectiva práctica, la segmentación jerárquica de HDBSCAN se percibe como más aplicable a la gestión del territorio.

8.3 Perfiles de sostenibilidad y reporte de agrupamientos generados

A partir del análisis realizado sobre los grupos generados por los algoritmos, y teniendo en cuenta la validación interna y la evaluación de los expertos en donde HDBSCAN tiene un mejor desempeño frente a K-means, se construyen los perfiles de sostenibilidad para cada grupo resultante, considerando los capitales evaluados, en especial el capital físico de suelo, las variedades cultivadas y los ingresos derivados de la actividad productiva.

Clúster 1 (n=49): Este es el grupo más representativo y diverso en cuanto a texturas. Predominan los suelos Franco Arcillo Arenosos (FArA) con 19 predios, seguido de Franco Arcillosos (FAr) y Francos Arenosos (FA). Esta diversidad sugiere un grupo con potencial de producción (rph = 410.59 kg/ha), pero con necesidades de manejo diferenciado por subzonas teniendo en cuenta de que cada clase textural requiere de un manejo puntual. En cuanto a las variedades, se observa predominio de híbridos (22 predios) y clonales (12 predios), mientras que, en términos económicos, el 88% de los productores reporta ingresos por debajo de un salario mínimo. En términos de sostenibilidad, es aconsejable el manejo específico de cada unidad productiva de acuerdo con sus condiciones de suelo, como la incorporación de materia orgánica, cobertura vegetal y fertilización adaptada a las condiciones y variedad cultivada.

Clúster 0 (n=13): Agrupa los predios con mayor rendimiento (557.69 kg/ha), suelos con textura FAr y FA en su mayoría, y una frecuencia de riego de 10 días. Este grupo representa condiciones óptimas tanto por textura como por prácticas de manejo. Las variedades predominantes son híbridos, y los ingresos son más altos con respecto a los demás, con nueve predios entre 1 y 2 SMLV. Para mantener su sostenibilidad, se sugiere: seguimiento nutricional periódico, mantenimiento de coberturas, monitoreo periódico de los agroambientales y prácticas que fortalezcan la resiliencia del sistema.

Clúster -1 (n=7): Corresponde a los valores atípicos, asociados con condiciones agronómicas y socioeconómicas adversas. Tienen los suelos más arcillosos, con predominio de FArA, el rendimiento más bajo (314.29 kg/ha), mayor edad de cultivo (50.86 años) y un pH subóptimo (7.29). Todas las

unidades productivas reportan ingresos menores a un salario mínimo. Para mejorar sus condiciones, se propone: renovación paulatina del cultivo, aplicación de enmiendas orgánicas, técnicas de manejo de suelo para, y un acompañamiento institucional focalizado, es el grupo que según este análisis es el más vulnerable del municipio de Baraya,

Clúster 4 (n=13): Presenta mayor proporción de suelos arenosos (Ar), con alta permeabilidad y baja capacidad de retención de nutrientes. El rendimiento promedio es de 354.62 kg/ha, y las variedades más comunes son híbridos (6 predios) y CCN-51. Todos los productores reportan ingresos inferiores a un salario mínimo. Se recomienda: uso intensivo de fertilización fraccionada, incorporación de materia orgánica, y monitoreo constante de la humedad del suelo para mejorar la eficiencia en el uso del agua.

Clúster 3 (n=13): Con texturas FArA y FAr, rendimiento de 395.1 kg/ha, y edad promedio de 19.8 años. Se encuentra en etapa de transición hacia un desarrollo eficiente, por esta razón el manejo debe estar fundamentado en el acompañamiento en términos de nutrición. Predominan híbridos y criollos. El 100% de los productores se encuentra por debajo de un salario mínimo. Se recomienda: fortalecimiento técnico, acompañamiento agronómico, fertilización racional y monitoreo de indicadores productivos.

Clúster 2 (n=8): Predominan suelos arcillosos (Ar) y francos arcillosos (FAr), con un rendimiento promedio de 430.2 kg/ha. Las variedades más comunes son híbridos y clonales. Todos los productores están en la categoría de ingresos inferiores a un salario mínimo. Se requiere: mejoramiento del drenaje, prácticas de conservación de suelo y fertilización adaptada.

Estos perfiles de sostenibilidad permiten concluir que las unidades productivas del sector cacaocultor requieren estrategias de manejo diferenciadas, si se desea avanzar hacia un modelo verdaderamente sostenible. Las diferencias en condiciones edáficas, prácticas agronómicas, materiales vegetales y niveles de ingreso evidencian que no es viable aplicar soluciones homogéneas a problemáticas estructurales que son profundamente heterogéneas.

Este tipo de caracterización técnica y contextual representa un insumo clave para la gestión del territorio basada en datos reales, permitiendo identificar oportunidades y necesidades específicas para cada grupo de productores. En ese sentido, este trabajo propone un enfoque metodológico como alternativa hacia una agricultura de precisión enfocada en la toma de decisiones informadas, orientadas no solo a mejorar la productividad, sino también a garantizar la sostenibilidad ecológica, económica y social del subsector cacaotero.

De esta forma, se establece un marco de análisis replicable para otros territorios y cultivos, donde la segmentación inteligente y la evaluación multidimensional permitan construir rutas técnicas más precisas, equitativas y adaptadas a la realidad de cada territorio.

9 VISUALIZACIÓN INTERACTIVA

Un aspecto clave de este proyecto es poder visualizar de manera interactiva los resultados generados por la clusterización, ya que estos han organizado las unidades productivas en grupos que comparten características similares en términos agronómicos, sociales o económicos. La estrategia de visualización debe contemplar un formato que permita consultar los datos según cada patrón o perfil de sostenibilidad identificado. Esta solución parte de la filosofía del *GeoBigData*, que, además de las cuatro "V" clásicas (valor, variedad, velocidad y veracidad), incorpora una quinta "V" asociada a la visualización. Al contar con datos georreferenciados, se puede relacionar cada variable, cada patrón detectado o cada característica con una unidad productiva específica. Esta ventaja metodológica posibilita diseñar estrategias de visualización orientadas a la toma de decisiones en el ámbito productivo en este caso, el sector cacaocultor, de modo que la gestión se base en el estado real del territorio.

