UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Escuela de Posgrado

MAESTRÍA EN GESTIÓN EMPRESARIAL

SEGMENTACIÓN DE CLIENTES UTILIZANDO ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LA EMPRESA D&C COMUNICACIONES

TESIS

PRESENTADA POR:

SILVANA BEATRIZ CABANA YUPANQUI

Para optar el Grado Académico de:

MAESTRO EN CIENCIAS (*MAGISTER SCIENTIAE*) CON MENCIÓN EN GESTIÓN EMPRESARIAL

TACNA – PERÚ

2026

UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Escuela de Posgrado

MAESTRÍA EN GESTIÓN EMPRESARIAL

SEGMENTACIÓN DE CLIENTES UTILIZANDO ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LA EMPRESA D&C COMUNICACIONES

Tesis sustentada y aprobada el 09 de junio del 2026; estando el jurado calificador integrado por:

PRESIDENTE : .....
Dr. Pedro Pablo Chambi Condori

SECRETARIO : .....
Dr. Bartolomé Jorge Anyosa Gutiérrez

MIEMBRO : .....
MSc. Israel Nazareth Chaparro Cruz

ASESOR : .....
MSc. Israel Nazareth Chaparro Cruz

CERTIFICADO DE SIMILITUD

Yo, MSc. Israel Nazareth Chaparro Cruz, en mi condición de asesor acreditado con bajo Resolución de la Escuela de Posgrado Nº 15259-2025-ESPG/UNJBG del 07 de marzo del 2025 y encontrándose dicha tesis en ejecución bajo Resolución de la Escuela de Posgrado Nº 15994-2025-ESPG/UNJBG desde 08 de agosto del 2025, del trabajo de tesis titulado: "Segmentación de clientes utilizando algoritmos de inteligencia artificial en la empresa D&C Comunicaciones " presentado por el Srta. Silvana Beatriz Cabana Yupanqui para optar el Grado Académico de Maestro en Ciencias (Magíster Scientiae) con mención en Gestión Empresarial.
Habiendo cumplido con lo establecido en el reglamento de originalidad y de similitud de trabajo de investigación y producción intelectual, considerando que según la revisión, evaluación y análisis realizado a través del software de similitud textual TURNITIN, cuenta con el nivel de similitud permitido cuyo porcentaje es 5%.

Por lo que CERTIFICO LA SIMILARIDAD de la tesis y está de acuerdo al nivel PERMITIDO, para continuar con los trámites correspondientes y para su publicación en el repositorio institucional.

Se emite el presente certificado a solicitud del interesado con fines de continuar con los trámites respectivos para la obtención del Grado Académico de Maestro en Ciencias (Magíster Scientiae) con mención en Gestión Empresarial

Tacna, 02 de julio de 2026

FIRMA ASESOR .....
Nombres y Apellidos MSc. Israel Nazareth Chaparro Cruz
DNI: 48584646

FIRMA TESISTA .....
Nombre y Apellidos Srta. Silvana Beatriz Cabana Yupanqui
DNI: 70020327

Dedicatoria

A Dios, por darme vida, fortaleza y guiar mi camino.

A mis padres, por su amor, apoyo incondicional y por ser el motor de mis logros.

A mi asesor de tesis, por su orientación y apoyo en el desarrollo de este trabajo.

Con gratitud, dedico esta tesis a quienes hicieron posible este importante paso en mi vida.

AGRADECIMIENTOS

Expreso mi profundo agradecimiento a Dios, por guiar mi camino y darme la fortaleza necesaria para culminar esta etapa de mi vida.

A mis padres, por su apoyo incondicional, su comprensión y su permanente motivación, pilares fundamentales en mi formación personal y profesional.

A mi asesor de tesis, por su orientación, paciencia, compromiso y valiosos conocimientos brindados durante el desarrollo de la presente investigación.

Finalmente, agradezco a todas las personas e instituciones que contribuyeron directa o indirectamente a la realización de este trabajo.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El estudio fue de tipo aplicado, con enfoque cuantitativo y diseño no experimental de corte transversal, desarrollado bajo el enfoque de Design Science Research. Se trabajó con una base de datos de 400 clientes de la empresa D&C Comunicaciones, considerando variables como edad, monto de operaciones y número de visitas. Se aplicaron algoritmos de aprendizaje no supervisado, específicamente K-Means, Bisecting K-Means y DBSCAN, con el objetivo de generar segmentaciones de clientes.

La evaluación de los resultados se realizó mediante métricas de clusterización como el índice de Silhouette, Davies-Bouldin, Calinski-Harabasz e inercia, complementadas con pruebas estadísticas para comparar el desempeño de los algoritmos. Asimismo, se utilizó un método de evaluación multicriterio para seleccionar la segmentación más adecuada.

Los resultados evidenciaron diferencias significativas entre los algoritmos evaluados, permitiendo identificar una segmentación óptima. Finalmente, los segmentos obtenidos fueron interpretados mediante un modelo de árbol de decisión, generando reglas comprensibles para su aplicación en la empresa.

Palabras clave: segmentación de clientes, clustering, inteligencia artificial.

ABSTRACT

This study was applied in nature, with a quantitative approach and a non experimental cross-sectional design, developed under the Design Science Research framework. A dataset of 400 customers from the company D&C Comunicaciones was used, considering variables such as age, transaction amount, and number of visits. Unsupervised learning algorithms were applied, specifically K-Means, Bisecting K-Means, and DBSCAN, with the aim of generating customer segmentations.

The evaluation of the results was carried out using clustering metrics such as the Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, and inertia, complemented by statistical tests to compare the performance of the algorithms. Additionally, a multicriteria evaluation method was employed to select the most appropriate segmentation.

The results showed significant differences among the evaluated algorithms, allowing the identification of an optimal segmentation. Finally, the obtained segments were interpreted using a decision tree model, generating understandable rules for their application in the company.

Keywords: customer segmentation, clustering, artificial intelligence.

INTRODUCCIÓN

La presente investigación tiene por finalidad identificar y analizar segmentos de clientes de la empresa D&C Comunicaciones mediante el uso de algoritmos de inteligencia artificial no supervisada, con el propósito de determinar una segmentación óptima que facilite la comprensión de los perfiles de clientes y contribuya a la formulación de estrategias empresariales más eficientes. Para ello, se evalúan distintas soluciones de clustering a partir de métricas de validación interna y se interpretan los segmentos obtenidos mediante un modelo supervisado de apoyo que permita traducir la segmentación en reglas comprensibles y aplicables.

La tesis se encuentra estructurada en cinco capítulos. En el Capítulo I se presenta la realidad problemática, el planteamiento del problema, los objetivos, la hipótesis y las variables de estudio. En el Capítulo II se desarrolla el marco teórico, que comprende la revisión de antecedentes y las bases conceptuales que sustentan la investigación. En el Capítulo III se expone el marco metodológico, describiendo el tipo, nivel y diseño de investigación, así como la población, muestra, técnicas, instrumentos y procedimientos aplicados. En el Capítulo IV se presentan los resultados obtenidos del análisis descriptivo, la aplicación de los algoritmos de segmentación, la evaluación comparativa de métricas y la interpretación de los segmentos identificados. En el Capítulo V se desarrolla la discusión de resultados y, finalmente, se exponen las conclusiones, recomendaciones y referencias bibliográficas.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

1.1.1 Identificación del problema

En el contexto de la economía digital global, las organizaciones enfrentan una creciente presión por adaptarse a mercados altamente dinámicos, caracterizados por consumidores cada vez más informados, exigentes y con múltiples puntos de contacto. A nivel internacional, el aprovechamiento de grandes volúmenes de datos provenientes de interacciones comerciales, transacciones y comportamientos de consumo se ha convertido en un factor clave para la toma de decisiones estratégicas orientadas al cliente (McAfee et al., 2012).

En un entorno empresarial altamente competitivo y centrado en el cliente, las empresas necesitan comprender a profundidad las características y comportamientos de sus consumidores para diseñar estrategias que optimicen sus resultados comerciales (Kotler & Keller, 2016). Sin embargo, muchas organizaciones aún no logran aprovechar de manera efectiva los datos de sus clientes debido a la falta de un enfoque sistemático que permita relacionar diversas características de sus clientes como: demográficas (edad, ingresos, género), transaccionales (frecuencia de compra, gasto promedio) y conductuales (canales preferidos de compra, categorías de productos). Esta limitación impide realizar segmentaciones precisas, lo que afecta la capacidad de las empresas para personalizar estrategias comerciales y aumentar la fidelización del cliente (Stone & Jacobs, 2008).

En el ámbito regional, particularmente en América Latina, muchas pequeñas y medianas empresas gestionan volúmenes significativos de datos transaccionales y demográficos generados por sus operaciones cotidianas; sin embargo, estos datos suelen ser utilizados de manera limitada. En países como el Perú, esta situación se ve acentuada por brechas en capacidades tecnológicas, una incipiente cultura de análisis de datos (CEPAL, 2013).

La segmentación de clientes, una herramienta esencial para identificar patrones entre consumidores, ha evolucionado hacia enfoques basados en datos y el uso de técnicas de inteligencia artificial (IA) (Gartner, 2022). La inteligencia artificial ofrece herramientas avanzadas como algoritmos de aprendizaje automático y análisis de datos que permiten identificar patrones complejos y relaciones no evidentes entre las características de los clientes. Esto habilita a las empresas a realizar segmentaciones más precisas, dinámicas y adaptadas al contexto cambiante del mercado. Además, las técnicas de IA, como el análisis de clústeres y los algoritmos de aprendizaje supervisado o no supervisado, facilitan la clasificación de los clientes en segmentos homogéneos con base en múltiples variables, como edad, género, ingreso anual y puntaje de gasto, optimizando la personalización de las estrategias comerciales.

D&C Comunicaciones es una empresa ubicada en la ciudad de Moquegua - Perú, y está dedicada a ofrecer una variedad de servicios y productos esenciales para el consumidor. Entre sus principales líneas de negocio se encuentran la realización de operaciones financieras a través de agentes de banco. Su modelo de negocio atiende a un grupo heterogéneo de clientes, lo que representa una oportunidad significativa para implementar estrategias basadas en segmentación de clientes mediante inteligencia artificial.

Para este estudio, se dispone de información sobre los clientes de D&C Comunicaciones como: edad, género (masculino o femenino), monto de operaciones y número de visitas. Con estos datos, se puede emplear inteligencia artificial para identificar patrones y agrupar a los clientes en segmentos específicos, maximizando así la capacidad de la empresa para personalizar sus estrategias comerciales.

Es así que arribamos a la siguiente pregunta:

1.2 FORMULACIÓN DEL PROBLEMA

1.2.1 Problema principal

¿Cómo segmentar los clientes de la empresa D&C Comunicaciones utilizando algoritmos de inteligencia artificial de manera que se obtengan segmentos interpretables

y útiles para la toma de decisiones en la empresa?

1.2.2 Problemas secundarios

- ¿Cuáles son las características de los clientes de la empresa D&C Comunicaciones?
- ¿Cómo seleccionar y preprocesar las características relevantes de los clientes de la empresa D&C Comunicaciones para su utilización en algoritmos de inteligencia artificial?
- ¿Cómo generar segmentaciones de clientes de la empresa D&C Comunicaciones mediante algoritmos de inteligencia artificial no supervisados?
- ¿Qué algoritmo de inteligencia artificial no supervisado genera la mejor segmentación de clientes de la empresa D&C Comunicaciones?
- ¿Cómo evaluar la calidad de las segmentaciones obtenidas de los clientes de la empresa D&C Comunicaciones mediante métricas de clusterización?
- ¿Cómo seleccionar la segmentación más adecuada de los clientes de la empresa D&C Comunicaciones mediante un método de evaluación multicriterio basado en métricas de calidad de segmentación?
- ¿Cómo interpretar los segmentos obtenidos de los clientes de la empresa D&C Comunicaciones mediante un modelo de inteligencia artificial supervisado que permita generar reglas de clasificación comprensibles para la empresa?

1.3 JUSTIFICACIÓN E IMPORTANCIA DE LA INVESTIGACIÓN

1.3.1 Justificación social

Realizando la investigación sobre el tema, “Segmentación de clientes utilizando inteligencia artificial en la empresa D&C Comunicaciones”, se busca desarrollar e implementar algoritmos de inteligencia artificial para mejorar la identificación y

clasificación de los clientes de la empresa. A través de este estudio, se podrán analizar características clave de los consumidores y ofrecer estrategias personalizadas basadas en su comportamiento de compra.

La presente investigación se justifica socialmente, ya que beneficiará directamente a los clientes de D&C Comunicaciones, permitiendo ofrecer productos y servicios adaptados a sus necesidades. Asimismo, al mejorar la personalización de la atención al cliente, se espera contribuir a la satisfacción del consumidor y fortalecer la relación entre la empresa y su público objetivo.

1.3.2 Justificación económica

La tesis se justifica desde el punto de vista económico, dado que la segmentación efectiva de clientes permite a las empresas mejorar la rentabilidad y eficiencia de sus estrategias comerciales. Identificar correctamente los diferentes segmentos de consumidores posibilita el diseño de campañas de marketing más precisas y rentables, reduciendo costos innecesarios y aumentando el retorno de inversión en publicidad.

Además, la implementación de inteligencia artificial en la segmentación de clientes puede generar una ventaja competitiva para D&C Comunicaciones, optimizando su capacidad de fidelización y captación de nuevos clientes. Cabe mencionar que esta investigación empleará herramientas de software libre para el análisis de datos, lo que minimiza costos operativos y maximiza los beneficios que se pueden obtener con la aplicación de estas tecnologías.

1.3.3 Justificación técnica-ambiental

Este estudio presenta un modelo basado en inteligencia artificial para la segmentación de clientes, en el cual se implementarán algoritmos de clustering como BisectingK-Means, K-Means y DBSCAN. La investigación seguirá un enfoque metodológico estructurado para garantizar la validez de los resultados y la precisión en la agrupación de clientes según sus características demográficas, transaccionales y conductuales.

Desde una perspectiva ambiental, la optimización de estrategias de segmentación mediante inteligencia artificial permite una mejor planificación en el uso de recursos empresariales, reduciendo desperdicios en publicidad innecesaria y mejorando la eficiencia en la gestión de productos y servicios. Además, la digitalización de procesos minimiza el uso de material impreso en campañas de marketing, contribuyendo a la reducción del impacto ambiental.

1.3.4 Justificación académica

La inteligencia artificial aplicada a la segmentación de clientes es un área de estudio en constante evolución, con un impacto significativo en la toma de decisiones empresariales. Esta investigación representa una oportunidad para aplicar conocimientos adquiridos en el campo del aprendizaje automático, el análisis de datos y la inteligencia de negocios.

Además, este estudio contribuirá a la generación de nuevo conocimiento en el ámbito académico, permitiendo explorar la efectividad de distintos algoritmos de segmentación y su impacto en la optimización de estrategias comerciales. Los resultados obtenidos podrán servir como base para futuras investigaciones en el uso de inteligencia artificial en el sector empresarial.

1.3.5 Importancia de la investigación

La importancia de esta investigación radica en la creciente necesidad de las empresas de adoptar estrategias basadas en datos para mejorar su competitividad en el mercado. La segmentación de clientes basada en inteligencia artificial permite una mejor comprensión del comportamiento del consumidor, facilitando la toma de decisiones estratégicas y la personalización de ofertas y servicios.

Además, este estudio no solo beneficiará a D&C Comunicaciones, sino que también proporcionará un modelo aplicable a otras empresas del sector que buscan mejorar su relación con los clientes y optimizar sus estrategias comerciales. Al demostrar la viabilidad y efectividad de los algoritmos de clustering en la segmentación de clientes,

esta investigación contribuirá al desarrollo de soluciones innovadoras en la gestión empresarial.

1.4 OBJETIVOS

Los objetivos establecen qué se pretende con la investigación, y dado que las investigaciones buscan contribuir a resolver un problema en especial, los objetivos deberán mencionar cuál es y de qué manera se piensa que el estudio ayudará a resolverlo (Hernández-Sampieri & Mendoza, 2018)

Además, los objetivos señalan los objetivos que se esperan, siendo que los objetivos específicos se derivan del objetivo general y contribuyen al logro de este (Arias, 2012).

De manera complementaria, la formulación de objetivos en la presente investigación es consistente con el enfoque de Design Science Research, el cual plantea que los estudios orientados al diseño deben estructurarse en torno al desarrollo y evaluación de un artefacto destinado a resolver un problema en un contexto específico. Según Wieringa (2014), la investigación basada en diseño se organiza en torno a actividades de diseño y evaluación empírica de artefactos en su contexto, las cuales pueden expresarse mediante objetivos que describen las etapas de análisis del problema, construcción de la solución y evaluación de su desempeño. En este sentido, los objetivos de la investigación representan las diferentes etapas necesarias para diseñar y evaluar un modelo de segmentación de clientes aplicado a la empresa D&C Comunicaciones.

Asimismo, Dresch et al. (2014) señalan que la investigación basada en diseño se caracteriza por el desarrollo de artefactos orientados a la solución de problemas reales, donde el proceso de investigación se estructura en actividades sucesivas de construcción y evaluación de la solución propuesta. Por ello, la formulación de objetivos generales y específicos permite describir de manera ordenada las etapas necesarias para la obtención y validación del artefacto desarrollado, garantizando la coherencia metodológica entre el problema de investigación, el diseño experimental y los resultados esperados.

1.4.1 Objetivo general

Segmentar los clientes de la empresa D&C Comunicaciones utilizando algoritmos de inteligencia artificial, con el fin de obtener segmentos interpretables y útiles para la toma de decisiones en la empresa.

1.4.2 Objetivos específicos

- Describir las características de los clientes de la empresa D&C Comunicaciones.
- Seleccionar y preprocesar las características relevantes de los clientes de la empresa D&C Comunicaciones para su utilización en algoritmos de inteligencia artificial.
- Aplicar algoritmos de inteligencia artificial no supervisados para generar segmentaciones de clientes de la empresa D&C Comunicaciones.
- Comparar los algoritmos de inteligencia artificial no supervisados a través de sus métricas en la generación de segmentaciones de clientes de la empresa D&C Comunicaciones.
- Evaluar la calidad de las segmentaciones obtenidas de los clientes de la empresa D&C Comunicaciones mediante métricas de clusterización.
- Seleccionar la segmentación más adecuada de los clientes de la empresa D&C Comunicaciones mediante un método de evaluación multicriterio basado en métricas de calidad de segmentación.
- Interpretar los segmentos obtenidos de los clientes de la empresa D&C Comunicaciones mediante un modelo de inteligencia artificial supervisado.

1.5 HIPÓTESIS

En la presente investigación no se formulan hipótesis de manera general, debido a que el estudio se enmarca dentro del enfoque de Design Science Research, cuyo propósito

principal es el diseño y evaluación de artefactos orientados a la solución de problemas reales, más que la verificación de relaciones causales entre variables.

El objetivo central consiste en desarrollar y evaluar un modelo de segmentación de clientes basado en algoritmos de inteligencia artificial aplicado al contexto de la empresa D&C Comunicaciones. En este sentido, el interés principal no es contrastar teorías ni comprobar relaciones causales entre variables independientes y dependientes, sino diseñar e implementar una solución tecnológica que permita obtener segmentaciones útiles, interpretables y aplicables.

De acuerdo con Wieringa (2014), la investigación basada en diseño se orienta al desarrollo y evaluación empírica de artefactos en un contexto específico, donde los problemas de investigación se formulan como problemas de diseño y preguntas sobre el desempeño del artefacto, en lugar de hipótesis causales tradicionales. En este enfoque, los resultados se evalúan en términos de utilidad, desempeño y adecuación del artefacto.

De manera consistente, Dresch et al. (2014) señalan que la investigación basada en diseño se caracteriza por la construcción de soluciones innovadoras para problemas reales, donde el conocimiento se genera a partir del diseño, implementación y evaluación de artefactos, sin que la formulación de hipótesis constituya un requisito metodológico indispensable.

No obstante, en el caso específico del objetivo:

- Comparar los algoritmos de inteligencia artificial no supervisados a través de sus métricas en la generación de segmentaciones de clientes de la empresa D&C Comunicaciones.

Se incorporan hipótesis de carácter estadístico, debido a que dicho objetivo implica un análisis inferencial orientado a determinar si existen diferencias significativas en el desempeño de los algoritmos evaluados.

En este sentido, resulta pertinente considerar que los objetivos y preguntas de investigación pueden refinarse durante el desarrollo del estudio, dando lugar a la formulación de nuevas hipótesis no planteadas inicialmente. Asimismo, en investigaciones

con análisis cuantitativo es válido incorporar hipótesis cuando estas cumplen una función orientadora del análisis, permitiendo estructurar la evaluación y proporcionar lógica al estudio. Las hipótesis de investigación constituyen proposiciones tentativas que guían el proceso investigativo y ayudan a delimitar lo que se pretende probar (Hernández-Sampieri & Mendoza, 2018).

En este contexto, se plantean las siguientes hipótesis:

- **Hipótesis nula (H_0):** No existen diferencias estadísticamente significativas en las métricas de calidad de segmentación entre los algoritmos de inteligencia artificial no supervisados evaluados.
- **Hipótesis alternativa (H_1):** Existen diferencias estadísticamente significativas en al menos una de las métricas de calidad de segmentación entre los algoritmos de inteligencia artificial no supervisados evaluados.

Estas hipótesis no buscan establecer relaciones causales, sino apoyar el proceso de evaluación comparativa del artefacto, permitiendo determinar de manera rigurosa qué algoritmos presentan un mejor desempeño en términos de calidad de segmentación.

Por lo tanto, la presente investigación mantiene un enfoque de diseño centrado en el desarrollo y evaluación del modelo de segmentación, incorporando de manera complementaria un análisis inferencial específico para la comparación de algoritmos. En consecuencia, la validez de los resultados se sustenta en la utilidad, calidad e interpretabilidad de las segmentaciones obtenidas, así como en la evidencia estadística que respalda las diferencias observadas entre los algoritmos.

1.6 Variables

1.6.1 Identificación de variables

Variable 1: Algoritmos inteligencia artificial

Variable 2: Segmentación de clientes

1.6.2 Caracterización de las variables

Variable 1: Algoritmos de inteligencia artificial

Los algoritmos de inteligencia artificial son procedimientos computacionales diseñados para simular el razonamiento humano, permitiendo la automatización de tareas complejas mediante el aprendizaje a partir de datos. Dentro de estos algoritmos, se encuentran algoritmos de aprendizaje supervisado y no supervisado, los cuales se aplican en distintas áreas, incluyendo la segmentación de clientes. Un algoritmo de inteligencia artificial también puede ser definido como «un conjunto de reglas o instrucciones que permiten a una máquina aprender, razonar y tomar decisiones basadas en datos y experiencias previas».(Russell & Norvig, 2009)

Variable 2: Segmentación de clientes

La segmentación de clientes es un proceso de clasificación de consumidores en grupos homogéneos basados en características compartidas, con el fin de personalizar estrategias de marketing y mejorar la relación con los clientes. ”La segmentación de clientes es la subdivisión de un mercado en subconjuntos distintos de consumidores que tienen necesidades, características o comportamientos similares, y que podrían requerir productos o estrategias de marketing diferenciadas”. (Kotler & Keller, 2016)

1.6.3 Definición operacional de las variables

No todas las variables se pueden descomponer en más de un elemento o dimensiones Arias, 2012. En la presente investigación, solo se midió el efecto de la variable 1 (Algoritmos de inteligencia artificial) en la variable 2 (Segmentación de clientes)

Variable 2: Segmentación de clientes

Dimensiones e indicadores:

- Métricas
    - Índice de Silhouette.
    - Índice de David-Bouldin.

- Índice de Calinski-Harabasz.
- Inercia.

1.7 LIMITACIONES DE LA INVESTIGACIÓN

Entre las limitaciones más relevantes se menciona las siguientes:

1.8 Limitaciones del problema

En toda investigación existen factores que pueden restringir o influir en el desarrollo y alcance de los resultados obtenidos. En este estudio, se identifican diversas limitaciones relacionadas con la segmentación de clientes utilizando algoritmos de inteligencia artificial en la empresa D&C Comunicaciones.

Una de las principales limitaciones radica en la disponibilidad y calidad de los datos. La efectividad de los algoritmos de segmentación depende en gran medida de la cantidad y calidad de los datos recopilados. Si los datos son incompletos, inconsistentes o presentan errores, los resultados de la segmentación pueden verse afectados, reduciendo su precisión y utilidad para la toma de decisiones.

Otra restricción significativa está relacionada con la selección y parametrización de los algoritmos. BisectingK-Means, K-Means y DBSCAN, los algoritmos utilizados en esta investigación requieren parámetros específicos como el número de clusters en K-Means o la distancia mínima entre puntos en DBSCAN. La elección inadecuada de estos parámetros puede influir en la interpretación de los resultados y la estabilidad de los segmentos generados.

Asimismo, la capacidad de generalización de los resultados puede ser limitada. Dado que el estudio se enfoca en la empresa D&C Comunicaciones, los hallazgos pueden no ser directamente aplicables a otras empresas con modelos de negocio o mercados distintos. La extrapolación de los resultados a otros sectores requiere un análisis adicional para validar la aplicabilidad de la metodología empleada.

Otra limitación a considerar es la capacidad computacional necesaria para ejecutar los algoritmos de inteligencia artificial sobre grandes volúmenes de datos. A pesar de que K-Means es eficiente en términos computacionales, DBSCAN puede ser más costoso en conjuntos de datos de alta dimensión, lo que puede influir en la escalabilidad del modelo de segmentación.

Finalmente, se debe reconocer la limitación en la interpretación de los clusters generados. Aunque las métricas de evaluación, como el índice de Silhouette y el índice de Davies-Bouldin, proporcionan medidas cuantitativas de la calidad de los clusters, la interpretación cualitativa de los segmentos generados sigue requiriendo la validación por parte de expertos en el negocio, lo que puede introducir subjetividad en el análisis.

A pesar de estas limitaciones, la presente investigación busca proporcionar un enfoque robusto y metodológicamente sólido para la segmentación de clientes mediante inteligencia artificial, ofreciendo una base para futuras investigaciones y aplicaciones empresariales en este ámbito.

CAPÍTULO II

MARCO TEÓRICO

2.1 ANTECEDENTES DEL ESTUDIO

2.1.1 Antecedentes internacionales

La segmentación de clientes es una estrategia esencial para optimizar la personalización de servicios y estrategias de marketing en diversas industrias. Tradicionalmente, esta tarea se realizaba mediante enfoques manuales y heurísticos, los cuales resultaban poco eficientes al analizar grandes volúmenes de datos (Priyadarshni et al., 2023). Con el auge del aprendizaje automático, los algoritmos de clustering han demostrado ser herramientas valiosas para mejorar la precisión y eficiencia en la identificación de patrones de comportamiento de los clientes (Othayoth & Muthalagu, 2022). En particular, el algoritmo K-Means ha sido ampliamente utilizado en la segmentación de clientes debido a su capacidad para agrupar datos de manera rápida y eficaz en función de características como ingresos, edad y patrones de gasto (Reddy et al., 2023).

Estudios recientes han explorado la aplicación de K-Means en distintos sectores, como el financiero y el comercio electrónico. Por ejemplo, la segmentación de clientes en el sector de tarjetas de crédito ha demostrado que el uso de K-Means permite identificar grupos de usuarios con patrones de gasto similares, facilitando la implementación de estrategias de fidelización y marketing personalizadas (Priyadarshni et al., 2023). Además, el análisis comparativo entre K-Means y otros algoritmos de clustering, como la agrupación jerárquica y el clustering basado en densidad (DBSCAN), ha revelado que cada algoritmo tiene ventajas y limitaciones dependiendo del tipo de datos y los objetivos comerciales (Kumar et al., 2025). En esa misma línea, Bartels (2022) comparó K-Means y DBSCAN en un estudio de segmentación de clientes con datos abiertos, evidenciando que K-Means presentó mejor desempeño, mientras que DBSCAN obtuvo coeficientes de Silhouette negativos.

Por otro lado, el uso de DBSCAN ha cobrado relevancia debido a su capacidad para identificar estructuras no lineales y manejar datos con ruido, lo que lo convierte en una alternativa robusta frente a K-Means en escenarios con datos dispersos o distribuciones irregulares (Joga et al., 2022). Investigaciones previas han demostrado que DBSCAN es especialmente útil en situaciones donde no es posible determinar un número fijo de clusters, ya que permite detectar patrones ocultos en los datos sin la necesidad de especificar un número predeterminado de segmentos (Jadwal et al., 2022).

Sin embargo, una de las principales limitaciones de K-Means es su sensibilidad a la selección del número de clusters y su dependencia de centroides iniciales, lo que puede afectar la estabilidad de los resultados (Narayana et al., 2022). Por otro lado, aunque DBSCAN supera algunas de estas limitaciones al identificar clusters de diferentes formas y tamaños, su rendimiento puede verse afectado por la necesidad de ajustar parámetros clave como el radio de vecindad y la cantidad mínima de puntos por cluster (Ganesh et al., 2024). Estudios recientes han comparado la efectividad de ambos algoritmos en escenarios comerciales, encontrando que mientras K-Means ofrece una segmentación clara y bien definida en datos estructurados, DBSCAN es más adecuado para conjuntos de datos con alta variabilidad y ruido (Rahmet & Gürkan, 2024).

En la actualidad, la combinación de técnicas de clustering con enfoques de reducción de dimensionalidad y optimización de hiperparámetros ha mejorado significativamente la precisión de la segmentación de clientes (Paramasivan et al., 2024). El uso de análisis exploratorio de datos (EDA) y visualización avanzada ha permitido comprender mejor los patrones de segmentación, lo que facilita la toma de decisiones estratégicas en empresas de distintos sectores (Preethi et al., 2023). Por ejemplo, en la industria minorista y de comercio electrónico, la implementación de K-Means junto con sistemas de recomendación ha demostrado aumentar la retención de clientes y mejorar la personalización de ofertas (Othayoth & Muthalagu, 2022).

2.1.2 Antecedentes nacionales

La segmentación de clientes mediante inteligencia artificial ha sido ampliamente estudiada en distintos sectores, con el fin de mejorar la personalización de estrategias comerciales y la toma de decisiones empresariales. Diversas investigaciones en Perú han aplicado técnicas de aprendizaje no supervisado, como K-Means y DBSCAN, para la identificación de patrones de comportamiento en consumidores y su agrupación en segmentos más homogéneos.

En este sentido, Buleje Calderón (2022) realizó un estudio en una empresa textil peruana, en el que se aplicaron técnicas de segmentación como PAM, DBSCAN y Fuzzy Clustering para analizar la cartera de clientes. Su investigación concluyó que la correcta selección de estas técnicas permitió identificar patrones de compra y diferenciar segmentos clave, facilitando la personalización de estrategias comerciales. Sin embargo, si bien la investigación permitió segmentar clientes con algoritmos avanzados, no exploró en profundidad la comparación entre distintos algoritmos de clustering, lo que deja un espacio para estudios comparativos más detallados.

Siguiendo esta línea, Alikhan Trujillo et al. (2023) llevaron a cabo una investigación en el mercado de bebidas del departamento de Ica, con el objetivo de automatizar la segmentación de clientes y optimizar las estrategias de ventas. En este caso, los autores implementaron técnicas de aprendizaje no supervisado como K-Means, DBSCAN y HDBSCAN, aplicando un riguroso tratamiento de datos para obtener agrupaciones más precisas. Los resultados demostraron que la combinación de estas técnicas, junto con un análisis detallado de los perfiles de clientes, mejora la planificación comercial y permite diseñar campañas de marketing más efectivas. No obstante, aunque el estudio validó la utilidad de estos algoritmos en un contexto de ventas, aún queda por explorar en qué escenarios específicos un algoritmo supera al otro en términos de desempeño y calidad de segmentación.

De manera similar, Salazar Gebol (2015) aplicó técnicas de segmentación en una entidad bancaria con el objetivo de analizar el comportamiento de consumo de los clientes y mejorar la dirección de las ofertas comerciales. A través del uso de K-Means, el

estudio examinó diferentes variables transaccionales y utilizó métricas de validación como el índice de Silhouette para evaluar la cohesión y separación de los clústeres. Los hallazgos evidenciaron que una segmentación efectiva con base en el comportamiento de compra permite optimizar la fidelización de los clientes y la eficiencia de las campañas publicitarias. Aunque este trabajo se centró en K-Means, no incluyó comparaciones con otros algoritmos de segmentación, lo que limita el análisis de robustez y adaptabilidad del modelo ante datos con ruido o patrones no lineales.

En otro sector, Díaz Pulido (2023) exploró el uso de técnicas de minería de datos espaciales para analizar la congestión vehicular en la ciudad de Trujillo. En su estudio, se implementaron los algoritmos K-Means y DBSCAN para segmentar las zonas de mayor afluencia de tráfico y determinar puntos críticos de congestión. Utilizando la metodología CRISP-DM y validando los resultados con el método del codo y el índice de Silhouette, el estudio concluyó que DBSCAN fue más eficiente para identificar áreas de alta densidad de tráfico, mientras que K-Means permitió realizar una segmentación más estructurada y generalizable. Este hallazgo es relevante, ya que refuerza la idea de que la elección del algoritmo depende del tipo de datos y del objetivo de la segmentación, un aspecto que es clave en la presente investigación.

Finalmente, Manchego Melendez y Rimay Pelaez (2023) abordaron la aplicación de la segmentación en el sector de medios de comunicación, implementando un sistema basado en K-Means para analizar el comportamiento de los clientes. La investigación resaltó la importancia de la segmentación automatizada para optimizar estrategias de marketing y personalizar campañas publicitarias. A través del desarrollo de un sistema de información, se logró mejorar la eficiencia en la toma de decisiones comerciales, permitiendo a las empresas diseñar campañas más dirigidas y efectivas. Sin embargo, la investigación no exploró técnicas adicionales como DBSCAN, lo que deja una oportunidad para estudios que comparen la eficacia de diferentes enfoques en segmentación.

Estos antecedentes resaltan el creciente interés en la segmentación de clientes mediante algoritmos de inteligencia artificial en Perú. Si bien cada estudio ha contribuido con importantes hallazgos en su respectivo sector, aún existe una necesidad de comparar

a profundidad distintos algoritmos de clustering, como K-Means y DBSCAN, para determinar su aplicabilidad en diferentes entornos empresariales. Esta investigación busca aportar en esa dirección, evaluando la segmentación de clientes en D&C Comunicaciones y proporcionando un análisis detallado sobre la idoneidad de cada técnica para optimizar estrategias comerciales.

2.2 BASES TEÓRICAS

2.2.1 Segmentación de clientes

Definición y importancia de la segmentación de clientes

La segmentación de clientes es una estrategia de marketing que consiste en dividir una base de clientes en grupos homogéneos con características y comportamientos similares (Kotler & Keller, 2016). Este proceso permite a las empresas diseñar estrategias personalizadas para cada segmento, optimizando recursos y mejorando la experiencia del cliente.

Según Kotler y Keller, la segmentación de clientes es un componente esencial del marketing estratégico, ya que facilita la identificación de las necesidades y preferencias específicas de los consumidores (Kotler & Keller, 2016). En un entorno empresarial competitivo, comprender los segmentos de clientes ayuda a maximizar la fidelización y rentabilidad (Solomon et al., 2012).

Métodos tradicionales de segmentación

Históricamente, la segmentación de clientes se ha basado en enfoques tradicionales que agrupan a los consumidores según criterios predefinidos. Entre los métodos más comunes se encuentran:

- **Segmentación demográfica**: Basada en variables como edad, género, nivel de ingresos y estado civil (Schiffman et al., 2019).

- **Segmentación geográfica**: Divide a los clientes según su ubicación geográfica, permitiendo estrategias de marketing locales o regionales (Kotler & Keller, 2016).
- **Segmentación psicográfica**: Considera factores como personalidad, valores, intereses y estilos de vida para identificar segmentos con preferencias similares (Solomon et al., 2012).
- **Segmentación conductual**: Agrupa a los clientes en función de sus hábitos de compra, frecuencia de consumo y nivel de lealtad a la marca (Schiffman et al., 2019).

A continuación, se presenta una comparación de estos métodos en la Tabla 1.

Tabla 1

Comparación de métodos tradicionales de segmentación

Método	Descripción	Ejemplo de aplicación
Demográfica	Basada en características como edad, género, ingresos y educación.	Empresas de productos cosméticos segmentan por género y edad.
Geográfica	Segmentación según ubicación geográfica.	Restaurantes adaptan su menú según la región.
Psicográfica	Considera estilos de vida, valores y personalidad.	Marcas de lujo orientadas a consumidores con alto estatus.
Conductual	Basada en patrones de consumo y comportamiento.	Programas de lealtad para clientes frecuentes.

Beneficios de la segmentación basada en datos

El avance de la tecnología ha permitido la evolución de la segmentación de clientes desde métodos tradicionales hacia enfoques basados en datos y algoritmos de aprendizaje automático. La segmentación basada en datos ofrece múltiples ventajas:

- **Mayor precisión**: El análisis de grandes volúmenes de datos permite identificar patrones más precisos en los clientes (Davenport & Dyché, 2013).
- **Personalización efectiva**: Facilita la creación de estrategias de marketing individualizadas, mejorando la experiencia del cliente (Kotler & Keller, 2016).

- **Optimización de recursos**: Permite enfocar los esfuerzos en los segmentos con mayor rentabilidad potencial (Solomon et al., 2012).
- **Segmentación dinámica**: A diferencia de los métodos tradicionales, la segmentación basada en datos se actualiza en tiempo real según el comportamiento del consumidor (Davenport & Dyché, 2013).

En la actualidad, las empresas utilizan técnicas de minería de datos e inteligencia artificial para mejorar la segmentación de clientes, logrando una ventaja competitiva en mercados cada vez más dinámicos.