9.1 Definición de herramienta de visualización

Entre las múltiples herramientas de visualización que admiten datos geoespaciales e integran mapas interactivos, se realizó una vigilancia tecnológica y se seleccionó *ArcGIS Dashboards*. Esta es una herramienta de visualización y monitoreo geoespacial que permite integrar mapas, gráficos y widgets en un único panel interactivo. Para este caso particular, se alojaron las unidades productivas de cacao de Baraya (Huila) en conjunto con las variables que describen el sector. Sobre un mapa base satelital, cada parcela se representa como un polígono georreferenciado cuyos colores de cada marcador en forma de punto indican el clúster de pertenencia según el análisis previo. El *dashboard* incluye indicadores numéricos (por ejemplo, área sembrada o número de predios), gráficos de barras (texturas de suelo) y mapas de calor (temperatura IDEAM), todos ellos vinculados: al filtrar un clúster o condición de suelo, el mapa y los gráficos se actualizan simultáneamente para reflejar únicamente las unidades productivas que cumplen ese criterio. Además, los pop-ups espaciales muestran al instante datos detallados de cada predio. Esta solución explora la dinámica del territorio y la comparación de resultados de clustering, de modo que el usuario final puede acceder a información útil y en tiempo real sobre todo el sector cacaocultor y los resultados de las clasificaciones obtenidas.

9.2 Estrategia de visualización e integración con mapa interactivo

En la Figura 21 se muestra en forma de captura de pantalla la interfaz gráfica del panel de control, el cual está disponible en línea y puede ser consultado con todas sus funcionalidades en este link " Dashboard " . Además, la demostración en video sobre el uso y funcionamiento de la herramienta puede visualizarse en el siguiente enlace " Video USO_HERRAMEINTA " . Este dashboard ha sido desarrollado para alojar y visualizar los resultados de la clusterización de las unidades productivas de cacao en el municipio de Baraya (Huila). Entre sus componentes principales se incluyen:

- **Clústeres generados por HDBSCAN** (validados por expertos), que agrupan las unidades según patrones agronómicos, sociales y económicos.
- **Variables clave de cada unidad productiva**:

    1. **pH del suelo.**
    2. **Texturas de suelo.**

3. **Condición de suelo**: clasificada en “desafiante”, “óptima” o “ideal”.
4. **Temperaturas medias IDEAM**: representadas mediante mapas de calor para evaluar el contexto climático.

- **Información espacial** de cada predio, en formato de polígonos georreferenciados que permiten vincular todas las capas de atributos.
- **Widgets interactivos** que facilitan la filtración por clúster, por variables agronómicas específicas o por criterios de sostenibilidad.

El objetivo de este panel de control es generar una vista dinámica del territorio, de modo que los diferentes actores del sector puedan explorar y comparar resultados, priorizar intervenciones y diseñar estrategias de manejo basadas en la realidad espacial y productiva de cada unidad productiva.

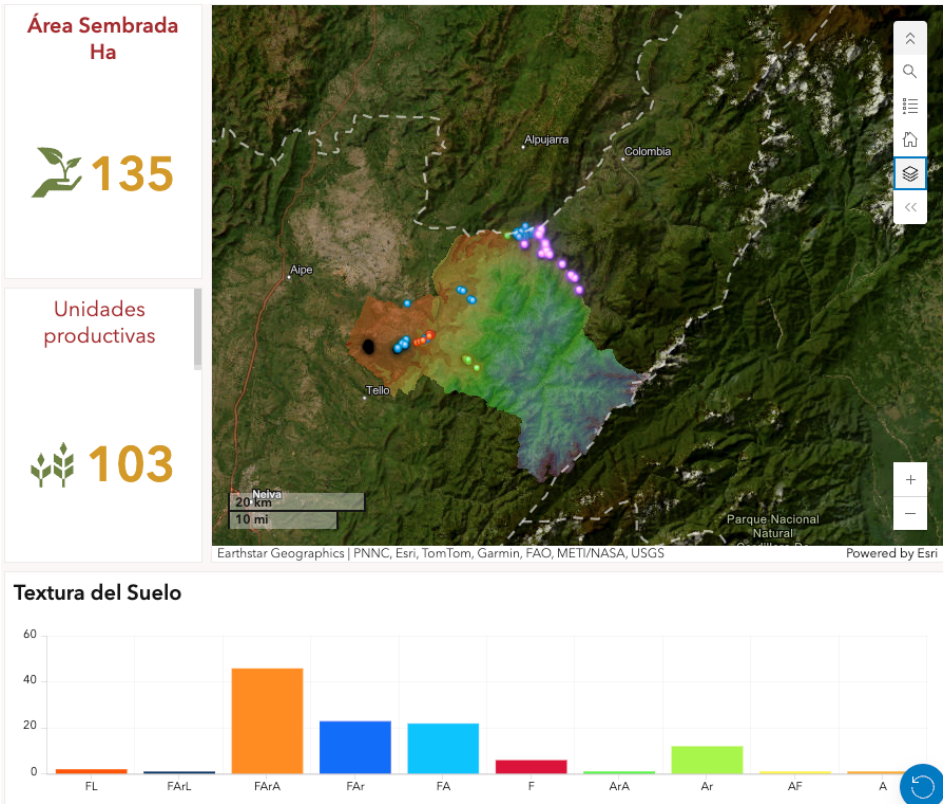


Figura 21. Dashboard de Unidades Productivas y Características de Suelo en Baraya (Huila).

Este mapa interactivo permite que, al hacer clic en cualquier punto, se despliegue la información específica de cada unidad productiva. Por ejemplo, en la Figura 21 se observan puntos de diferentes colores; cada color corresponde a un grupo distinto. Asimismo, en la Figura 22 se muestra cómo el panel de control puede filtrar las unidades según criterios específicos. Si el usuario selecciona el *pop-pup* de capas en “Suelos Óptimos, los predios que cumplen ese criterio se resaltan con un color determinado, permitiendo identificar rápidamente qué parcelas presentan suelos con alguna condición diferencial.

La textura del suelo se clasifica en diez categorías: Arenoso (A), Arenoso franco (AF), Arcilloso (Ar), Arcillo arenoso (ArA), Franco (F), Franco arenoso (FA), Franco arcilloso (FAr), Franco arcillo arenoso (FArA), Franco arcillo limoso (FArL) y Franco limoso (FL). Mediante el gráfico “Textura del Suelo” se muestra cuántos predios de cada clúster pertenecen a cada categoría. Con base en esa textura y en el valor de pH, cada unidad productiva también se clasifica dentro de alguna de estas etiquetas:

- **Desafiante (Ar, FAr)**: suelos pesados y de drenaje limitado.
- **Óptima (FArA, FA)**: suelos con buen equilibrio entre drenaje y retención de agua.
- **Ideal (FArA con pH ~ 6)**: condiciones casi perfectas para el cultivo de cacao.

Cuando se activa esta capa en el mapa, los polígonos son resaltados en rojo, amarillo o verde según esta clasificación, y es posible filtrar directamente por cualquiera de ellas, ver Figura 22.

Por otra parte, a cada predio se le asocia la temperatura media anual de la estación IDEAM más cercana. En el mapa principal, un mapa de calor colorea los polígonos según rangos de temperatura (< 22 °C en azul claro, 22–24 °C en verde y >= 30 °C en rojo), lo que permite identificar rápidamente microclimas favorables o desfavorables para el cultivo.

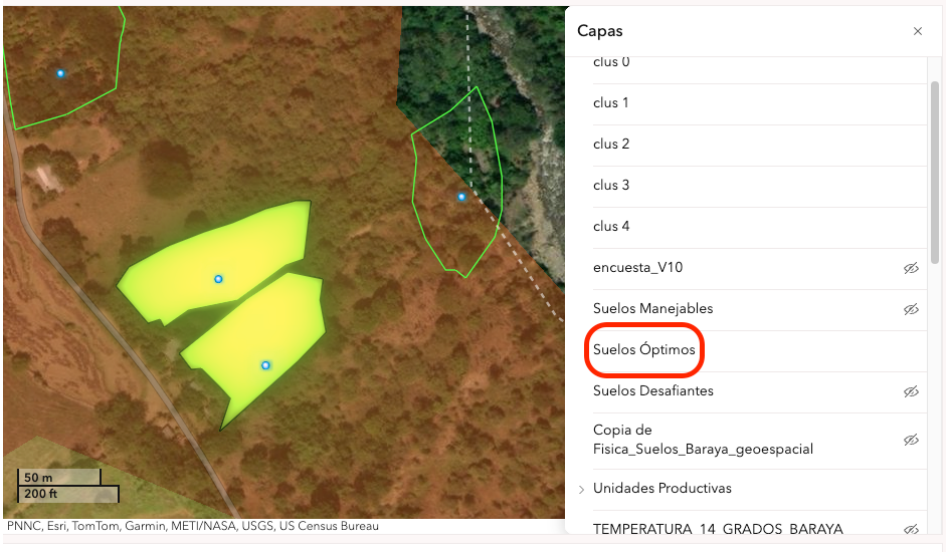


Figura 22. Manejo de capas del geovisor.

Esta metodología desarrollada resulta versátil porque permite cruzar múltiples variables y generar información adicional que aporta un verdadero valor a los datos, brindando un conocimiento del territorio a escala predial. Con ello es posible implementar estrategias de agricultura de precisión altamente específicas, adaptadas a las condiciones de cada productor. Además, al concentrar toda esta información en un solo lugar, se facilita la incorporación de datos de otros sectores por ejemplo, parámetros climáticos históricos o proyecciones meteorológicas, lo que abre nuevas posibilidades para analizar cómo fenómenos como El Niño u La Niña afectan a cada unidad productiva en particular. Al combinar estos datos con técnicas de inteligencia artificial, es factible modelar y anticipar en el tiempo qué parcelas son más susceptibles a eventos extremos, como sequías o inundaciones, y planificar así acciones preventivas y de mitigación de forma precisa, solo por nombrar un caso de usó.

10 CONCLUSIONES Y TRABAJOS FUTUROS

10.1 Conclusiones

Este trabajo muestra que la integración de herramientas de ciencia de datos y aprendizaje automático en el sector cacaocultor del departamento del Huila permite comprender de forma más profunda la complejidad de los sistemas agrícolas. Aunque se emplearon modelos estadísticos y computacionales, su verdadero valor radica en los fundamentos físicos y matemáticos que los sustentan, como lo evidencian las bases teóricas de los métodos de reducción de dimensionalidad estocástica y los modelos de clustering aplicados. La combinación de técnicas de análisis no supervisado, reducción topológica de datos y visualización geoespacial permitió transformar información dispersa en conocimiento útil para la gestión productiva y la sostenibilidad del cultivo de cacao.

El primer resultado clave fue la generacion de una base de datos integral que articula información de diversas fuentes como imágenes multiespectrales, análisis fisicoquímicos de suelo y encuestas socio-agronómicas. Esta estructura permitió describir el territorio de forma multidimensional y aportar, por primera vez en el departamento un repositorio integrado y estandarizado de esta naturaleza, habilitando la aplicación de técnicas no supervisadas. Dado el elevado número y la heterogeneidad de variables, la reducción de dimensionalidad no lineal, en particular UMAP, resultó más adecuada que PCA para preservar vecindades y formas globales; ello confirmó la naturaleza no lineal y estocástica del fenómeno estudiado y permitió revelar patrones que las proyecciones lineales no alcanzan a capturar.

Sobre estas representaciones compactas se compararon enfoques de clustering. Los resultados evidencian que HDBSCAN se ajusta mejor que K-Means a la estructura de los datos agrícolas ya que no exige fijar a priori el número de grupos, detecta outliers como ruido y modela agrupamientos guiados por densidad, lo que se alinea con la distribución espacial y la variabilidad intrínseca del cultivo. Las métricas internas y la persistencia de clúster mostraron agrupaciones estables y coherentes con la realidad productiva que aportan segmentaciones con significado agronómico.

La interpretación conjunta de los clústeres reveló que la productividad no depende solo de la ubicación, sino de combinaciones de factores agronómicos, edáficos, climáticos y socioeconómicos. La visualización interactiva integró todo el conocimiento generado en un entorno único, facilitando leer los patrones espaciales, explorar diferencias entre grupos y reconocer unidades con necesidades específicas. Este recurso convierte el análisis técnico en una herramienta operativa para productores y tomadores de decisión, favoreciendo manejo diferenciado, focalización de asistencia y uso más eficiente de recursos.

Finalmente, el trabajo sienta bases para una transformación digital y sostenible del sector: ampliar series temporales e integrar datos de estaciones meteorológicas y satélite permitirá construir modelos predictivos (riesgos, rendimiento, sanidad) y sistemas de alerta temprana. Con ello, se fortalecen la resiliencia, la competitividad y la gestión basada en evidencia del cacao en el Huila, avanzando hacia una agricultura precisa, inteligente y orientada a la conservación.

10.2 Trabajos futuros

Este proyecto forma parte de una iniciativa macro que abarca siete municipios del Huila, dentro de los cuales Baraya es uno de los casos de aplicación. La metodología de clusterización desarrollada se replica en los demás municipios, permitiendo identificar patrones espaciales y seleccionar predios representativos para análisis más detallados. Esta herramienta ha sido presentada en Colombia 4.0 como una solución Agrotech de alto impacto y actualmente es considerada por la Secretaría de Agricultura del Huila dentro de su estrategia de transformación digital. A futuro, se proyecta integrar modelos predictivos y agentes de inteligencia artificial que for4talezcan el panel de control existente, así como la elaboración de un artículo científico que documente la metodología y sus principales resultados.