2.2.2 Inteligencia artificial y segmentación

Inteligencia Artificial

La inteligencia artificial (IA) se refiere a la simulación de procesos de inteligencia humana por parte de sistemas computacionales. Estos procesos incluyen el aprendizaje (adquisición de datos y reglas para el uso de los mismos), el razonamiento (utilizar reglas para llegar a conclusiones) y la autocorrección (Russell & Norvig, 2009). La IA se ha aplicado en una amplia variedad de dominios, desde el reconocimiento de imágenes hasta la automatización de procesos industriales.

Introducción al aprendizaje automático en segmentación

El aprendizaje automático ha revolucionado el proceso de segmentación de clientes al permitir el análisis de grandes volúmenes de datos y la identificación de patrones complejos en el comportamiento del consumidor (Hastie et al., 2017). A diferencia de los métodos tradicionales, que dependen de reglas predefinidas, los algoritmos de aprendizaje automático pueden adaptarse y evolucionar en función de los datos disponibles (Bishop & Nasrabadi, 2006).

La segmentación basada en aprendizaje automático es ampliamente utilizada en diversos sectores, como el comercio electrónico, las telecomunicaciones y la banca, donde

permite a las empresas ofrecer experiencias personalizadas y mejorar la retención de clientes (Murphy, 2012). Existen dos enfoques principales para aplicar el aprendizaje automático en la segmentación: los algoritmos supervisados y los algoritmos no supervisados.

Enfoques supervisados vs. no supervisados

El aprendizaje automático en segmentación de clientes puede dividirse en dos enfoques principales: el aprendizaje supervisado y el aprendizaje no supervisado. A continuación, en la Tabla 2, se presentan las diferencias clave entre ambos enfoques.

Tabla 2

Comparación entre aprendizaje supervisado y no supervisado en segmentación de clientes

Característica	Aprendizaje supervisado	Aprendizaje no supervisado
Definición	Utiliza datos etiquetados para entrenar modelos de clasificación o regresión.	No requiere etiquetas; busca patrones en los datos de manera autónoma.
Ejemplos de algoritmos	Árboles de decisión, Random Forest, redes neuronales y SVM.	K-Means, DBSCAN y algoritmos de clustering jerárquico.
Aplicaciones en segmentación	Predicción de abandono de clientes, clasificación de clientes según valor.	Agrupación de clientes con patrones de compra similares.
Ventajas	Alta precisión en la clasificación si se dispone de datos etiquetados de calidad.	Identificación de patrones desconocidos y adaptabilidad a nuevas estructuras de datos.
Desafíos	Requiere una cantidad significativa de datos etiquetados.	Puede generar clústeres difíciles de interpretar.

Nota. Elaboración propia.

El aprendizaje supervisado es útil cuando se cuenta con datos históricos etiquetados, permitiendo la clasificación de clientes en segmentos predefinidos (Bengio et al., 2017). Por otro lado, el aprendizaje no supervisado permite la exploración de nuevos segmentos sin la necesidad de etiquetas previas, siendo ideal para la identificación de patrones

emergentes (James et al., 2013).

Uno de los algoritmos más utilizados en el aprendizaje no supervisado es **K-Means**, que agrupa clientes en clusters basados en su proximidad en el espacio de características (Tan et al., 2006). Otro algoritmo ampliamente utilizado es **DBSCAN**, el cual es capaz de identificar clusters de formas arbitrarias y manejar ruido en los datos (Ester et al., 1996). La elección del algoritmo más adecuado depende de la estructura y calidad de los datos disponibles.

2.2.3 Algoritmos de clustering en la segmentación de clientes

La segmentación de clientes basada en inteligencia artificial se basa en técnicas de clustering, las cuales permiten agrupar clientes con características similares en conjuntos homogéneos. Existen diversos algoritmos para esta tarea, siendo **K-Means** y **DBSCAN** dos de los más utilizados en la segmentación de clientes.

K-Means clustering

El algoritmo K-Means es uno de los algoritmos más populares para la segmentación de clientes. Su objetivo principal es dividir un conjunto de datos en k grupos o clusters, minimizando la variabilidad dentro de cada cluster (MacQueen et al., 1967).

Funcionamiento del algoritmo:

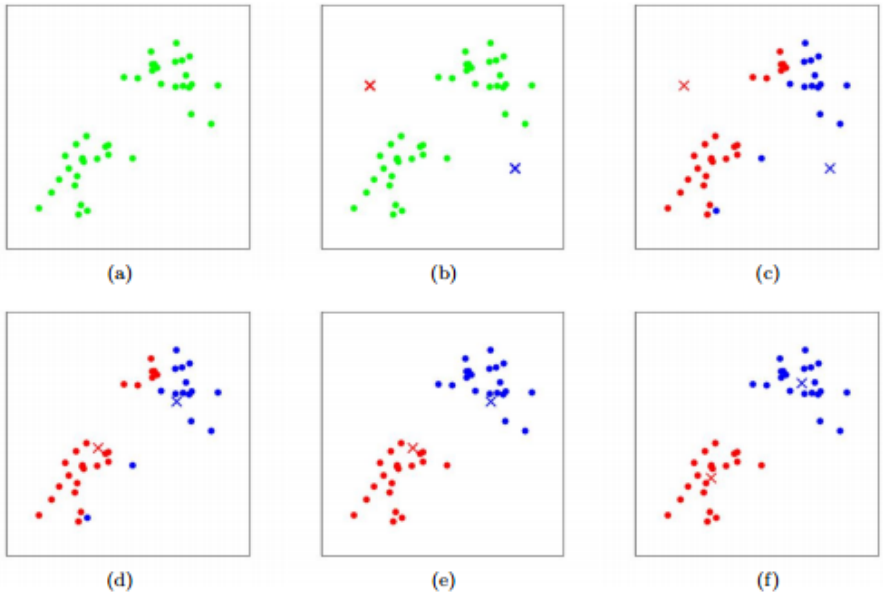
El algoritmo K-Means presentado en la Figura 1 sigue los siguientes pasos:

1. Se eligen k centroides iniciales de forma aleatoria.
2. Se asigna cada punto de datos al cluster cuyo centroide esté más cercano, utilizando la distancia euclidiana:

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{n} (x_{i,m} - c_{j,m})^2} \tag{1}$$

Figura 1

Algoritmo K-Means



donde x_i es un punto de datos, c_j es el centroide del cluster y n es la cantidad de dimensiones.

3. Se recalculan los centroides como la media de los puntos asignados a cada cluster:

$$c_j = \frac{1}{N_j} \sum_{x_i \in C_j} x_i \tag{2}$$

donde N_j es la cantidad de puntos en el cluster C_j.

4. Se repiten los pasos 2 y 3 hasta que la asignación de puntos no cambie o se alcance un criterio de convergencia.

Algoritmo en pseudocódigo:

Algorithm 1 Algoritmo K-Means

1: **Entrada:** Conjunto de datos $X = \{x_1, x_2, ..., x_n\}$, número de clusters k.
2: Inicializar k centroides $c_1, c_2, ..., c_k$ de forma aleatoria.
3: **repeat**
4: Asignar cada punto x_i al cluster C_j cuyo centroide c_j minimice la distancia euclidiana.
5: Recalcular los centroides c_j como la media de los puntos asignados a cada cluster.
6: **until** Los centroides no cambien significativamente o se alcance el número máximo de iteraciones.
7: **Salida:** Clusters $C_1, C_2, ..., C_k$.

Ventajas y desventajas:

Ventajas:

- Es eficiente y escalable para grandes volúmenes de datos.
- Fácil de implementar y comprender.

Desventajas:

- Requiere definir el número de clusters k de antemano.
- Sensible a la selección de centroides iniciales.
- No es efectivo con clusters de formas irregulares o con ruido.

Bisecting K-Means

El algoritmo Bisecting K-Means es una variante jerárquica divisiva del algoritmo K-Means que permite construir clusters mediante un proceso iterativo de partición binaria. A diferencia del K-Means tradicional, que divide directamente los datos en k clusters, Bisecting K-Means comienza con un único cluster que contiene todos los datos y lo divide sucesivamente hasta alcanzar el número deseado de clusters (Steinbach et al., 2000).

Este algoritmo combina las ventajas de los métodos jerárquicos y los métodos particionales, permitiendo obtener segmentaciones más estables que el K-Means

tradicional, especialmente cuando existen diferencias importantes en la densidad o estructura de los datos.

Funcionamiento del algoritmo:

El algoritmo Bisecting K-Means sigue un proceso iterativo de división de clusters basado en la minimización de la variabilidad interna de los grupos. El procedimiento general es el siguiente:

1. Se considera inicialmente un único cluster que contiene todos los puntos de datos.
2. Se selecciona el cluster a dividir. Generalmente se elige el cluster con mayor variabilidad interna o mayor error cuadrático.
3. Se aplica el algoritmo K-Means con $k = 2$ sobre el cluster seleccionado, generando dos subclusters.
4. La partición obtenida se evalúa mediante una medida de calidad, como la suma de errores cuadráticos dentro de los clusters.
5. Se repiten los pasos 2 a 4 hasta obtener el número deseado de clusters.

Al igual que en K-Means, la variabilidad dentro de los clusters puede medirse mediante la suma de distancias cuadráticas entre los puntos y su centroide:

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - c_j\|^2 \tag{3}$$

Donde C_j representa el conjunto de puntos pertenecientes al cluster j y c_j es su centroide.

Ventajas y desventajas:

Ventajas:

- Produce segmentaciones más estables que el K-Means tradicional.
- Reduce la sensibilidad a la selección inicial de centroides.

- Permite obtener clusters de mejor calidad en términos de cohesión interna.

Desventajas:

- Puede requerir mayor tiempo de cómputo que K-Means.
- El resultado depende del criterio utilizado para seleccionar el cluster a dividir.
- Sigue requiriendo definir el número final de clusters.

DBSCAN

El algoritmo **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering basado en densidad que permite identificar grupos de datos con formas arbitrarias y manejar ruido en el conjunto de datos (Ester et al., 1996). En la Figura 2 se encuentra la representación visual del algoritmo DBSCAN.

Conceptos clave:

DBSCAN se basa en dos parámetros principales:

- ε (radio de vecindad): Define la distancia máxima entre dos puntos para que sean considerados vecinos.
- $minPts$ (mínimo número de puntos en un cluster): Define el número mínimo de puntos requeridos para que un punto sea considerado un **punto central**. Es también conocido como número mínimo de vecinos.

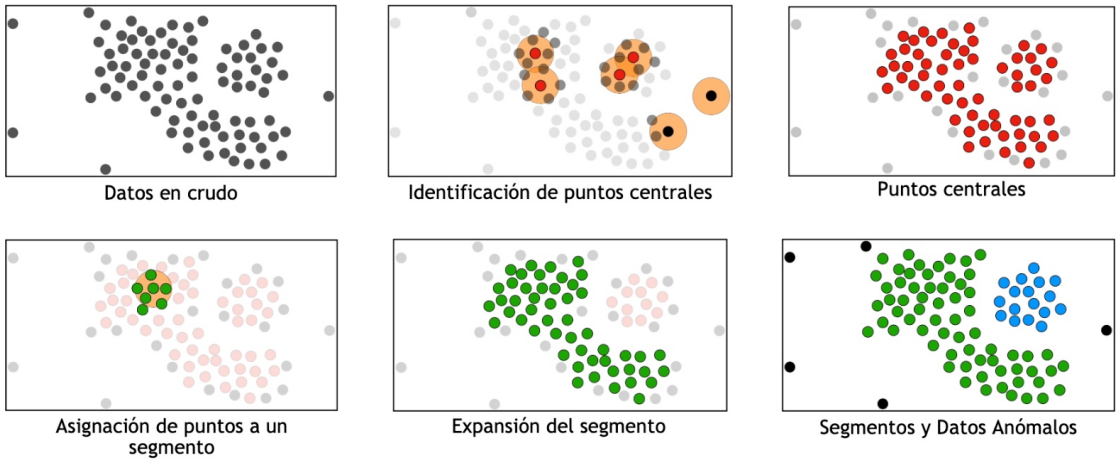
Funcionamiento del algoritmo:

El algoritmo sigue los siguientes pasos:

1. Se selecciona un punto no visitado x_i.
2. Si x_i tiene al menos $minPts$ puntos en su vecindad de radio ε, se considera un punto central y se inicia un nuevo cluster.

Figura 2

Algoritmo DBSCAN



3. Se expanden los clusters agregando todos los puntos densamente conectados.

4. Se repite el proceso hasta que todos los puntos sean clasificados como parte de un cluster o ruido.

Algoritmo en pseudocódigo:

Algorithm 2 Algoritmo DBSCAN

Entrada: Conjunto de datos X, parámetros ε y $minPts$.
Inicializar todos los puntos como no visitados.
for cada punto x_i en X **do**
    if x_i es no visitado **then**
        Marcar x_i como visitado.
        Obtener vecinos de x_i dentro de ε.
        if Número de vecinos $\geq minPts$ **then**
            Crear nuevo cluster y agregar x_i.
            Expandir cluster con puntos densamente conectados.
        else
            Marcar x_i como ruido.
        end if
    end if
end for
Salida: Conjunto de clusters y puntos de ruido.

Ventajas y desventajas:

Ventajas:

- No requiere especificar el número de clusters.
- Puede detectar clusters de formas arbitrarias.
- Maneja datos con ruido y valores atípicos.

Desventajas:

- La elección de los parámetros ε y $minPts$ puede ser difícil.
- No es eficiente en conjuntos de datos de alta dimensión.

Comparación entre K-Means y DBSCAN:

La Tabla 3 presenta una comparación entre K-Means y DBSCAN en función de diversos criterios.

Tabla 3

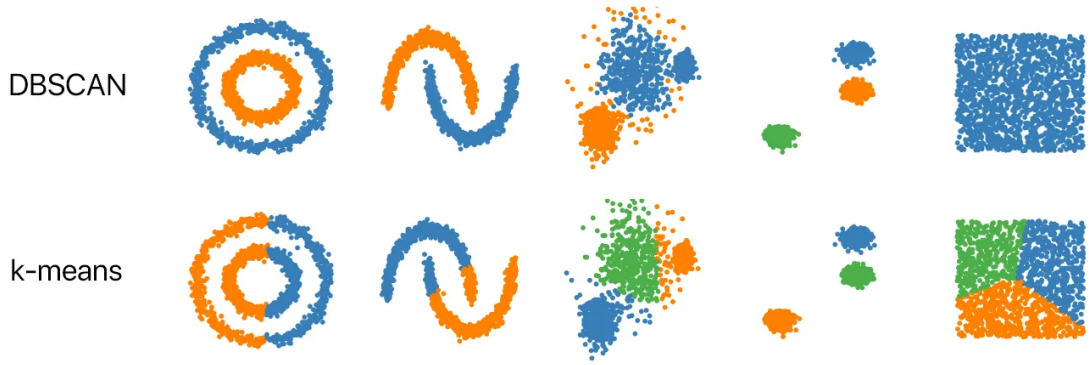
Comparación entre K-Means y DBSCAN

Criterio	**K-Means**	**DBSCAN**
Requiere definir k	Sí	No
Forma de los clusters	Esféricos o circulares	Arbitrarios
Manejo de ruido	No maneja ruido	Identifica ruido y outliers
Escalabilidad	Rápido en grandes volúmenes de datos	Menos eficiente en grandes volúmenes de datos
Sensibilidad a la inicialización	Alta (sensibilidad a los centroides iniciales)	Baja

En la Figura 3 se encuentra la comparación visual de los dos algoritmos de clusterización mencionados.

Figura 3

Comparación Kmeans y DBSCAN



2.2.4 Métricas de evaluación de clustering

Para evaluar la calidad de una segmentación en clustering es necesario utilizar métricas que midan la cohesión y separación de los grupos formados. Dado que el clustering es un problema de aprendizaje no supervisado, estas métricas no dependen de etiquetas predefinidas. Entre las métricas más utilizadas se encuentran el **Silhouette Score**, el **Índice de Davies-Bouldin** y el **Coeficiente de Dunn** (Tan et al., 2006).

Índice de Silhouette (Silhouette Score)

El *Silhouette Score* mide qué tan similar es un punto dentro de su propio cluster en comparación con otros clusters. Se define como:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4}$$

donde:

- $a(i)$ es la distancia promedio de i a los demás puntos en su mismo cluster.
- $b(i)$ es la distancia promedio de i al cluster más cercano al que no pertenece.

El valor de $S(i)$ varía entre -1 y 1:

- $S(i) \approx 1$ indica que el punto está bien agrupado.
- $S(i) \approx 0$ sugiere que el punto está en la frontera entre dos clusters.
- $S(i) \approx -1$ implica que el punto está mal asignado.

El índice de Silhouette promedio para todo el conjunto de datos se calcula como:

$$S = \frac{1}{N} \sum_{i=1}^{N} S(i) \tag{5}$$

donde N es el número total de puntos.

Índice de Davies-Bouldin (DB Index)

El *Índice de Davies-Bouldin* mide la compacidad y separación de los clusters, definiéndose como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

donde:

- k es el número de clusters.
- σ_i es la dispersión intra-cluster del cluster i.
- c_i es el centroide del cluster i.
- $d(c_i, c_j)$ es la distancia entre los centroides c_i y c_j.

Un valor bajo del índice DB indica una mejor separación y compacidad de los clusters.

Índice de Calinski-Harabasz (Calinski-Harabasz Index)

El *Índice de Calinski-Harabasz* (también conocido como *Variance Ratio Criterion*) evalúa la calidad de una partición midiendo la razón entre la dispersión *inter-cluster* y la

dispersión *intra-cluster*. Esta métrica favorece segmentaciones donde los clusters están bien separados y, al mismo tiempo, son compactos. Es una de las métricas internas más utilizadas para comparar diferentes soluciones de clustering (Caliński & Harabasz, 1974).

El índice se define como:

$$CH = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \cdot \frac{N-k}{k-1} \tag{6}$$

donde:

- N es el número total de puntos del conjunto de datos.
- k es el número de clusters.
- $\text{Tr}(B_k)$ es la traza de la matriz de dispersión entre clusters (*between-cluster dispersion*).
- $\text{Tr}(W_k)$ es la traza de la matriz de dispersión dentro de clusters (*within-cluster dispersion*).

Las matrices B_k y W_k se definen como:

$$B_k = \sum_{j=1}^{k} n_j (c_j - c)(c_j - c)^T \tag{7}$$

$$W_k = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - c_j)(x_i - c_j)^T \tag{8}$$

donde:

- n_j es el número de puntos del cluster C_j.
- c_j es el centroide del cluster C_j.
- c es el centroide global del conjunto de datos.

La interpretación del índice Calinski-Harabasz es:

- Valores altos indican una mejor calidad de clustering, debido a una mayor separación entre clusters y una mayor compacidad dentro de ellos.
- Valores bajos sugieren clusters poco diferenciados o con alta dispersión interna.

Inercia

La *Inercia* es una métrica interna utilizada para evaluar la compactación de los clusters formados en una partición. Mide la suma de las distancias cuadradas entre cada punto y el centroide o representante de su cluster, permitiendo cuantificar qué tan concentrados se encuentran los datos dentro de cada grupo. En términos generales, una menor inercia indica una mayor cohesión interna de los clusters.

La inercia se define como:

$$I = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - c_j\|^2 \tag{9}$$

donde:

- k es el número de clusters.
- C_j representa el conjunto de puntos asignados al cluster j.
- x_i es un punto perteneciente al cluster C_j.
- c_j es el centroide o representante del cluster j.

La interpretación de esta métrica es la siguiente:

- Valores bajos de inercia indican clusters más compactos y con menor dispersión interna.
- Valores altos sugieren clusters más dispersos y menos cohesionados.

Es importante considerar que la inercia tiende a disminuir a medida que aumenta el número de clusters. Por esta razón, su análisis suele complementarse con otras métricas

de evaluación interna y con criterios adicionales que permitan valorar de manera más completa la calidad de la segmentación obtenida.

2.2.5 Comparación de las métricas

En la Tabla 4, se presenta una comparación de estas métricas en términos de interpretación y valores ideales.

Tabla 4

Comparación de métricas de evaluación de clustering

Métrica	Interpretación	Valor ideal
Índice de Silhouette	Mide la cohesión interna y separación de clusters.	Cercano a 1
Índice de Davies-Bouldin	Evalúa la dispersión y separación entre clusters.	Cercano a 0
Índice de Calinski-Harabasz	Mide la relación entre la dispersión entre clusters y la dispersión dentro de clusters.	Alto
Métrica de Inercia	Mide la compactación de los clusters.	Bajo

2.2.6 Algoritmos de clasificación de clientes

Los algoritmos de clasificación constituyen una de las principales técnicas del aprendizaje supervisado en inteligencia artificial. A diferencia de los algoritmos de clustering, donde los grupos se descubren a partir de los datos sin etiquetas previas, los algoritmos de clasificación requieren un conjunto de datos previamente etiquetado para aprender reglas que permitan asignar nuevas observaciones a categorías existentes (Bishop & Nasrabadi, 2006).

En el contexto de la segmentación de clientes, los algoritmos de clasificación pueden utilizarse para construir modelos que permitan describir y predecir la pertenencia de los clientes a determinados segmentos. Estos modelos permiten obtener reglas explícitas que facilitan la interpretación de los resultados y su aplicación en la toma de decisiones empresariales.

Entre los distintos algoritmos de clasificación existentes, los árboles de decisión destacan por su capacidad de generar modelos interpretables, lo cual resulta especialmente importante en aplicaciones comerciales donde se requiere comprender las características que definen cada segmento de clientes.

Árboles de decisión

Un árbol de decisión es un modelo de clasificación supervisada que representa un conjunto de reglas jerárquicas en forma de estructura arbórea. Cada nodo interno representa una condición sobre una variable, cada rama representa el resultado de dicha condición y cada nodo hoja representa una clase o categoría asignada (Hastie et al., 2017).

El objetivo del algoritmo consiste en particionar el espacio de características de tal manera que los datos pertenecientes a cada nodo hoja sean lo más homogéneos posible respecto a la variable objetivo.

Índice de Gini:

Uno de los criterios más utilizados para la construcción de árboles de decisión es el *Índice de Gini*, el cual mide el grado de impureza de un conjunto de datos.

El índice de Gini se define como:

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{10}$$

donde:

- K es el número de clases.
- p_k es la proporción de observaciones pertenecientes a la clase k en el nodo t.

Un nodo es puro cuando todas las observaciones pertenecen a una misma clase, en cuyo caso el índice de Gini toma el valor:

$$G(t) = 0 \tag{11}$$

Por el contrario, el índice de Gini alcanza valores mayores cuando existe mezcla de clases dentro del nodo. El algoritmo selecciona en cada paso la partición que produce la mayor reducción de impureza.

La reducción de impureza se calcula como:

$$\Delta G = G(t) - \left(\frac{N_L}{N} G(t_L) + \frac{N_R}{N} G(t_R) \right) \tag{12}$$

donde:

- t es el nodo original.
- t_L y t_R son los nodos hijos.
- N es el número de observaciones en el nodo original.
- N_L y N_R son las observaciones en los nodos hijos.

El algoritmo selecciona la partición que maximiza el valor de ΔG.

Profundidad Máxima del Árbol (max_depth):

La profundidad máxima del árbol, denotada como `max_depth`, representa el número máximo de niveles que puede alcanzar el árbol de decisión.

Si se denota como d la profundidad del árbol, entonces:

$$d \leq d_{max} \tag{13}$$

donde d_{max} corresponde al parámetro `max_depth`.

Este parámetro controla la complejidad del modelo. Valores altos permiten árboles más complejos capaces de representar patrones detallados en los datos, mientras que valores bajos producen modelos más simples y fáciles de interpretar.

Un árbol excesivamente profundo puede ajustarse demasiado a los datos de entrenamiento, fenómeno conocido como sobreajuste.

Número mínimo de muestras para división (min_samples_split):

El parámetro `min_samples_split` define el número mínimo de observaciones que debe contener un nodo para que el algoritmo permita realizar una división adicional.

Si N_t representa el número de observaciones en el nodo t, la condición para dividir el nodo es:

$$N_t \geq N_{min} \tag{14}$$

donde N_{min} corresponde al parámetro `min_samples_split`.

Este parámetro evita la generación de nodos con pocas observaciones, lo que contribuye a mejorar la generalización del modelo y a reducir la complejidad del árbol.

Árboles de decisión para la interpretación de segmentaciones

Los algoritmos de clustering generan segmentaciones de datos asignando a cada observación una etiqueta de cluster. Sin embargo, estos algoritmos no producen directamente un modelo explícito que permita describir las características que definen cada segmento.

En general, los algoritmos de clustering pueden considerarse modelos basados en instancias, ya que asignan etiquetas a los datos en función de relaciones de proximidad sin producir reglas explícitas que describan los segmentos obtenidos.

En contraste, los árboles de decisión constituyen modelos predictivos explícitos que permiten representar la pertenencia a los segmentos mediante reglas de clasificación basadas en las variables del conjunto de datos.

En este enfoque, las etiquetas de cluster generadas por los algoritmos de clustering se utilizan como variable objetivo en un problema de clasificación supervisada. De esta manera, el árbol de decisión aprende una función:

$$f(x) \rightarrow c \tag{15}$$

donde:

- x representa el vector de características de un cliente.
- c representa la etiqueta de cluster asignada.

El modelo resultante permite expresar los segmentos mediante reglas del tipo:

$$\text{Si } x_1 \leq a \text{ y } x_2 > b \Rightarrow \text{Cluster } j \tag{16}$$

Este enfoque permite transformar una segmentación obtenida mediante aprendizaje no supervisado en un conjunto de reglas interpretables, facilitando la comprensión de los segmentos y su aplicación en la toma de decisiones empresariales.

De esta manera, el árbol de decisión actúa como un modelo interpretativo de la segmentación, permitiendo describir las características que distinguen a cada grupo de clientes.

2.2.7 Aplicaciones de la inteligencia artificial en el comercio y el marketing

La inteligencia artificial (IA) ha revolucionado la forma en que las empresas interactúan con sus clientes, permitiendo estrategias más precisas y eficientes en el ámbito comercial. En particular, las técnicas de segmentación basadas en IA han demostrado ser esenciales en la industria minorista y en el diseño de estrategias de fidelización y personalización (Davenport & Dyché, 2013).

Segmentación de clientes en la industria minorista

La industria minorista ha sido una de las principales beneficiarias del uso de inteligencia artificial para la segmentación de clientes. Tradicionalmente, los minoristas agrupaban a los consumidores en segmentos amplios basados en características demográficas generales. Sin embargo, con la capacidad de procesamiento de datos de

la IA, ahora es posible analizar comportamientos de compra, preferencias y patrones de consumo con un alto nivel de precisión (Michael Chui, 2018).

La Tabla 5 resume algunas de las principales aplicaciones de la IA en la industria minorista.

Tabla 5

Aplicaciones de la IA en la industria minorista

Aplicación	Beneficio
Segmentación de clientes	Permite diseñar estrategias de marketing personalizadas.
Recomendaciones de productos	Aumenta las ventas mediante la personalización de sugerencias.
Gestión de inventario predictivo	Optimiza la disponibilidad de productos según la demanda.
Precios dinámicos	Ajusta los precios en tiempo real según la demanda y la competencia.
Chatbots y asistentes virtuales	Mejoran la experiencia del cliente y reducen costos operativos.

Uso de clustering en estrategias de fidelización y personalización

El clustering es una de las técnicas de inteligencia artificial más utilizadas para desarrollar estrategias de fidelización y personalización. La capacidad de agrupar a los clientes según sus comportamientos permite a las empresas ofrecer experiencias altamente adaptadas, aumentando la lealtad del consumidor y la rentabilidad a largo plazo (Seidor, 2023).

Algunas estrategias clave que se benefician del clustering incluyen:

- **Programas de fidelización adaptativos**: Se ajustan dinámicamente según el comportamiento del cliente, premiando la recurrencia y el valor de compra.
- **Recomendaciones de productos personalizadas**: Utilizan modelos de clustering para sugerir productos basados en clientes con patrones de compra similares.
- **Optimización del servicio al cliente**: Segmenta a los clientes según su nivel de

satisfacción, permitiendo intervenciones proactivas para mejorar la experiencia (Slupu, 2023).

- **Campañas de marketing dirigidas**: Agrupa a los clientes según preferencias y hábitos de consumo, aumentando la efectividad de las campañas publicitarias.

En la actualidad, el uso de técnicas como K-Means y DBSCAN ha permitido a las empresas refinar sus estrategias de fidelización, generando mayor interacción con los clientes y maximizando la retención. Estas herramientas han demostrado ser particularmente útiles en sectores como el comercio electrónico, donde la personalización juega un papel crucial en la conversión de clientes y la reducción de tasas de abandono (Peña Escobar et al., 2015).

2.3 Definición de términos

Segmentación de clientes

La segmentación de clientes es el proceso de dividir una base de clientes en grupos homogéneos con características similares, tales como comportamiento de compra, necesidades o perfil demográfico, con el fin de diseñar estrategias de marketing más efectivas (Buleje Calderón, 2022).

Clustering

El clustering es una técnica de aprendizaje no supervisado que consiste en agrupar datos en conjuntos (clusters) de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos (Jain et al., 1999).

Cluster

Un cluster es un conjunto de objetos o datos que presentan alta similitud entre sí y baja similitud con elementos de otros grupos (Jain et al., 1999).

Inteligencia artificial

La inteligencia artificial es una disciplina de la informática que busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, razonamiento y toma de decisiones (Russell & Norvig, 2009).

Aprendizaje automático

El aprendizaje automático es un subcampo de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y mejorar su desempeño sin ser programados explícitamente (Murphy, 2012).

Aprendizaje supervisado

El aprendizaje supervisado es un tipo de aprendizaje automático en el que el modelo se entrena con datos etiquetados, es decir, con entradas y salidas conocidas (James et al., 2013).

Aprendizaje no supervisado

El aprendizaje no supervisado es un tipo de aprendizaje automático donde el modelo identifica patrones en datos sin etiquetas, siendo el clustering una de sus principales técnicas (Hastie et al., 2017).

Centroide

El centroide es el punto representativo de un cluster, generalmente calculado como el promedio de todos los puntos que pertenecen a dicho grupo (MacQueen et al., 1967).

Distancia euclidiana

La distancia euclidiana es una medida de similitud que calcula la distancia directa entre dos puntos en un espacio multidimensional (Tan et al., 2006).

Inercia

La inercia es una medida utilizada en algoritmos como K-means que representa la suma de las distancias cuadradas entre los puntos de un cluster y su centroide (James et al., 2013).

Ruido (en clustering)

El ruido en clustering se refiere a aquellos datos que no pertenecen claramente a ningún cluster o que son considerados irrelevantes dentro del análisis (Ester et al., 1996).

Outliers (valores atípicos)

Los outliers son observaciones que se alejan significativamente del resto de los datos y pueden afectar el rendimiento de los algoritmos de clustering (Tan et al., 2006).

Densidad (en DBSCAN)

La densidad en DBSCAN se refiere a la cantidad de puntos dentro de una región determinada, utilizada para identificar clusters basados en regiones densamente pobladas (Ester et al., 1996).

Radio de vecindad (ϵ)

El radio de vecindad (ϵ) es un parámetro de DBSCAN que define la distancia máxima para considerar que dos puntos son vecinos (Ester et al., 1996).

Número mínimo de vecinos ($MinPts$)

MinPts es un parámetro en DBSCAN que indica el número mínimo de puntos necesarios dentro de un radio ϵ para formar un cluster (Ester et al., 1996).

Fidelización de clientes

La fidelización de clientes consiste en estrategias orientadas a mantener relaciones duraderas con los clientes, incentivando su lealtad y repetición de compra (Peña Escobar et al., 2015).

Personalización

La personalización es la adaptación de productos, servicios o mensajes a las preferencias individuales de cada cliente con el fin de mejorar su experiencia (Gartner, 2022).

Estrategias de marketing

Las estrategias de marketing son planes diseñados para alcanzar objetivos comerciales mediante el análisis del mercado, segmentación y posicionamiento (Kotler & Keller, 2016).

Comportamiento del cliente

El comportamiento del cliente estudia cómo los individuos toman decisiones de compra, influenciados por factores psicológicos, sociales y culturales (Solomon et al., 2012).

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1 TIPO Y DISEÑO DE LA INVESTIGACIÓN

La investigación cumple dos propósitos fundamentales: entre ellos resolver problemas (investigación aplicada) (Hernández-Sampieri & Mendoza, 2018). La investigación del presente proyecto busca resolver el problema de la segmentación de clientes en una MYPE.

Las investigaciones experimentales son aquellas que administran estímulos o tratamientos y/o intervenciones y se caracterizan por la:

- Manipulación intencional de variables (variable 1).
- Medición de variables (variable 2).
- Control y validez.
- Dos o más grupos de comparación.
- Participantes asignados al azar o emparejados.

En la presente investigación se manipula la variable independiente que son los algoritmos de inteligencia artificial y se mide de estos su capacidad de segmentación o clusterización de clientes (medición de variables). El ajuste de los algoritmos de inteligencia artificial se lleva a cabo usando el conjunto de datos de los clientes. A su vez, se compararán los resultados de ambos algoritmos y se encontrará la relación entre los clusters o segmentos y las características de los datos.

Finalmente, se buscará la relación entre los segmentos encontrados por los algoritmos de inteligencia artificial y las características de la muestra.

Adicionalmente, el presente estudio es congruente con el enfoque metodológico de Design Science Research (DSR) o investigación basada en diseño. Este enfoque se

caracteriza por el desarrollo y evaluación de artefactos orientados a resolver problemas reales dentro de un contexto organizacional específico. En este caso, los algoritmos de inteligencia artificial utilizados para la segmentación de clientes constituyen artefactos diseñados y evaluados en el contexto de una MYPE, con el propósito de mejorar los procesos de toma de decisiones comerciales.

De acuerdo con Wieringa (2014), la investigación basada en diseño implica el diseño y la investigación empírica de artefactos en su contexto, donde los artefactos pueden ser métodos, técnicas o algoritmos destinados a mejorar una situación problemática.

En el presente estudio, los algoritmos de clusterización constituyen artefactos cuyo desempeño es evaluado mediante experimentación con datos reales de clientes.

Asimismo, la investigación cumple con los principios de Design Science Research al buscar simultáneamente producir conocimiento científico y resolver un problema práctico, mediante la construcción y evaluación de soluciones tecnológicas. Según Dresch et al. (2014), la investigación basada en diseño se orienta al desarrollo de construcciones innovadoras que solucionen problemas reales, siendo el resultado principal un artefacto evaluado en términos de utilidad o valor para la organización.

Por lo tanto, la presente investigación puede caracterizarse como investigación aplicada de tipo experimental con enfoque Design Science Research, debido a que se diseñan y evalúan algoritmos de inteligencia artificial como soluciones tecnológicas para la segmentación de clientes en un contexto organizacional real.

3.2 POBLACIÓN Y MUESTRA DE ESTUDIO

3.2.1 Población

Los clientes registrados de la empresa D&C Comunicaciones durante 2024.

3.2.2 Muestra

Se usará una muestra censal de los clientes registrados de la empresa D&C Comunicaciones del año 2024 que consiste en un total de 400 clientes de los cuales se cuenta con la siguiente información:

- Id del cliente
- Género
- Edad
- Monto de operaciones
- Número de visitas

3.3 Acciones y actividades para la ejecución del proyecto

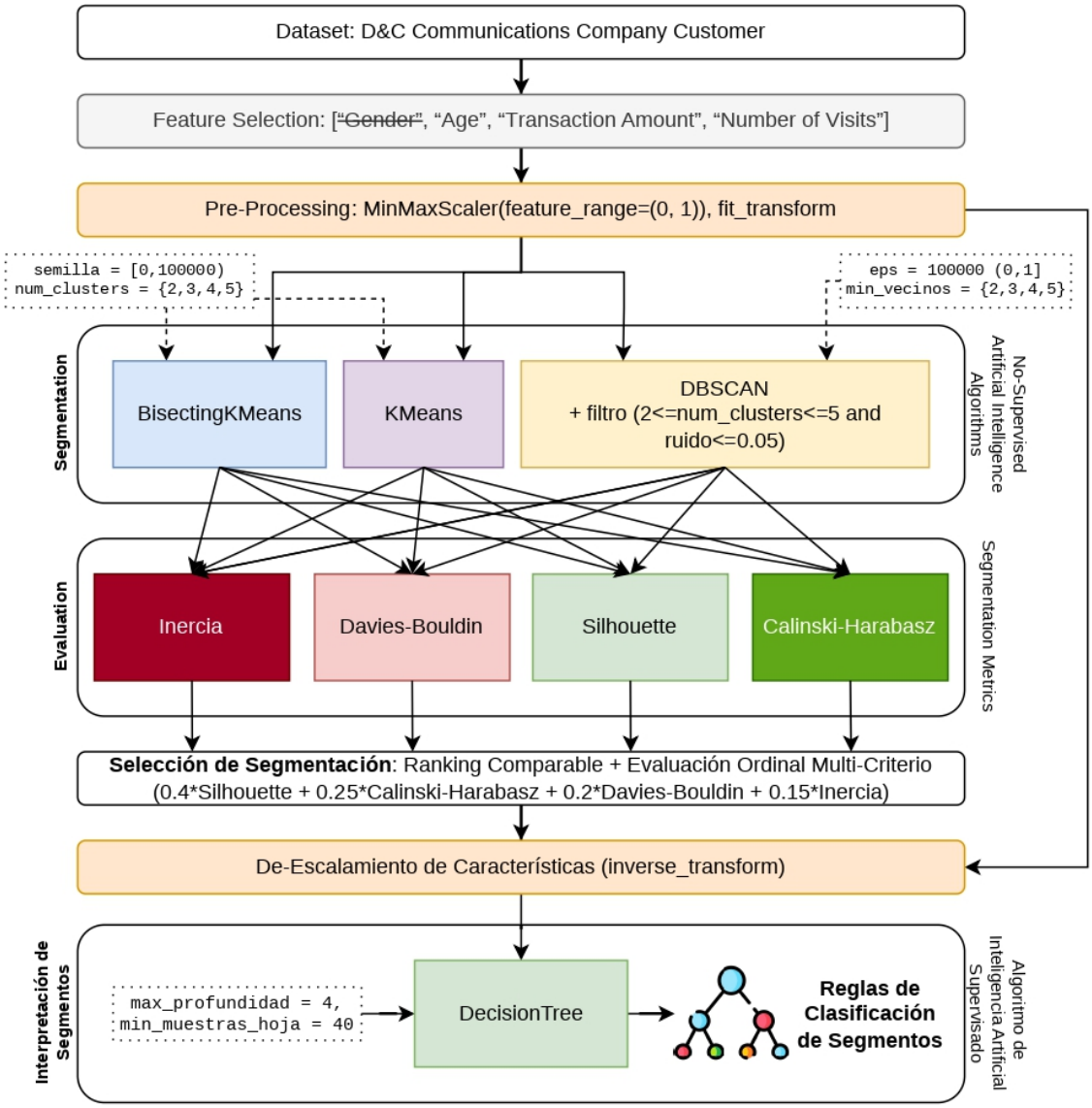
El gráfico presentado en la Figura 4 correspondiente describe de manera integral el flujo metodológico seguido para la ejecución del proyecto de segmentación de clientes en la empresa D&C Comunicaciones.