11 REFERENCIAS BIBLIOGRÁFICAS

[1] «Analisis-cadena-de-valor-del-cacao-en-el-Huila.pdf». Accedido: 5 de diciembre de 2024. [En línea]. Disponible en: https://www.cchuila.org/wp-content/uploads/Analisis-cadena-de-valor-del-cacao-en-el-Huila.pdf

[2] Fedecacao, «Colombia reportó en 2024 producción histórica de cacao», Sitefedecacao. Accedido: 7 de octubre de 2025. [En línea]. Disponible en: https://www.fedecacao.com.co/post/colombia-reportó-en-2024-producción-histórica-de-cacao

[3] R. Central, «ALERTA SANITARIA EN HUILA POR PLAGAS EN CULTIVOS DE CACAO», RAP-E Región Central. Accedido: 12 de mayo de 2025. [En línea]. Disponible en: https://regioncentralrape.gov.co/alerta-sanitaria-en-huila-por-plagas-en-cultivos-de-cacao/

[4] «Programa mundial del censo agropecuario 2020».

[5] BBVA, «"Agrotech", soluciones digitales y desarrollo rural: sembrar oportunidades contra el vacío poblacional», BBVA NOTICIAS. Accedido: 17 de noviembre de 2024. [En línea]. Disponible en: https://www.bbva.com/es/innovacion/agrotech-soluciones-digitales-y-desarrollo-rural-sembrar-oportunidades-contra-el-vacio-poblacional/

[6] M. Castro Franco y M. Domenech, *Agro Big Data: El próximo desafío.* 2014.

[7] M. L. Quintero R y K. M. Díaz Morales, «El mercado mundial del cacao», *Agroalimentaria*, vol. 9, n.º 18, pp. 47-59, ene. 2004.

[8] S. User, «Agroforestería - INFOR», Agroforestería. Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://agroforesteria.infor.cl

[9] «cacao-en-sistemas-agroforestales.pdf». Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://mocca.org/wp-content/uploads/2021/08/cacao-en-sistemas-agroforestales.pdf

[10] «Cacao: Propiedades, Origen, Beneficios, Qué es e Información». Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://cuidateplus.marca.com/alimentacion/diccionario/cacao.html

[11] «Cacao - Concepto, origen, historia y propiedades», https://concepto.de/. Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://concepto.de/cacao/

[12] «Alimentación y agricultura sostenibles», Food and Agriculture Organization of the United Nations. Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: http://www.fao.org/sustainability/es/

[13] BBVA, «¿Qué es la agricultura de precisión? La gestión digital del campo», BBVA NOTICIAS. Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://www.bbva.com/es/sostenibilidad/que-es-la-agricultura-de-precision-la-gestion-digital-del-campo/

[14] L. Bonnaire Rivera, B. Montoya Bonilla, y F. Obando-Vidal, «Procesamiento de imágenes multiespectrales captadas con drones para evaluar el índice de vegetación de diferencia normalizada en plantaciones de café variedad Castillo», *Cienc. Tecnol. Agropecu.*, vol. 22, n.º 1, abr. 2021, doi: 10.21930/rcta.vol22_num1_art:1578.

[15] «Índices de vegetación en agricultura de precisión - Agromática». Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://www.agromatica.es/indices-de-vegetacion/

[16] H. Abd-El Monsef, S. E. Smith, D. L. Rowland, y N. Abd El Rasol, «Using multispectral imagery to extract a pure spectral canopy signature for predicting peanut maturity», *Comput. Electron. Agric.*, vol. 162, pp. 561-572, jul. 2019, doi: 10.1016/j.compag.2019.04.028.

[17] «Las-imagenes.-Pixeles-y-tamanos.pdf». Accedido: 7 de octubre de 2025. [En línea]. Disponible en: https://perio.unlp.edu.ar/catedras/wp-content/uploads/sites/125/2022/04/Las-imagenes.-Pixeles-y-tamanos.pdf

[18] «Qué es un píxel—ArcGIS Pro | Documentación». Accedido: 14 de abril de 2025. [En línea]. Disponible en: https://pro.arcgis.com/es/pro-app/latest/help/data/imagery/what-s-in-a-pixel.htm

[19] «1.4: Física del Suelo», LibreTexts Español. Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://espanol.libretexts.org/Geociencias/Ciencia_del_Suelo/Excavando_en_suelos_canadienses%3A_una_introducci%C3%B3n_a_la_ciencia_del_suelo/01%3A_Excavando/1.04%3A_F%C3%ADsica_del_Suelo

[20] «1.5: Química del Suelo», LibreTexts Español. Accedido: 20 de noviembre de 2024. [En línea]. Disponible en: https://espanol.libretexts.org/Geociencias/Ciencia_del_Suelo/Excavando_en_suelos_canadienses%3A_una_introducci%C3%B3n_a_la_ciencia_del_suelo/01%3A_Excavando/1.05%3A_Qu%C3%ADmica_del_Suelo

[21] E. Said Mohamed, Aa. Belal, S. Kotb Abd-Elmabod, M. A. El-Shirbeny, A. Gad, y M. B. Zahran, «Smart farming for improving agricultural management», *Egypt. J. Remote Sens. Space Sci.*, vol. 24, n.º 3, pp. 971-981, dic. 2021, doi: 10.1016/j.ejrs.2021.08.007.

[22] S. Wolfert, D. Goense, y C. A. G. Sørensen, «A Future Internet Collaboration Platform for Safe and Healthy Food from Farm to Fork», en *2014 Annual SRII Global Conference*, abr. 2014, pp. 266-273. doi: 10.1109/SRII.2014.47.

[23] P. Kumar *et al.*, «Role of artificial intelligence, sensor technology, big data in agriculture: next-generation farming», en *Bioinformatics in Agriculture*, P. Sharma, D. Yadav, y R. K. Gaur, Eds., Academic Press, 2022, pp. 625-639. doi: 10.1016/B978-0-323-89778-5.00035-0.

[24] «Aprendizaje automático: descripción general | Temas ScienceDirect». Accedido: 26 de junio de 2024. [En línea]. Disponible en: https://www.sciencedirect.com/topics/computer-science/machine-learning

[25] «¿Qué es el aprendizaje supervisado? | IBM». Accedido: 26 de junio de 2024. [En línea]. Disponible en: https://www.ibm.com/es-es/topics/supervised-learning

[26] «¿Qué es el aprendizaje no supervisado? | IBM». Accedido: 26 de junio de 2024. [En línea]. Disponible en: https://www.ibm.com/es-es/topics/unsupervised-learning

[27] G. Zhang, C. D. Carrasco, K. Winsler, B. Bahle, F. Cong, y S. J. Luck, «Assessing the effectiveness of spatial PCA on SVM-based decoding of EEG data», *NeuroImage*, vol. 293, p. 120625, jun. 2024, doi: 10.1016/j.neuroimage.2024.120625.

[28] J. Triana-Martinez, A. Álvarez-Meza, y G. Castellanos-Dominguez, «Enhancing agricultural data interpretability and visualization with TabNet-driven feature extraction and Local Biplots», *Results Eng.*, vol. 27, p. 106672, sep. 2025, doi: 10.1016/j.rineng.2025.106672.