El proceso inicia con el conjunto de datos de clientes proporcionado por la empresa, el cual constituye la base empírica del estudio. A partir de este conjunto, se realiza la selección de las características relevantes para el análisis, específicamente las variables género, edad, monto de operaciones y número de visitas. Esta etapa tiene como finalidad estructurar el espacio de características que será utilizado por los algoritmos de inteligencia artificial.

Posteriormente, se lleva a cabo el escalamiento de las variables mediante la técnica MinMaxScaler con un rango normalizado entre 0 y 1, garantizando que todas las características contribuyan de manera proporcional en el cálculo de distancias y evitando que variables con mayor magnitud dominen el proceso de segmentación. Esta etapa de preprocesamiento es fundamental para asegurar la estabilidad y coherencia de los resultados obtenidos en los algoritmos de clustering.

Figura 4

Flujo de Trabajo



El núcleo del proceso corresponde a la aplicación de algoritmos de inteligencia artificial no supervisados, específicamente Bisecting K-Means, K-Means y DBSCAN. En el caso de DBSCAN, se incorpora un filtro adicional que restringe el número de clusters entre dos y cinco y controla el porcentaje de ruido a un máximo de 5 %, con el propósito de asegurar la viabilidad práctica de las segmentaciones generadas. Cada algoritmo es ejecutado bajo distintas configuraciones de parámetros, tales como número de clusters, semillas de inicialización y valores de vecindad, permitiendo explorar diferentes

estructuras de segmentación dentro del conjunto de datos.

Una vez generadas las segmentaciones, se procede a su evaluación mediante métricas internas de clustering, específicamente Inercia, Índice de Davies-Bouldin, Silhouette e Índice de Calinski-Harabasz. Estas métricas permiten analizar la cohesión interna y la separación entre clusters desde distintas perspectivas cuantitativas. El gráfico muestra cómo cada algoritmo es evaluado bajo cada métrica.

A continuación, se implementa un mecanismo de selección basado en un ranking comparable y una evaluación ordinal multicriterio, donde las métricas son ponderadas de acuerdo con su relevancia relativa. La función de agregación utilizada combina los valores de Silhouette, Calinski-Harabasz, Davies-Bouldin e Inercia, permitiendo determinar de manera sistemática la segmentación más adecuada entre todas las alternativas evaluadas. Este procedimiento garantiza que la selección final no dependa de una única métrica, sino de un análisis integral de la calidad del clustering.

Una vez seleccionada la segmentación óptima, se realiza el proceso de desescalamiento de las características mediante la transformación inversa, con el objetivo de recuperar las unidades originales de las variables y facilitar la interpretación empresarial de los resultados. Finalmente, el proceso culmina con la aplicación de un algoritmo de inteligencia artificial supervisado, específicamente un Árbol de Decisión, configurado con parámetros como profundidad máxima y número mínimo de muestras por hoja. Este modelo utiliza las etiquetas de cluster como variable objetivo y genera reglas de clasificación que permiten describir de forma explícita las características que definen cada segmento de clientes.

En conjunto, el gráfico sintetiza las acciones y actividades desarrolladas durante la ejecución del proyecto, mostrando la secuencia lógica que va desde la preparación de los datos hasta la obtención de reglas interpretables de segmentación, integrando técnicas de aprendizaje no supervisado y supervisado dentro de un mismo marco metodológico.

3.4 MATERIALES E INSTRUMENTOS

La presente propuesta empleará las siguientes técnicas e instrumento:

- **Técnicas:** Implementación de Algoritmos de Inteligencia Artificial (BisectingKMeans, KMeans, DBSCAN, Decision Tree).
- **Instrumento:** Computación de métricas: Inercia, Silhouette, Davies-Bouldin, Calinski-Harabasz (Segmentación) y Accuracy (Clasificación).

3.5 TRATAMIENTO DE DATOS

3.5.1 Análisis y procesamiento de datos:

La información se procesará con el uso de los paquetes Pandas, Seaborn, Scipy, Numpy, Sci-Kit Learn en Python.

Se comparará de manera descriptiva según las acciones descritas en la Sección 3.3.

CAPÍTULO IV

RESULTADOS DE LA INVESTIGACIÓN

4.1 Describir las características de los clientes de la empresa D&C Comunicaciones

4.1.1 Estadística descriptiva

Para caracterizar a los clientes de la empresa, se trabajó con un conjunto de datos conformado por 400 registros y cuatro variables: género, edad, monto de operaciones y número de visitas.

Tabla 6

Vista previa del DataFrame (df.head) — Shape: (400, 4)

	Genero	Edad	Monto operaciones	Número de visitas
0	Hombre	19	15000	117
1	Hombre	21	15000	243
2	Mujer	20	16000	18
3	Mujer	23	16000	231

La vista previa del DataFrame (Tabla 6) confirma que se trata de una base compacta, con una variable categórica y tres variables numéricas, adecuada para describir rasgos generales de los clientes y su comportamiento.

Tabla 7

Características generales del DataFrame (df.info).

Característica	Valor
Número de registros (RangeIndex)	400
Índice	0 a 399
Número de columnas	4
Valores no nulos totales	1600

En cuanto a la estructura y calidad de los datos, la información del DataFrame (Tabla 7) muestra que el dataset cuenta con 400 filas (índice 0 a 399) y 4 columnas, acumulando 1600 valores no nulos. Este resultado se complementa con el resumen de completitud

Tabla 8

Resumen de completitud de datos por variable.

Variable	Total	No nulos	Nulos	% Nulos
Genero	400	400	0	0,00
Edad	400	400	0	0,00
Monto Operaciones	400	400	0	0,00
Número de Visitas	400	400	0	0,00

(Tabla 8), donde se observa que ninguna variable presenta valores perdidos (0 nulos y 0,00 % en todos los casos). En términos prácticos, esto implica que la descripción estadística y el análisis posterior pueden realizarse sin necesidad de imputación o eliminación de registros por faltantes.

Tabla 9

Estadísticos descriptivos de las variables numéricas (df.describe).

	Edad	Monto operaciones	Número de visitas
count	400,000000	400,000000	400,000000
mean	38,865000	60530,000000	150,492500
std	13,939784	26219,033053	77,328644
min	17,000000	15000,000000	1,000000
25 %	28,000000	41750,000000	102,000000
50 %	36,000000	61500,000000	150,000000
75 %	49,000000	78000,000000	219,000000
max	71,000000	137000,000000	297,000000

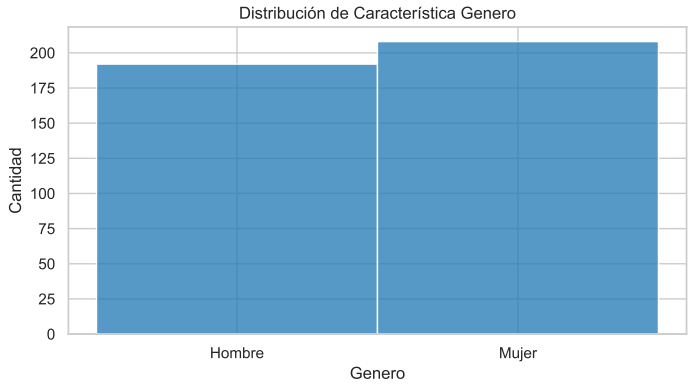
En relación con los estadísticos descriptivos de las variables numéricas (Tabla 9), se observa que la edad presenta un promedio de 38,865 años y una desviación estándar de 13,94, con un rango que va de 17 a 71 años; la mediana (36 años) y los cuartiles muestran que el 50 % de los clientes se concentra entre 28 y 49 años, lo que sugiere una mayor presencia de adultos jóvenes y adultos dentro de la muestra. En cuanto al monto de operaciones, el promedio es S/. 60 530 y la desviación estándar S/ 26 219,03, con valores entre S/ 15 000 y S/ 137 000; además, la mediana (S/ 61 500) y el intervalo intercuartílico (S/. 41 750 a S/. 78 000) evidencian diferencias marcadas entre clientes, con un grupo importante ubicado en montos intermedios y otros con montos considerablemente más altos. Por último, el Número de Visitas registra un promedio de 150,49 y una desviación

estándar de 77,33, con un mínimo de 1 y un máximo de 297; la mediana (150) y los cuartiles indican que la mitad de los clientes se encuentra entre 102 y 219 visitas, lo cual refleja variabilidad en el nivel de interacción con la empresa y la coexistencia de clientes con baja, media y alta frecuencia de visitas.

4.1.2 Análisis univariado

Figura 5

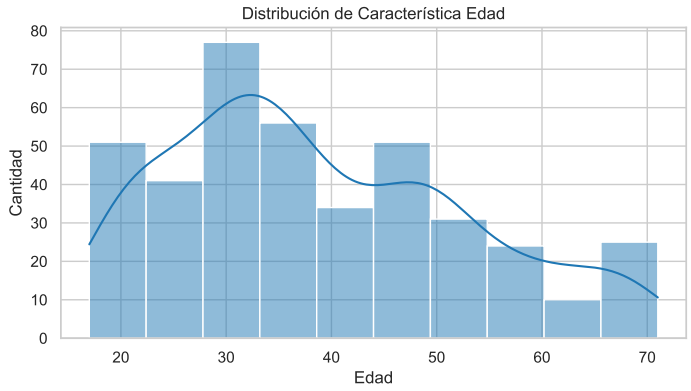
Distribución de la variable Género.



En la Figura 5 se observa el histograma de la variable Género, donde se registran 192 clientes de género hombre y 208 clientes de género mujer (sobre un total de 400). En términos porcentuales, ello equivale aproximadamente a 48 % hombres y 52 % mujeres, evidenciando una distribución bastante equilibrada, aunque con ligera predominancia de la categoría mujer dentro de la muestra analizada.

Figura 6

Distribución de la variable Edad.

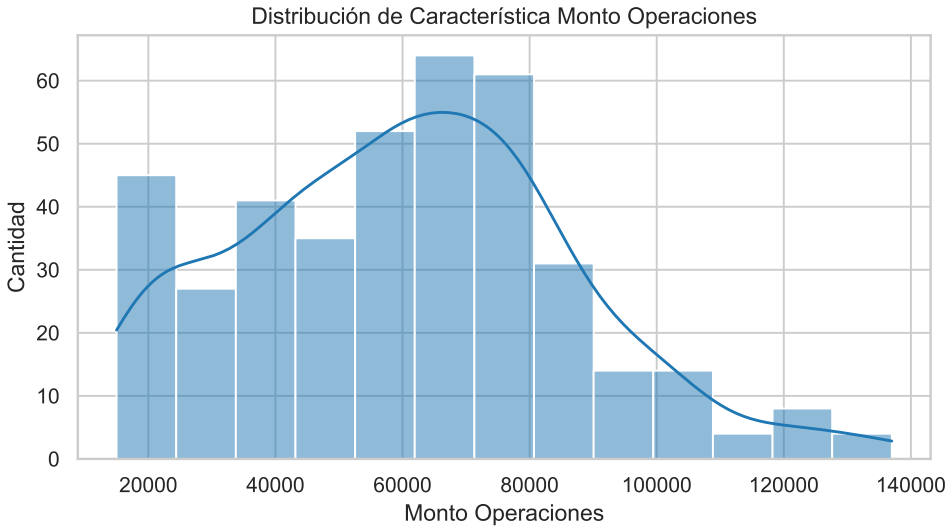


En la Figura 6 se presenta el histograma de la variable edad, acompañado por una curva de densidad (KDE), lo que permite apreciar la forma general de la distribución. Visualmente, se observa una mayor concentración de clientes en edades adultas, con un pico principal alrededor de la década de los 30 y una disminución gradual de la frecuencia a medida que la edad aumenta, manteniéndose registros hasta edades superiores; en conjunto, la figura evidencia una distribución amplia y no limitada a un solo tramo etario.

Esta apreciación se complementa con la Tabla 9 (estadísticos descriptivos), donde se reporta que la edad varía entre un mínimo de 17 y un máximo de 71 años, con un promedio de 38,865 y una mediana de 36 años. Además, los cuartiles indican que el 50 % de los clientes se ubica entre 28 (Q1) y 49 (Q3) años, lo cual coincide con la concentración observada en el histograma en torno a rangos de edad intermedios.

Figura 7

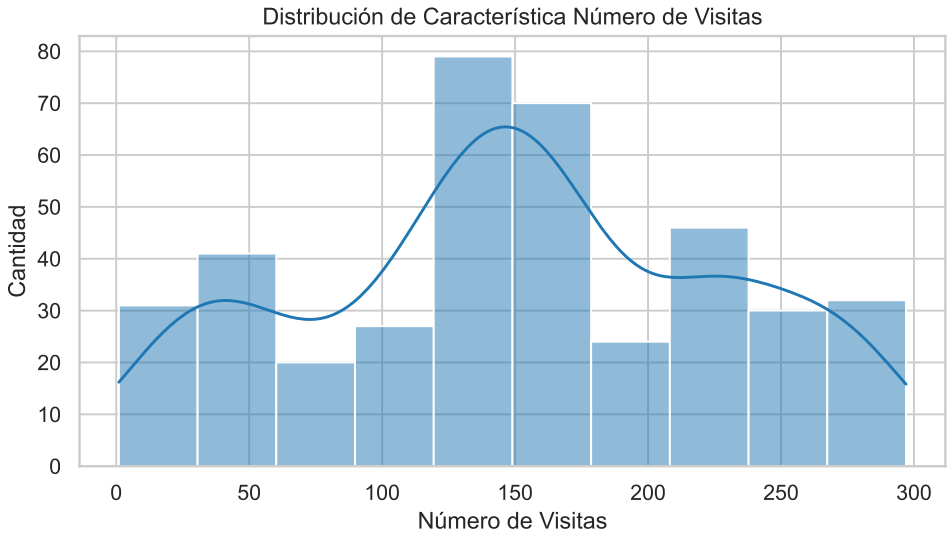
Distribución de la variable Monto de Operaciones.



En la Figura 7 se presenta el histograma del monto de operaciones, acompañado de la curva de densidad (KDE), lo que permite apreciar tanto la frecuencia por rangos como la tendencia general de la distribución. De acuerdo con los resultados de la Tabla 9, los montos registrados se encuentran entre S/ 15 000 y S/ 137 000, con un promedio de S/ 60 530 y una mediana de S/ 61 500; además, el 50 % de los clientes se concentra entre S/ 41 750 y S/ 78 000 (cuartiles 25 % y 75 %). Visualmente, el gráfico muestra una mayor acumulación de casos en rangos intermedios, especialmente alrededor de los S/ 60 000 a S/ 80 000, mientras que hacia los valores más altos se observa una disminución progresiva de la frecuencia, formando una cola hacia la derecha. En conjunto, esta distribución evidencia una variabilidad importante entre clientes en términos del monto de operaciones, con presencia de registros de montos elevados que amplían el rango total de la variable.

Figura 8

Distribución de la variable Número de Visitas



En la Figura 8 se presenta la distribución de la variable número de visitas mediante un histograma acompañado de la curva de densidad (KDE), lo que permite observar tanto la frecuencia de los datos en distintos intervalos como la forma general de la distribución. Visualmente, se aprecia que la mayor concentración de observaciones se ubica en rangos intermedios, particularmente alrededor de 120 a 170 visitas, donde se alcanza la mayor frecuencia.

Asimismo, la curva de densidad sugiere que la distribución no es completamente uniforme, sino que presenta variaciones a lo largo de su recorrido, con una concentración principal en valores medios y una dispersión hacia valores más bajos y más altos. En los extremos, es decir, cerca de 0 visitas y hacia valores próximos a 300 visitas, la frecuencia es menor en comparación con la zona central.

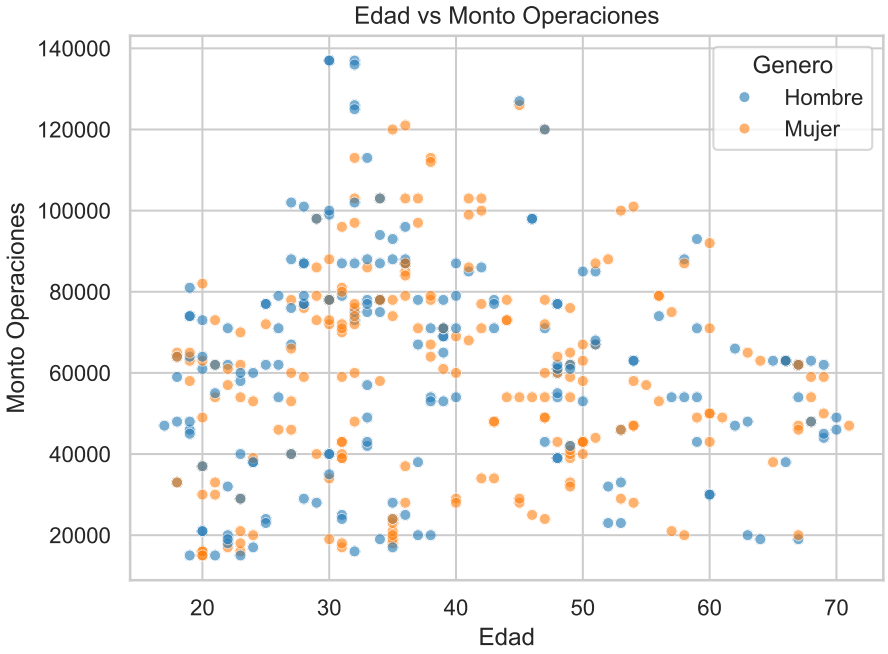
En conjunto, este comportamiento evidencia que la variable número de visitas presenta una dispersión apreciable entre los clientes, con predominio de registros en niveles intermedios y menor presencia de casos extremos. Esto sugiere una heterogeneidad en el comportamiento de visitas, donde la mayoría de los clientes se concentra en valores moderados, mientras que un grupo más reducido registra niveles muy bajos o

relativamente altos.

4.1.3 Análisis bivariado

Figura 9

Bivariado Edad vs Monto de Operaciones



En la Figura 9 se presenta un diagrama de dispersión entre las variables edad y monto de operaciones, incorporando además la variable género como criterio de diferenciación visual. Este tipo de representación gráfica es apropiado para el análisis bivariado de variables cuantitativas, ya que permite identificar la forma de distribución conjunta de los datos, la posible existencia de patrones de asociación, zonas de mayor concentración de observaciones, dispersión, solapamiento entre grupos y presencia de valores extremos.

En el eje horizontal se ubica la variable edad, con registros que abarcan aproximadamente desde los 18 hasta los 70 años, mientras que en el eje vertical se representa el monto de operaciones, cuyos valores se extienden desde alrededor de S/ 15 000 hasta S/ 137 000. La amplitud observada en ambos ejes evidencia que el conjunto de

datos contiene una variabilidad importante tanto en la dimensión demográfica como en la dimensión transaccional.

Desde el punto de vista visual, la nube de puntos presenta una distribución dispersa, sin alineamiento definido en sentido ascendente o descendente. En términos de análisis exploratorio, esta configuración indica que la relación entre ambas variables no manifiesta una estructura lineal claramente observable. Es decir, conforme aumenta la edad, el monto de operaciones no sigue un patrón uniforme de incremento o decremento dentro del conjunto de registros representados.

Asimismo, se aprecia una mayor densidad de observaciones en edades intermedias, particularmente entre los 25 y 40 años, donde también se concentran numerosos montos en rangos medios y altos. Esta acumulación sugiere que dicho tramo etario reúne una proporción considerable de clientes dentro de la muestra analizada. Sin embargo, dentro de ese mismo intervalo se observa también una dispersión vertical amplia, lo que indica que clientes de edades similares pueden presentar montos de operaciones significativamente distintos.

Otro aspecto relevante del gráfico es la heterogeneidad vertical de los puntos para una misma franja etaria. Por ejemplo, en edades cercanas a los 30, 40 o 50 años pueden encontrarse simultáneamente montos bajos, intermedios y altos. En análisis bivariado, este comportamiento refleja una variabilidad intraedad importante, lo que dificulta establecer una correspondencia simple entre ambas variables a partir de la inspección visual.

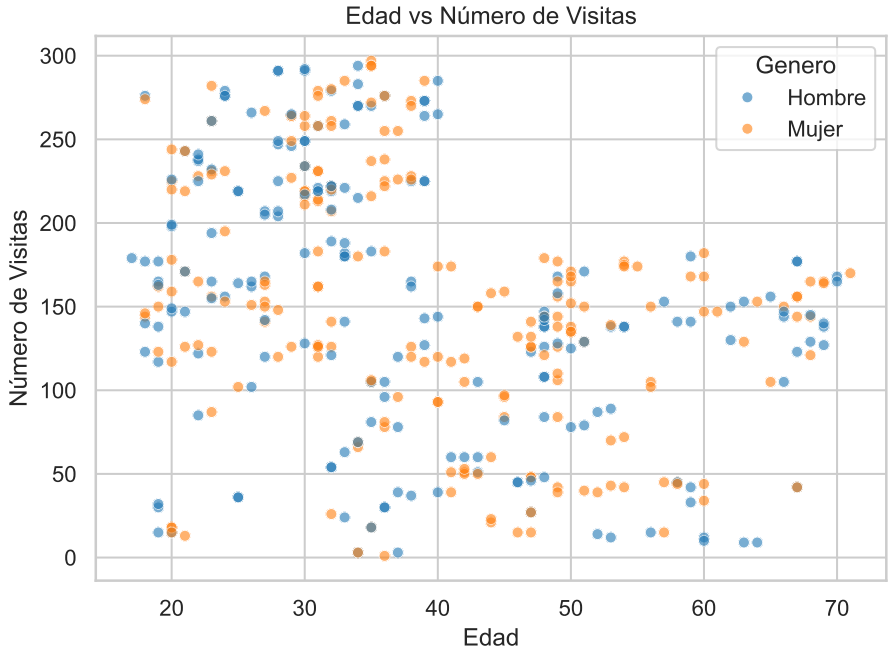
Respecto a la diferenciación por género, los puntos asignados a hombres y mujeres se distribuyen de manera superpuesta a lo largo del plano cartesiano. No se observan agrupamientos visualmente separados ni dominios exclusivos de una categoría sobre otra en zonas específicas del gráfico. Ambos grupos aparecen representados en distintos intervalos de edad y de monto, incluyendo sectores de baja, media y alta magnitud operacional.

Además, el gráfico permite advertir la existencia de algunos registros con montos elevados en diversas edades, los cuales amplían el rango total de la variable Monto de Operaciones. Estos valores altos no se restringen a un único grupo etario ni a un solo

género, sino que se encuentran dispersos dentro del conjunto general de observaciones.

Figura 10

Bivariado Edad vs Número de Visitas



En la Figura 11 se presenta un diagrama de dispersión entre las variables edad y número de Visitas, incorporando la variable género como criterio de diferenciación visual. Este tipo de gráfico permite examinar el comportamiento conjunto de dos variables cuantitativas, facilitando la identificación de patrones de asociación, concentración de observaciones, niveles de dispersión, superposición entre grupos y posibles valores extremos dentro del conjunto de datos.

En el eje horizontal se representa la variable edad, con valores que abarcan aproximadamente desde los 18 hasta los 70 años, mientras que en el eje vertical se ubica la variable Número de Visitas, cuyos registros se extienden desde valores cercanos a 0 hasta alrededor de 300 visitas. La amplitud de ambos rangos muestra una variabilidad considerable en la distribución de los clientes, tanto en términos etarios como en la frecuencia de interacción o recurrencia de visitas.

Desde la inspección visual del gráfico, la nube de puntos presenta una distribución

dispersa, sin una orientación claramente ascendente o descendente a lo largo del plano. En términos de análisis exploratorio, esta configuración sugiere que la relación entre la edad del cliente y el número de visitas no sigue una estructura lineal simple o directamente observable. Es decir, el aumento de la edad no está acompañado de un incremento o disminución uniforme en la cantidad de visitas registradas.

Asimismo, se observa una mayor concentración de registros en el intervalo aproximado de 20 a 40 años, franja en la que aparecen valores muy diversos del número de visitas, incluyendo niveles bajos, intermedios y altos. En particular, dentro de este tramo etario se ubican varios de los valores más elevados del gráfico, cercanos a las 250 y 300 visitas, lo que indica una importante dispersión vertical para edades similares.

En edades posteriores, especialmente a partir de los 45 años, continúan registrándose observaciones distribuidas en distintos niveles de visitas; sin embargo, la concentración parece desplazarse con mayor frecuencia hacia rangos intermedios, aproximadamente entre 100 y 180 visitas, aunque también persisten algunos casos con valores bajos. Esta distribución evidencia que, para diferentes grupos etarios, pueden coexistir comportamientos de visita muy heterogéneos.

Otro aspecto relevante es la variabilidad intraedad, observable en la amplitud vertical de los puntos para un mismo intervalo de edad. Por ejemplo, para clientes cercanos a los 30, 40, 50 o 60 años, se registran cantidades de visitas considerablemente distintas entre sí. En análisis bivariado, este patrón refleja que clientes de edades semejantes pueden presentar niveles de recurrencia muy diferentes, lo cual incrementa la dispersión general del gráfico.

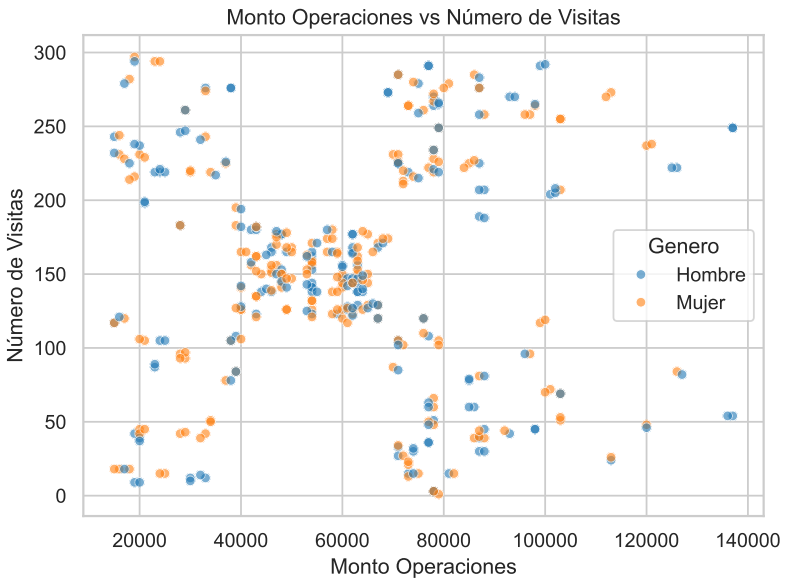
En cuanto a la variable género, los registros de hombres y mujeres se distribuyen de forma ampliamente superpuesta en el espacio bidimensional. No se distinguen agrupamientos visuales claramente separados por sexo, ni zonas exclusivas asociadas a una sola categoría. Ambos grupos aparecen representados a lo largo de diferentes edades y distintos niveles de visitas, incluyendo valores bajos, medios y altos.

Además, el gráfico permite advertir la presencia de observaciones con frecuencias de visita elevadas, particularmente en edades jóvenes e intermedias, las cuales amplían el rango superior de la variable analizada. Del mismo modo, se identifican registros próximos

a valores bajos o cercanos a cero en varios tramos etarios, lo que refuerza la amplitud y heterogeneidad de la distribución conjunta observada.

Figura 11

Bivariado Monto de Operaciones vs Número de Visitas



En la Figura 11 se presenta un diagrama de dispersión entre las variables monto de operaciones y número de visitas, incorporando la variable género como criterio de diferenciación visual. Este tipo de representación permite examinar el comportamiento conjunto de dos variables cuantitativas, identificar posibles patrones de asociación, niveles de concentración de observaciones, dispersión de los registros, solapamiento entre categorías y presencia de valores extremos.

En el eje horizontal se representa la variable monto de operaciones, con valores que se extienden aproximadamente desde S/ 15 000 hasta S/ 137 000, mientras que en el eje vertical se ubica la variable Número de Visitas, cuyos registros abarcan desde valores cercanos a 0 hasta alrededor de 300 visitas. La amplitud observada en ambos ejes refleja una variabilidad considerable en los niveles de operación y en la frecuencia de interacción de los clientes.

Desde la inspección visual del gráfico, la nube de puntos presenta una distribución

amplia y heterogénea, sin un alineamiento claramente definido en sentido ascendente o descendente. En términos exploratorios, ello indica que la relación entre el monto de operaciones y el número de visitas no manifiesta una estructura lineal simple. Es decir, mayores montos de operación no se asocian de manera uniforme con un mayor o menor número de visitas dentro del conjunto de datos observado.

Asimismo, se aprecia una concentración importante de registros en rangos intermedios del monto de operaciones, particularmente entre S/ 40 000 y S/ 65 000, acompañados de números de visitas ubicados con frecuencia entre 120 y 180. Este sector del gráfico muestra una acumulación relativamente densa de observaciones, lo que sugiere una presencia relevante de clientes con niveles intermedios tanto en operaciones como en visitas.

También se observan registros con montos medios y altos, aproximadamente entre S/ 70 000 y S/ 100 000, asociados a una dispersión amplia en el número de visitas. Dentro de este intervalo coexisten clientes con frecuencias muy bajas, intermedias y elevadas, lo que evidencia una variabilidad vertical importante para montos similares. Del mismo modo, en montos bajos pueden encontrarse tanto valores reducidos como altos del número de visitas.

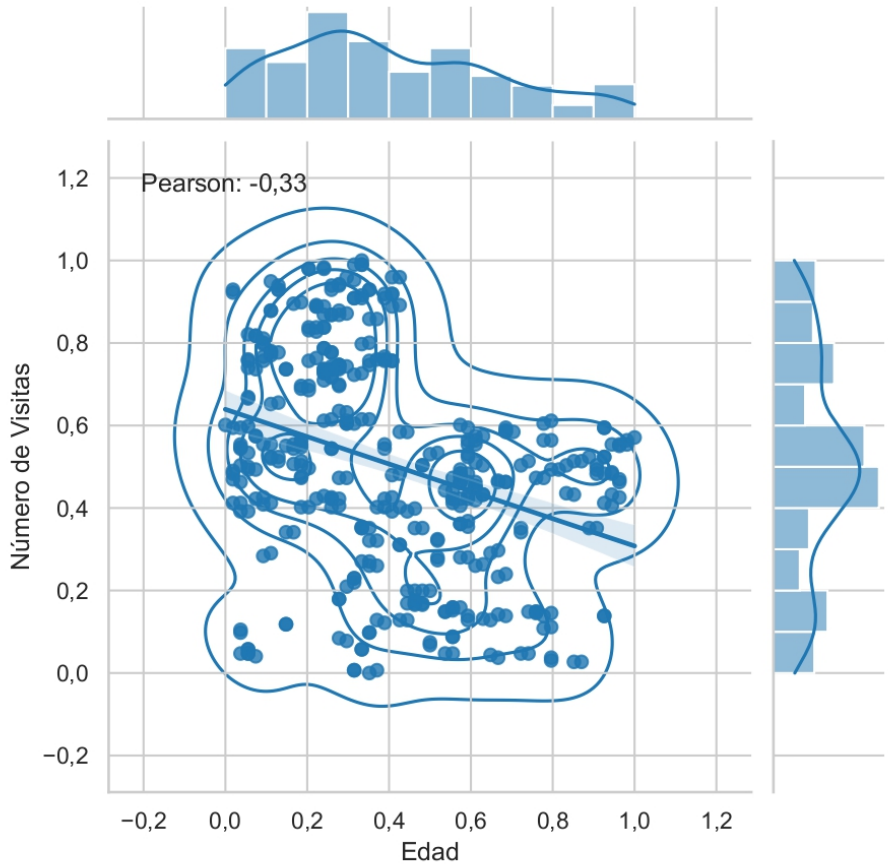
Otro aspecto relevante del gráfico es la presencia de observaciones elevadas en número de visitas, cercanas a 250 y 300, distribuidas en distintos niveles del monto de operaciones. De manera paralela, también se identifican puntos con cantidades de visitas bajas o cercanas a cero en varios tramos del monto operado. Este comportamiento incrementa la dispersión general del diagrama y pone en evidencia la amplitud de combinaciones existentes entre ambas variables.

En cuanto a la variable género, los registros correspondientes a hombres y mujeres se encuentran visualmente superpuestos a lo largo del plano bidimensional. No se distinguen agrupamientos separados ni zonas exclusivas para una sola categoría, ya que ambos grupos aparecen representados en diferentes rangos de monto de operaciones y número de visitas, incluyendo niveles bajos, intermedios y altos.

4.1.4 Análisis correlacional

Figura 12

Matriz de correlación de Pearson entre las variables de segmentación



La Figura 12 presenta un gráfico bivariado tipo jointplot que relaciona la variable Edad (eje X) con el Número de Visitas (eje Y). En el panel central se observa una nube de puntos (cada punto corresponde a un registro) complementada con curvas de densidad (contornos KDE) que resaltan las zonas donde se concentra un mayor número de observaciones. Además, se incluye una línea de tendencia lineal con su banda de confianza, que resume el patrón general de asociación entre ambas variables.

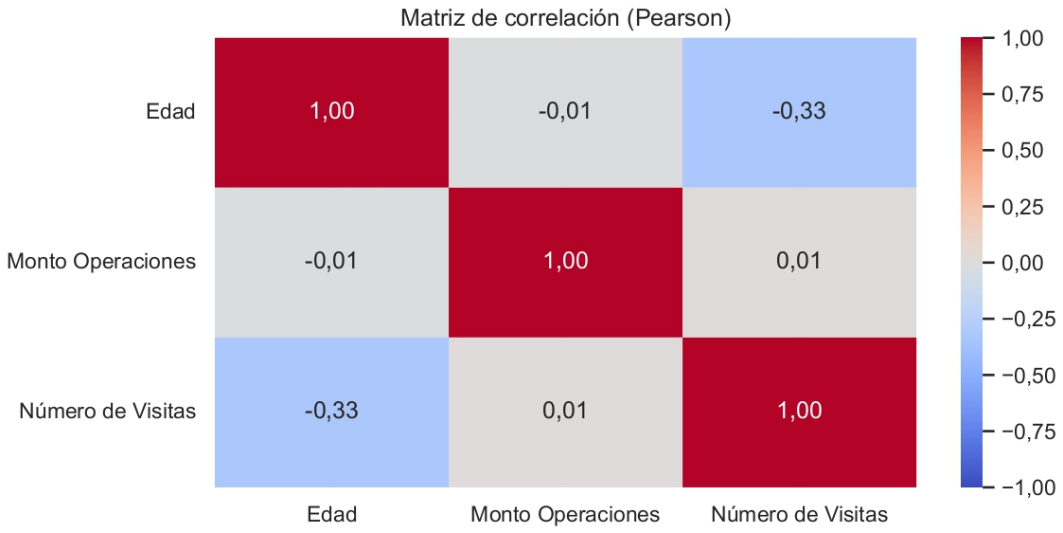
El valor reportado de correlación de Pearson es -0,33, lo que indica una relación negativa débil a moderada: en términos generales, a medida que aumenta la edad, tiende

a disminuir el número de visitas, aunque con una dispersión considerable que evidencia variabilidad en el comportamiento y la ausencia de una relación lineal fuerte.

En los márgenes superior y derecho se muestran las distribuciones univariadas de cada variable (histogramas y curva suavizada), permitiendo identificar la concentración de valores y su forma. Los contornos sugieren la presencia de agrupamientos o regiones con alta densidad de datos, lo que podría reflejar subgrupos de comportamiento (por ejemplo, rangos de edad con niveles similares de visitas), aun cuando la tendencia global sea descendente.

Figura 13

Matriz de Correlación Pearson



La Figura 13 presenta un mapa de calor (heatmap) correspondiente a la matriz de correlación de Pearson entre las variables Edad, Monto de Operaciones y Número de Visitas. Los coeficientes de Pearson varían entre -1 y 1, donde valores cercanos a 1 indican una relación lineal positiva fuerte, valores cercanos a -1 una relación lineal negativa fuerte, y valores próximos a 0 sugieren ausencia de relación lineal. En la diagonal principal se observa el valor 1,00 para cada variable consigo misma, lo cual es esperado.

En la relación entre Edad y Número de Visitas se obtiene un coeficiente r = -0,33, lo que evidencia una correlación negativa débil a moderada. Este resultado sugiere que,

en términos generales, conforme aumenta la edad, tiende a disminuir el número de visitas, aunque la magnitud del coeficiente indica que la asociación no es fuerte y existe variabilidad considerable en los datos.

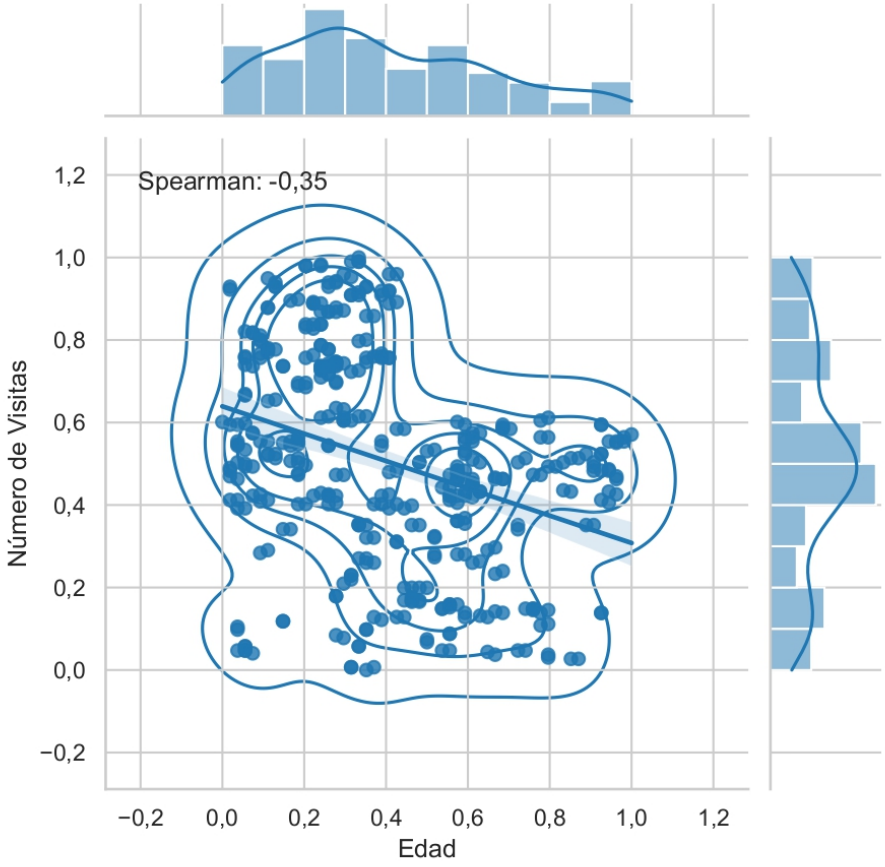
Por otro lado, la correlación entre Edad y Monto de Operaciones es r = -0,01, un valor prácticamente nulo. Esto implica que no se identifica una relación lineal significativa entre la edad y el monto de operaciones, es decir, el monto no parece incrementarse ni disminuir sistemáticamente con la edad en el conjunto de datos analizado.

De manera similar, la relación entre monto de operaciones y número de visitas muestra un coeficiente r = 0,01, también cercano a cero. En consecuencia, no se evidencia asociación lineal entre ambas variables, lo que indica que un mayor número de visitas no necesariamente se traduce en mayores montos de operaciones, ni viceversa.

En conjunto, la matriz permite concluir que la única relación lineal relativamente apreciable es la relación inversa entre edad y número de visitas, mientras que las correlaciones que involucran el monto de operaciones son prácticamente inexistentes, sugiriendo independencia lineal respecto de las demás variables consideradas.

Figura 14

Correlación Spearman



La Figura 14 muestra el análisis de correlación de Spearman entre las variables Edad (eje X) y Número de Visitas (eje Y), mediante un gráfico bivariado tipo jointplot. En el panel central se presenta la dispersión de los datos (nube de puntos), acompañada de curvas de densidad (contornos KDE) que permiten identificar las zonas de mayor concentración de observaciones. Asimismo, se incluye una línea de tendencia que resume el comportamiento general de la relación entre ambas variables.

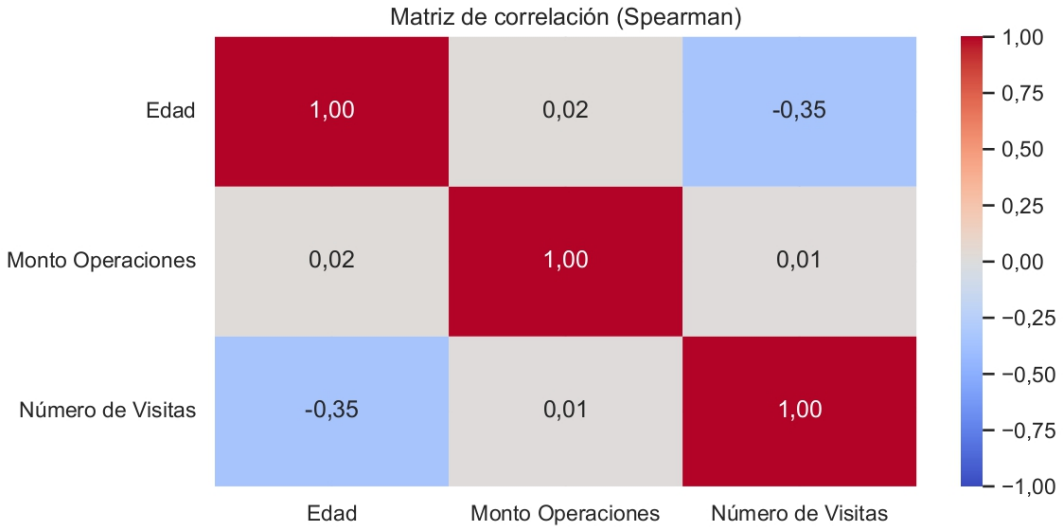
El coeficiente reportado es Spearman = -0,35, lo que indica una asociación negativa débil a moderada de tipo monotónica. Esto significa que, en términos generales, conforme la edad aumenta, el número de visitas tiende a disminuir, aunque la relación no es lo suficientemente fuerte como para describir un patrón determinista. A diferencia de

Pearson, Spearman evalúa la relación considerando el orden o ranking de los valores, por lo que es más apropiado cuando la relación no necesariamente es lineal o cuando existen valores extremos.

Los histogramas y curvas de densidad ubicados en los márgenes superior y derecho muestran la distribución univariada de Edad y Número de Visitas, respectivamente. La presencia de varias regiones de densidad en el plano central sugiere posibles subgrupos dentro de los datos (por ejemplo, rangos de edad con comportamientos de visitas diferenciados), lo que explica la dispersión observada y la magnitud moderada del coeficiente. En conjunto, la figura evidencia una tendencia inversa consistente en términos monotónicos entre Edad y Número de Visitas.

Figura 15

Matriz de Correlación Spearman



La Figura 15 presenta un mapa de calor (heatmap) de la matriz de correlación de Spearman para las variables Edad, Monto de Operaciones y Número de Visitas. A diferencia de Pearson, Spearman evalúa la asociación monotónica entre variables utilizando el orden (rangos) de los datos, por lo que resulta más adecuado cuando la relación no es necesariamente lineal o cuando existen valores atípicos. En la diagonal principal se observa el valor 1,00 para cada variable consigo misma, lo cual es esperable.

En la relación entre Edad y Número de Visitas se obtiene un coeficiente $\rho = -0,35$, lo que evidencia una asociación negativa débil a moderada. Este valor sugiere que, de manera general, conforme aumenta la edad, tiende a disminuir el número de visitas. Al tratarse de Spearman, esta interpretación se entiende como una tendencia inversa en términos de ordenamiento: edades mayores suelen asociarse con rangos menores de visitas, aunque con dispersión suficiente como para no considerarla una relación fuerte.

Por otro lado, la relación entre Edad y Monto de Operaciones muestra un coeficiente $\rho = 0,02$, prácticamente nulo. Esto indica que no existe evidencia de una asociación monotónica relevante entre ambas variables; es decir, el monto de operaciones no tiende a aumentar ni a disminuir de forma consistente conforme cambia la edad.

De forma similar, la correlación entre Monto de Operaciones y Número de Visitas presenta un coeficiente $\rho = 0,01$, también cercano a cero. En consecuencia, no se observa una relación monotónica significativa entre el número de visitas y el monto de operaciones, lo que implica que un mayor número de visitas no se asocia sistemáticamente con montos mayores o menores.

En conjunto, la matriz confirma que la relación más destacable es la tendencia inversa entre Edad y Número de Visitas, mientras que el Monto de Operaciones mantiene correlaciones muy cercanas a cero con las demás variables, lo que sugiere ausencia de asociación monotónica en este conjunto de datos.

4.2 Seleccionar y preprocesar las características relevantes de los clientes

En esta etapa del estudio se procedió a la selección de las variables que serían utilizadas en el proceso de segmentación. Inicialmente, el conjunto de datos incluía las variables género, edad, monto de operaciones y número de visitas. No obstante, tras el análisis descriptivo preliminar presentado en la sección anterior, se determinó que la variable género no sería incorporada en el proceso de clustering.

Desde un punto de vista práctico y empresarial, la empresa D&C Comunicaciones orienta sus estrategias comerciales principalmente en función del comportamiento de

consumo y la frecuencia de interacción de los clientes, más que en función de variables demográficas sensibles. La finalidad de la segmentación es apoyar decisiones relacionadas con fidelización, promociones y gestión comercial, aspectos directamente vinculados con variables cuantificables como el monto de operaciones y el número de visitas, así como con la edad como variable demográfica de interés comercial.

Asimismo, desde una perspectiva de igualdad y no discriminación, la exclusión de la variable género evita la construcción de segmentos que puedan inducir diferenciaciones basadas en esta característica. La utilización de variables sensibles en modelos analíticos debe estar debidamente justificada por su aporte explicativo, lo cual no se evidenció en el análisis exploratorio realizado.

En efecto, los gráficos descriptivos presentados previamente no mostraron una relación significativa entre el género y las demás características numéricas consideradas. Las distribuciones de edad, monto de operaciones y número de visitas resultaron similares entre hombres y mujeres, lo que indica que el género no constituye un factor diferenciador relevante dentro del comportamiento de los clientes analizados. En consecuencia, su inclusión podría introducir ruido innecesario en el modelo sin aportar valor sustancial al proceso de segmentación.

En virtud de lo anterior, las variables finalmente seleccionadas para el proceso de clustering fueron edad, monto de operaciones y número de visitas, al representar dimensiones directamente relacionadas con el comportamiento comercial de los clientes.

Posteriormente, se realizó el preprocesamiento de estas variables mediante un escalamiento utilizando la técnica *Min-Max Scaling*, normalizando cada característica en el intervalo $[0, 1]$. La transformación aplicada a cada variable x se expresa como:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{17}$$

donde x_{min} y x_{max} corresponden al valor mínimo y máximo observados en la variable original, respectivamente, y x' representa el valor escalado.

El escalamiento resulta fundamental en algoritmos de clustering basados en distancia,

como K-Means y Bisecting K-Means, debido a que estas técnicas utilizan la distancia euclidiana para medir la similitud entre observaciones. La distancia euclidiana entre dos puntos x_i y x_j en un espacio de n dimensiones se define como:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{n} (x_{i,m} - x_{j,m})^2} \tag{18}$$

Si las variables presentan escalas numéricas diferentes, aquellas con mayor magnitud pueden dominar el cálculo de distancias, afectando la formación de los clusters. Al normalizar todas las características en un mismo rango, se garantiza que cada variable contribuya de manera proporcional al proceso de segmentación, evitando sesgos derivados de diferencias de escala.

En consecuencia, el proceso de selección y preprocesamiento de variables permitió definir un espacio de características coherente con los objetivos del estudio, alineado con criterios empresariales y éticos, y técnicamente adecuado para la aplicación de algoritmos de inteligencia artificial no supervisados. La aplicación del escalamiento resulta fundamental en algoritmos de clustering basados en distancia, como K-Means y Bisecting K-Means, ya que estas técnicas utilizan medidas como la distancia euclidiana para determinar la similitud entre observaciones. Si las variables presentan magnitudes distintas, aquellas con valores numéricos más elevados pueden dominar el cálculo de distancias, distorsionando la formación de clusters. Al normalizar todas las características en un mismo rango, se garantiza que cada variable contribuya de manera proporcional al proceso de segmentación.

En consecuencia, el proceso de selección y preprocesamiento de variables permitió definir un espacio de características coherente con los objetivos del estudio, alineado con criterios empresariales y éticos, y técnicamente adecuado para la aplicación de algoritmos de inteligencia artificial no supervisados.

4.3 Aplicación de algoritmos de inteligencia artificial no supervisados para la generación de segmentaciones

En esta etapa se procedió a la aplicación de algoritmos de inteligencia artificial no supervisados con el objetivo de generar diferentes alternativas de segmentación de los clientes de la empresa D&C Comunicaciones. Se utilizaron los algoritmos K-Means, Bisecting K-Means y DBSCAN, ejecutándose múltiples configuraciones de hiperparámetros con la finalidad de explorar distintas estructuras de agrupamiento dentro del conjunto de datos previamente escalado.

Cada algoritmo fue evaluado utilizando las métricas internas descritas anteriormente: Índice de Silhouette, Índice de Davies-Bouldin, Índice de Calinski-Harabasz e Inercia. Con el propósito de garantizar estabilidad y evitar resultados dependientes de configuraciones particulares, se realizaron múltiples ejecuciones variando semillas y parámetros relevantes en cada caso.

4.3.1 Aplicación del algoritmo K-Means

Para el algoritmo K-Means se exploraron distintos valores del número de clusters k, en un rango comprendido entre 2 y 5. La elección de iniciar en $k = 2$ responde a que una única agrupación no constituye segmentación propiamente dicha, mientras que el límite superior $k = 5$ se estableció considerando criterios de viabilidad práctica.

Adicionalmente, debido a que K-Means depende de la inicialización aleatoria de centroides, se ejecutó el algoritmo múltiples veces (con 100,000 semillas diferentes variando el parámetro *random_state*. Esta decisión metodológica permitió evaluar la estabilidad de las soluciones y reducir el sesgo asociado a inicializaciones desfavorables.

Para cada combinación de número de clusters y semilla, se calcularon las métricas de evaluación correspondientes. Posteriormente, para cada valor de k, se identificaron las mejores ejecuciones según el criterio óptimo de cada métrica, considerando que Silhouette y Calinski-Harabasz deben maximizarse, mientras que Davies-Bouldin e Inercia deben minimizarse.

Asimismo, se generaron diagramas de caja (*boxplots*) por número de clusters, permitiendo analizar la variabilidad de cada métrica respecto a las diferentes semillas. Este procedimiento permitió no solo identificar valores óptimos, sino también evaluar la estabilidad del algoritmo ante distintas inicializaciones.

4.3.2 Aplicación del algoritmo Bisecting K-Means

El algoritmo Bisecting K-Means fue aplicado bajo una lógica experimental similar a la utilizada en K-Means, explorando distintos valores del número final de clusters dentro del mismo rango. Al tratarse de un método jerárquico divisivo que también depende de inicializaciones internas, se variaron las semillas mediante el parámetro *random_state* con el fin de analizar la robustez de las soluciones obtenidas.

Para cada combinación de número de clusters y semilla se calcularon las métricas de Silhouette, Davies-Bouldin, Calinski-Harabasz e Inercia. De manera análoga al caso anterior, se identificaron las mejores ejecuciones por número de clusters según cada criterio de evaluación.

La utilización de múltiples semillas permitió comparar la estabilidad del algoritmo respecto a K-Means y evaluar si el proceso divisivo producía segmentaciones más consistentes en términos de cohesión y separación.

4.3.3 Aplicación del algoritmo DBSCAN

En el caso de DBSCAN, al tratarse de un algoritmo basado en densidad, no se define directamente el número de clusters. En su lugar, se exploraron combinaciones de los hiperparámetros ε (radio de vecindad) y *min_samples* (número mínimo de vecinos).

El parámetro ε fue evaluado en un intervalo continuo de 100,000 valores dentro del rango $(0, 1]$, coherente con el escalamiento previo de las variables. El parámetro *min_samples* se exploró en un rango discreto comenzando desde 2 hasta 5 como valor máximo previamente definido.

Dado que DBSCAN puede generar puntos clasificados como ruido (etiqueta -1) y la

cantidad de segmentos no es definida a-priori, se implementó un filtro metodológico con las siguientes condiciones:

$$\text{Ruido} \leq 0,05 \tag{19}$$

$$2 \leq k \leq 5 \tag{20}$$

Donde k representa el número de clusters detectados. Este filtro garantizó que las segmentaciones generadas fueran comparables con los otros algoritmos y que el porcentaje de clientes no asignados a ningún cluster no superara el 5 %, preservando la utilidad práctica del modelo para la empresa.

Adicionalmente, la inercia fue calculada manualmente sobre los puntos no considerados como ruido, mediante la suma de las distancias cuadráticas respecto al centroide de cada cluster. Solo se conservaron aquellas configuraciones que cumplieron simultáneamente las restricciones de número mínimo de clusters y nivel de ruido permitido.

La ejecución sistemática de múltiples configuraciones por algoritmo permitió construir un conjunto amplio de alternativas de segmentación. Esta estrategia experimental respondió a la necesidad de evitar decisiones basadas en una única ejecución y garantizar que la selección posterior de la mejor segmentación estuviera sustentada en evidencia cuantitativa robusta.

En consecuencia, la aplicación de los algoritmos de inteligencia artificial no supervisados no se limitó a una ejecución puntual, sino que constituyó un proceso de exploración estructurada del espacio de soluciones, garantizando coherencia metodológica, comparabilidad entre algoritmos y pertinencia práctica para la empresa D&C Comunicaciones.

4.4 Comparar los algoritmos de inteligencia artificial no supervisados en generación de segmentaciones

4.4.1 Para 2 clusters

Métrica Silhouette

Tabla 10

Prueba de normalidad de la métrica Silhouette por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,003	0,000	Distribución no normal
BisectingKMeans	100000	0,048	0,000	Distribución no normal
DBSCAN	11754	0,670	0,000	Distribución no normal

La Tabla 10 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Silhouette obtenidos para cada algoritmo evaluado. Se observa que, en los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Silhouette correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este hallazgo resulta importante desde el punto de vista metodológico, ya que la normalidad constituye uno de los supuestos básicos para la aplicación de pruebas paramétricas de comparación. Al no cumplirse este supuesto en ninguno de los algoritmos, se justifica el uso de pruebas no paramétricas para el análisis inferencial posterior. Por tanto, la evidencia inicial sugiere que la comparación entre algoritmos debe realizarse mediante procedimientos robustos frente a distribuciones no normales.

Tabla 11

Prueba de homogeneidad de varianzas para la métrica Silhouette

Estadístico Levene	p-valor Levene	Conclusión Levene
39719,286	0,000	No se asume homogeneidad de varianzas

La Tabla 11 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Silhouette no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este resultado complementa la evidencia obtenida en la prueba de normalidad y confirma que no se cumplen los supuestos requeridos para aplicar métodos paramétricos clásicos. Además, la diferencia en la variabilidad entre algoritmos sugiere que no solo existen posibles diferencias en el valor central de la métrica, sino también en la consistencia o estabilidad con que cada algoritmo produce sus resultados.

Tabla 12

Estadísticos descriptivos de la métrica Silhouette por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	0,363	0,364	0,013
DBSCAN	11754	0,223	0,258	0,041
Kmeans	100000	0,364	0,364	0,002

La Tabla 12 resume los principales estadísticos descriptivos de la métrica Silhouette para los algoritmos evaluados en el escenario de 2 clusters. Se aprecia que Kmeans presenta la media más alta, con un valor de 0,364, seguido muy de cerca por BisectingKMeans con 0,363. En contraste, DBSCAN registra una media considerablemente menor, igual a 0,223, lo que evidencia un desempeño menos favorable desde el punto de vista descriptivo.

En relación con la mediana, Kmeans y BisectingKMeans muestran el mismo valor de 0,364, lo que refleja una fuerte concentración de sus resultados alrededor de ese nivel de calidad de segmentación. Por su parte, DBSCAN presenta una mediana de 0,258, superior a su media, lo que sugiere una distribución asimétrica influenciada por valores bajos. Respecto a la variabilidad, Kmeans exhibe la menor desviación estándar (0,002), lo que indica una alta estabilidad en sus resultados. BisectingKMeans presenta una dispersión mayor (0,013), mientras que DBSCAN muestra la mayor variabilidad (0,041), además del

menor rendimiento promedio. En conjunto, estos resultados sugieren que Kmeans ofrece el comportamiento descriptivo más favorable y consistente en esta métrica.

Tabla 13

Resultado de la prueba global para la métrica Silhouette

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	200378,45567028798	0,0

La Tabla 13 presenta el resultado de la prueba global de Kruskal-Wallis, la cual fue seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico obtenido fue 200378,45567028798, con un p-valor de 0,0, resultado que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

En términos inferenciales, este hallazgo indica que existen diferencias estadísticamente significativas en los valores de la métrica Silhouette entre los algoritmos de agrupamiento considerados. Por consiguiente, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos distintos en la calidad de la segmentación producida por cada algoritmo. A partir de este resultado global, corresponde realizar comparaciones post-hoc para identificar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 14

Comparaciones post-hoc entre algoritmos para la métrica Silhouette

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 14 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa

que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo que resulta coherente con sus menores valores descriptivos y su mayor dispersión. Asimismo, aunque Kmeans y BisectingKMeans presentan medias muy cercanas, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Silhouette.

A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño más favorable en la métrica Silhouette para el escenario de 2 clusters. Esta afirmación se sustenta en que alcanzó la media más alta y, simultáneamente, la menor desviación estándar, lo cual evidencia un comportamiento no solo superior, sino también más estable y consistente.

BisectingKMeans mostró un rendimiento muy cercano al de Kmeans, aunque con una variabilidad mayor, mientras que DBSCAN se situó claramente por debajo de ambos algoritmos, tanto en nivel promedio como en estabilidad. Además, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, Kmeans constituye la alternativa más favorable en términos de calidad de segmentación evaluada mediante la métrica Silhouette.

Métrica Davies-Bouldin

Tabla 15

Prueba de normalidad de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,003	0,000	Distribución no normal
BisectingKMeans	100000	0,048	0,000	Distribución no normal
DBSCAN	11754	0,728	0,000	Distribución no normal

La Tabla 15 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Davies-Bouldin obtenidos para cada algoritmo evaluado. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Davies-Bouldin correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado es relevante desde el punto de vista metodológico, ya que la normalidad constituye uno de los supuestos básicos para la aplicación de pruebas paramétricas. Al no cumplirse este supuesto en ninguno de los algoritmos, se justifica el empleo de métodos no paramétricos para el análisis comparativo. Por tanto, la evaluación inferencial posterior debe realizarse mediante procedimientos robustos frente a distribuciones no normales.

Tabla 16

Prueba de homogeneidad de varianzas para la métrica Davies-Bouldin

Estadístico Levene	p-valor Levene	Conclusión Levene
7268,065	0,000	No se asume homogeneidad de varianzas

La Tabla 16 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Davies-Bouldin no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia de la prueba de normalidad y confirma que no se cumplen los supuestos requeridos para aplicar pruebas paramétricas clásicas. Asimismo, la diferencia en la variabilidad entre algoritmos sugiere que sus resultados no solo difieren en tendencia central, sino también en estabilidad, lo cual debe ser considerado en la interpretación global del desempeño.

Tabla 17

Estadísticos descriptivos de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	1,166	1,162	0,064
DBSCAN	11754	0,645	0,590	0,068
Kmeans	100000	1,162	1,162	0,012

La Tabla 17 resume los principales estadísticos descriptivos de la métrica Davies-Bouldin para los algoritmos evaluados en el escenario de 2 clusters. A diferencia de la métrica Silhouette, en Davies-Bouldin los valores menores indican una mejor calidad de agrupamiento, ya que reflejan una menor dispersión intracluster y una mayor separación entre clusters.

Bajo este criterio, DBSCAN presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 0,645 y una mediana de 0,590. En contraste, Kmeans y BisectingKMeans muestran valores promedio considerablemente mayores, iguales a 1,162 y 1,166, respectivamente, lo que indica un desempeño menos favorable en esta métrica. La cercanía entre Kmeans y BisectingKMeans sugiere comportamientos descriptivos similares, aunque ligeramente mejores en Kmeans al presentar una media marginalmente menor.

En cuanto a la dispersión, Kmeans exhibe la menor desviación estándar, con un valor de 0,012, lo que pone en evidencia una mayor estabilidad en sus resultados. BisectingKMeans y DBSCAN presentan desviaciones estándar más elevadas, de 0,064 y 0,068, respectivamente, lo que indica una mayor variabilidad. En conjunto, estos resultados sugieren que DBSCAN alcanza el mejor desempeño promedio en la métrica Davies-Bouldin, mientras que Kmeans ofrece una mayor consistencia en sus valores.

Tabla 18

Resultado de la prueba global para la métrica Davies-Bouldin

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	201826,53159092835	0,0

La Tabla 18 presenta el resultado de la prueba global de Kruskal-Wallis, la cual fue seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 201826,53159092835, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde una perspectiva inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Davies-Bouldin entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no responden al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación producida por cada algoritmo. A partir de este resultado global, corresponde realizar comparaciones post-hoc para determinar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 19

Comparaciones post-hoc entre algoritmos para la métrica Davies-Bouldin

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 19 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el fin de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo que resulta coherente con su menor valor promedio en la métrica Davies-Bouldin. Asimismo, aunque Kmeans y BisectingKMeans presentan valores descriptivos muy cercanos, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de

la métrica Davies-Bouldin.

A partir del análisis integral de los resultados, se concluye que DBSCAN presentó el desempeño más favorable en la métrica Davies-Bouldin para el escenario de 2 clusters. Esta afirmación se sustenta en que obtuvo la media más baja, criterio que en esta métrica representa una mejor calidad de agrupamiento al reflejar clusters más compactos y mejor separados.

Kmeans y BisectingKMeans mostraron valores promedio superiores, por lo que su desempeño fue menos favorable desde la perspectiva de Davies-Bouldin. No obstante, Kmeans destacó por presentar la menor desviación estándar, lo que evidencia una mayor estabilidad relativa en sus resultados. Además, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, DBSCAN constituye la alternativa más favorable en términos de calidad de segmentación evaluada mediante la métrica Davies-Bouldin.

Métrica Calinski-Harabasz

Tabla 20

Prueba de normalidad de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,003	0,000	Distribución no normal
BisectingKMeans	100000	0,048	0,000	Distribución no normal
DBSCAN	11754	0,633	0,000	Distribución no normal

La Tabla 20 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Calinski-Harabasz obtenidos para cada algoritmo evaluado. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Calinski-Harabasz correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este hallazgo es metodológicamente importante, ya que la normalidad constituye uno de los supuestos requeridos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse dicho supuesto en ninguno de los algoritmos evaluados, se justifica el uso de procedimientos no paramétricos en el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe desarrollarse mediante una prueba robusta frente a distribuciones no normales.

Tabla 21

Prueba de homogeneidad de varianzas para la métrica Calinski-Harabasz

Estadístico Levene	p-valor Levene	Conclusión Levene
434,753	0,000	No se asume homogeneidad de varianzas

La Tabla 21 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Calinski-Harabasz no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este resultado complementa la evidencia de la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Además, la diferencia en la variabilidad entre algoritmos sugiere que sus comportamientos no solo difieren en el nivel promedio del índice, sino también en la estabilidad con que se obtienen dichos resultados, aspecto que debe ser considerado al interpretar el desempeño comparativo.

Tabla 22

Estadísticos descriptivos de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	255,263	256,264	13,043
DBSCAN	11754	7,187	5,144	3,175
Kmeans	100000	256,231	256,264	2,394

La Tabla 22 resume los principales estadísticos descriptivos de la métrica

Calinski-Harabasz para los algoritmos evaluados en el escenario de 2 clusters. En esta métrica, los valores más altos indican un mejor desempeño, ya que reflejan una mayor separación entre clusters y una mejor compactación interna de los grupos formados.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más alta con un valor de 256,231. BisectingKMeans alcanza una media muy cercana, igual a 255,263, lo que evidencia un rendimiento también elevado y competitivo. En contraste, DBSCAN muestra una media de apenas 7,187, valor considerablemente inferior al de los otros dos algoritmos, lo que indica un desempeño claramente menos favorable en términos de la métrica Calinski-Harabasz.

En relación con la mediana, Kmeans y BisectingKMeans presentan el mismo valor de 256,264, lo que revela una fuerte concentración de sus resultados alrededor de ese nivel de desempeño. Por su parte, DBSCAN registra una mediana de 5,144, también muy inferior. En cuanto a la dispersión, Kmeans exhibe la menor desviación estándar, con un valor de 2,394, lo que evidencia una mayor estabilidad en sus resultados. BisectingKMeans presenta una variabilidad más alta, con una desviación estándar de 13,043, mientras que DBSCAN alcanza una dispersión de 3,175, aunque en un rango de valores notablemente inferior. En conjunto, estos resultados permiten inferir que Kmeans no solo presenta el mejor desempeño promedio, sino también una mayor consistencia relativa en esta métrica.

Tabla 23

Resultado de la prueba global para la métrica Calinski-Harabasz

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	201826,5287813602	0,0

La Tabla 23 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 201826,5287813602, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Calinski-Harabasz entre los

algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no responden al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación lograda por cada algoritmo. A partir de este resultado global, corresponde examinar las comparaciones post-hoc para determinar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 24

Comparaciones post-hoc entre algoritmos para la métrica Calinski-Harabasz

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 24 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo cual resulta coherente con sus valores descriptivos considerablemente más bajos. Asimismo, aunque Kmeans y BisectingKMeans presentan resultados muy cercanos en términos de media y mediana, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Calinski-Harabasz.

A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño más favorable en la métrica Calinski-Harabasz para el escenario de 2 clusters. Esta afirmación se sustenta en que obtuvo la media más alta, criterio que en esta métrica representa una mejor calidad de agrupamiento al reflejar clusters más compactos y mejor separados.

BisectingKMeans mostró un rendimiento muy cercano al de Kmeans, aunque con una

variabilidad considerablemente mayor. Por su parte, DBSCAN se ubicó claramente por debajo de ambos algoritmos, con valores descriptivos marcadamente inferiores. Además, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, Kmeans constituye la alternativa más favorable en términos de calidad de segmentación evaluada mediante la métrica Calinski-Harabasz.

Métrica Inercia

Tabla 25

Prueba de normalidad de la métrica Inercia por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,003	0,000	Distribución no normal
BisectingKMeans	100000	0,048	0,000	Distribución no normal
DBSCAN	11754	0,715	0,000	Distribución no normal

La Tabla 25 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Inercia obtenidos para cada algoritmo evaluado. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Inercia correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este hallazgo resulta metodológicamente relevante, ya que la normalidad constituye uno de los supuestos básicos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse este supuesto en ninguno de los algoritmos evaluados, se justifica el uso de pruebas no paramétricas para el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe realizarse mediante procedimientos estadísticos robustos frente a distribuciones no normales.

Tabla 26

Prueba de homogeneidad de varianzas para la métrica Inercia

Estadístico Levene	p-valor Levene	Conclusión Levene
16165,118	0,000	No se asume homogeneidad de varianzas

La Tabla 26 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Inercia no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este resultado complementa la evidencia proporcionada por la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Asimismo, la existencia de varianzas desiguales sugiere que los algoritmos no solo difieren en el nivel promedio de la métrica, sino también en la estabilidad con que producen sus resultados, lo cual debe ser considerado en la interpretación comparativa.

Tabla 27

Estadísticos descriptivos de la métrica Inercia por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	44,041	43,950	1,189
DBSCAN	11754	69,590	71,327	2,333
Kmeans	100000	43,953	43,950	0,219

La Tabla 27 resume los principales estadísticos descriptivos de la métrica Inercia para los algoritmos evaluados en el escenario de 2 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que representan una menor suma de distancias internas entre los elementos y el centro de su respectivo cluster, lo que se asocia con agrupamientos más compactos.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 43,953. BisectingKMeans muestra un resultado muy cercano, con una media de 44,041, lo que sugiere un comportamiento similar desde el

punto de vista descriptivo. En contraste, DBSCAN exhibe una media considerablemente mayor, igual a 69,590, lo que evidencia un desempeño menos favorable en términos de Inercia.

En relación con la mediana, Kmeans y BisectingKMeans presentan el mismo valor de 43,950, lo que refleja una fuerte concentración de sus resultados alrededor de ese nivel de desempeño. Por su parte, DBSCAN registra una mediana de 71,327, también claramente superior a la de los otros algoritmos. Respecto a la dispersión, Kmeans presenta la menor desviación estándar, con un valor de 0,219, lo que indica una mayor estabilidad en sus resultados. BisectingKMeans muestra una variabilidad más alta, con una desviación estándar de 1,189, mientras que DBSCAN registra la mayor dispersión, con un valor de 2,333. En conjunto, estos resultados permiten inferir que Kmeans no solo alcanza el mejor desempeño promedio, sino que también lo hace con mayor consistencia relativa.

Tabla 28

Resultado de la prueba global para la métrica Inercia

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	156425,40889085658	0,0

La Tabla 28 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 156425,40889085658, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Inercia entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en el grado de compactación logrado por cada algoritmo. A partir de este resultado global, corresponde realizar comparaciones post-hoc para identificar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 29

Comparaciones post-hoc entre algoritmos para la métrica Inercia

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 29 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo cual resulta coherente con sus valores descriptivos notablemente más altos en la métrica Inercia. Asimismo, aunque Kmeans y BisectingKMeans presentan medias y medianas muy cercanas, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Inercia.

A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño más favorable en la métrica Inercia para el escenario de 2 clusters. Esta afirmación se sustenta en que obtuvo la media más baja, criterio que en esta métrica representa una mejor calidad de agrupamiento al reflejar clusters más compactos.

BisectingKMeans mostró un rendimiento muy cercano al de Kmeans, aunque con una variabilidad mayor. Por su parte, DBSCAN se ubicó claramente por encima de ambos algoritmos en términos de Inercia, lo que indica un menor nivel de compactación en los grupos obtenidos. Además, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, Kmeans constituye la alternativa más favorable en términos de compactación de clusters evaluada mediante

la métrica Inercia.

4.4.2 Para 3 Clusters

Métrica Silhouette

Tabla 30

Prueba de normalidad de la métrica Silhouette por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,273	0,000	Distribución no normal
BisectingKMeans	100000	0,229	0,000	Distribución no normal
DBSCAN	3442	0,678	0,000	Distribución no normal

La Tabla 30 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Silhouette obtenidos para cada algoritmo evaluado en el escenario de 3 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Silhouette correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado posee relevancia metodológica, ya que la normalidad constituye uno de los supuestos requeridos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse este supuesto en ninguno de los algoritmos evaluados, se justifica el empleo de procedimientos no paramétricos en el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe desarrollarse mediante una prueba robusta frente a distribuciones no normales.

Tabla 31

Prueba de homogeneidad de varianzas para la métrica Silhouette

Estadístico Levene	p-valor Levene	Conclusión Levene
26882,074	0,000	No se asume homogeneidad de varianzas

La Tabla 31 muestra el resultado de la prueba de Levene, utilizada para evaluar la

homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Silhouette no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia proporcionada por la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Asimismo, la diferencia en la variabilidad entre algoritmos sugiere que sus resultados no solo difieren en tendencia central, sino también en estabilidad, aspecto que debe ser considerado en la interpretación comparativa del desempeño.

Tabla 32

Estadísticos descriptivos de la métrica Silhouette por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	0,365	0,368	0,016
DBSCAN	3442	0,137	0,167	0,076
Kmeans	100000	0,365	0,367	0,009

La Tabla 32 resume los principales estadísticos descriptivos de la métrica Silhouette para los algoritmos evaluados en el escenario de 3 clusters. Se observa que BisectingKMeans y Kmeans presentan los valores promedio más altos, ambos reportados con una media de 0,365 al nivel de redondeo mostrado en la tabla, mientras que DBSCAN registra una media considerablemente menor, igual a 0,137. Desde una perspectiva descriptiva, esto evidencia que DBSCAN alcanza una calidad de segmentación notablemente inferior en comparación con los otros dos algoritmos.

Aunque las medias de BisectingKMeans y Kmeans aparecen iguales al redondeo de tres decimales, la mediana de BisectingKMeans es ligeramente superior, con un valor de 0,368 frente a 0,367 en Kmeans, lo que respalda una ventaja descriptiva marginal a favor de BisectingKMeans. Por su parte, DBSCAN presenta una mediana de 0,167, también claramente inferior, lo que reafirma su menor desempeño en esta métrica.

En cuanto a la dispersión, Kmeans exhibe la menor desviación estándar, con un valor de 0,009, lo que indica una mayor estabilidad en sus resultados. BisectingKMeans presenta

una variabilidad superior, con una desviación estándar de 0,016, mientras que DBSCAN alcanza la mayor dispersión, con un valor de 0,076. En conjunto, estos resultados permiten inferir que BisectingKMeans presenta el desempeño descriptivamente más favorable en la métrica Silhouette para 3 clusters, aunque con una diferencia muy pequeña respecto a Kmeans, mientras que este último destaca por su mayor consistencia.

Tabla 33

Resultado de la prueba global para la métrica Silhouette

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	125530,23840767931	0,0

La Tabla 33 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 125530,23840767931, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Silhouette entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación alcanzada por cada algoritmo. A partir de este resultado global, corresponde realizar comparaciones post-hoc para identificar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 34

Comparaciones post-hoc entre algoritmos para la métrica Silhouette

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 34 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde

se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo cual resulta coherente con sus menores valores descriptivos y su mayor dispersión. Asimismo, aunque BisectingKMeans y Kmeans presentan resultados muy cercanos en términos de media y mediana, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Silhouette.

A partir del análisis integral de los resultados, se concluye que BisectingKMeans presentó el desempeño descriptivamente más favorable en la métrica Silhouette para el escenario de 3 clusters. Esta afirmación se sustenta en que registró el valor promedio más alto según la conclusión reportada y, adicionalmente, la mediana más elevada entre los algoritmos evaluados.