[29] M. Macho Stadler, *Topologia General*. Universidad del Pa´ıs Vasco–Euskal Herriko Unibertsitatea. [En línea]. Disponible en: https://www.ehu.eus/~mtwmastm/TopoGralMana.pdf

[30] «GIS for Agriculture | Precision Agriculture & Farm Management». Accedido: 25 de octubre de 2025. [En línea]. Disponible en: https://www.esri.com/en-us/industries/agriculture/overview

[31] A. Dina Nur y F. Achmad, «Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) Approach for Identifying Potential Villages in Buleleng Regency». Accedido: 25 de octubre de 2025. [En línea]. Disponible en: https://www.researchgate.net/publication/390743298_A_Hierarchical_Density-Based_Spatial_Clustering_of_Applications_with_Noise_HDBSCAN_Approach_for_Identifying_Poten

tial_Villages_in_Buleleng_Regency
[32] N. K. Mogurampally, K. V. S. S. Chaitanya, y B. A. Bambah, «Quantum Entanglement in Coupled Lossy Waveguides Using SU(2) and SU(1, 1) Thermo-Algebras», *J. Mod. Phys.*, vol. 06, n.º 11, pp. 1554-1571, 2015, doi: 10.4236/jmp.2015.611158.
[33] K.-W. Wong, G. A. M. Dreschhoff, y H. J. N. Jungner, «Breaking SU(3) Symmetry and Baryon Masses», *J. Mod. Phys.*, vol. 06, n.º 11, pp. 1492-1497, 2015, doi: 10.4236/jmp.2015.611153.
[34] W. Guo *et al.*, «Develop agricultural planting structure prediction model based on machine learning: The aging of the population has prompted a shift in the planting structure toward food crops», *Comput. Electron. Agric.*, vol. 221, p. 108941, jun. 2024, doi: 10.1016/j.compag.2024.108941.
[35] Mustaqimah, Devianti, A. A. Munawar, y S. Sufardi, «Capability of short Vis-NIR band tandem with machine learning to rapidly predict NPK content in tropical farmland: A case study of Aceh Province agricultural soil dry land, Indonesia», *Case Stud. Chem. Environ. Eng.*, vol. 9, p. 100711, jun. 2024, doi: 10.1016/j.cscee.2024.100711.
[36] C. Leon-Moreno *et al.*, «The First National Survey of Cadmium in Cacao Farm Soil in Colombia», *Agronomy*, vol. 11, n.º 4, Art. n.º 4, abr. 2021, doi: 10.3390/agronomy11040761.
[37] D. Bravo *et al.*, «First national mapping of cadmium in cacao beans in Colombia», *Sci. Total Environ.*, vol. 954, p. 176398, dic. 2024, doi: 10.1016/j.scitotenv.2024.176398.
[38] T. J. Fontalvo Herrera, M. A. Vega Hernández, F. Mejía Zambrano, T. J. Fontalvo Herrera, M. A. Vega Hernández, y F. Mejía Zambrano, «Método de clustering e inteligencia artificial para clasificar y proyectar delitos violentos en Colombia», *Rev. Científica Gen. José María Córdova*, vol. 21, n.º 42, pp. 550-572, jun. 2023, doi: 10.21830/19006586.1117.
[39] «Procesar imágenes de Altum con Agisoft Metashape». Accedido: 9 de febrero de 2025. [En línea]. Disponible en: https://knowledge.wingtra.com/es/procesar-imagenes-de-altum-con-agisoft-metashape
[40] «Altum-PT - Sensores para drones», AgEagle Aerial Systems Inc. Accedido: 9 de febrero de 2025. [En línea]. Disponible en: https://ageagle.com/es/drone-sensors/altum-pt/
[41] «Fundamentos_29.pdf». Accedido: 27 de febrero de 2025. [En línea]. Disponible en: https://evidenciasenpediatria.es/files/41-13363-RUTA/Fundamentos_29.pdf
[42] «Pandas - get_dummies() method», GeeksforGeeks. Accedido: 27 de febrero de 2025. [En línea]. Disponible en: https://www.geeksforgeeks.org/python-pandas-get_dummies-method/
[43] «TFG-Quiles Ruiz, Francisco Javier.pdf». Accedido: 27 de febrero de 2025. [En línea]. Disponible en: https://dspace.umh.es/bitstream/11000/32779/1/TFG-Quiles%20Ruiz%2c%20Francisco%20Javier.pdf
[44] N. Alonso Escudero, «Uso de los clústeres en la minería de datos. Creación de una aplicación web con R para clasificar o predecir datos reales.», jul. 2023, Accedido: 27 de febrero de 2025. [En línea]. Disponible en: https://gredos.usal.es/handle/10366/156797
[45] B. Lantz, *Machine Learning with R: Learn techniques for building and improving machine learning models, from data preparation to model tuning, evaluation, and working with big data*. Packt Publishing Ltd, 2023.
[46] L. McInnes, J. Healy, y J. Melville, «UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction», 18 de septiembre de 2020, *arXiv*: arXiv:1802.03426. doi: 10.48550/arXiv.1802.03426.
[47] J. Healy y L. McInnes, «Uniform manifold approximation and projection», *Nat. Rev. Methods Primer*, vol. 4, n.º 1, pp. 1-15, nov. 2024, doi: 10.1038/s43586-024-00363-x.
[48] «PDF». Accedido: 11 de marzo de 2025. [En línea]. Disponible en: https://dl.ebooksworld.ir/books/Machine.Learning.with.PyTorch.and.Scikit-

Learn.Sebastian.Raschka.Packt.9781801819312.EBooksWorld.ir.pdf
[49] P. M. Pattison, J. Y. Tsao, G. C. Brainard, y B. Bugbee, «LEDs for photons, physiology and food», *Nature*, vol. 563, n.º 7732, pp. 493-500, nov. 2018, doi: 10.1038/s41586-018-0706-x.
[50] E. Fooladgar y C. Duwig, «A new post-processing technique for analyzing high-dimensional combustion data», *Combust. Flame*, vol. 191, pp. 226-238, may 2018, doi: 10.1016/j.combustflame.2018.01.014.
[51] I. De Zarzà, J. De Curtò, y C. T. Calafate, «UMAP for Geospatial Data Visualization», *Procedia Comput. Sci.*, vol. 225, pp. 1661-1671, 2023, doi: 10.1016/j.procs.2023.10.155.
[52] «Selección de parámetros para HDBSCAN* — documentación de hdbscan 0.8.1». Accedido: 18 de marzo de 2025. [En línea]. Disponible en: https://hdbscan.readthedocs.io/en/latest/parameter_selection.html