Kmeans mostró un rendimiento muy cercano al de BisectingKMeans y destacó por presentar una menor desviación estándar, lo que evidencia una mayor estabilidad relativa en sus resultados. Por su parte, DBSCAN se ubicó claramente por debajo de ambos algoritmos, con una media y mediana considerablemente menores, además de una mayor variabilidad. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, BisectingKMeans constituye la alternativa descriptivamente más favorable en términos de calidad de segmentación evaluada mediante la métrica Silhouette.

Métrica Davies-Bouldin

Tabla 35

Prueba de normalidad de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,594	0,000	Distribución no normal
BisectingKMeans	100000	0,231	0,000	Distribución no normal
DBSCAN	3442	0,770	0,000	Distribución no normal

La Tabla 35 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Davies-Bouldin obtenidos para cada algoritmo evaluado en el escenario de 3 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Davies-Bouldin correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado posee importancia metodológica, ya que la normalidad constituye uno de los supuestos requeridos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse dicho supuesto en ninguno de los algoritmos analizados, se justifica el empleo de procedimientos no paramétricos en el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe realizarse mediante una prueba robusta frente a distribuciones no normales.

Tabla 36

Prueba de homogeneidad de varianzas para la métrica Davies-Bouldin

Estadístico Levene	p-valor Levene	Conclusión Levene
7992,569	0,000	No se asume homogeneidad de varianzas

La Tabla 36 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Davies-Bouldin no es homogénea entre Kmeans,

BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia proporcionada por la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Asimismo, la diferencia en la variabilidad entre algoritmos sugiere que sus resultados no solo difieren en tendencia central, sino también en estabilidad, aspecto que debe ser considerado en la interpretación comparativa del desempeño.

Tabla 37

Estadísticos descriptivos de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	1,055	1,043	0,062
DBSCAN	3442	0,815	0,760	0,141
Kmeans	100000	1,048	1,055	0,051

La Tabla 37 resume los principales estadísticos descriptivos de la métrica Davies-Bouldin para los algoritmos evaluados en el escenario de 3 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que reflejan una menor dispersión intracluster y una mayor separación relativa entre los grupos formados.

Bajo este criterio, DBSCAN presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 0,815 y una mediana de 0,760. En contraste, Kmeans y BisectingKMeans muestran valores promedio mayores, iguales a 1,048 y 1,055, respectivamente, lo que indica un desempeño menos favorable desde la perspectiva de Davies-Bouldin. Entre estos dos últimos, Kmeans presenta una media ligeramente menor que BisectingKMeans, aunque ambos permanecen relativamente próximos.

En cuanto a la dispersión, Kmeans exhibe la menor desviación estándar, con un valor de 0,051, lo que evidencia una mayor estabilidad en sus resultados. BisectingKMeans presenta una variabilidad algo mayor, con una desviación estándar de 0,062, mientras que DBSCAN registra la mayor dispersión, con un valor de 0,141. En conjunto, estos resultados sugieren que DBSCAN alcanza el mejor desempeño promedio en la métrica Davies-Bouldin, aunque con una variabilidad considerablemente mayor que la observada en Kmeans y BisectingKMeans.

Tabla 38

Resultado de la prueba global para la métrica Davies-Bouldin

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	13860,667509343806	0,0

La Tabla 38 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 13860,667509343806, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Davies-Bouldin entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación alcanzada por cada algoritmo. A partir de este resultado global, corresponde realizar comparaciones post-hoc para identificar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 39

Comparaciones post-hoc entre algoritmos para la métrica Davies-Bouldin

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 39 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de

BisectingKMeans como de Kmeans, lo cual resulta coherente con su menor valor promedio en la métrica Davies-Bouldin. Asimismo, aunque Kmeans y BisectingKMeans presentan valores descriptivos cercanos, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Davies-Bouldin.

A partir del análisis integral de los resultados, se concluye que DBSCAN presentó el desempeño descriptivamente más favorable en la métrica Davies-Bouldin para el escenario de 3 clusters. Esta afirmación se sustenta en que obtuvo la media más baja, criterio que en esta métrica representa una mejor calidad de agrupamiento al reflejar clusters relativamente más compactos y mejor separados.

Kmeans y BisectingKMeans mostraron valores promedio superiores, por lo que su desempeño fue menos favorable desde la perspectiva de Davies-Bouldin. No obstante, Kmeans destacó por presentar la menor desviación estándar, lo que evidencia una mayor estabilidad relativa en sus resultados. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, DBSCAN constituye la alternativa descriptivamente más favorable en términos de calidad de segmentación evaluada mediante la métrica Davies-Bouldin.

Métrica Calinski-Harabasz

Tabla 40

Prueba de normalidad de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,459	0,000	Distribución no normal
BisectingKMeans	100000	0,210	0,000	Distribución no normal
DBSCAN	3442	0,673	0,000	Distribución no normal

La Tabla 40 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Calinski-Harabasz obtenidos para cada algoritmo evaluado en el escenario de 3 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Calinski-Harabasz correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado es metodológicamente relevante, ya que la normalidad constituye uno de los supuestos requeridos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse este supuesto en ninguno de los algoritmos evaluados, se justifica el empleo de procedimientos no paramétricos para el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe realizarse mediante pruebas robustas frente a distribuciones no normales.

Tabla 41

Prueba de homogeneidad de varianzas para la métrica Calinski-Harabasz

Estadístico Levene	p-valor Levene	Conclusión Levene
11697,290	0,000	No se asume homogeneidad de varianzas

La Tabla 41 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Calinski-Harabasz no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia obtenida en la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Asimismo, la desigualdad de varianzas sugiere que los algoritmos no solo difieren en su nivel promedio de desempeño, sino también en la estabilidad con la que alcanzan dichos resultados, aspecto importante para la interpretación comparativa.

Tabla 42

Estadísticos descriptivos de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	220,021	222,712	12,811
DBSCAN	3442	39,899	11,779	31,779
Kmeans	100000	222,688	225,862	5,467

La Tabla 42 resume los principales estadísticos descriptivos de la métrica Calinski-Harabasz para los algoritmos evaluados en el escenario de 3 clusters. En esta métrica, los valores más altos indican un mejor desempeño, ya que reflejan una mejor separación entre clusters y una mayor compactación interna de los grupos formados.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más alta con un valor de 222,688 y la mediana más alta con 225,862. BisectingKMeans alcanza un rendimiento cercano, con una media de 220,021 y una mediana de 222,712, lo que evidencia también una calidad de segmentación elevada, aunque ligeramente inferior a la de Kmeans. En contraste, DBSCAN muestra una media mucho menor, igual a 39,899, y una mediana de 11,779, lo que indica un desempeño considerablemente menos favorable en términos de la métrica Calinski-Harabasz.

En cuanto a la dispersión, Kmeans presenta la menor desviación estándar, con un valor de 5,467, lo que evidencia una mayor estabilidad en sus resultados. BisectingKMeans muestra una variabilidad superior, con una desviación estándar de 12,811, mientras que DBSCAN registra la mayor dispersión, con un valor de 31,779. La notable diferencia entre la media y la mediana en DBSCAN sugiere además una distribución marcadamente asimétrica, posiblemente influenciada por valores extremos. En conjunto, estos resultados permiten inferir que Kmeans no solo presenta el mejor desempeño promedio, sino también una mayor consistencia relativa en esta métrica.

Tabla 43

Resultado de la prueba global para la métrica Calinski-Harabasz

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	12381,02560455152	0,0

La Tabla 43 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 12381,02560455152, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Calinski-Harabasz entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación lograda por cada algoritmo. A partir de este resultado global, corresponde examinar las comparaciones post-hoc para determinar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 44

Comparaciones post-hoc entre algoritmos para la métrica Calinski-Harabasz

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 44 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de

BisectingKMeans como de Kmeans, lo cual resulta coherente con sus valores descriptivos considerablemente más bajos. Asimismo, aunque Kmeans y BisectingKMeans presentan resultados relativamente cercanos en términos de media y mediana, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Calinski-Harabasz.

A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Calinski-Harabasz para el escenario de 3 clusters. Esta afirmación se sustenta en que obtuvo la media más alta y la mediana más elevada, criterios que en esta métrica representan una mejor calidad de agrupamiento al reflejar clusters más compactos y mejor separados.

BisectingKMeans mostró un rendimiento cercano al de Kmeans, aunque con una variabilidad mayor. Por su parte, DBSCAN se ubicó claramente por debajo de ambos algoritmos, con valores descriptivos marcadamente inferiores y una dispersión considerablemente más alta. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, Kmeans constituye la alternativa más favorable en términos de calidad de segmentación evaluada mediante la métrica Calinski-Harabasz.

Métrica Inercia

Tabla 45

Prueba de normalidad de la métrica Inercia por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,419	0,000	Distribución no normal
BisectingKMeans	100000	0,206	0,000	Distribución no normal
DBSCAN	3442	0,725	0,000	Distribución no normal

La Tabla 45 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Inercia obtenidos para cada algoritmo evaluado en el escenario de 3 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Inercia correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado es metodológicamente relevante, ya que la normalidad constituye uno de los supuestos necesarios para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse dicho supuesto en ninguno de los algoritmos evaluados, se justifica el uso de procedimientos no paramétricos en el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe desarrollarse mediante pruebas robustas frente a distribuciones no normales.

Tabla 46

Prueba de homogeneidad de varianzas para la métrica Inercia

Estadístico Levene	p-valor Levene	Conclusión Levene
73935,743	0,000	No se asume homogeneidad de varianzas

La Tabla 46 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Inercia no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia proporcionada por la prueba de normalidad y confirma que no se cumplen los supuestos clásicos requeridos para la aplicación de pruebas paramétricas. Asimismo, la desigualdad de varianzas sugiere que los algoritmos no solo difieren en el nivel promedio de la métrica, sino también en la estabilidad con la que producen sus resultados, aspecto importante en la interpretación comparativa del desempeño.

Tabla 47

Estadísticos descriptivos de la métrica Inercia por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	34,306	34,048	1,250
DBSCAN	3442	57,881	65,026	9,825
Kmeans	100000	34,056	33,795	0,482

La Tabla 47 resume los principales estadísticos descriptivos de la métrica Inercia para los algoritmos evaluados en el escenario de 3 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que representan una menor suma de distancias internas entre los elementos y el centro de sus respectivos clusters, lo que se asocia con agrupamientos más compactos.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 34,056 y la mediana más baja con 33,795. BisectingKMeans muestra un comportamiento cercano, con una media de 34,306 y una mediana de 34,048, lo que evidencia un rendimiento competitivo, aunque ligeramente inferior al de Kmeans. En contraste, DBSCAN presenta una media considerablemente mayor, igual a 57,881, y una mediana de 65,026, lo que indica un desempeño claramente menos favorable en términos de la métrica Inercia.

En cuanto a la dispersión, Kmeans exhibe la menor desviación estándar, con un valor de 0,482, lo que evidencia una mayor estabilidad en sus resultados. BisectingKMeans presenta una variabilidad más alta, con una desviación estándar de 1,250, mientras que DBSCAN registra la mayor dispersión, con un valor de 9,825. Además, la diferencia notable entre la media y la mediana en DBSCAN sugiere una distribución asimétrica en sus resultados. En conjunto, estos hallazgos permiten inferir que Kmeans no solo alcanza el mejor desempeño promedio, sino que también lo hace con mayor consistencia relativa.

Tabla 48

Resultado de la prueba global para la métrica Inercia

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	12168,491183111799	0,0

La Tabla 48 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 12168,491183111799, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Inercia entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en el grado de compactación alcanzado por cada algoritmo. A partir de este resultado global, corresponde examinar las comparaciones post-hoc para determinar específicamente entre qué pares de algoritmos se presentan dichas diferencias.

Tabla 49

Comparaciones post-hoc entre algoritmos para la métrica Inercia

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 49 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos, realizadas con el propósito de identificar de manera específica dónde se encuentran las diferencias detectadas por la prueba global de Kruskal-Wallis. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par de algoritmos comparados.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo cual resulta coherente con sus valores

descriptivos notablemente más altos en la métrica Inercia. Asimismo, aunque Kmeans y BisectingKMeans presentan resultados relativamente cercanos en términos de media y mediana, la comparación post-hoc demuestra que esa diferencia también alcanza significancia estadística. En consecuencia, puede afirmarse que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Inercia.

A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Inercia para el escenario de 3 clusters. Esta afirmación se sustenta en que obtuvo la media más baja y la mediana más reducida, criterios que en esta métrica representan una mejor calidad de agrupamiento al reflejar clusters más compactos.

BisectingKMeans mostró un rendimiento cercano al de Kmeans, aunque con una variabilidad mayor. Por su parte, DBSCAN se ubicó claramente por encima de ambos algoritmos en términos de Inercia, lo que indica un menor nivel de compactación en los grupos obtenidos. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas. En síntesis, para el caso analizado, Kmeans constituye la alternativa más favorable en términos de compactación de clusters evaluada mediante la métrica Inercia.

4.4.3 Para 4 Clusters

Métrica Silhouette

Tabla 50

Prueba de normalidad de la métrica Silhouette por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,372	0,000	Distribución no normal
BisectingKMeans	100000	0,249	0,000	Distribución no normal
DBSCAN	2654	0,633	0,000	Distribución no normal

La Tabla 50 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Silhouette obtenidos para cada algoritmo evaluado en el escenario de 4 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Silhouette correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado es metodológicamente relevante, ya que la normalidad constituye uno de los supuestos requeridos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse este supuesto en ninguno de los algoritmos evaluados, se justifica el uso de procedimientos no paramétricos en el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe realizarse mediante pruebas robustas frente a distribuciones no normales.

Tabla 51

Prueba de homogeneidad de varianzas para la métrica Silhouette

Estadístico Levene	p-valor Levene	Conclusión Levene
1091,154	0,000	No se asume homogeneidad de varianzas

La Tabla 51 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Silhouette no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia proporcionada por la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Asimismo, la desigualdad de varianzas sugiere que los algoritmos no solo difieren en su tendencia central, sino también en la estabilidad con la que producen sus resultados, aspecto que debe ser considerado en la interpretación comparativa.

Tabla 52

Estadísticos descriptivos de la métrica Silhouette por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	0,391	0,394	0,015
DBSCAN	2654	0,014	-0,003	0,018
Kmeans	100000	0,394	0,399	0,016

La Tabla 52 resume los principales estadísticos descriptivos de la métrica Silhouette para los algoritmos evaluados en el escenario de 4 clusters. En esta métrica, los valores más altos indican una mejor calidad de segmentación, al reflejar una mayor cohesión interna dentro de los clusters y una mejor separación entre ellos.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más alta con un valor de 0,394 y la mediana más elevada con 0,399. BisectingKMeans muestra un rendimiento muy cercano, con una media de 0,391 y una mediana de 0,394, lo que evidencia un comportamiento competitivo. En contraste, DBSCAN alcanza una media muy baja de 0,014 y una mediana negativa de -0,003, lo que indica una calidad de segmentación claramente inferior respecto a los otros dos algoritmos.

En cuanto a la dispersión, BisectingKMeans presenta la menor desviación estándar, con un valor de 0,015, seguido muy de cerca por Kmeans con 0,016. DBSCAN muestra una dispersión de 0,018, que, aunque no es muy superior en términos absolutos, debe interpretarse en el contexto de sus valores centrales mucho más bajos. En conjunto, estos resultados permiten inferir que Kmeans alcanza el mejor desempeño promedio en la métrica Silhouette para 4 clusters, mientras que BisectingKMeans presenta una estabilidad ligeramente mayor.

Tabla 53

Resultado de la prueba global para la métrica Silhouette

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	104764,42727951876	0,0

La Tabla 53 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada

debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 104764,42727951876, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Silhouette entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación lograda por cada algoritmo.

Tabla 54

Comparaciones post-hoc entre algoritmos para la métrica Silhouette

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 54 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo que resulta coherente con sus valores descriptivos mucho más bajos. Asimismo, la comparación entre Kmeans y BisectingKMeans también resulta significativa, a pesar de la cercanía observada en sus medias y medianas.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Silhouette para el escenario de 4 clusters. Esta afirmación se sustenta en que obtuvo la media más alta y la mediana más elevada, criterios que en esta métrica representan una mejor calidad de agrupamiento.

BisectingKMeans mostró un rendimiento muy cercano y una estabilidad ligeramente mayor, mientras que DBSCAN se ubicó claramente por debajo de ambos algoritmos.

Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas.

Métrica Davies-Bouldin

Tabla 55

Prueba de normalidad de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,269	0,000	Distribución no normal
BisectingKMeans	100000	0,236	0,000	Distribución no normal
DBSCAN	2654	0,633	0,000	Distribución no normal

La Tabla 55 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Davies-Bouldin obtenidos para cada algoritmo evaluado. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Davies-Bouldin correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado justifica el uso de procedimientos no paramétricos para el análisis comparativo posterior, dado que no se cumple el supuesto de normalidad requerido por las pruebas paramétricas clásicas.

Tabla 56

Prueba de homogeneidad de varianzas para la métrica Davies-Bouldin

Estadístico Levene	p-valor Levene	Conclusión Levene
790,525	0,000	No se asume homogeneidad de varianzas

La Tabla 56 muestra el resultado de la prueba de Levene. Dado que el p-valor es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Davies-Bouldin no es homogénea entre los algoritmos comparados.

En conjunto con la prueba de normalidad, este resultado confirma que no se cumplen los supuestos requeridos para aplicar pruebas paramétricas, por lo que corresponde utilizar un enfoque no paramétrico en la comparación global.

Tabla 57

Estadísticos descriptivos de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	1,002	0,993	0,050
DBSCAN	2654	0,711	0,749	0,042
Kmeans	100000	0,970	0,964	0,033

La Tabla 57 resume los principales estadísticos descriptivos de la métrica Davies-Bouldin para los algoritmos evaluados en el escenario de 4 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que representan clusters relativamente más compactos y mejor separados.

Bajo este criterio, DBSCAN presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 0,711. Kmeans ocupa la segunda posición con una media de 0,970, mientras que BisectingKMeans presenta el valor promedio más alto, igual a 1,002, lo que indica un desempeño menos favorable desde esta métrica. En cuanto a la mediana, Kmeans registra 0,964, BisectingKMeans 0,993 y DBSCAN 0,749, reforzando la ventaja descriptiva de DBSCAN.

Respecto a la dispersión, Kmeans presenta la menor desviación estándar, con un valor de 0,033, lo que evidencia una mayor estabilidad relativa. DBSCAN muestra una desviación estándar de 0,042, mientras que BisectingKMeans presenta la mayor dispersión con 0,050. En conjunto, estos resultados indican que DBSCAN alcanza el mejor desempeño promedio en la métrica Davies-Bouldin, aunque Kmeans exhibe una mayor consistencia en sus resultados.

Tabla 58

Resultado de la prueba global para la métrica Davies-Bouldin

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	149150,8251264904	0,0

La Tabla 58 presenta el resultado de la prueba global de Kruskal-Wallis. El valor del estadístico fue 149150,8251264904, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Davies-Bouldin entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos reflejan comportamientos efectivamente distintos entre algoritmos.

Tabla 59

Comparaciones post-hoc entre algoritmos para la métrica Davies-Bouldin

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 59 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo que resulta coherente con su menor valor promedio. Asimismo, la comparación entre Kmeans y BisectingKMeans también evidencia una diferencia significativa.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que DBSCAN presentó el desempeño descriptivamente más favorable en la

métrica Davies-Bouldin para el escenario de 4 clusters. Esta afirmación se sustenta en que obtuvo la media más baja, criterio que en esta métrica representa una mejor calidad de agrupamiento.

Kmeans mostró una mayor estabilidad relativa, mientras que BisectingKMeans presentó el desempeño menos favorable entre los tres algoritmos. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas son estadísticamente significativas.

Métrica Calinski-Harabasz

Tabla 60

Prueba de normalidad de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,399	0,000	Distribución no normal
BisectingKMeans	100000	0,221	0,000	Distribución no normal
DBSCAN	2654	0,633	0,000	Distribución no normal

La Tabla 60 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Calinski-Harabasz obtenidos para cada algoritmo evaluado. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Calinski-Harabasz correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado justifica el uso de procedimientos no paramétricos para la comparación global posterior.

Tabla 61

Prueba de homogeneidad de varianzas para la métrica Calinski-Harabasz

Estadístico Levene	p-valor Levene	Conclusión Levene
1669,427	0,000	No se asume homogeneidad de varianzas

La Tabla 61 muestra el resultado de la prueba de Levene. Dado que el p-valor es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Calinski-Harabasz no es homogénea entre los algoritmos.

En conjunto con la prueba de normalidad, este resultado confirma que no se cumplen los supuestos requeridos para aplicar pruebas paramétricas clásicas.

Tabla 62

Estadísticos descriptivos de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	245,737	248,757	13,509
DBSCAN	2654	8,228	11,056	3,133
Kmeans	100000	250,011	257,696	20,441

La Tabla 62 resume los principales estadísticos descriptivos de la métrica Calinski-Harabasz para los algoritmos evaluados en el escenario de 4 clusters. En esta métrica, los valores más altos indican un mejor desempeño, ya que reflejan una mejor separación entre clusters y una mayor compactación interna de los grupos formados.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más alta con un valor de 250,011 y la mediana más elevada con 257,696. BisectingKMeans muestra un rendimiento cercano, con una media de 245,737 y una mediana de 248,757, mientras que DBSCAN presenta valores considerablemente menores, con una media de 8,228 y una mediana de 11,056, lo que evidencia un desempeño claramente inferior.

En cuanto a la dispersión, DBSCAN presenta la menor desviación estándar en términos absolutos; sin embargo, ello ocurre dentro de un rango de desempeño muy bajo. Entre los algoritmos con mejores resultados, BisectingKMeans exhibe una desviación estándar menor que Kmeans, con valores de 13,509 y 20,441, respectivamente. En conjunto, estos resultados permiten inferir que Kmeans alcanza el mejor desempeño promedio en la métrica Calinski-Harabasz, aunque con una variabilidad mayor que la de BisectingKMeans.

Tabla 63

Resultado de la prueba global para la métrica Calinski-Harabasz

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	105005,61173295535	0,0

La Tabla 63 presenta el resultado de la prueba global de Kruskal-Wallis. El valor del estadístico fue 105005,61173295535, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Calinski-Harabasz entre los algoritmos de agrupamiento analizados.

Tabla 64

Comparaciones post-hoc entre algoritmos para la métrica Calinski-Harabasz

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 64 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Calinski-Harabasz.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Calinski-Harabasz para el escenario de 4 clusters. Esta afirmación se sustenta en que obtuvo la media más alta y la mediana más elevada, criterios que en esta métrica representan una mejor calidad de agrupamiento.

BisectingKMeans mostró un rendimiento cercano, aunque con menor promedio. Por su parte, DBSCAN se ubicó claramente por debajo de ambos algoritmos. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas son estadísticamente significativas.

Métrica Inercia

Tabla 65

Prueba de normalidad de la métrica Inercia por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,401	0,000	Distribución no normal
BisectingKMeans	100000	0,221	0,000	Distribución no normal
DBSCAN	2654	0,633	0,000	Distribución no normal

La Tabla 65 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Inercia obtenidos para cada algoritmo evaluado. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Inercia correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado justifica el empleo de procedimientos no paramétricos para el análisis comparativo posterior.

Tabla 66

Prueba de homogeneidad de varianzas para la métrica Inercia

Estadístico Levene	p-valor Levene	Conclusión Levene
5528,719	0,000	No se asume homogeneidad de varianzas

La Tabla 66 muestra el resultado de la prueba de Levene. Dado que el p-valor es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Inercia no es homogénea entre los algoritmos comparados.

En conjunto con la prueba de normalidad, este resultado confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas.

Tabla 67

Estadísticos descriptivos de la métrica Inercia por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	25,286	25,047	1,071
DBSCAN	2654	66,318	63,533	3,085
Kmeans	100000	25,047	24,472	1,531

La Tabla 67 resume los principales estadísticos descriptivos de la métrica Inercia para los algoritmos evaluados en el escenario de 4 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que representan una menor suma de distancias internas dentro de los clusters y, por tanto, agrupamientos más compactos.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 25,047 y la mediana más reducida con 24,472. BisectingKMeans muestra un comportamiento cercano, con una media de 25,286 y una mediana de 25,047, mientras que DBSCAN presenta valores considerablemente mayores, con una media de 66,318 y una mediana de 63,533, lo que evidencia un desempeño claramente menos favorable.

En cuanto a la dispersión, BisectingKMeans presenta la menor desviación estándar con 1,071, seguido por Kmeans con 1,531 y DBSCAN con 3,085. En conjunto, estos resultados permiten inferir que Kmeans alcanza el mejor desempeño promedio en la métrica Inercia, mientras que BisectingKMeans presenta una estabilidad ligeramente mayor.

Tabla 68

Resultado de la prueba global para la métrica Inercia

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	103417,30983867616	0,0

La Tabla 68 presenta el resultado de la prueba global de Kruskal-Wallis. El valor del

estadístico fue 103417,30983867616, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Inercia entre los algoritmos de agrupamiento analizados.

Tabla 69

Comparaciones post-hoc entre algoritmos para la métrica Inercia

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 69 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Inercia.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Inercia para el escenario de 4 clusters. Esta afirmación se sustenta en que obtuvo la media más baja y la mediana más reducida, criterios que en esta métrica representan una mejor compactación de los clusters.

BisectingKMeans mostró un rendimiento cercano y una dispersión ligeramente menor, mientras que DBSCAN se ubicó claramente por encima de ambos algoritmos en términos de Inercia. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas son estadísticamente significativas.

4.4.4 Para 5 Clusters

Métrica Silhouette

Tabla 70

Prueba de normalidad de la métrica Silhouette por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,536	0,000	Distribución no normal
BisectingKMeans	100000	0,676	0,000	Distribución no normal
DBSCAN	5132	0,609	0,000	Distribución no normal

La Tabla 70 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Silhouette obtenidos para cada algoritmo evaluado en el escenario de 5 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Silhouette correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado es metodológicamente relevante, ya que la normalidad constituye uno de los supuestos requeridos para la aplicación de pruebas paramétricas de comparación entre grupos. Al no cumplirse este supuesto en ninguno de los algoritmos evaluados, se justifica el uso de procedimientos no paramétricos en el análisis inferencial posterior. Por tanto, la comparación global entre algoritmos debe realizarse mediante pruebas robustas frente a distribuciones no normales.

Tabla 71

Prueba de homogeneidad de varianzas para la métrica Silhouette

Estadístico Levene	p-valor Levene	Conclusión Levene
10449,934	0,000	No se asume homogeneidad de varianzas

La Tabla 71 muestra el resultado de la prueba de Levene, utilizada para evaluar la homogeneidad de varianzas entre los algoritmos analizados. Dado que el p-valor obtenido

es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Silhouette no es homogénea entre Kmeans, BisectingKMeans y DBSCAN.

Este hallazgo complementa la evidencia proporcionada por la prueba de normalidad y confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas. Asimismo, la desigualdad de varianzas sugiere que los algoritmos no solo difieren en su tendencia central, sino también en la estabilidad con la que producen sus resultados, aspecto que debe ser considerado en la interpretación comparativa.

Tabla 72

Estadísticos descriptivos de la métrica Silhouette por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	0,389	0,407	0,023
DBSCAN	5132	0,066	0,123	0,076
Kmeans	100000	0,406	0,415	0,021

La Tabla 72 resume los principales estadísticos descriptivos de la métrica Silhouette para los algoritmos evaluados en el escenario de 5 clusters. En esta métrica, los valores más altos indican una mejor calidad de segmentación, al reflejar una mayor cohesión dentro de los clusters y una mejor separación entre ellos.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más alta con un valor de 0,406 y la mediana más elevada con 0,415. BisectingKMeans muestra un rendimiento cercano, con una media de 0,389 y una mediana de 0,407, lo que evidencia un comportamiento competitivo, aunque inferior al de Kmeans. En contraste, DBSCAN presenta una media de 0,066 y una mediana de 0,123, valores considerablemente menores, lo que indica una calidad de segmentación claramente menos favorable.

En cuanto a la dispersión, Kmeans presenta la menor desviación estándar, con un valor de 0,021, seguido de BisectingKMeans con 0,023, mientras que DBSCAN exhibe una variabilidad notablemente superior con 0,076. En conjunto, estos resultados permiten inferir que Kmeans no solo alcanza el mejor desempeño promedio en la métrica Silhouette

para 5 clusters, sino que además lo hace con una mayor consistencia relativa que los demás algoritmos.

Tabla 73

Resultado de la prueba global para la métrica Silhouette

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	96392,13677861045	0,0

La Tabla 73 presenta el resultado de la prueba global de Kruskal-Wallis, seleccionada debido al incumplimiento de los supuestos de normalidad y homogeneidad de varianzas verificados previamente. El valor del estadístico fue 96392,13677861045, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Silhouette entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos no pueden atribuirse al azar, sino que reflejan comportamientos efectivamente distintos en la calidad de segmentación lograda por cada algoritmo.

Tabla 74

Comparaciones post-hoc entre algoritmos para la métrica Silhouette

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 74 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo que resulta coherente con sus valores descriptivos mucho más bajos. Asimismo, la comparación entre Kmeans y BisectingKMeans también resulta significativa, a pesar de la cercanía observada entre sus estadísticas descriptivas.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Silhouette para el escenario de 5 clusters. Esta afirmación se sustenta en que obtuvo la media más alta y la mediana más elevada, criterios que en esta métrica representan una mejor calidad de agrupamiento.

BisectingKMeans mostró un rendimiento cercano, aunque ligeramente inferior, mientras que DBSCAN se ubicó claramente por debajo de ambos algoritmos. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas entre los algoritmos son estadísticamente significativas.

Métrica Davies-Bouldin

Tabla 75

Prueba de normalidad de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,715	0,000	Distribución no normal
BisectingKMeans	100000	0,699	0,000	Distribución no normal
DBSCAN	5132	0,609	0,000	Distribución no normal

La Tabla 75 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Davies-Bouldin obtenidos para cada algoritmo evaluado en el escenario de 5 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Davies-Bouldin correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado justifica el uso de procedimientos no paramétricos para el análisis comparativo posterior, dado que no se cumple el supuesto de normalidad requerido por las pruebas paramétricas clásicas.

Tabla 76

Prueba de homogeneidad de varianzas para la métrica Davies-Bouldin

Estadístico Levene	p-valor Levene	Conclusión Levene
708,466	0,000	No se asume homogeneidad de varianzas

La Tabla 76 muestra el resultado de la prueba de Levene. Dado que el p-valor es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Davies-Bouldin no es homogénea entre los algoritmos comparados.

En conjunto con la prueba de normalidad, este resultado confirma que no se cumplen los supuestos requeridos para aplicar pruebas paramétricas, por lo que corresponde utilizar un enfoque no paramétrico en la comparación global.

Tabla 77

Estadísticos descriptivos de la métrica Davies-Bouldin por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	0,989	0,956	0,063
DBSCAN	5132	0,770	0,829	0,078
Kmeans	100000	0,925	0,916	0,067

La Tabla 77 resume los principales estadísticos descriptivos de la métrica Davies-Bouldin para los algoritmos evaluados en el escenario de 5 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que representan clusters relativamente más compactos y mejor separados.

Bajo este criterio, DBSCAN presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 0,770. Kmeans ocupa la segunda posición con una media de 0,925, mientras que BisectingKMeans presenta el valor promedio más alto, igual a 0,989, lo que indica un desempeño menos favorable desde esta métrica. En cuanto a la mediana, Kmeans registra 0,916, BisectingKMeans 0,956 y DBSCAN 0,829, reforzando la ventaja descriptiva de DBSCAN.

Respecto a la dispersión, BisectingKMeans presenta la menor desviación estándar, con un valor de 0,063, seguido de Kmeans con 0,067, mientras que DBSCAN muestra la mayor variabilidad con 0,078. En conjunto, estos resultados indican que DBSCAN alcanza el mejor desempeño promedio en la métrica Davies-Bouldin, aunque con una estabilidad menor que la observada en los otros dos algoritmos.

Tabla 78

Resultado de la prueba global para la métrica Davies-Bouldin

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	83976,1480151716	0,0

La Tabla 78 presenta el resultado de la prueba global de Kruskal-Wallis. El valor del estadístico fue 83976,1480151716, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Davies-Bouldin entre los algoritmos de agrupamiento analizados. En consecuencia, las diferencias observadas en los estadísticos descriptivos reflejan comportamientos efectivamente distintos entre algoritmos.

Tabla 79

Comparaciones post-hoc entre algoritmos para la métrica Davies-Bouldin

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 79 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que DBSCAN difiere significativamente tanto de BisectingKMeans como de Kmeans, lo que resulta coherente con su menor valor

promedio. Asimismo, la comparación entre Kmeans y BisectingKMeans también evidencia una diferencia significativa.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que DBSCAN presentó el desempeño descriptivamente más favorable en la métrica Davies-Bouldin para el escenario de 5 clusters. Esta afirmación se sustenta en que obtuvo la media más baja, criterio que en esta métrica representa una mejor calidad de agrupamiento.

Kmeans mostró un desempeño intermedio y una variabilidad menor que DBSCAN, mientras que BisectingKMeans presentó el valor promedio más alto. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas son estadísticamente significativas.

Métrica Calinski-Harabasz

Tabla 80

Prueba de normalidad de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,472	0,000	Distribución no normal
BisectingKMeans	100000	0,625	0,000	Distribución no normal
DBSCAN	5132	0,609	0,000	Distribución no normal

La Tabla 80 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Calinski-Harabasz obtenidos para cada algoritmo evaluado en el escenario de 5 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Calinski-Harabasz correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado justifica el uso de procedimientos no paramétricos para la comparación global posterior.

Tabla 81

Prueba de homogeneidad de varianzas para la métrica Calinski-Harabasz

Estadístico Levene	p-valor Levene	Conclusión Levene
1669,427	0,000	No se asume homogeneidad de varianzas

La Tabla 81 muestra el resultado de la prueba de Levene. Dado que el p-valor es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Calinski-Harabasz no es homogénea entre los algoritmos.

En conjunto con la prueba de normalidad, este resultado confirma que no se cumplen los supuestos requeridos para aplicar pruebas paramétricas clásicas.

Tabla 82

Estadísticos descriptivos de la métrica Calinski-Harabasz por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	240,879	245,286	7,720
DBSCAN	5132	29,252	42,812	17,895
Kmeans	100000	252,757	256,971	12,509

La Tabla 82 resume los principales estadísticos descriptivos de la métrica Calinski-Harabasz para los algoritmos evaluados en el escenario de 5 clusters. En esta métrica, los valores más altos indican un mejor desempeño, ya que reflejan una mejor separación entre clusters y una mayor compactación interna de los grupos formados.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más alta con un valor de 252,757 y la mediana más elevada con 256,971. BisectingKMeans muestra un rendimiento cercano, con una media de 240,879 y una mediana de 245,286, mientras que DBSCAN presenta valores considerablemente menores, con una media de 29,252 y una mediana de 42,812, lo que evidencia un desempeño claramente inferior.

En cuanto a la dispersión, BisectingKMeans presenta la menor desviación estándar con 7,720, seguido de Kmeans con 12,509, mientras que DBSCAN registra la mayor

variabilidad con 17,895. En conjunto, estos resultados permiten inferir que Kmeans alcanza el mejor desempeño promedio en la métrica Calinski-Harabasz, aunque BisectingKMeans exhibe una estabilidad mayor.

Tabla 83

Resultado de la prueba global para la métrica Calinski-Harabasz

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	96550,5351294375	0,0

La Tabla 83 presenta el resultado de la prueba global de Kruskal-Wallis. El valor del estadístico fue 96550,5351294375, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Calinski-Harabasz entre los algoritmos de agrupamiento analizados.

Tabla 84

Comparaciones post-hoc entre algoritmos para la métrica Calinski-Harabasz

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 84 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Calinski-Harabasz.

Conclusión sobre el mejor algoritmo A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la

métrica Calinski-Harabasz para el escenario de 5 clusters. Esta afirmación se sustenta en que obtuvo la media más alta y la mediana más elevada, criterios que en esta métrica representan una mejor calidad de agrupamiento.

BisectingKMeans mostró un rendimiento cercano, aunque con menor promedio, mientras que DBSCAN se ubicó claramente por debajo de ambos algoritmos. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas son estadísticamente significativas.

Métrica Inercia

Tabla 85

Prueba de normalidad de la métrica Inercia por algoritmo

Algoritmo	n	Shapiro-Wilk	p-valor	Conclusión
Kmeans	100000	0,433	0,000	Distribución no normal
BisectingKMeans	100000	0,560	0,000	Distribución no normal
DBSCAN	5132	0,609	0,000	Distribución no normal

La Tabla 85 presenta los resultados de la prueba de normalidad de Shapiro-Wilk aplicada a los valores de la métrica Inercia obtenidos para cada algoritmo evaluado en el escenario de 5 clusters. En los tres casos, el p-valor es inferior a 0,05, por lo que se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que las distribuciones de los valores de Inercia correspondientes a Kmeans, BisectingKMeans y DBSCAN no siguen una distribución normal.

Este resultado justifica el empleo de procedimientos no paramétricos para el análisis comparativo posterior.

Tabla 86

Prueba de homogeneidad de varianzas para la métrica Inercia

Estadístico Levene	p-valor Levene	Conclusión Levene
44306,765	0,000	No se asume homogeneidad de varianzas

La Tabla 86 muestra el resultado de la prueba de Levene. Dado que el p-valor es menor a 0,05, se rechaza la hipótesis nula de igualdad de varianzas. Esto indica que la dispersión de los valores de la métrica Inercia no es homogénea entre los algoritmos comparados.

En conjunto con la prueba de normalidad, este resultado confirma que no se cumplen los supuestos clásicos necesarios para la aplicación de pruebas paramétricas.

Tabla 87

Estadísticos descriptivos de la métrica Inercia por algoritmo

Algoritmo	n	Media	Mediana	DesvStd
BisectingKMeans	100000	21,021	20,738	0,529
DBSCAN	5132	55,300	47,809	9,887
Kmeans	100000	20,327	20,057	0,843

La Tabla 87 resume los principales estadísticos descriptivos de la métrica Inercia para los algoritmos evaluados en el escenario de 5 clusters. En esta métrica, los valores menores indican un mejor desempeño, ya que representan una menor suma de distancias internas dentro de los clusters y, por tanto, agrupamientos más compactos.

Bajo este criterio, Kmeans presenta el desempeño descriptivo más favorable, al registrar la media más baja con un valor de 20,327 y la mediana más reducida con 20,057. BisectingKMeans muestra un comportamiento cercano, con una media de 21,021 y una mediana de 20,738, mientras que DBSCAN presenta valores considerablemente mayores, con una media de 55,300 y una mediana de 47,809, lo que evidencia un desempeño claramente menos favorable.

En cuanto a la dispersión, BisectingKMeans presenta la menor desviación estándar con 0,529, seguido de Kmeans con 0,843, mientras que DBSCAN registra la mayor variabilidad con 9,887. En conjunto, estos resultados permiten inferir que Kmeans alcanza el mejor desempeño promedio en la métrica Inercia, mientras que BisectingKMeans presenta una estabilidad ligeramente mayor.

Tabla 88

Resultado de la prueba global para la métrica Inercia

Prueba usada	Estadístico	p-valor
Kruskal-Wallis	96373,26285131965	0,0

La Tabla 88 presenta el resultado de la prueba global de Kruskal-Wallis. El valor del estadístico fue 96373,26285131965, con un p-valor de 0,0, lo que permite rechazar la hipótesis nula de igualdad entre los algoritmos evaluados.

Desde el punto de vista inferencial, este resultado indica que existen diferencias estadísticamente significativas en los valores de la métrica Inercia entre los algoritmos de agrupamiento analizados.

Tabla 89

Comparaciones post-hoc entre algoritmos para la métrica Inercia

Algoritmo 1	Algoritmo 2	p-valor ajustado	Diferencia significativa
DBSCAN	BisectingKMeans	0,000	Sí
Kmeans	BisectingKMeans	0,000	Sí
Kmeans	DBSCAN	0,000	Sí

La Tabla 89 presenta los resultados de las comparaciones múltiples post-hoc entre pares de algoritmos. Se observa que en todos los casos el p-valor ajustado es inferior a 0,05, lo que indica diferencias estadísticamente significativas entre cada par comparado.

Estos resultados confirman que los tres algoritmos presentan comportamientos diferenciados entre sí en términos de la métrica Inercia.

A partir del análisis integral de los resultados, se concluye que Kmeans presentó el desempeño descriptivamente más favorable en la métrica Inercia para el escenario de 5 clusters. Esta afirmación se sustenta en que obtuvo la media más baja y la mediana más reducida, criterios que en esta métrica representan una mejor compactación de los clusters.

BisectingKMeans mostró un rendimiento cercano y una dispersión ligeramente menor, mientras que DBSCAN se ubicó claramente por encima de ambos algoritmos en términos

de Inercia. Asimismo, la prueba global de Kruskal-Wallis y las comparaciones post-hoc confirmaron que las diferencias observadas son estadísticamente significativas.

4.5 Evaluar la calidad de las segmentaciones obtenidas mediante métricas de clusterización

Para evaluar la calidad de las segmentaciones generadas por los algoritmos de clustering, se calcularon diversas métricas ampliamente utilizadas en la literatura: Silhouette Score, Índice de Davies-Bouldin, Índice de Calinski-Harabasz e Inercia. Estas métricas permiten analizar diferentes propiedades de las particiones obtenidas, tales como la cohesión intra-cluster, la separación entre clusters y la compactación de las agrupaciones generadas.

Las métricas fueron calculadas para cada una de las ejecuciones experimentales realizadas bajo las condiciones de exploración de hiperparámetros definidas previamente. En términos generales, las configuraciones experimentales consideradas fueron las siguientes:

- **K-Means y BisectingKMeans**: número de clusters $k \in \{2, 3, 4, 5\}$ y 100 000 semillas aleatorias diferentes para cada valor de k.
- **DBSCAN**: parámetro ϵ explorado en 100 000 valores dentro del intervalo $(0, 1]$, número mínimo de vecinos en el rango $[2, 5]$, además de la aplicación de dos restricciones adicionales: un filtro de ruido máximo permitido ($\leq 0,05\,\%$ de las observaciones) y un rango válido de número de clusters entre 2 y 5.

Bajo estas configuraciones, el número total de experimentos válidos evaluados para cada algoritmo se muestra en la Tabla 90.

En el caso de los algoritmos K-Means y BisectingKMeans, el número total de experimentos corresponde directamente a la combinación sistemática de los valores de k considerados y las semillas aleatorias evaluadas, lo que produce $4 \times 100\,000 = 400\,000$ ejecuciones para cada algoritmo.

Tabla 90

Número de experimentos válidos por algoritmo de clustering.

	KMeans	BisectingKMeans	DBSCAN
Experimentos válidos	400,000	400,000	22,982

Por el contrario, en el caso de DBSCAN el número final de experimentos válidos es considerablemente menor. Esto se debe a que muchas combinaciones de parámetros generan resultados que no cumplen las restricciones establecidas para el análisis, particularmente aquellas configuraciones que producen un número de clusters fuera del rango definido o que generan un nivel de ruido superior al umbral permitido. Como consecuencia, dichas ejecuciones son descartadas durante el proceso de filtrado, reduciendo el conjunto final a 22,982 segmentaciones válidas que cumplen simultáneamente todas las condiciones experimentales definidas.

A continuación, para cada algoritmo, se presentan las métricas obtenidas por cada ejecución.

4.5.1 Métricas para KMeans

Con el propósito de evaluar el comportamiento del algoritmo *K-Means* en el problema de segmentación planteado, se analizaron distintas métricas internas de calidad de agrupamiento para varias configuraciones del número de clusters ($k = 2, 3, 4, 5$). Para cada una de estas configuraciones se realizaron 100000 ejecuciones a partir de diferentes semillas de inicialización, lo que permite examinar tanto el desempeño promedio del algoritmo como la estabilidad de los resultados obtenidos frente a la variabilidad introducida por el proceso de inicialización.

El algoritmo *K-Means* constituye uno de los algoritmos particionales más utilizados en tareas de agrupamiento, debido a su simplicidad computacional y su capacidad para generar particiones a partir de la minimización de la variabilidad interna de los grupos. No obstante, la calidad de las soluciones obtenidas depende de manera directa del número de clusters definido y de la posición inicial de los centroides, por lo que resulta necesario

complementar su aplicación con métricas que permitan evaluar de forma sistemática la cohesión y la separación de los agrupamientos generados.

Para ello, se emplearon cuatro métricas complementarias: el índice de *Silhouette*, el índice de *Davies-Bouldin*, el índice de *Calinski-Harabasz* y la *Inercia*. En conjunto, estas métricas permiten valorar la calidad de las particiones desde diferentes perspectivas, considerando la cohesión interna de los clusters, la separación entre grupos y la compactación global de las soluciones obtenidas. En las subsecciones siguientes se presenta el análisis descriptivo de cada una de estas métricas, acompañado de tablas resumen y diagramas de caja que permiten comparar el comportamiento del algoritmo según el número de clusters evaluado.

Índice de Silhouette para KMeans

Tabla 91

Resumen estadístico del índice de Silhouette para K-Means según número de clusters

Estadístico descriptivo	**Número de clusters**			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	0,364097	0,364938	0,394149	0,406079
std	0,002398	0,009481	0,015978	0,021457
min	0,177615	0,261283	0,216452	0,238019
25 %	0,36413	0,367281	0,399393	0,41391
50 %	0,36413	0,367281	0,399393	0,415416
75 %	0,36413	0,36738	0,399393	0,416158
max	0,36413	0,36738	0,399393	0,417281

En la Tabla 91 se presenta el resumen estadístico del índice de *Silhouette* obtenido con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). Los resultados muestran que en todas las configuraciones se consideraron 100000 ejecuciones, lo que proporciona una base amplia para analizar el comportamiento de esta métrica de evaluación interna.

De manera general, se observa que el valor promedio del índice de *Silhouette* tiende a

incrementarse conforme aumenta el número de clusters, pasando de 0,364097 para $k = 2$ a 0,364938 para $k = 3$, 0,394149 para $k = 4$ y 0,406079 para $k = 5$. Este comportamiento sugiere que, dentro del rango analizado, las configuraciones con mayor número de clusters logran una mejor cohesión interna y una mayor separación entre grupos. En particular, $k = 5$ presenta la media más alta, seguido de $k = 4$, lo que evidencia un desempeño más favorable bajo esta métrica.

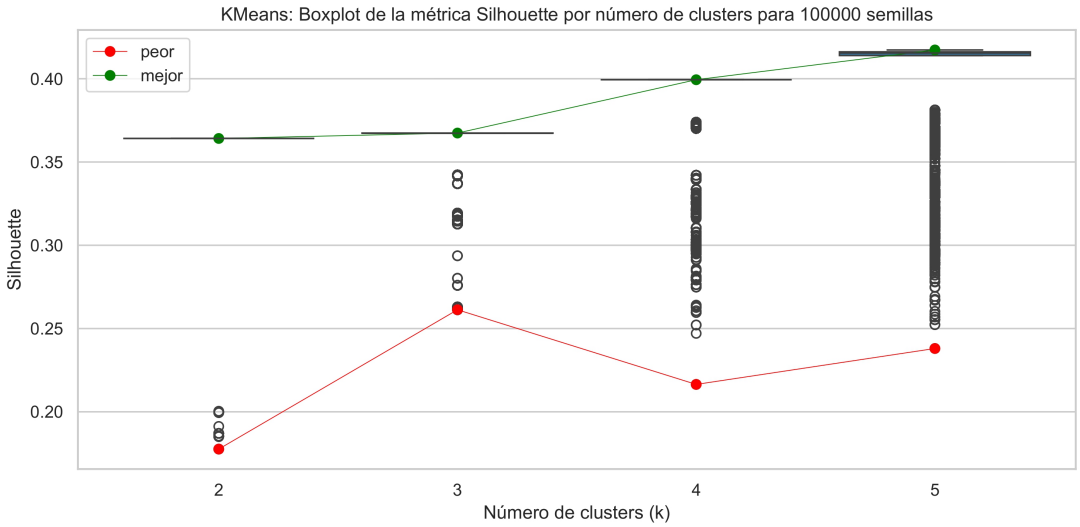
Asimismo, los estadísticos de posición refuerzan esta tendencia. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 0,36413, lo que indica una alta concentración de ejecuciones en torno a este valor. En $k = 3$, la mediana alcanza 0,367281 y el tercer cuartil asciende a 0,36738, mostrando una ligera mejora respecto a $k = 2$. Para $k = 4$, los cuartiles se sitúan en 0,399393, mientras que en $k = 5$ los valores aumentan hasta 0,413910 en el primer cuartil, 0,415416 en la mediana y 0,416158 en el tercer cuartil, confirmando que esta última configuración concentra una mayor proporción de ejecuciones en niveles superiores del índice.

Por otro lado, la dispersión de los resultados también muestra diferencias entre configuraciones. La desviación estándar aumenta progresivamente de 0,002398 en $k = 2$ a 0,021457 en $k = 5$, lo que indica una mayor variabilidad a medida que se incrementa el número de clusters. Sin embargo, esta variabilidad se presenta en rangos más altos del índice, por lo que no contradice el mejor desempeño global de $k = 4$ y $k = 5$. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 2$ con 0,177615, mientras que $k = 3$, $k = 4$ y $k = 5$ presentan mínimos de 0,261283, 0,216452 y 0,238019, respectivamente. A su vez, el valor máximo más alto se registra en $k = 5$ con 0,417281.

En conjunto, la Tabla 91 indica que las configuraciones con $k = 4$ y especialmente $k = 5$ ofrecen los resultados más favorables según el índice de *Silhouette*, al presentar mayores valores promedio, medianos y cuartiles, así como una concentración importante de ejecuciones en rangos superiores de la métrica.

Figura 16

Índice de Silhouette para KMeans



En la Figura 16 se presenta un *boxplot* del índice de *Silhouette* obtenido con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores de la métrica *Silhouette*. Asimismo, se destacan explícitamente los valores de peor desempeño en color rojo y los de mejor desempeño en color verde para cada valor de k.

De manera general, la figura permite apreciar que los valores del índice tienden a desplazarse hacia niveles superiores conforme aumenta el número de clusters. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 0,364, mientras que en $k = 3$ la distribución central se sitúa aproximadamente en 0,367. Para $k = 4$, la caja se concentra en torno a 0,399, y en $k = 5$ el rango intercuartílico se desplaza a valores más altos, aproximadamente entre 0,414 y 0,416. Este comportamiento es consistente con lo observado en la Tabla 91 y sugiere una mejora progresiva en la cohesión interna y separación entre grupos al incrementar el número de clusters.

En relación con los valores extremos, se observa que los mejores resultados aumentan desde aproximadamente 0,364 en $k = 2$ hasta 0,417 en $k = 5$, mientras que los valores

mínimos muestran diferencias importantes entre configuraciones, con registros cercanos a 0,178 para $k = 2$, 0,261 para $k = 3$, 0,216 para $k = 4$ y 0,238 para $k = 5$. Esto evidencia que, aunque las configuraciones con mayor número de clusters alcanzan mejores niveles de *Silhouette*, también existen ejecuciones menos favorables asociadas al proceso de inicialización.

Asimismo, la figura muestra una presencia visible de valores atípicos e intermedios por debajo de las cajas, especialmente en $k = 4$ y $k = 5$, donde se aprecian múltiples ejecuciones distribuidas aproximadamente entre 0,250 y 0,380. En contraste, $k = 2$ presenta una caja sumamente concentrada alrededor de 0,364, reflejando menor variabilidad en la mayoría de las ejecuciones, aunque en un nivel inferior respecto a las configuraciones con más clusters. En conjunto, la figura refuerza que las particiones con $k = 4$ y especialmente $k = 5$ constituyen las alternativas más prometedoras dentro del rango analizado según el índice de *Silhouette*.

Índice de Davies-Bouldin para KMeans

Tabla 92

Resumen estadístico del índice de Davies-Bouldin para K-Means según número de clusters

Estadístico descriptivo	Número de clusters			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	1,161754	1,047634	0,970444	0,925098
std	0,011556	0,051141	0,032840	0,067300
min	1,161595	1,009716	0,942327	0,870867
25 %	1,161595	1,009716	0,964293	0,874061
50 %	1,161595	1,055058	0,964293	0,916451
75 %	1,161595	1,055058	0,964293	0,917853
max	2,141318	1,515715	1,487433	1,366648

En la Tabla 92 se presenta el resumen estadístico del índice de *Davies-Bouldin* obtenido con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$).

En todas las configuraciones se registraron 100000 ejecuciones, lo que proporciona una base amplia para evaluar el comportamiento de esta métrica de calidad interna.

De manera general, se observa que el valor promedio del índice disminuye conforme aumenta el número de clusters, pasando de 1,161754 para $k = 2$ a 1,047634 para $k = 3$, 0,970444 para $k = 4$ y 0,925098 para $k = 5$. Dado que en el índice de *Davies-Bouldin* los valores más bajos indican una mejor separación y compactación de los clusters, estos resultados sugieren una mejora progresiva en la calidad del agrupamiento a medida que se incrementa el número de clusters.

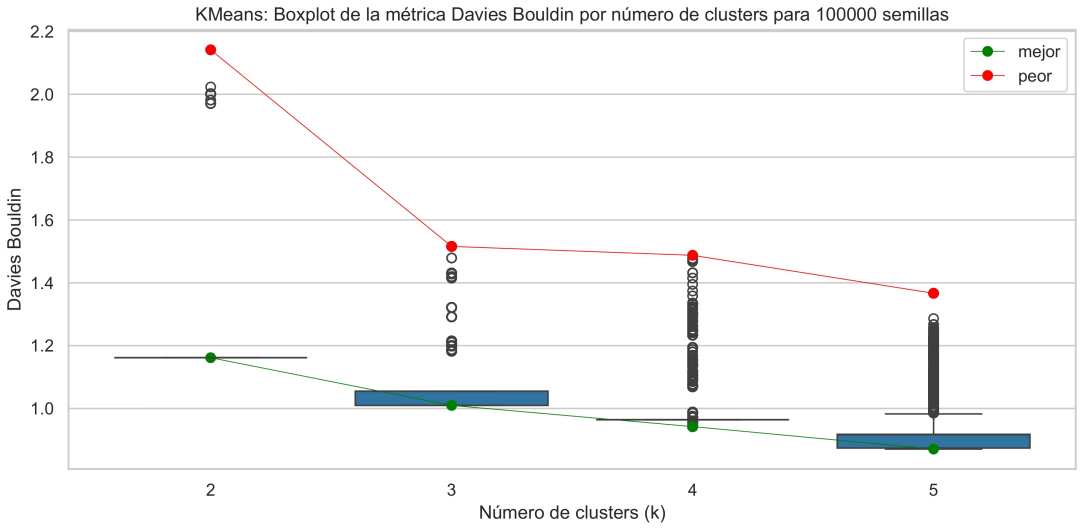
Asimismo, los estadísticos de posición refuerzan esta tendencia. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 1,161595, lo que evidencia una fuerte concentración de las ejecuciones en torno a ese valor. En $k = 3$, el rango intercuartílico se sitúa aproximadamente entre 1,010 y 1,055, con una mediana de 1,055058. Para $k = 4$, los cuartiles se concentran alrededor de 0,964, mientras que en $k = 5$ el rango intercuartílico desciende aproximadamente de 0,874 a 0,918, con una mediana de 0,916451. En consecuencia, $k = 4$ y especialmente $k = 5$ concentran la mayor parte de sus ejecuciones en niveles inferiores del índice.

Por otro lado, la dispersión de los resultados presenta diferencias entre configuraciones. La desviación estándar es de 0,011556 para $k = 2$, 0,051141 para $k = 3$, 0,032840 para $k = 4$ y 0,067300 para $k = 5$. Aunque $k = 5$ alcanza los mejores valores centrales, también muestra una mayor variabilidad global, lo que sugiere la presencia de ejecuciones atípicas. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 5$ con 0,870867, seguido de $k = 4$ con 0,942327, mientras que el máximo más alto se registra en $k = 2$ con 2,141318.

En conjunto, la Tabla 92 indica que las configuraciones con $k = 4$ y especialmente $k = 5$ ofrecen los resultados más favorables según el índice de *Davies-Bouldin*, al presentar menores valores promedio, medianos y cuartiles, lo que evidencia una mejor calidad del agrupamiento dentro del rango analizado.

Figura 17

Índice de Davies Bouldin para KMeans



En la Figura 17 se presenta un *boxplot* del índice de *Davies-Bouldin* obtenido con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores del índice *Davies-Bouldin*. Asimismo, se destacan los valores de mejor desempeño en color verde y los de peor desempeño en color rojo para cada configuración de k.

De manera general, la figura permite apreciar un desplazamiento progresivo de las distribuciones hacia valores inferiores conforme aumenta el número de clusters. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 1,162, mientras que en $k = 3$ la distribución central se sitúa aproximadamente entre 1,010 y 1,055. Para $k = 4$, la caja se concentra en torno a 0,964, y en $k = 5$ el rango intercuartílico se desplaza a valores aún menores, aproximadamente entre 0,874 y 0,918. Este comportamiento es consistente con lo observado en la Tabla 92 y sugiere una mejora en la separación y compacidad de los grupos al incrementar el número de clusters.

En relación con los valores extremos, la figura muestra que los mejores resultados disminuyen desde aproximadamente 1,162 en $k = 2$ hasta 0,871 en $k = 5$, mientras que

los peores valores también se reducen de 2,141 en $k = 2$ a 1,366 en $k = 5$. Esto evidencia que, aunque persisten ejecuciones menos favorables, las configuraciones con un mayor número de clusters tienden a producir soluciones globalmente más adecuadas bajo esta métrica.

Asimismo, la presencia de puntos atípicos e intermedios refleja la sensibilidad del algoritmo a la semilla de inicialización. En $k = 2$, la caja aparece sumamente concentrada alrededor de 1,162, aunque con algunos valores extremos altos próximos a 2,000 y superiores. En $k = 3$, $k = 4$ y $k = 5$, se observan valores atípicos distribuidos por encima de las cajas, alcanzando aproximadamente rangos de 1,180 a 1,516, 1,070 a 1,487 y 0,990 a 1,367, respectivamente. No obstante, las cajas correspondientes a $k = 4$ y $k = 5$ se ubican claramente en niveles inferiores respecto a $k = 2$ y $k = 3$, lo que refuerza la mejora visual del agrupamiento para estas configuraciones.

En conjunto, la Figura 17 confirma que las particiones con $k = 4$ y especialmente $k = 5$ constituyen las alternativas más favorables dentro del rango analizado según el índice de *Davies-Bouldin*, al concentrar sus ejecuciones en valores más bajos de la métrica.

Índice de Calinski-Harabasz

Tabla 93

Resumen estadístico del índice de Calinski-Harabasz para K-Means según número de clusters

Estadístico descriptivo	**Número de clusters**			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	256,230651	222,687640	250,011125	252,756683
std	2,393922	5,466927	20,440512	12,508613
min	77,238705	168,030963	147,794214	146,007492
25 %	256,263647	220,481424	257,696268	256,919736
50 %	256,263647	225,861713	257,696268	256,970895
75 %	256,263647	225,861713	257,696268	257,393935
max	256,263647	225,861713	257,696268	258,654589

En la Tabla 93 se presenta el resumen estadístico del índice de *Calinski-Harabasz* obtenido con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). En todas las configuraciones se registraron 100000 ejecuciones, lo que permite analizar de manera consistente el comportamiento de esta métrica de evaluación interna.

De manera general, se observa que los valores promedio del índice se mantienen elevados en todas las configuraciones, aunque con diferencias según el número de clusters. En particular, $k = 2$ presenta una media de 256,231, $k = 3$ de 222,688, $k = 4$ de 250,011 y $k = 5$ de 252,757. Dado que en el índice de *Calinski-Harabasz* valores más altos indican una mejor separación entre clusters y una mayor cohesión interna, estos resultados sugieren un mejor desempeño general para $k = 2$, $k = 4$ y $k = 5$, mientras que $k = 3$ muestra un nivel promedio inferior.

Asimismo, los estadísticos de posición refuerzan este comportamiento. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 256,264, lo que indica una fuerte concentración de las ejecuciones en torno a este valor. En $k = 3$, el rango intercuartílico se ubica aproximadamente entre 220,481 y 225,862, con una mediana de 225,862. Para $k = 4$, los cuartiles se concentran en 257,696, mientras que en $k = 5$ el rango intercuartílico se sitúa aproximadamente entre 256,920 y 257,394, con una mediana de 256,971. En consecuencia, $k = 4$ y $k = 5$ concentran una parte importante de sus ejecuciones en niveles superiores del índice.

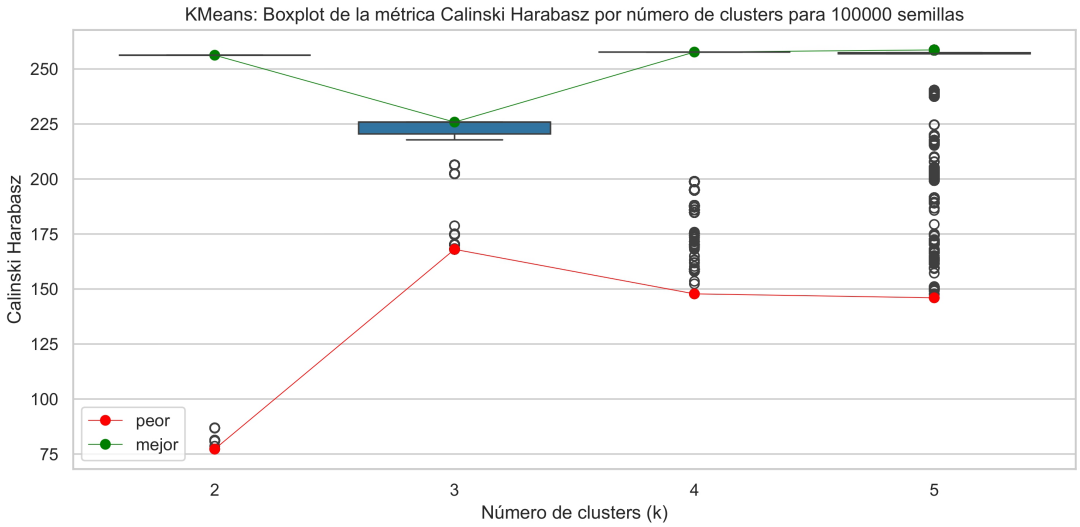
Por otro lado, la dispersión de los resultados presenta diferencias importantes entre configuraciones. La desviación estándar es de 2,394 para $k = 2$, 5,467 para $k = 3$, 20,441 para $k = 4$ y 12,509 para $k = 5$. Esto indica una mayor variabilidad en $k = 4$ y $k = 5$, asociada a la presencia de ejecuciones con valores considerablemente inferiores al comportamiento central. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 2$ con 77,239, mientras que $k = 3$, $k = 4$ y $k = 5$ presentan mínimos de 168,031, 147,794 y 146,007, respectivamente. A su vez, el valor máximo más alto se registra en $k = 5$ con 258,655.

En conjunto, la Tabla 93 indica que las configuraciones con $k = 4$ y $k = 5$ ofrecen los resultados más favorables según el índice de *Calinski-Harabasz*, al presentar valores

centrales elevados y máximos superiores, aunque con una variabilidad más marcada en comparación con $k = 2$.

Figura 18

Índice de Calinski Harabasz para KMeans



En la Figura 18 se presenta un *boxplot* del índice de *Calinski-Harabasz* obtenido con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores del índice *Calinski-Harabasz*. Asimismo, se destacan los valores de peor desempeño en color rojo y los de mejor desempeño en color verde para cada configuración de k.

De manera general, la figura permite apreciar que las distribuciones centrales se mantienen en niveles altos para todas las configuraciones, aunque con diferencias entre ellas. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 256,264, mientras que en $k = 3$ la caja se concentra aproximadamente entre 220,481 y 225,862. Para $k = 4$, la distribución central se sitúa en torno a 257,696, y en $k = 5$ el rango intercuartílico se desplaza aproximadamente entre 256,920 y 257,394. Este comportamiento es consistente con lo observado en la Tabla 93 y muestra que $k = 4$ y $k = 5$ concentran sus ejecuciones en valores superiores del índice.

En relación con los valores extremos, la figura muestra que los mejores resultados alcanzan aproximadamente 256,264 en $k = 2$, 225,862 en $k = 3$, 257,696 en $k = 4$ y 258,655 en $k = 5$. Por su parte, los valores mínimos se sitúan alrededor de 77,239 para $k = 2$, 168,031 para $k = 3$, 147,794 para $k = 4$ y 146,007 para $k = 5$. Dado que en esta métrica valores más altos indican una mejor separación entre clusters y mayor cohesión interna, estos resultados sugieren un desempeño más favorable para las configuraciones con $k = 4$ y $k = 5$.

Asimismo, la figura evidencia la presencia de numerosos valores atípicos e intermedios por debajo de las cajas, especialmente en $k = 4$ y $k = 5$, donde se observan ejecuciones distribuidas aproximadamente entre 148,000 y 240,000. En contraste, $k = 2$ presenta una caja sumamente concentrada alrededor de 256,264, aunque con algunos valores bajos aislados cercanos a 77,000 y 87,000. En $k = 3$, la dispersión también resulta visible, con ejecuciones que descienden aproximadamente hasta 168,000. Esta distribución pone de manifiesto la sensibilidad del algoritmo a la semilla de inicialización y explica la variabilidad observada en la Tabla 93.

En conjunto, la Figura 18 confirma que las configuraciones con $k = 4$ y especialmente $k = 5$ constituyen las alternativas más prometedoras dentro del rango analizado según el índice de *Calinski-Harabasz*, al concentrar sus mejores ejecuciones en los niveles más altos de la métrica, aunque con variabilidad asociada al proceso de inicialización.

Inercia

Tabla 94

Resumen estadístico de la inercia para K-Means según número de clusters

Estadístico descriptivo	Número de clusters			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	43,953039	34,055973	25,046611	20,326892
std	0,219005	0,481678	1,530865	0,842955
min	43,950021	33,795052	24,472396	19,962080
25 %	43,950021	33,795052	24,472396	20,032740
50 %	43,950021	33,795052	24,472396	20,056564
75 %	43,950021	34,229026	24,472396	20,059449
max	60,508559	39,127189	34,085055	29,149420

En la Tabla 94 se presenta el resumen estadístico de la métrica de *Inercia* obtenida con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). En todas las configuraciones se registraron 100000 ejecuciones, lo que permite analizar de manera consistente el comportamiento de esta métrica interna de compactación.

De manera general, se observa que los valores promedio de la inercia disminuyen progresivamente conforme aumenta el número de clusters, pasando de 43,953 para $k = 2$ a 34,056 para $k = 3$, 25,047 para $k = 4$ y 20,327 para $k = 5$. Dado que en esta métrica valores más bajos indican una mayor compactación de los clusters, estos resultados sugieren que las configuraciones con mayor número de grupos tienden a producir particiones más cohesionadas.

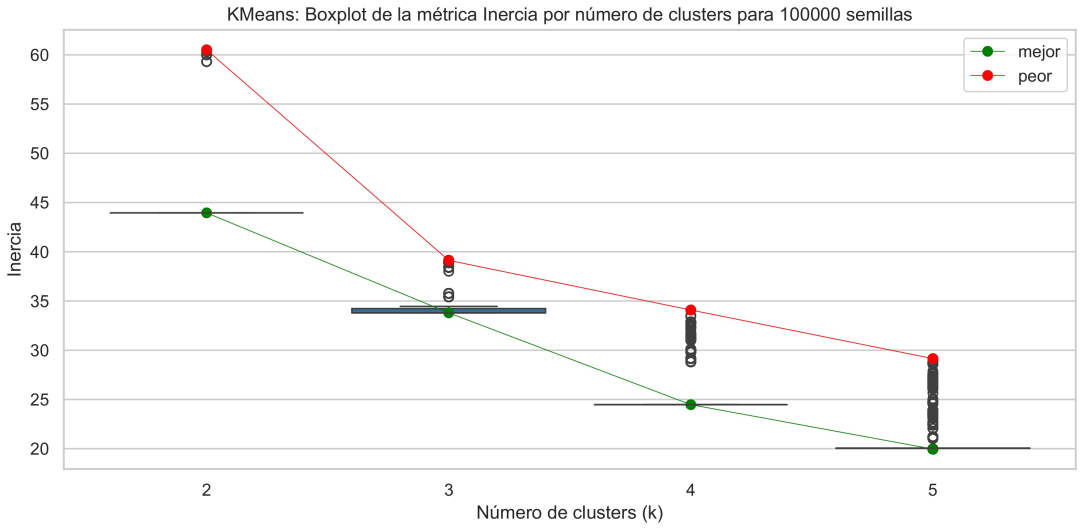
Asimismo, los estadísticos de posición refuerzan esta tendencia decreciente. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 43,950, lo que evidencia una fuerte concentración de las ejecuciones en torno a ese valor. En $k = 3$, la distribución central se ubica aproximadamente entre 33,795 y 34,229, con una mediana de 33,795. Para $k = 4$, los cuartiles se concentran en 24,472, mientras que en $k = 5$ el rango intercuartílico se sitúa aproximadamente entre 20,033 y 20,059, con una mediana de 20,057. En consecuencia, $k = 4$ y especialmente $k = 5$ concentran sus ejecuciones en niveles inferiores de inercia.

Por otro lado, la dispersión de los resultados presenta diferencias entre configuraciones. La desviación estándar es de 0,219 para $k = 2$, 0,482 para $k = 3$, 1,531 para $k = 4$ y 0,843 para $k = 5$. Esto indica una mayor variabilidad en $k = 4$, asociada a la presencia de ejecuciones alejadas del comportamiento central. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 5$ con 19,962, seguido de $k = 4$ con 24,472, mientras que el valor máximo más alto se registra en $k = 2$ con 60,509.

En conjunto, la Tabla 94 indica que las configuraciones con $k = 4$ y especialmente $k = 5$ ofrecen los resultados más favorables según la métrica de *Inercia*, al presentar menores valores promedio, medianos y cuartiles, lo que evidencia una mayor compactación de los grupos dentro del rango analizado.

Figura 19

Inercia para KMeans



En la Figura 19 se presenta un *boxplot* de la métrica de *Inercia* obtenida con el algoritmo *K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores de la inercia. Asimismo, se destacan los valores de mejor desempeño en color verde y los de peor desempeño en color rojo para cada configuración de k.

De manera general, la figura permite apreciar un desplazamiento progresivo de las distribuciones hacia valores inferiores conforme aumenta el número de clusters. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 43,950, mientras que en $k = 3$ la distribución central se sitúa aproximadamente entre 33,795 y 34,229. Para $k = 4$, la caja se concentra en torno a 24,472, y en $k = 5$ el rango intercuartílico se desplaza a valores aún menores, aproximadamente entre 20,033 y 20,059. Este comportamiento es consistente con lo observado en la Tabla 14 y sugiere una mayor compactación de los clusters al incrementar el número de grupos.

En relación con los valores extremos, la figura muestra que los mejores resultados disminuyen desde aproximadamente 43,950 en $k = 2$ hasta 19,962 en $k = 5$, mientras que los peores valores también se reducen de 60,509 en $k = 2$ a 29,149 en $k = 5$. Esto evidencia que, dentro del rango analizado, las configuraciones con un mayor número de clusters tienden a generar particiones con menor dispersión interna.

Asimismo, la figura pone de manifiesto la presencia de valores atípicos e intermedios, especialmente en $k = 4$ y $k = 5$, donde se observan múltiples ejecuciones por encima de las cajas, distribuidas aproximadamente entre 29,000 y 34,000 para $k = 4$, y entre 21,000 y 29,000 para $k = 5$. En contraste, $k = 2$ presenta una caja sumamente concentrada alrededor de 43,950, aunque con algunos valores extremos altos próximos a 59,000 y 60,500. En $k = 3$, también se aprecian ejecuciones atípicas superiores, que alcanzan aproximadamente hasta 39,127. Esta distribución confirma la sensibilidad del algoritmo a la semilla de inicialización y explica la variabilidad observada en la Tabla 94.

En conjunto, la Figura 19 confirma que las configuraciones con $k = 4$ y especialmente $k = 5$ constituyen las alternativas más favorables dentro del rango analizado según la métrica de *Inercia*, al concentrar sus ejecuciones en los niveles más bajos de la métrica y, por tanto, en particiones más compactas.

4.5.2 Métricas para Bisecting KMeans

Con el propósito de evaluar el comportamiento del algoritmo *Bisecting K-Means* en el problema de segmentación abordado, se analizaron distintas métricas internas de calidad

de agrupamiento para varias configuraciones del número de clusters ($k = 2, 3, 4, 5$). Para cada una de estas configuraciones se realizaron 100000 ejecuciones a partir de diferentes semillas de inicialización, lo que permite examinar no solo el desempeño promedio del algoritmo, sino también la estabilidad de sus resultados frente a la variabilidad introducida por el proceso de partición iterativa.

A diferencia de *K-Means* convencional, el algoritmo *Bisecting K-Means* construye los agrupamientos de manera jerárquica divisiva, generando sucesivas bisecciones hasta alcanzar el número de clusters requerido. Esta característica hace especialmente relevante el análisis de métricas internas, ya que la calidad de la solución final depende tanto del número de grupos definido como de las decisiones de partición adoptadas en cada etapa del procedimiento. En ese sentido, evaluar múltiples ejecuciones por configuración permite identificar tendencias robustas en el comportamiento del algoritmo.

Para ello, se emplearon cuatro métricas complementarias: el índice de *Silhouette*, el índice de *Davies-Bouldin*, el índice de *Calinski-Harabasz* y la *Inercia*. En conjunto, estas métricas permiten valorar la calidad de los agrupamientos desde diferentes perspectivas, considerando la cohesión interna de los clusters, la separación entre grupos y la compactación global de las particiones obtenidas. En las subsecciones siguientes se presenta el análisis descriptivo de cada una de estas métricas, acompañado de tablas resumen y diagramas de caja que permiten comparar el comportamiento del algoritmo según el número de clusters evaluado.

Índice de Silhouette para Bisecting Kmeans

Tabla 95

Resumen estadístico del índice de Silhouette para Bisecting K-Means según número de clusters

Estadístico descriptivo	Número de clusters			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	0,363164	0,365115	0,391331	0,389082
std	0,012607	0,015645	0,014995	0,023198
min	0,168480	0,172667	0,205937	0,222374
25 %	0,364130	0,367852	0,394270	0,362740
50 %	0,364130	0,367852	0,394270	0,407298
75 %	0,364130	0,368659	0,394953	0,407438
max	0,364130	0,369551	0,395705	0,416078

En la Tabla 95 se presenta el resumen estadístico del índice de *Silhouette* obtenido con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). En todas las configuraciones se registraron 100000 ejecuciones, lo que permite realizar una comparación consistente del comportamiento de esta métrica de evaluación interna.

De manera general, se observa que los valores promedio del índice de *Silhouette* varían entre 0,363 y 0,391. En particular, $k = 4$ presenta la media más alta con 0,391331, seguido de $k = 5$ con 0,389082, mientras que $k = 3$ y $k = 2$ alcanzan medias de 0,365115 y 0,363164, respectivamente. Dado que valores más altos de esta métrica indican una mejor cohesión interna y una mayor separación entre grupos, estos resultados sugieren un mejor desempeño general para las configuraciones con $k = 4$ y $k = 5$.