12 ANEXOS

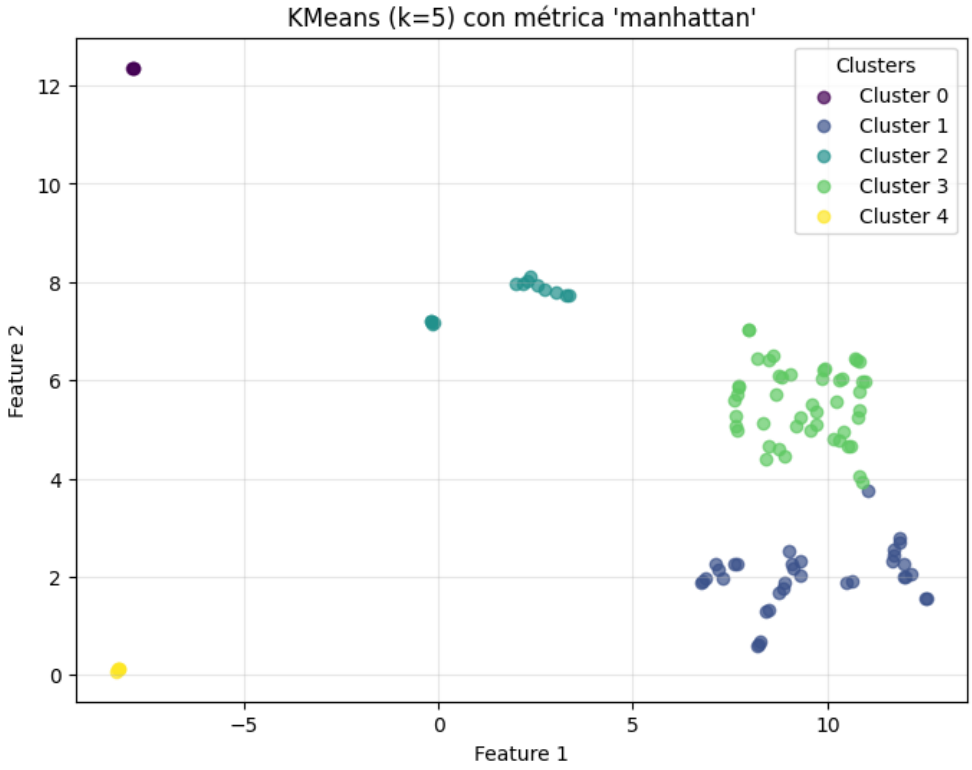
Anexo A Diccionario de variables

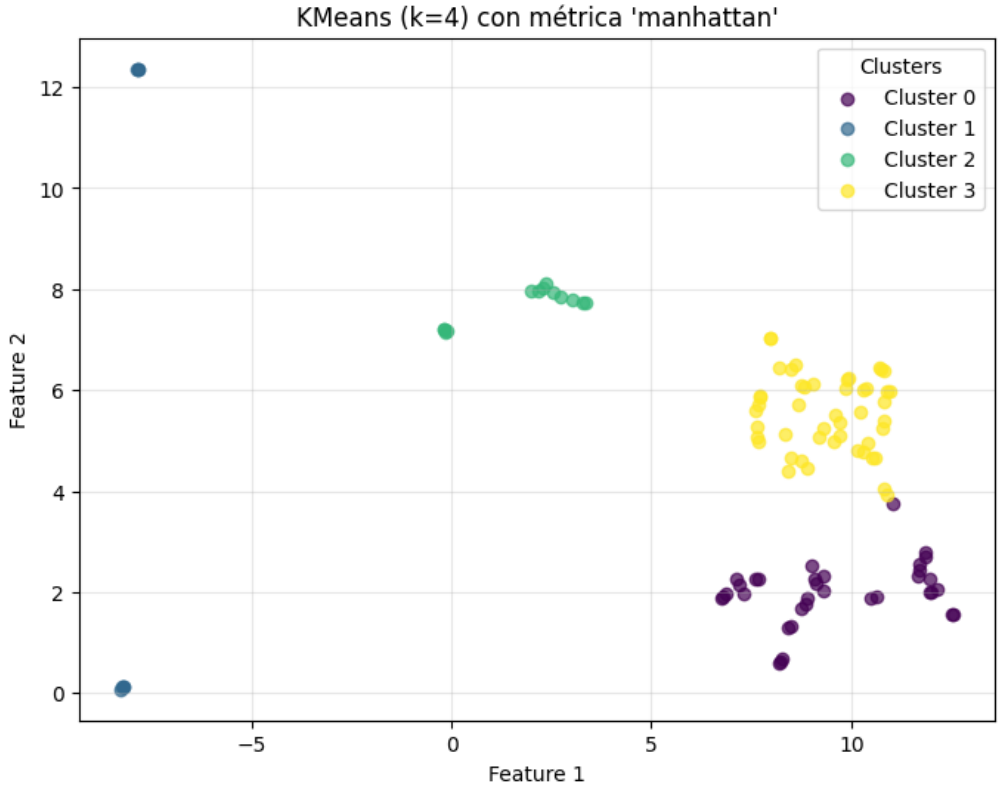
Código	Descripción
id	identificación
chum_1	En años. Cuanta experiencia tiene en el manejo de cultivo del cacao cultivo.
chum_2	A su consideración que tanto conocimiento tiene del cultivo de caco.
chum_3	A su consideración en su área de influencia que tanta experiencia tiene la mano de obra para las labore del cultivo de cacao.
chum_4	Aproximadamente a cuantos eventos de capacitación en tenar relacionados con el cacao asiste al año.
chum_5	Tiene usted conocimiento de la oferta de programas de formación formal afines con la producción de cacao.
chum_6	Considera usted que en su cultivo hay la posibilidad de dar continuidad en el tiempo a la actividad.
cfis_1	El predio posee tenencia legal
cfis_2	como considera que es la calidad de la vivienda (en términos de infraestructura)
cfis_3	Como es el acceso al predio desde el casco urbano
cfis_4	El predio tiene acceso por transporte fluvial
cfis_5	El predio tiene acceso a transporte público desde el casco urbano.
cfis_6	Como es la calidad del transporte público desde el casco urbano hasta el predio.
cfis_7	Posee acceso a agua como servicio público en el predio de cacao (ESP, concesión o distrito de riego).
cfis_8	Cuál es la frecuencia de acceso al servicio de agua
cfis_9	El agua a la que tiene acceso es apta para el consumo humano
cfis_10	posee acceso a fuentes de energía eléctrica pública (ESP)
cfis_11	cuál es la frecuencia al acceso de la energía eléctrica
cfis_12	posee acceso a servicios públicos de telecomunicaciones (teléfono, internet y TV)
cfis_13	Como es la calidad de la señal de telefonía celular
cfis_14	Posee acceso al suministro público de gas natural
cfin_1	Cuál es el valor estimado de los todos los ingresos mensuales del núcleo familia