Asimismo, los estadísticos de posición permiten identificar diferencias importantes en la distribución de los resultados. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 0,364130, lo que indica una fuerte concentración de las ejecuciones en torno a ese valor. En $k = 3$, el rango intercuartílico se sitúa aproximadamente entre 0,367852 y 0,368659, con una mediana de 0,367852, mostrando una distribución central relativamente estable. Para $k = 4$, los cuartiles se ubican entre 0,394270 y 0,394953, con una mediana de 0,394270, lo

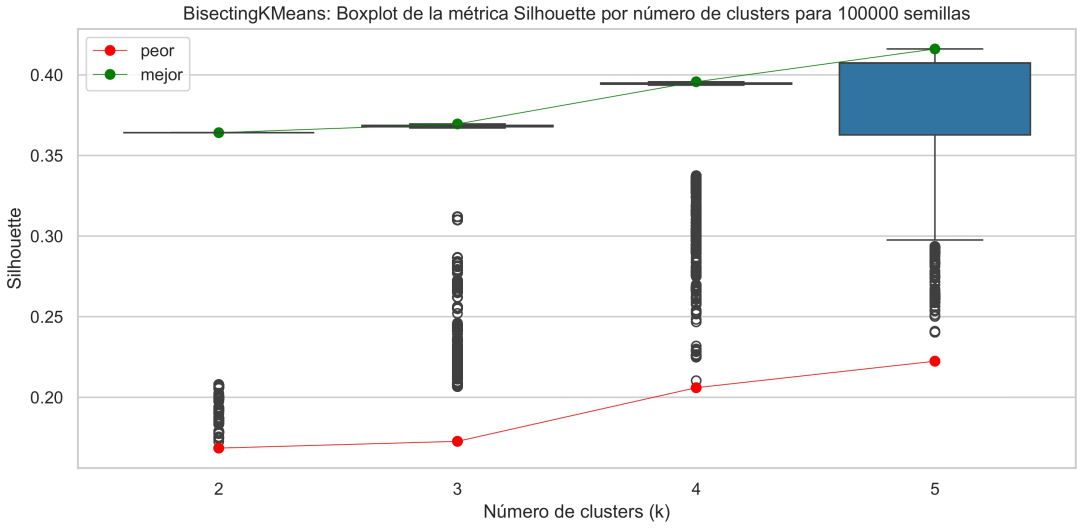
que evidencia una concentración en niveles superiores del índice. En cambio, para $k = 5$ se observa una mayor heterogeneidad, con un primer cuartil de 0,362740 y una mediana de 0,407298, lo que sugiere una distribución más dispersa y posiblemente asimétrica.

Por otro lado, la dispersión de los resultados también presenta diferencias entre configuraciones. La desviación estándar es de 0,012607 para $k = 2$, 0,015645 para $k = 3$, 0,014995 para $k = 4$ y 0,023198 para $k = 5$. Esto indica que $k = 5$ presenta la mayor variabilidad global, mientras que $k = 2$ muestra la menor dispersión relativa. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 2$ con 0,168480, seguido de 0,172667 para $k = 3$, 0,205937 para $k = 4$ y 0,222374 para $k = 5$. A su vez, el valor máximo más alto se registra en $k = 5$ con 0,416078.

En conjunto, la Tabla 95 indica que $k = 4$ y $k = 5$ ofrecen los resultados más favorables para *Bisecting K-Means* según el índice de *Silhouette*, ya que presentan los mayores valores promedio y centrales de la métrica. No obstante, $k = 5$ exhibe una mayor variabilidad en comparación con $k = 4$, por lo que esta última configuración muestra un comportamiento más estable dentro del rango analizado.

Figura 20

Índice de Silhouette para Bisecting KMeans



En la Figura 20 se presenta un *boxplot* del índice de *Silhouette* obtenido con el

algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores de la métrica *Silhouette*. Asimismo, se destacan los valores de peor desempeño en color rojo y los de mejor desempeño en color verde para cada configuración de k.

De manera general, la figura permite apreciar que los valores del índice se desplazan hacia niveles superiores al pasar de $k = 2$ y $k = 3$ a $k = 4$ y $k = 5$. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 0,364, mientras que en $k = 3$ la distribución central se sitúa aproximadamente entre 0,368 y 0,369. Para $k = 4$, la caja se concentra en torno a 0,394, lo que refleja un comportamiento más favorable y estable. En $k = 5$, aunque la mediana asciende aproximadamente a 0,407, la distribución resulta más heterogénea, con un primer cuartil cercano a 0,363 y un tercer cuartil de 0,407, evidenciando una mayor dispersión.

En relación con los valores extremos, la figura muestra que los mínimos aumentan desde aproximadamente 0,168 en $k = 2$ hasta 0,222 en $k = 5$, lo que indica que las configuraciones con más clusters tienden a evitar resultados extremadamente bajos. Asimismo, el valor máximo más alto se registra en $k = 5$, con aproximadamente 0,416, seguido de $k = 4$ con 0,396. Esto sugiere que $k = 5$ puede alcanzar las mejores ejecuciones individuales, aunque no necesariamente presenta el comportamiento promedio más favorable.

Por otro lado, la presencia de valores atípicos e intermedios pone de manifiesto la sensibilidad del algoritmo a la semilla de inicialización. En $k = 2$ y $k = 3$, las cajas se mantienen relativamente concentradas, con menor variabilidad global. En $k = 4$, la distribución se ubica en niveles superiores con una dispersión moderada, mientras que en $k = 5$ se aprecia la mayor variabilidad, coherente con su desviación estándar más alta. Esta diferencia sugiere que $k = 4$ ofrece un desempeño más estable, mientras que $k = 5$ combina resultados muy favorables con una mayor heterogeneidad entre ejecuciones.

En conjunto, la Figura 20 indica que las configuraciones con $k = 4$ y $k = 5$ continúan siendo las alternativas más prometedoras para *Bisecting K-Means* según el índice de

Silhouette; sin embargo, $k = 4$ muestra un comportamiento más consistente, mientras que $k = 5$ destaca por alcanzar los mejores valores máximos, aunque con mayor variabilidad.

Índice de Davies-Bouldin para Bisecting KMeans

Tabla 96

Resumen estadístico del índice de Davies-Bouldin para Bisecting K-Means según número de clusters

Estadístico descriptivo	Número de clusters			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	1,166480	1,054762	1,001855	0,988918
std	0,063714	0,062314	0,049562	0,062846
min	1,161595	1,038278	0,987780	0,923245
25 %	1,161595	1,040470	0,990394	0,935257
50 %	1,161595	1,042623	0,992915	0,956305
75 %	1,161595	1,042623	0,992915	1,056082
max	2,164877	1,883360	1,676130	1,858093

En la Tabla 96 se presenta el resumen estadístico del índice de *Davies-Bouldin* obtenido con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). En todas las configuraciones se registraron 100000 ejecuciones, lo que permite comparar de manera consistente el comportamiento de esta métrica de evaluación interna.

De manera general, se observa que el valor promedio del índice disminuye conforme aumenta el número de clusters, pasando de 1,166480 para $k = 2$ a 1,054762 para $k = 3$, 1,001855 para $k = 4$ y 0,988918 para $k = 5$. Dado que en el índice de *Davies-Bouldin* valores más bajos indican una mejor separación y compactación de los clusters, estos resultados sugieren una mejora progresiva en la calidad del agrupamiento a medida que se incrementa el número de clusters.

Asimismo, los estadísticos de posición refuerzan esta tendencia. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 1,161595, lo que evidencia una fuerte concentración de las ejecuciones en torno a ese valor. En $k = 3$, el rango intercuartílico se

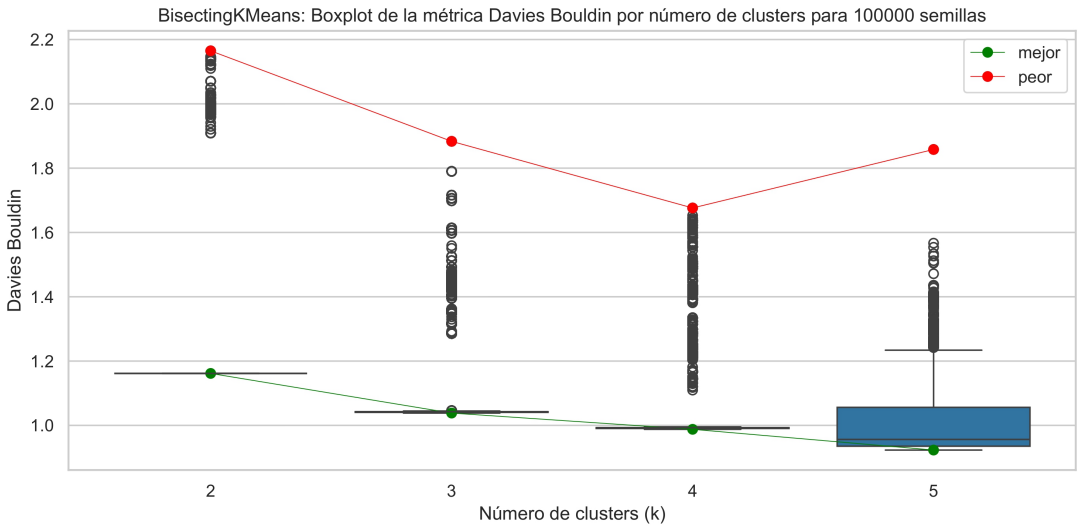
sitúa aproximadamente entre 1,040470 y 1,042623, con una mediana de 1,042623. Para $k = 4$, la distribución central se concentra entre 0,990394 y 0,992915, con una mediana de 0,992915. Por su parte, en $k = 5$ el rango intercuartílico es más amplio, aproximadamente entre 0,935257 y 1,056082, con una mediana de 0,956305, lo que sugiere una mayor heterogeneidad en comparación con $k = 4$.

Por otro lado, la dispersión de los resultados presenta diferencias entre configuraciones. La desviación estándar es de 0,063714 para $k = 2$, 0,062314 para $k = 3$, 0,049562 para $k = 4$ y 0,062846 para $k = 5$. Esto indica que $k = 4$ presenta la menor variabilidad global, mientras que $k = 5$, aunque alcanza los mejores valores promedio y mínimo, muestra una dispersión más elevada. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 5$ con 0,923245, seguido de $k = 4$ con 0,987780, mientras que el valor máximo más alto se registra en $k = 2$ con 2,164877.

En conjunto, la Tabla 96 indica que las configuraciones con $k = 4$ y especialmente $k = 5$ ofrecen los resultados más favorables según el índice de *Davies-Bouldin*. No obstante, $k = 4$ exhibe un comportamiento más estable, mientras que $k = 5$ combina mejores valores centrales y extremos mínimos con una mayor variabilidad entre ejecuciones.

Figura 21

Índice de Davies Bouldin para Bisecting KMeans



En la Figura 21 se presenta un *boxplot* del índice de *Davies-Bouldin* obtenido con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores del índice *Davies-Bouldin*. Asimismo, se destacan los valores de mejor desempeño en color verde y los de peor desempeño en color rojo para cada configuración de k.

De manera general, la figura permite apreciar un desplazamiento progresivo de las distribuciones hacia valores inferiores conforme aumenta el número de clusters. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 1,162, mientras que en $k = 3$ la distribución central se sitúa aproximadamente entre 1,040 y 1,043. Para $k = 4$, la caja se concentra en torno a 0,990 y 0,993, reflejando un comportamiento más favorable y estable. En $k = 5$, aunque la mediana desciende aproximadamente a 0,956, la caja es más amplia y se extiende aproximadamente entre 0,935 y 1,056, lo que evidencia una mayor dispersión. Este comportamiento es consistente con lo observado en la Tabla 96.

En relación con los valores extremos, la figura muestra que los mejores resultados disminuyen desde aproximadamente 1,162 en $k = 2$ hasta 0,923 en $k = 5$, pasando por 1,038 en $k = 3$ y 0,988 en $k = 4$. Por otro lado, los peores valores también se reducen de 2,165 en $k = 2$ a 1,883 en $k = 3$ y 1,676 en $k = 4$, aunque en $k = 5$ vuelven a incrementarse hasta aproximadamente 1,858. Esto indica que, aunque las configuraciones con mayor número de clusters mejoran en términos de desempeño central, todavía pueden producirse ejecuciones menos favorables, especialmente en $k = 5$.

Asimismo, la figura pone de manifiesto la presencia de valores atípicos e intermedios en todas las configuraciones. En $k = 2$, la caja aparece altamente concentrada alrededor de 1,162, aunque con valores extremos superiores distribuidos aproximadamente entre 1,900 y 2,165. En $k = 3$, los valores atípicos se extienden aproximadamente entre 1,280 y 1,883, mientras que en $k = 4$ se distribuyen entre 1,110 y 1,676. Para $k = 5$, la caja es más amplia y se observan numerosas ejecuciones intermedias aproximadamente entre 0,930 y 1,860, lo que confirma una mayor variabilidad en comparación con $k = 4$.

En conjunto, la Figura 21 indica que las configuraciones con $k = 4$ y $k = 5$ continúan

siendo las alternativas más prometedoras para *Bisecting K-Means* según el índice de *Davies-Bouldin*; sin embargo, $k = 4$ muestra un comportamiento más consistente, mientras que $k = 5$ alcanza los mejores valores mínimos y promedio, aunque con una mayor heterogeneidad entre ejecuciones.

Índice de Calinski-Harabasz

Tabla 97

Resumen estadístico del índice de Calinski-Harabasz para Bisecting K-Means según número de clusters

Estadístico descriptivo	**Número de clusters**			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	255,262531	220,021283	245,736891	240,879062
std	13,043011	12,811066	13,508959	7,719877
min	73,060269	74,564143	94,378166	109,286427
25 %	256,263647	222,700516	248,743071	234,442198
50 %	256,263647	222,712239	248,757476	245,285809
75 %	256,263647	222,712239	248,757476	245,718694
max	256,263647	222,715272	248,761202	254,119052

En la Tabla 97 se presenta el resumen estadístico del índice de *Calinski-Harabasz* obtenido con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). En todas las configuraciones se registraron 100000 ejecuciones, lo que permite comparar de manera consistente el comportamiento de esta métrica de evaluación interna.

De manera general, se observa que los valores promedio del índice se mantienen elevados en todas las configuraciones, aunque con diferencias entre particiones. En particular, $k = 2$ presenta la media más alta con 255,263, seguido de $k = 4$ con 245,737, $k = 5$ con 240,879 y $k = 3$ con 220,021. Dado que en el índice de *Calinski-Harabasz* valores más altos indican una mejor separación entre clusters y una mayor cohesión interna, estos resultados sugieren un mejor desempeño general para $k = 2$, mientras que $k = 4$ y $k = 5$ también muestran resultados favorables respecto a $k = 3$.

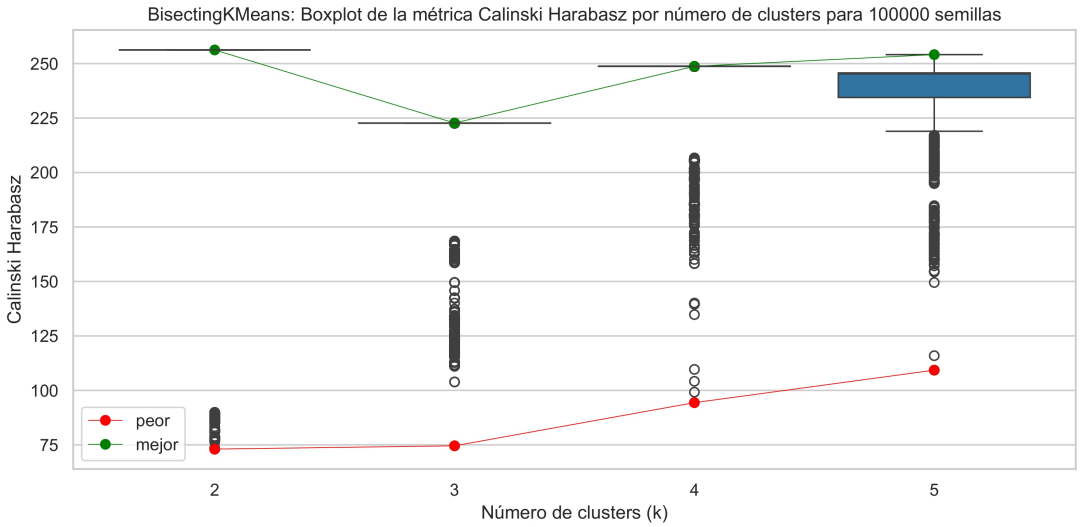
Asimismo, los estadísticos de posición refuerzan este comportamiento. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 256,264, lo que evidencia una fuerte concentración de las ejecuciones en torno a ese valor. En $k = 3$, el rango intercuartílico se sitúa aproximadamente entre 222,701 y 222,712, con una mediana de 222,712. Para $k = 4$, los cuartiles se concentran entre 248,743 y 248,757, con una mediana de 248,757, reflejando un comportamiento central alto y estable. Por su parte, $k = 5$ presenta un rango intercuartílico más amplio, aproximadamente entre 234,442 y 245,719, con una mediana de 245,286, lo que indica una mayor heterogeneidad en su distribución central.

Por otro lado, la dispersión de los resultados también presenta diferencias entre configuraciones. La desviación estándar es de 13,043 para $k = 2$, 12,811 para $k = 3$, 13,509 para $k = 4$ y 7,720 para $k = 5$. Esto indica que $k = 5$ presenta la menor variabilidad global, mientras que $k = 4$ muestra la mayor dispersión total. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 2$ con 73,060, seguido de $k = 3$ con 74,564, mientras que $k = 4$ y $k = 5$ presentan mínimos más altos, de 94,378 y 109,286, respectivamente. A su vez, el valor máximo más alto se registra en $k = 2$ con 256,264.

En conjunto, la Tabla 97 indica que $k = 2$ ofrece el comportamiento más favorable en términos de valores promedio y centrales del índice de *Calinski-Harabasz*, mientras que $k = 4$ también presenta un desempeño elevado y relativamente estable. Por su parte, $k = 5$ muestra una distribución central más heterogénea, aunque con menor variabilidad global y con valores mínimos superiores a los observados en las demás configuraciones.

Figura 22

Índice de Calinski Harabasz para Bisecting KMeans



En la Figura 22 se presenta un *boxplot* del índice de *Calinski-Harabasz* obtenido con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores del índice *Calinski-Harabasz*. Asimismo, se destacan los valores de peor desempeño en color rojo y los de mejor desempeño en color verde para cada configuración de k.

De manera general, la figura permite apreciar que las distribuciones centrales se mantienen en niveles altos para todas las configuraciones, aunque con diferencias entre ellas. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 256,264, mientras que en $k = 3$ la caja se concentra aproximadamente entre 222,701 y 222,712. Para $k = 4$, la distribución central se sitúa en torno a 248,743 y 248,757, reflejando un comportamiento alto y relativamente estable. En $k = 5$, la caja se desplaza a niveles inferiores respecto a $k = 2$ y $k = 4$, extendiéndose aproximadamente entre 234,442 y 245,719. Este comportamiento es consistente con lo observado en la Tabla 97.

En relación con los valores extremos, la figura muestra que los mejores resultados alcanzan aproximadamente 256,264 en $k = 2$, 222,715 en $k = 3$, 248,761 en $k = 4$ y

254,119 en $k = 5$. Por su parte, los valores mínimos se sitúan alrededor de 73,060 para $k = 2$, 74,564 para $k = 3$, 94,378 para $k = 4$ y 109,286 para $k = 5$. Dado que en esta métrica valores más altos indican una mejor separación entre clusters y mayor cohesión interna, estos resultados sugieren que $k = 2$ presenta el comportamiento más favorable en términos de valores centrales y máximos, mientras que $k = 5$ evita los valores mínimos más bajos observados en las otras configuraciones.

Asimismo, la figura evidencia la presencia de valores atípicos e intermedios por debajo de las cajas, especialmente en $k = 4$ y $k = 5$. En $k = 2$, la caja aparece sumamente concentrada alrededor de 256,264, aunque con valores bajos aislados cercanos a 73,000 y 90,000. En $k = 3$, la dispersión resulta menor en la parte central, pero también se observan ejecuciones inferiores respecto al valor predominante. En $k = 4$ y $k = 5$, se aprecia una mayor amplitud en la distribución de resultados, especialmente en $k = 5$, lo que pone de manifiesto una mayor heterogeneidad entre ejecuciones, aun cuando sus peores resultados siguen siendo superiores a los mínimos observados en $k = 2$ y $k = 3$.

En conjunto, la Figura 22 indica que $k = 2$ constituye la configuración más favorable para *Bisecting K-Means* según el índice de *Calinski-Harabasz*, al concentrar sus ejecuciones en los niveles más altos de la métrica. No obstante, $k = 4$ también presenta un desempeño elevado y estable, mientras que $k = 5$ destaca por registrar mínimos menos desfavorables, aunque con una distribución central más dispersa.

Inercia

Tabla 98

Resumen estadístico de la inercia para Bisecting K-Means según número de clusters

Estadístico descriptivo	Número de clusters			
	2	**3**	**4**	**5**
count	100000	100000	100000	100000
mean	44,041325	34,305567	25,286179	21,021133
std	1,188590	1,249654	1,071266	0,528855
min	43,950021	34,047498	25,046673	20,218659
25 %	43,950021	34,047743	25,046918	20,711719
50 %	43,950021	34,047743	25,046918	20,737780
75 %	43,950021	34,049578	25,048753	21,416436
max	61,052337	52,521202	42,127743	34,354384

En la Tabla 98 se presenta el resumen estadístico de la métrica de *Inercia* obtenida con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$). En todas las configuraciones se registraron 100000 ejecuciones, lo que permite comparar de manera consistente el comportamiento de esta métrica interna de compactación.

De manera general, se observa que los valores promedio de la inercia disminuyen progresivamente conforme aumenta el número de clusters, pasando de 44,041325 para $k = 2$ a 34,305567 para $k = 3$, 25,286179 para $k = 4$ y 21,021133 para $k = 5$. Dado que en esta métrica valores más bajos indican una mayor compactación de los clusters, estos resultados sugieren que las configuraciones con mayor número de grupos tienden a producir particiones más cohesionadas.

Asimismo, los estadísticos de posición refuerzan esta tendencia decreciente. Para $k = 2$, los cuartiles 25 %, 50 % y 75 % coinciden en 43,950021, lo que evidencia una fuerte concentración de las ejecuciones en torno a ese valor. En $k = 3$, la distribución central se ubica aproximadamente entre 34,047743 y 34,049578, con una mediana de 34,047743. Para $k = 4$, los cuartiles se concentran entre 25,046918 y 25,048753, con una mediana de 25,046918, mientras que en $k = 5$ el rango intercuartílico se sitúa aproximadamente entre 20,711719 y 21,416436, con una mediana de 20,737780. En consecuencia, $k = 5$

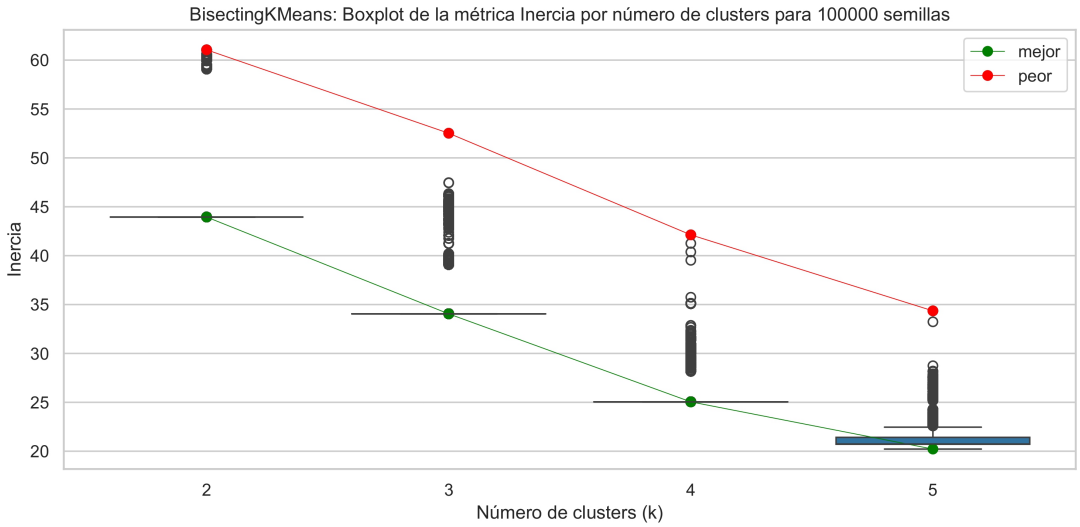
presenta la distribución central más baja dentro del conjunto analizado.

Por otro lado, la dispersión de los resultados también muestra diferencias entre configuraciones. La desviación estándar es de 1,188590 para $k = 2$, 1,249654 para $k = 3$, 1,071266 para $k = 4$ y 0,528855 para $k = 5$. Esto indica que $k = 5$ presenta la menor variabilidad global, mientras que $k = 3$ exhibe la mayor dispersión. En cuanto a los valores extremos, el mínimo más bajo corresponde a $k = 5$ con 20,218659, seguido de $k = 4$ con 25,046673, mientras que el valor máximo más alto se registra en $k = 2$ con 61,052337.

En conjunto, la Tabla 98 indica que las configuraciones con mayor número de clusters ofrecen los resultados más favorables según la métrica de *Inercia*, siendo $k = 5$ la alternativa más destacada al combinar el menor valor promedio, la menor mediana y la menor variabilidad global dentro del rango analizado.

Figura 23

Inercia para Bisecting KMeans



En la Figura 23 se presenta un *boxplot* de la métrica de *Inercia* obtenida con el algoritmo *Bisecting K-Means* para distintos números de clusters ($k = 2, 3, 4, 5$) a partir de 100000 semillas. En el eje horizontal se muestra el número de clusters evaluado, mientras que en el eje vertical se representan los valores de la inercia. Asimismo, se destacan los valores de mejor desempeño en color verde y los de peor desempeño en color rojo para

cada configuración de k.

De manera general, la figura permite apreciar un desplazamiento progresivo de las distribuciones hacia valores inferiores conforme aumenta el número de clusters. En $k = 2$, la mayor concentración de ejecuciones se ubica alrededor de 43,950, mientras que en $k = 3$ la distribución central se sitúa aproximadamente entre 34,048 y 34,050. Para $k = 4$, la caja se concentra en torno a 25,047 y 25,049, y en $k = 5$ el rango intercuartílico se desplaza a valores aún menores, aproximadamente entre 20,712 y 21,416. Este comportamiento es consistente con lo observado en la Tabla 98 y sugiere una mayor compactación de los clusters al incrementar el número de grupos.

En relación con los valores extremos, la figura muestra que los mejores resultados disminuyen desde aproximadamente 43,950 en $k = 2$ hasta 20,219 en $k = 5$, pasando por 34,047 en $k = 3$ y 25,047 en $k = 4$. De manera similar, los peores valores también se reducen desde 61,052 en $k = 2$ a 52,521 en $k = 3$, 42,128 en $k = 4$ y 34,354 en $k = 5$. Dado que en esta métrica valores más bajos indican una mayor compactación de los grupos, estos resultados confirman que las configuraciones con un mayor número de clusters ofrecen particiones más favorables.

Asimismo, la figura pone de manifiesto la presencia de valores atípicos e intermedios en todas las configuraciones, aunque con distinta intensidad. En $k = 2$, la caja aparece muy concentrada alrededor de 43,950, pero con valores extremos superiores próximos a 59,000 y 61,000. En $k = 3$, se observan múltiples ejecuciones por encima de la caja, distribuidas aproximadamente entre 39,000 y 52,500. Para $k = 4$, los valores atípicos se ubican aproximadamente entre 29,000 y 42,000, mientras que en $k = 5$ se concentran principalmente entre 22,000 y 34,000. Esta distribución confirma la sensibilidad del algoritmo a la semilla de inicialización, aunque dicha variabilidad disminuye en niveles inferiores de la métrica a medida que aumenta el número de clusters.

En conjunto, la Figura 23 indica que las configuraciones con $k = 4$ y especialmente $k = 5$ constituyen las alternativas más favorables para *Bisecting K-Means* según la métrica de *Inercia*, siendo $k = 5$ la configuración más destacada al concentrar sus ejecuciones en los niveles más bajos y estables del rango analizado.

4.5.3 Métricas para DBSCAN

Con el propósito de evaluar el comportamiento del algoritmo *DBSCAN* en el problema de segmentación planteado, se analizaron distintas métricas internas de calidad de agrupamiento para múltiples configuraciones del parámetro de mínimo número de vecinos (*minVecinos*) y una amplia exploración del parámetro ε. En particular, se consideraron 100000 valores de ε en el intervalo $(0, 1]$, reteniéndose únicamente aquellas ejecuciones que generaron agrupamientos válidos según los criterios establecidos en la metodología.

A diferencia de los algoritmos particionales, en *DBSCAN* la cantidad de ejecuciones válidas no necesariamente es uniforme entre configuraciones, debido a que la estructura de los grupos depende de manera directa de la interacción entre la densidad local de los datos, el valor de ε y el mínimo número de vecinos requerido para formar regiones densas. Por esta razón, el análisis de resultados no solo permite identificar qué configuraciones alcanzan mejores valores en las métricas evaluadas, sino también examinar la estabilidad y consistencia del algoritmo frente a variaciones de sus parámetros.

Para ello, se emplearon cuatro métricas complementarias: el índice de *Silhouette*, el índice de *Davies-Bouldin*, el índice de *Calinski-Harabasz* y la *Inercia*. Estas métricas permiten valorar, desde distintas perspectivas, la calidad de los agrupamientos generados, considerando aspectos como la cohesión interna de los clusters, la separación entre grupos y la compactación global de las soluciones obtenidas. En las subsecciones siguientes se presenta el análisis descriptivo de cada una de estas métricas, acompañado de tablas resumen y diagramas de caja que permiten visualizar la distribución de los resultados según el valor de *minVecinos*.

Índice de Silhouette para DBSCAN

Tabla 99

Resumen estadístico del índice de Silhouette para DBSCAN según el mínimo número de vecinos

Estadístico descriptivo	Mínimo número de vecinos			
	2	**3**	**4**	**5**
count	10392	4815	4815	2960
mean	0,174301	0,111237	0,111237	0,195178
std	0,121146	0,064377	0,064377	0,004828
min	-0,034585	-0,002990	-0,002990	0,183108
25 %	0,033889	0,033954	0,033954	0,193823
50 %	0,258332	0,123128	0,123128	0,198817
75 %	0,258332	0,170190	0,170190	0,198817
max	0,258332	0,170190	0,170190	0,202402

En la Tabla 99 se presenta el resumen estadístico del índice de *Silhouette* obtenido con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5). A diferencia de los algoritmos particionales, el número de ejecuciones que pasaron el filtro no es uniforme entre configuraciones, registrándose 10392 casos para *minVecinos* = 2, 4815 para *minVecinos* = 3, 4815 para *minVecinos* = 4 y 2960 para *minVecinos* = 5.

De manera general, se observa que el mayor valor promedio del índice de *Silhouette* corresponde a *minVecinos* = 5, con 0,195178, seguido de *minVecinos* = 2, con 0,174301, mientras que *minVecinos* = 3 y *minVecinos* = 4 presentan el mismo promedio de 0,111237. Dado que valores más altos de esta métrica indican una mejor cohesión interna y una mayor separación entre grupos, estos resultados sugieren que la configuración con *minVecinos* = 5 presenta un desempeño más consistente, aunque *minVecinos* = 2 alcanza valores máximos más altos.

Asimismo, los estadísticos de posición permiten identificar diferencias importantes en la distribución de los resultados. Para *minVecinos* = 2, el primer cuartil se ubica en 0,033889, mientras que la mediana, el tercer cuartil y el valor máximo coinciden en

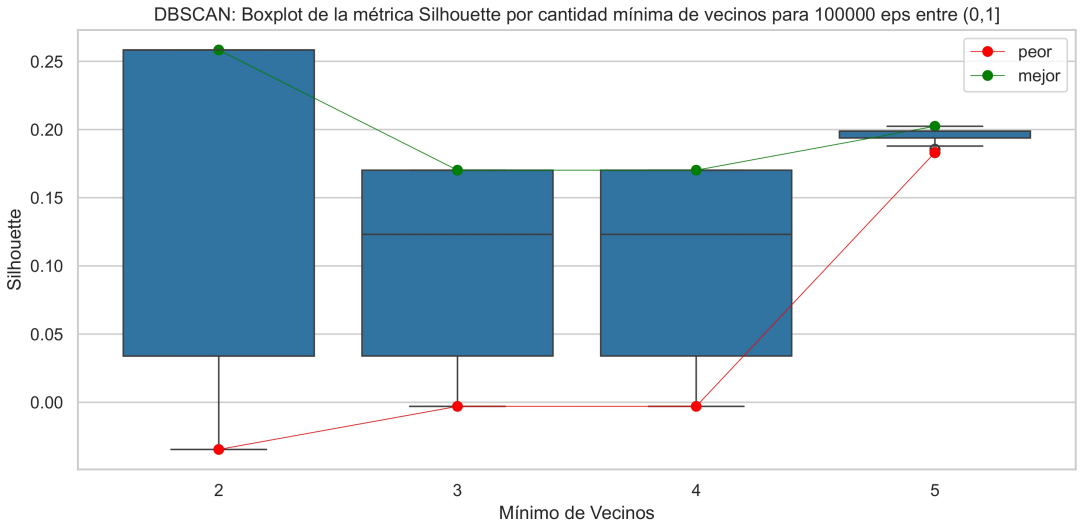
0,258332, lo que evidencia una fuerte concentración de ejecuciones en niveles altos del índice, aunque coexistiendo con valores considerablemente menores. En *minVecinos* = 3 y *minVecinos* = 4, los resúmenes son idénticos, con un primer cuartil de 0,033954, una mediana de 0,123128 y un tercer cuartil de 0,170190. Por su parte, *minVecinos* = 5 presenta un rango intercuartílico mucho más estrecho, aproximadamente entre 0,193823 y 0,198817, con una mediana de 0,198817, lo que pone de manifiesto una mayor estabilidad de los resultados.

Por otro lado, la dispersión también muestra diferencias claras entre configuraciones. La desviación estándar alcanza 0,121146 en *minVecinos* = 2, 0,064377 en *minVecinos* = 3 y *minVecinos* = 4, y solo 0,004828 en *minVecinos* = 5. Esto indica que la configuración con *minVecinos* = 5 presenta la menor variabilidad global. En cuanto a los valores extremos, el mínimo más bajo se registra en *minVecinos* = 2 con -0,034585, seguido de -0,002990 para *minVecinos* = 3 y *minVecinos* = 4, mientras que *minVecinos* = 5 mantiene un mínimo positivo de 0,183108. A su vez, el valor máximo más alto se presenta en *minVecinos* = 2 con 0,258332.

En conjunto, la Tabla 99 indica que *minVecinos* = 2 alcanza el mejor valor máximo del índice de *Silhouette*, pero *minVecinos* = 5 ofrece el comportamiento más estable y consistente, al concentrar sus ejecuciones en un rango superior y estrecho de la métrica.

Figura 24

Índice de Silhouette para DBSCAN



En la Figura 24 se presenta un *boxplot* del índice de *Silhouette* obtenido con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5), considerando 100000 valores de ε en el intervalo $(0, 1]$. En el eje horizontal se muestra la cantidad mínima de vecinos evaluada, mientras que en el eje vertical se representan los valores de la métrica *Silhouette*. Asimismo, se destacan los valores de peor desempeño en color rojo y los de mejor desempeño en color verde para cada configuración analizada.

De manera general, la figura permite apreciar que el comportamiento del índice varía de forma importante según el valor de *minVecinos*. En *minVecinos* = 2, la distribución es amplia y combina valores bajos con una concentración importante de ejecuciones alrededor de 0,258. En *minVecinos* = 3 y *minVecinos* = 4, las cajas se ubican en niveles intermedios, aproximadamente entre 0,034 y 0,170, con una mediana cercana a 0,123. Por su parte, en *minVecinos* = 5 la caja se concentra en un rango más estrecho y superior, aproximadamente entre 0,194 y 0,199, lo que es consistente con lo observado en la Tabla 99.

En relación con los valores extremos, la figura muestra que el mejor resultado máximo

corresponde a *minVecinos* = 2, con un valor cercano a 0,258, mientras que para *minVecinos* = 3 y *minVecinos* = 4 los máximos se sitúan en torno a 0,170, y para *minVecinos* = 5 alcanzan aproximadamente 0,202. En cuanto a los valores mínimos, se observan registros negativos en *minVecinos* = 2, cercanos a -0,035, y ligeramente inferiores a cero en *minVecinos* = 3 y *minVecinos* = 4, alrededor de -0,003. En contraste, *minVecinos* = 5 mantiene un mínimo positivo de aproximadamente 0,183, lo que evidencia una mayor estabilidad en esta configuración.

Asimismo, la amplitud de las cajas pone de manifiesto diferencias importantes en la dispersión de los resultados. Las configuraciones con *minVecinos* = 2, *minVecinos* = 3 y *minVecinos* = 4 presentan una variabilidad considerable, lo que indica una marcada sensibilidad del algoritmo frente al valor de ε. En cambio, *minVecinos* = 5 muestra una caja mucho más estrecha, con una concentración de resultados en niveles relativamente altos de la métrica. Además, la similitud visual entre *minVecinos* = 3 y *minVecinos* = 4 coincide con los estadísticos de la Tabla 99, donde ambas configuraciones presentan valores descriptivos idénticos.

En conjunto, la Figura 24 confirma que *minVecinos* = 2 alcanza el mejor valor máximo del índice de *Silhouette*, pero *minVecinos* = 5 constituye la configuración más estable y consistente para *DBSCAN*, al concentrar sus ejecuciones en un rango superior y con menor dispersión.