cfin_2	Cuál es el valor estimado de los ingresos solo relacionados con la producción del cacao
cfin_3	Tienen o han tenido acceso a créditos y/o prestamos con entidades legales de financiamiento .
cfin_4	Tienen o han tenido acceso a créditos y/o presentamos con entidades y o personas
cfin_5	Alguno de los miembros del núcleo familiar es pensionado
cfin_6	reciben algún tipo subsidios por parte del estado de carácter individual
cfin_7	reciben algún tipo subsidios por parte del estado a su actividad productiva (Producción de cacao)
cfin_8	Recibe algún tipo de ingreso económico externo
cnat_1	posee acceso a fuentes naturales de agua de buena calidad.
cnat_2	A su consideración. Como es la calidad del suelo del predio para el desarrollo actividades agropecuaria
cnat_3	A su consideración como es la calidad de aire del predio
cnat_4	Hay presencia de espacios para la conservación de la biodiversidad (bosques, parques naturales, reservas naturales)
csoc_1	pertenece a algún grupo asociativo de cacao (asociación)
csoc_2	cómo es la relación de confianza con la asociación
csoc_3	cómo es la relación de confianza con la comunidad cercana al predio de cacao
csoc_4	Recibe o ha recibido algún tipo de ayuda (económica, asesoría, asistencia técnica)
csoc_5	Reciben algún tipo de ayuda (económica, asesoría, asistencia técnica) de parte de entidades no gubernamentales
csoc_6	Cotiza salud y pensión
atp	área total predio [ha]
ctl	cantidad de lotes de cacao en el predio
acc	área cultivada en cacao [ha]
oacc	De requerirlo, posee otra área disponible para ampliar el cultivo de cacao
ec	edad del cultivo
vm	variedad mayoritaria
vs	variedad secundaria
vt	variedad terciaria

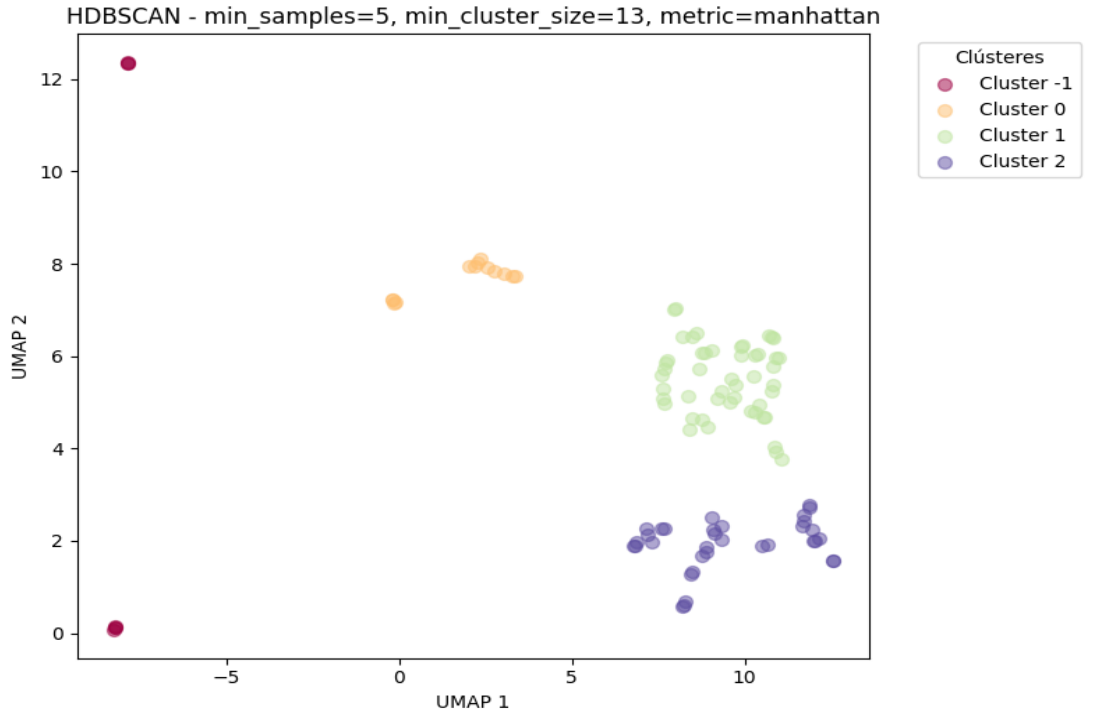
vc	variedad cuaternaria
vq	variedad quíntupla
rph	rendimiento por [kg/ha)
af	aplica fertilizante
tf	qué tipo de fertilizante
tfu	que fertilizante aplica (separe valores por coma)
fra	FRECUENCIA DE APLICACION
dta	dotación de agua
cdta	cuanto es el volumen en litros por segundo que le asignan
fdta	con que frecuencia le asignan turno (cada cuanto día)
arc	aplica riego en el cultivo
tsr	qué tipo de sistema de riego tiene
ppla	presencia de plaga
tplam	tipo de plaga mayoritaria
tplas	tipo de plaga secundaria
tplat	tipo de plaga terciaria
tplac	tipo de plaga cuaternaria
tplaq	tplaq
penf	presencia de enfermedades
tenfm	tipo de enfermedad mayoritaria
tenfs	tipo de enfermedad secundaria
tenft	tipo de enfermedad terciaria
tenfc	tipo de enfermedad cuaternaria
tenfq	tipo de enfermedad quíntupla
rpo	realiza podas
frp	frecuencia de podas
afun	aplica fungicidas
tfun	tipo de fungicida aplica (separe los tipos por comas)
apla	aplica plaguicidas
tpla	tipo de plaguicida aplica (separe los tipos por comas)
pstc_1	Que actividades de postcosecha realiza
pstc_2	Con que infraestructura cuenta para postcosecha
pstc_3	Qué nivel de transformación realiza
pstc_4	Donde realiza la venta del cacao
pstc_5	Cuenta con algún sello o distinción que realice su producto
pstc_6	En qué mes realiza la cosecha del cacao

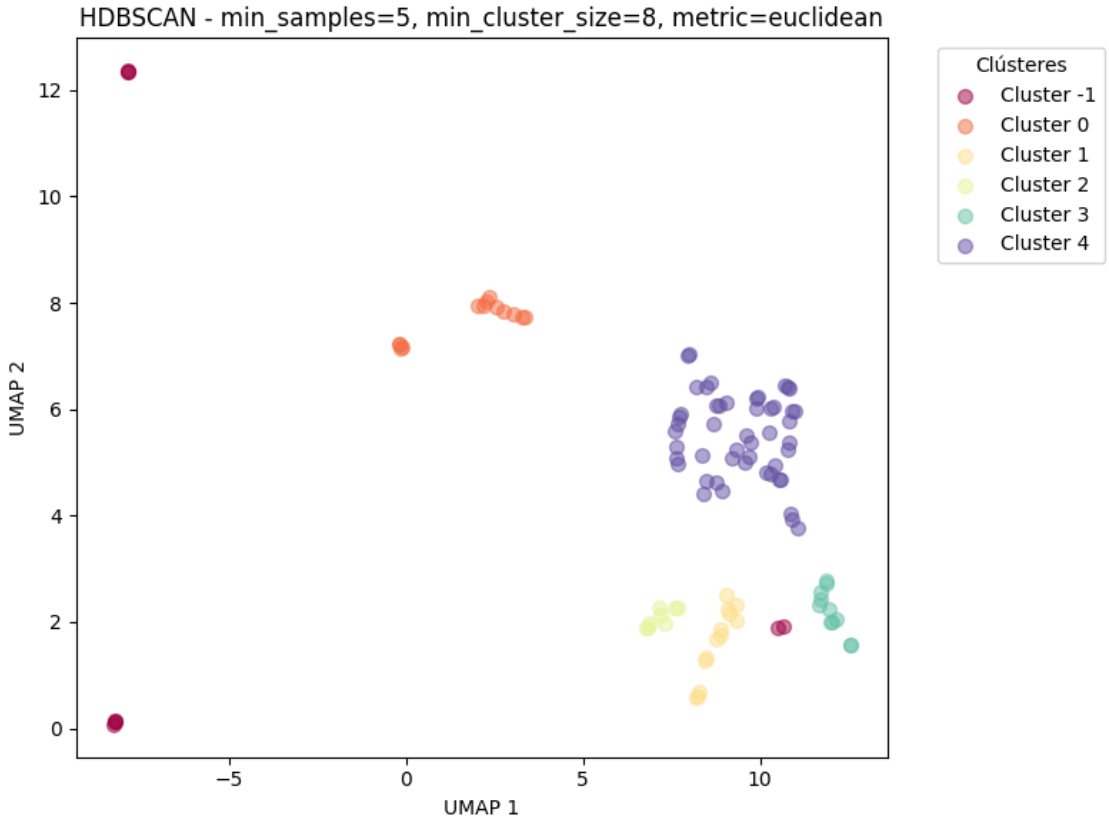
Anexo B Iteraciones K-MEANS





Anexo C Iteración H-DBSCAN





Anexo D Perfil de expertos

Nombre	Armando Torrente Trujillo
Nombre en citaciones	TORRENTE TRUJILLO, ARMANDO

12.1.1 Formación Académica

- **Doctorado** Universidad Nacional de Colombia - Sede Palmira
  Ciencias Agrarias
  Juliode1999 - Septiembrede 2003
  Caracteristicas Hidrodinámicas de suelos con alta saturación de magnesio en el Valle del Cauca
- **Maestría/Magister** Universidad Nacional de Colombia - Sede Palmira
  Magister En Suelos y Aguas
  Juliode1992 - Juliode 1994
  Metodología y extracción de sustancias húmicas de cuatro lombricompuestos
- **Especialización** UNIVERSIDAD DEL TOLIMA
  ESPECIALIZACION EN RIEGO
  Juliode1996 - Juniode 1997
  Gestión Ambiental para el Distrito de Riego el Juncal
- **Pregrado/Universitario** UNIVERSIDAD NACIONAL DE COLOMBIA SEDE BOGOTA
  INGENIERIA AGRICOLA
  Juliode1975 - Marzode 1983
  Diseño de un digestor anaeróbico para la producción de gas metano y biofertilizante en la Granja Marengo
- **Perfeccionamiento** INSTITUTO SUPERIOR DE CIENCIAS AGROPECUARIAS DE LA HABANA-CUBA
  Evaluacion y Operacion de Los Sistemas de Riego
  Juniode1997 - Juniode 1997
  Evaluación y operación de sistema de riego superficial para el cultivo de arroz
- **Perfeccionamiento** MINISTERIO DE RELACIONES EXTERIORES
  Planificacion de Redes de Riego a Presion
  Mayode1995 - Juliode 1995
  Planificación y diseño de un sistema de riego a presión para la producción de hortalizas
- **Perfeccionamiento** Centro Interamericano de Adecuacion de Tierras
  Curso de Riego y Drenaje
  Juniode1987 - Juliode 1987
  Diseño y evaluación de un sistema de riego y drenaje.
- **Secundario** Instituo Tecnico Industrial Centro Don Bosco
  Febrerode1969 - Noviembrede 1974

Nombre	Jorge Ivan Chavarro Diaz
Nombre en citaciones	CHAVARRO DIAZ, JORGE IVAN
Nacionalidad	Colombiana
Sexo	Masculino

12.1.2 Formación Académica

- **Maestría/Magister** Structuralia
  PROYECTOS DE HIDROGENO VERDE
  Marzode2023 - Mayode 2024
  CLUSTER AGROENERGÉTICO EN ZONAS MUERTAS DE COLOMBIA
- **Maestría/Magister** PONTIFICIA UNIVERSIDAD JAVERIANA
  Hidrosistemas
  Enerode2015 - Noviembrede 2017
  Diseño de un MEM's para el control ambiental
- **Pregrado/Universitario** UNIVERSIDAD SURCOLOMBIANA
  INGENIERIA AGRICOLA
  Agostode1999 - Mayode 2005
  Diagnostico y evaluacion del recurso hidrico del departamento del Huila

Anexo E Ficha técnica de materiales vegetales

Variedad	Origen	Rendimiento potencial (kg/ha)	Ciclo de maduración (días)	Resistencia a enfermedades	Perfil de sabor	Uso recomendado
CCN-51	Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Costa Rica	1 500 – 2 000	120 – 140	Alta: Moniliasis, Vascular	Fuerte, ligeramente amargo, bajo astringencia	Producción masiva, chocolate industrial
FEAR-5	Programa FEAR, Brasil	1 200 – 1 600	130 – 150	Media–Alta: Podredumbre negra, Escoba de bruja	Dulce, notas frutales	Chocolatería, fermentaciones controladas
ICS-95	Instituto de Cacao, Trinidad	1 000 – 1 400	140 – 160	Media: Escoba de bruja, Mal del brazo	Astringente, notas florales	Chocotaninos, blends con criollo
Criollo	Región amazónica de América Latina	500 – 800	150 – 180	Baja: Muy susceptible a plagas y enfermedades	Very fino, afrutado, muy fino en astringencia	Chocolates gourmet, alta calidad

Variedad	Origen	Rendimiento potencial (kg/ha)	Ciclo de maduración (días)	Resistencia a enfermedades	Perfil de sabor	Uso recomendado
Clonal	Clon seleccionado local	800 – 1 200	130 – 150	Variable según el clon; generalmente media	Perfil balanceado, moderada astringencia	Mezclas con criollo, producción media
Híbrido	Cruce entre variedades modernas y criollo	1 000 – 1 800	120 – 140	Alta: combinan resitencia de padres modernos	Versátil, notas intermedias	Producción comercial general
Trinitario	Híbrido natural (criollo × forastero)	1 000 – 1 500	140 – 160	Media: mejor que criollo, peor que forastero		