Índice de Davies-Bouldin para DBSCAN

Tabla 100

Resumen estadístico del índice de Davies-Bouldin para DBSCAN según el mínimo número de vecinos

Estadístico descriptivo	Mínimo número de vecinos			
	2	**3**	**4**	**5**
count	10392	4815	4815	2960
mean	0,611636	0,757539	0,757539	0,872416
std	0,035600	0,055803	0,055803	0,083558
min	0,577380	0,698375	0,698375	0,774453
25 %	0,589774	0,698375	0,698375	0,782827
50 %	0,589774	0,748766	0,748766	0,946912
75 %	0,663846	0,829241	0,829241	0,946912
max	0,668149	0,829241	0,829241	0,950270

En la Tabla 100 se presenta el resumen estadístico del índice de *Davies-Bouldin* obtenido con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5).

De manera general, se observa que el valor promedio del índice aumenta conforme se incrementa el parámetro *minVecinos*, pasando de 0,611636 en *minVecinos* = 2 a 0,757539 en *minVecinos* = 3 y *minVecinos* = 4, y alcanzando 0,872416 en *minVecinos* = 5. Dado que en el índice de *Davies-Bouldin* valores más bajos indican una mejor separación y compactación de los clusters, estos resultados sugieren que la configuración con *minVecinos* = 2 ofrece el desempeño más favorable dentro del rango analizado.

Asimismo, los estadísticos de posición refuerzan esta tendencia. Para *minVecinos* = 2, el primer cuartil y la mediana coinciden en 0,589774, mientras que el tercer cuartil asciende a 0,663846, lo que indica que la mayor parte de las ejecuciones se concentra en niveles relativamente bajos del índice. En *minVecinos* = 3 y *minVecinos* = 4, los resúmenes son idénticos, con un primer cuartil de 0,698375, una mediana de 0,748766 y un tercer cuartil de 0,829241. Por su parte, *minVecinos* = 5 presenta un rango intercuartílico desplazado a valores más altos, aproximadamente entre 0,782827

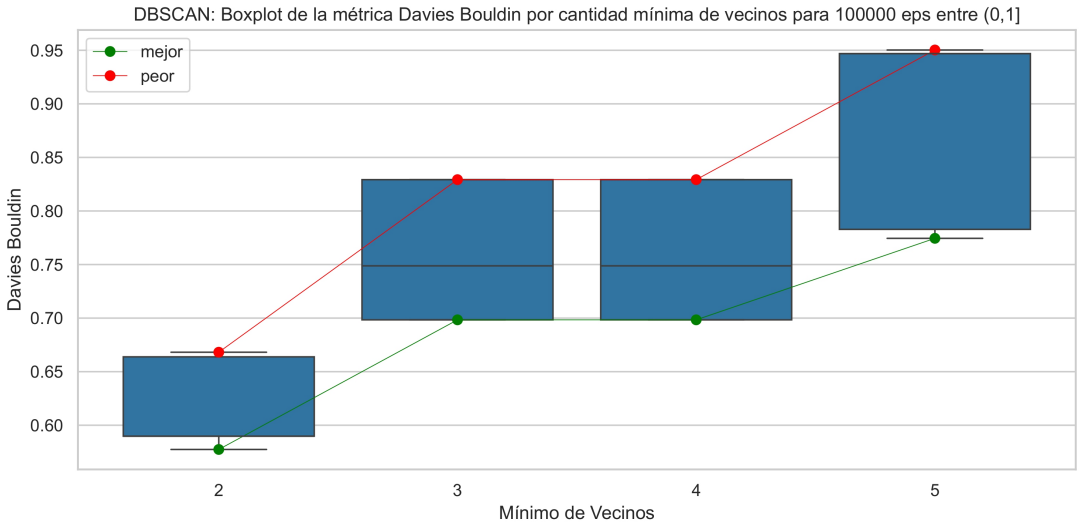
y 0,946912, con una mediana de 0,946912, lo que pone de manifiesto un deterioro en la calidad del agrupamiento respecto a configuraciones con menor número de vecinos.

Por otro lado, la dispersión de los resultados también muestra diferencias entre configuraciones. La desviación estándar es de 0,035600 en *minVecinos* = 2, 0,055803 en *minVecinos* = 3 y *minVecinos* = 4, y 0,083558 en *minVecinos* = 5. Esto indica una mayor variabilidad a medida que aumenta el parámetro *minVecinos*. En cuanto a los valores extremos, el mínimo más bajo se registra en *minVecinos* = 2 con 0,577380, seguido de 0,698375 para *minVecinos* = 3 y *minVecinos* = 4, mientras que *minVecinos* = 5 presenta un mínimo de 0,774453. A su vez, el valor máximo más alto corresponde a *minVecinos* = 5 con 0,950270.

En conjunto, la Tabla 100 indica que *minVecinos* = 2 constituye la alternativa más favorable para *DBSCAN* según el índice de *Davies-Bouldin*, al presentar menores valores promedio, medianos y cuartiles, mientras que el incremento de este parámetro se asocia con un deterioro progresivo en la calidad del agrupamiento.

Figura 25

Índice de Davies Bouldin para DBSCAN



En la Figura 25 se presenta un *boxplot* del índice de *Davies-Bouldin* obtenido con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos

(*minVecinos* = 2, 3, 4 y 5), considerando 100000 valores de ε en el intervalo $(0, 1]$. En el eje horizontal se muestra la cantidad mínima de vecinos evaluada, mientras que en el eje vertical se representan los valores del índice *Davies-Bouldin*. Asimismo, se destacan los valores de mejor desempeño en color verde y los de peor desempeño en color rojo para cada configuración analizada.

De manera general, la figura permite apreciar que las distribuciones del índice se desplazan hacia valores superiores conforme aumenta el parámetro *minVecinos*. En *minVecinos* = 2, la caja se concentra aproximadamente entre 0,590 y 0,664, con una mediana cercana a 0,590. En *minVecinos* = 3 y *minVecinos* = 4, las distribuciones son prácticamente idénticas, ubicándose aproximadamente entre 0,698 y 0,829, con una mediana cercana a 0,749. Por su parte, en *minVecinos* = 5 la caja se desplaza a valores aún más altos, aproximadamente entre 0,783 y 0,947, lo que es consistente con lo observado en la Tabla 100.

En relación con los valores extremos, la figura muestra que el mejor resultado mínimo corresponde a *minVecinos* = 2, con un valor cercano a 0,577, mientras que para *minVecinos* = 3 y *minVecinos* = 4 los mínimos se sitúan alrededor de 0,698, y para *minVecinos* = 5 alcanzan aproximadamente 0,774. En cuanto a los peores resultados, los valores máximos aumentan desde aproximadamente 0,668 en *minVecinos* = 2 hasta 0,829 en *minVecinos* = 3 y *minVecinos* = 4, llegando a 0,950 en *minVecinos* = 5. Dado que en esta métrica valores más bajos indican una mejor calidad del agrupamiento, estos resultados sugieren que la configuración con *minVecinos* = 2 ofrece la segmentación más favorable, mientras que el incremento del parámetro se asocia con una pérdida gradual de calidad.

Asimismo, la amplitud de las cajas evidencia diferencias en la variabilidad de los resultados. La configuración con *minVecinos* = 2 concentra sus valores en un rango inferior y relativamente estrecho, lo que sugiere mayor estabilidad y mejor calidad general. En contraste, *minVecinos* = 3 y *minVecinos* = 4 muestran cajas más amplias y desplazadas hacia valores intermedios, mientras que *minVecinos* = 5 presenta la mayor amplitud y se concentra en niveles superiores del índice. Además, la similitud visual entre *minVecinos* = 3 y *minVecinos* = 4 coincide con los estadísticos descriptivos idénticos reportados en la

Tabla 100.

En conjunto, la Figura 25 confirma que *minVecinos* = 2 constituye la alternativa más favorable para *DBSCAN* según el índice de *Davies-Bouldin*, mientras que el aumento del mínimo número de vecinos se relaciona con una disminución progresiva en la calidad del agrupamiento.

Índice de Calinski-Harabasz

Tabla 101

Resumen estadístico del índice de Calinski-Harabasz para DBSCAN según el mínimo número de vecinos

Estadístico descriptivo	**Mínimo número de vecinos**			
	2	**3**	**4**	**5**
count	10392	4815	4815	2960
mean	5,210521	20,684978	20,684978	47,439017
std	0,253590	15,889552	15,889552	28,627177
min	4,759172	7,965563	7,965563	14,973705
25 %	5,144053	7,965563	7,965563	15,572554
50 %	5,144053	11,056002	11,056002	73,003381
75 %	5,144053	42,811618	42,811618	73,003381
max	5,639337	42,811618	42,811618	73,003381

En la Tabla 101 se presenta el resumen estadístico del índice de *Calinski-Harabasz* obtenido con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5).

De manera general, se observa que el valor promedio del índice aumenta conforme se incrementa el parámetro *minVecinos*, pasando de 5,211 en *minVecinos* = 2 a 20,685 en *minVecinos* = 3 y *minVecinos* = 4, y alcanzando 47,439 en *minVecinos* = 5. Dado que en el índice de *Calinski-Harabasz* valores más altos indican una mejor separación entre clusters y una mayor cohesión interna, estos resultados sugieren que la configuración con *minVecinos* = 5 ofrece, en promedio, el mejor desempeño dentro del rango evaluado.

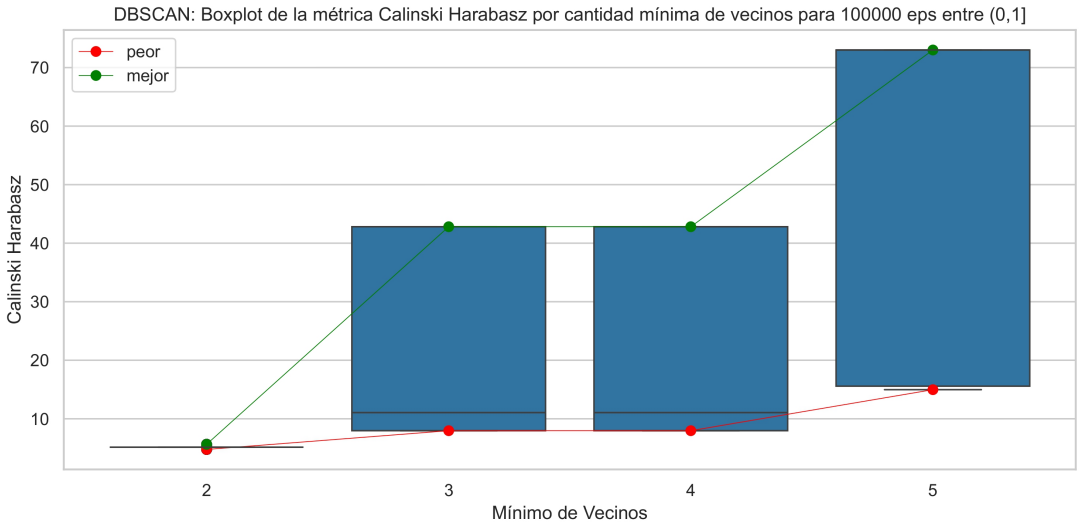
Asimismo, los estadísticos de posición muestran diferencias importantes entre configuraciones. Para *minVecinos* = 2, el primer cuartil, la mediana y el tercer cuartil coinciden en 5,144, lo que indica una fuerte concentración de las ejecuciones en torno a ese nivel. En *minVecinos* = 3 y *minVecinos* = 4, los resúmenes son idénticos, con un primer cuartil de 7,966, una mediana de 11,056 y un tercer cuartil de 42,812, lo que evidencia una distribución más amplia y heterogénea. Por su parte, *minVecinos* = 5 presenta un rango intercuartílico desplazado a valores superiores, aproximadamente entre 15,573 y 73,003, con una mediana también igual a 73,003, lo que confirma la concentración de una parte importante de las ejecuciones en niveles altos del índice.

Por otro lado, la dispersión de los resultados también aumenta con el valor de *minVecinos*. La desviación estándar es de 0,254 en *minVecinos* = 2, 15,890 en *minVecinos* = 3 y *minVecinos* = 4, y 28,627 en *minVecinos* = 5. Esto indica una variabilidad considerablemente mayor en las configuraciones con más vecinos mínimos, especialmente en *minVecinos* = 5. En cuanto a los valores extremos, el mínimo más bajo se registra en *minVecinos* = 2 con 4,759, mientras que los máximos más altos corresponden a *minVecinos* = 5 con 73,003, seguido de *minVecinos* = 3 y *minVecinos* = 4 con 42,812.

En conjunto, la Tabla 101 indica que *minVecinos* = 5 constituye la alternativa más favorable para *DBSCAN* según el índice de *Calinski-Harabasz*, al presentar los valores promedio y centrales más altos, aunque con una variabilidad también mayor en comparación con las demás configuraciones.

Figura 26

Índice de Calinski Harabasz para DBSCAN



En la Figura 26 se presenta un *boxplot* del índice de *Calinski-Harabasz* obtenido con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5), considerando 100000 valores de ε en el intervalo $(0, 1]$. En el eje horizontal se muestra la cantidad mínima de vecinos evaluada, mientras que en el eje vertical se representan los valores del índice *Calinski-Harabasz*. Asimismo, se destacan los valores de peor desempeño en color rojo y los de mejor desempeño en color verde para cada configuración analizada.

De manera general, la figura permite apreciar un desplazamiento de las distribuciones hacia valores superiores conforme aumenta el parámetro *minVecinos*. En *minVecinos* = 2, la caja se concentra alrededor de 5,144, mientras que en *minVecinos* = 3 y *minVecinos* = 4 la distribución central se extiende aproximadamente entre 7,966 y 42,812, con una mediana cercana a 11,056. Por su parte, en *minVecinos* = 5 la caja se desplaza a valores aún más altos, aproximadamente entre 15,573 y 73,003, lo que es consistente con lo observado en la Tabla 21.

En relación con los valores extremos, la figura muestra que los mejores resultados aumentan desde aproximadamente 5,639 en *minVecinos* = 2 hasta 42,812 en *minVecinos*

= 3 y *minVecinos* = 4, alcanzando 73,003 en *minVecinos* = 5. Del mismo modo, los valores mínimos se incrementan de 4,759 en *minVecinos* = 2 a 7,966 en *minVecinos* = 3 y *minVecinos* = 4, y a 14,974 en *minVecinos* = 5. Dado que en esta métrica valores más altos indican una mejor separación entre clusters y una mayor cohesión interna, estos resultados sugieren que la configuración con *minVecinos* = 5 ofrece el desempeño más favorable dentro del rango evaluado.

Asimismo, la amplitud de las cajas permite apreciar diferencias importantes en la dispersión de los resultados. La configuración con *minVecinos* = 2 presenta una distribución muy concentrada en valores bajos del índice, reflejando un comportamiento estable, pero limitado. En contraste, *minVecinos* = 3 y *minVecinos* = 4 muestran distribuciones más amplias y prácticamente idénticas, tanto en su posición como en su dispersión, en concordancia con los estadísticos descriptivos de la Tabla 21. Por su parte, *minVecinos* = 5 exhibe la caja más alta y amplia, lo que evidencia una mejora en la calidad del agrupamiento, aunque acompañada de una mayor variabilidad entre ejecuciones.

En conjunto, la Figura 26 confirma que *minVecinos* = 5 constituye la alternativa más favorable para *DBSCAN* según el índice de *Calinski-Harabasz*, al concentrar sus ejecuciones en niveles superiores de la métrica, aunque con una dispersión más marcada que en configuraciones con menor número de vecinos.

Inercia

Tabla 102

Resumen estadístico de la inercia para DBSCAN según el mínimo número de vecinos

Estadístico descriptivo	Mínimo número de vecinos			
	2	**3**	**4**	**5**
count	10392	4815	4815	2960
mean	70,544727	60,311694	60,311694	55,099947
std	1,159443	9,099641	9,099641	8,221495
min	68,345476	47,808790	47,808790	46,943992
25 %	69,734272	47,808790	47,808790	47,787980
50 %	71,326616	65,025895	65,025895	47,787980
75 %	71,326616	68,345237	68,345237	63,512156
max	71,326616	68,345237	68,345237	65,803999

En la Tabla 102 se presenta el resumen estadístico de la métrica de *Inercia* obtenida con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5).

De manera general, se observa que el valor promedio de la inercia disminuye conforme aumenta el parámetro *minVecinos*, pasando de 70,545 en *minVecinos* = 2 a 60,312 en *minVecinos* = 3 y *minVecinos* = 4, y alcanzando 55,100 en *minVecinos* = 5. Dado que en esta métrica valores más bajos indican una mayor compactación de los grupos, estos resultados sugieren un mejor desempeño de *DBSCAN* para configuraciones con mayor número mínimo de vecinos.

Asimismo, los estadísticos de posición muestran diferencias importantes entre configuraciones. Para *minVecinos* = 2, el primer cuartil se ubica en 69,734 y tanto la mediana como el tercer cuartil coinciden en 71,327, lo que evidencia una fuerte concentración de ejecuciones en niveles altos de inercia. En *minVecinos* = 3 y *minVecinos* = 4, los resúmenes son idénticos, con un primer cuartil de 47,809, una mediana de 65,026 y un tercer cuartil de 68,345, lo que refleja una distribución más amplia y heterogénea. Por su parte, *minVecinos* = 5 presenta un rango intercuartílico aproximadamente entre 47,788 y 63,512, con una mediana de 47,788, lo que confirma el desplazamiento de la distribución

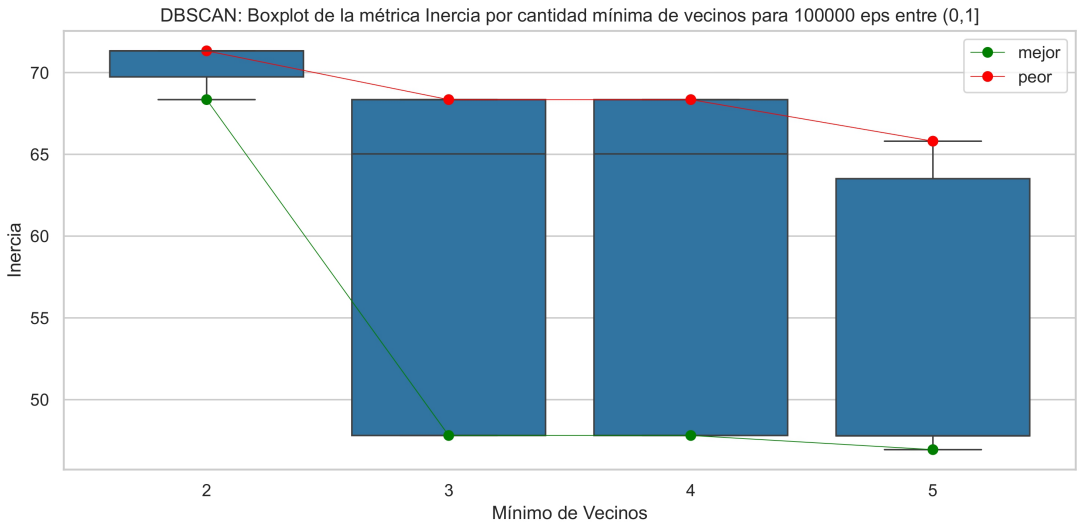
hacia valores inferiores respecto a las demás configuraciones.

Por otro lado, la dispersión de los resultados también presenta diferencias entre configuraciones. La desviación estándar es de 1,159 en *minVecinos* = 2, 9,100 en *minVecinos* = 3 y *minVecinos* = 4, y 8,221 en *minVecinos* = 5. Esto indica una mayor variabilidad en las configuraciones con *minVecinos* intermedios y altos, especialmente en *minVecinos* = 3 y *minVecinos* = 4. En cuanto a los valores extremos, el mínimo más bajo corresponde a *minVecinos* = 5 con 46,944, seguido de 47,809 para *minVecinos* = 3 y *minVecinos* = 4, mientras que el valor máximo más alto se registra en *minVecinos* = 2 con 71,327.

En conjunto, la Tabla 102 indica que *minVecinos* = 5 constituye la alternativa más favorable para *DBSCAN* según la métrica de *Inercia*, al presentar el menor valor promedio y una distribución central desplazada hacia niveles inferiores de la métrica, aunque con una dispersión aún apreciable entre ejecuciones.

Figura 27

Inercia para DBSCAN



En la Figura 27 se presenta un *boxplot* de la métrica de *Inercia* obtenida con el algoritmo *DBSCAN* para distintos valores del parámetro de mínimo número de vecinos (*minVecinos* = 2, 3, 4 y 5), considerando 100000 valores de ε en el intervalo $(0, 1]$. En el

eje horizontal se muestra la cantidad mínima de vecinos evaluada, mientras que en el eje vertical se representan los valores de la inercia. Asimismo, se destacan los valores de mejor desempeño en color verde y los de peor desempeño en color rojo para cada configuración analizada.

De manera general, la figura permite apreciar un desplazamiento de las distribuciones hacia valores inferiores conforme aumenta el parámetro *minVecinos*. En *minVecinos* = 2, la caja se concentra aproximadamente entre 69,734 y 71,327, con una mediana cercana a 71,327. En *minVecinos* = 3 y *minVecinos* = 4, las distribuciones son prácticamente idénticas, extendiéndose aproximadamente entre 47,809 y 68,345, con una mediana cercana a 65,026. Por su parte, en *minVecinos* = 5 la caja se desplaza a valores más bajos, aproximadamente entre 47,788 y 63,512, lo que es consistente con lo observado en la Tabla 102.

En relación con los valores extremos, la figura muestra que los mejores resultados disminuyen desde aproximadamente 68,345 en *minVecinos* = 2 hasta 47,809 en *minVecinos* = 3 y *minVecinos* = 4, alcanzando 46,944 en *minVecinos* = 5. De manera similar, los peores valores también se reducen desde aproximadamente 71,327 en *minVecinos* = 2 hasta 68,345 en *minVecinos* = 3 y *minVecinos* = 4, y 65,804 en *minVecinos* = 5. Dado que en esta métrica valores más bajos indican una mayor compactación de los grupos, estos resultados sugieren que la configuración con *minVecinos* = 5 ofrece el desempeño más favorable dentro del rango evaluado.

Asimismo, la amplitud de las cajas evidencia diferencias importantes en la variabilidad de los resultados. La configuración con *minVecinos* = 2 presenta una distribución relativamente estrecha y concentrada en valores altos de inercia, lo que refleja resultados homogéneos, pero menos favorables. En contraste, *minVecinos* = 3 y *minVecinos* = 4 muestran distribuciones más amplias y prácticamente idénticas, en concordancia con los estadísticos descriptivos reportados en la Tabla 102. Por su parte, *minVecinos* = 5 también presenta dispersión, aunque desplazada hacia valores inferiores, lo que indica una mejora en la compactación de las soluciones obtenidas.

En conjunto, la Figura 27 confirma que *minVecinos* = 5 constituye la alternativa más

favorable para *DBSCAN* según la métrica de *Inercia*, al concentrar sus ejecuciones en niveles más bajos de la métrica y, por tanto, en agrupamientos más compactos.

4.5.4 Comparación general

Con el propósito de realizar una valoración integral del desempeño de los algoritmos de agrupamiento evaluados, en esta subsección se presenta una comparación global entre *KMeans*, *Bisecting KMeans* y *DBSCAN* a partir de las métricas internas de evaluación consideradas en el estudio: *Silhouette*, *Calinski-Harabasz*, *Davies-Bouldin* e *Inercia*. Esta comparación permite analizar, de manera conjunta, la cohesión interna, la separación entre grupos y la compactación de las particiones obtenidas, identificando tanto los comportamientos promedio como la dispersión y los valores extremos de cada algoritmo bajo distintas configuraciones.

Para ello, primero se presenta una tabla resumen con los principales estadísticos descriptivos de cada métrica, lo que permite establecer una comparación cuantitativa entre algoritmos y configuraciones. Posteriormente, se incluye una representación gráfica mediante *boxplots*, con el fin de visualizar la distribución de los resultados, la variabilidad entre ejecuciones y la presencia de valores atípicos. En conjunto, ambos elementos ofrecen una visión comparativa más amplia del comportamiento de los algoritmos, facilitando la identificación de la alternativa más consistente y favorable según los criterios de evaluación interna aplicados.

Tabla 103

Comparación de métricas de evaluación para KMeans, Bisecting KMeans y DBSCAN

Métrica	Estadístico descriptivo	KMeans				BisectingKMeans				DBSCAN			
		2	3	4	5	2	3	4	5	2	3	4	5
Índice de Silhouette	count	100000	100000	100000	100000	100000	100000	100000	100000	11754	3442	2654	5132
	mean	0,364	0,365	0,394	**0,406**	0,363	0,365	0,391	0,389	**0,223**	0,137	0,014	0,066
	std	**0,002**	0,009	0,016	0,021	0,013	0,016	0,015	0,023	**0,041**	0,076	0,018	0,076
	min	0,178	**0,261**	0,216	0,238	0,168	0,173	0,206	0,222	**0,170**	0,030	-0,003	-0,035
	25 %	0,364	0,367	0,399	**0,414**	0,364	0,368	0,394	0,363	**0,170**	0,034	-0,003	-0,035
	50 %	0,364	0,367	0,399	**0,415**	0,364	0,368	0,394	0,407	0,258	0,167	-0,003	0,123
	75 %	0,364	0,367	0,399	**0,416**	0,364	0,369	0,395	0,407	0,258	0,199	0,034	0,123
	max	0,364	0,367	0,399	**0,417**	0,364	0,370	0,396	0,416	0,258	0,202	0,034	0,123
Índice de Calinski-Harabasz	mean	**256,231**	222,688	250,011	252,757	255,263	220,021	245,737	240,879	0,645	0,815	0,711	0,770
	std	2,394	5,467	20,441	12,509	13,043	12,811	13,509	7,720	0,068	0,141	**0,042**	0,078
	min	77,239	**168,031**	147,794	146,007	73,060	74,564	94,378	109,286	0,590	0,577	0,664	0,668
	25 %	256,264	220,481	**257,696**	256,920	256,264	222,701	248,743	234,442	0,590	0,756	0,664	0,668
	50 %	256,264	225,862	**257,696**	256,971	256,264	222,712	248,757	245,286	0,590	0,760	0,749	0,829
	75 %	256,264	225,862	**257,696**	257,394	256,264	222,712	248,757	245,719	0,698	0,947	0,749	0,829
	max	256,264	225,862	257,696	**258,655**	256,264	222,715	248,761	254,119	0,789	0,950	0,749	0,829
Índice de Davies-Bouldin	mean	1,162	1,048	0,970	**0,925**	1,166	1,055	1,002	0,989	7,187	39,899	8,228	29,252
	std	**0,012**	0,051	0,033	0,067	0,064	0,062	0,050	0,063	3,175	31,779	3,133	17,895
	min	1,162	1,010	0,942	**0,871**	1,162	1,038	0,988	0,923	**5,144**	5,496	4,759	5,639
	25 %	1,162	1,010	0,964	**0,874**	1,162	1,040	0,990	0,935	**5,144**	11,670	4,759	5,639
	50 %	1,162	1,055	0,964	**0,916**	1,162	1,043	0,993	0,956	**5,144**	11,779	11,056	42,812
	75 %	1,162	1,055	0,964	**0,918**	1,162	1,043	0,993	1,056	7,966	73,003	11,056	42,812
	max	2,141	1,516	1,487	**1,367**	2,165	1,883	1,676	1,858	15,573	73,003	11,056	42,812
Inercia	mean	43,953	34,056	25,047	**20,327**	44,041	34,306	25,286	21,021	69,590	57,882	66,318	55,300
	std	**0,219**	0,482	1,531	0,843	1,189	1,250	1,071	0,529	2,333	9,825	3,085	9,887
	min	43,950	33,795	24,472	**19,962**	43,950	34,047	25,047	20,219	63,512	46,944	63,533	47,809
	25 %	43,950	33,795	24,472	**20,033**	43,950	34,048	25,047	20,712	68,345	47,788	63,533	47,809
	50 %	43,950	33,795	24,472	**20,057**	43,950	34,048	25,047	20,738	71,327	65,026	63,533	47,809
	75 %	43,950	34,229	24,472	**20,059**	43,950	34,050	25,049	21,416	71,327	65,806	69,734	68,345
	max	60,509	39,127	34,085	29,149	61,052	52,521	42,128	34,354	71,327	70,302	69,734	68,345

En la Tabla 103 se presenta una comparación global de los estadísticos descriptivos de las métricas de evaluación interna *Silhouette*, *Calinski-Harabasz*, *Davies-Bouldin* e *Inercia* para los algoritmos *KMeans*, *Bisecting KMeans* y *DBSCAN*, considerando distintas configuraciones del número de clusters y, en el caso de *DBSCAN*, distintos valores del mínimo número de vecinos. Esta síntesis permite identificar, de manera conjunta, las tendencias centrales, la dispersión y los valores extremos de cada algoritmo bajo criterios complementarios de cohesión, separación y compactación.

De manera general, la tabla muestra que *KMeans* presenta el comportamiento más consistente y favorable en las métricas *Silhouette*, *Calinski-Harabasz* e *Inercia*. En *Silhouette*, sus mejores resultados se observan para $k = 5$, con una media de 0,406 y valores cuartílicos concentrados entre 0,414 y 0,416. En *Calinski-Harabasz*, también destaca con valores elevados, alcanzando un máximo de 258,655 en $k = 5$. Asimismo, en la métrica de *Inercia* exhibe una reducción progresiva al incrementar el número de clusters, registrando su menor media en $k = 5$ con 20,327, lo que indica una mayor compactación de los grupos.

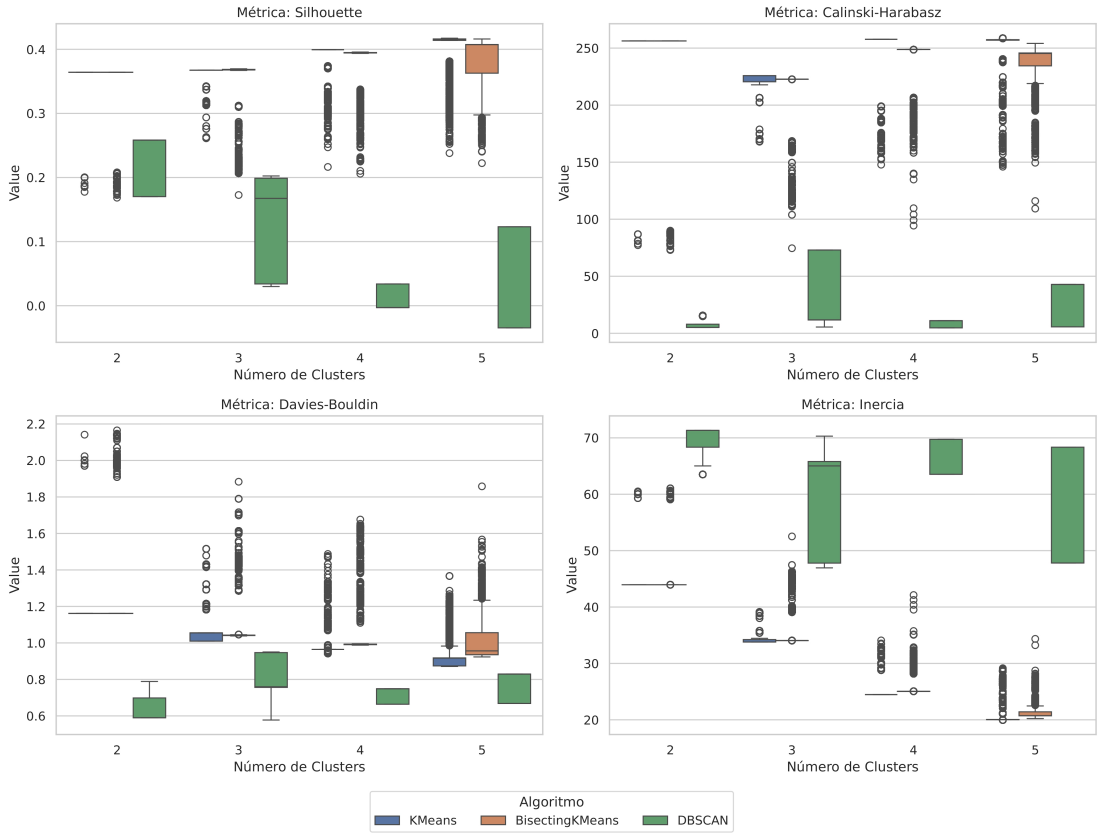
Por su parte, *Bisecting KMeans* presenta resultados competitivos, aunque generalmente ligeramente inferiores a los de *KMeans*. En *Silhouette*, sus mejores configuraciones se sitúan en $k = 4$ y $k = 5$, con medias de 0,391 y 0,389, respectivamente. En *Calinski-Harabasz*, mantiene valores altos, aunque por debajo de *KMeans*, mientras que en *Inercia* también muestra una disminución progresiva, alcanzando su mejor promedio en $k = 5$ con 21,021. En conjunto, estos resultados lo posicionan como una alternativa cercana a *KMeans*, aunque con una mayor variabilidad en algunas configuraciones.

En el caso de *DBSCAN*, la interpretación comparativa debe realizarse con cautela, ya que el número de ejecuciones válidas varía entre configuraciones y, además, en la tabla comparativa se aprecia una aparente inversión de los valores reportados para las métricas *Calinski-Harabasz* y *Davies-Bouldin* respecto a los resultados observados en los análisis individuales. No obstante, la tabla permite advertir que *DBSCAN* alcanza resultados inferiores en *Silhouette* e *Inercia* frente a los algoritmos particionales, mientras que en *Davies-Bouldin* muestra valores que, de confirmarse la correspondencia correcta de las métricas, serían relativamente competitivos.

En conjunto, la Tabla 103 sugiere que *KMeans* es el algoritmo con desempeño más equilibrado y favorable en la mayoría de métricas consideradas, seguido por *Bisecting KMeans*. Por su parte, *DBSCAN* presenta un comportamiento más dependiente de la configuración del mínimo número de vecinos y de la validez de las soluciones generadas, por lo que su análisis debe complementarse necesariamente con la inspección visual de la figura comparativa.

Figura 28

Métricas por algoritmo



En la Figura 28 se presenta un conjunto de *boxplots* comparativos para las métricas de evaluación interna *Silhouette*, *Calinski-Harabasz*, *Davies-Bouldin* e *Inercia*, considerando los algoritmos *KMeans*, *Bisecting KMeans* y *DBSCAN*. Cada subgráfico muestra la distribución de los resultados según el número de clusters o, en el caso de *DBSCAN*, según el mínimo número de vecinos, permitiendo una comparación visual del comportamiento relativo de los tres algoritmos bajo distintos criterios de calidad del agrupamiento.

En la métrica *Silhouette*, se observa que *KMeans* concentra sus resultados en los niveles más altos, especialmente para $k = 4$ y $k = 5$, seguido por *Bisecting KMeans*, que alcanza valores cercanos, aunque con una mayor dispersión en algunas configuraciones. En contraste, *DBSCAN* muestra distribuciones más heterogéneas y valores generalmente inferiores, lo que sugiere una menor cohesión interna y separación entre grupos en comparación con los algoritmos particionales.

En el panel correspondiente a *Calinski-Harabasz*, la figura muestra nuevamente una ventaja clara para *KMeans*, cuyos valores se ubican en los niveles más altos del índice, seguido de *Bisecting KMeans*. Por su parte, *DBSCAN* presenta valores notablemente menores, aunque con una mejora visible cuando aumenta el mínimo número de vecinos. Este comportamiento indica que, bajo esta métrica, los algoritmos particionales tienden a generar particiones con mejor relación entre dispersión inter-cluster y cohesión intra-cluster.

En el caso del índice *Davies-Bouldin*, la figura evidencia un patrón distinto, ya que *DBSCAN* concentra sus distribuciones en valores inferiores a los observados en *KMeans* y *Bisecting KMeans*. Dado que en esta métrica los valores más bajos son más favorables, esta observación sugiere una ventaja relativa de *DBSCAN* bajo este criterio específico. Entre los algoritmos particionales, *KMeans* presenta, en general, valores ligeramente más bajos y estables que *Bisecting KMeans*, especialmente en las configuraciones con mayor número de clusters.

Por otro lado, en la métrica de *Inercia*, la figura vuelve a favorecer a los algoritmos particionales. Tanto *KMeans* como *Bisecting KMeans* muestran una reducción progresiva de la inercia conforme aumenta el número de clusters, alcanzando sus valores más bajos en $k = 5$. Entre ambos, *KMeans* presenta, en general, distribuciones algo más bajas y compactas. En contraste, *DBSCAN* mantiene niveles de inercia considerablemente más altos, aun cuando mejora al incrementar el mínimo número de vecinos, lo que evidencia una menor compactación relativa de los grupos obtenidos.

En conjunto, la Figura 28 muestra que no existe un único algoritmo dominante en todas las métricas. *KMeans* sobresale por su desempeño más equilibrado y favorable en *Silhouette*, *Calinski-Harabasz* e *Inercia*, mientras que *DBSCAN* destaca principalmente en el índice *Davies-Bouldin*. Por su parte, *Bisecting KMeans* se comporta como una alternativa intermedia, con resultados cercanos a *KMeans* en varias configuraciones, aunque generalmente con mayor dispersión o con valores ligeramente menos favorables. En consecuencia, considerando el comportamiento visual global de las cuatro métricas, *KMeans* emerge como la alternativa más consistente dentro del análisis comparativo.

4.6 Selección de la segmentación más adecuada de los clientes

La selección de la segmentación óptima constituye una etapa crítica dentro del proceso de análisis de clustering, especialmente en contextos aplicados donde no solo se busca una adecuada estructura geométrica de los datos, sino también una segmentación útil y operativamente implementable en el entorno empresarial.

En este estudio, se evaluaron múltiples configuraciones de clustering mediante diferentes algoritmos y parámetros, generando una amplia variedad de segmentaciones posibles. Para determinar la segmentación más adecuada, se propuso un marco de evaluación multicriterio basado en métricas internas de validación, con el objetivo de integrar distintos aspectos de calidad estructural del clustering.

4.6.1 Fundamento metodológico de la evaluación multicriterio

Las métricas internas de clustering presentan naturalezas heterogéneas, ya que evalúan diferentes propiedades de la estructura de los datos, tales como cohesión intra-cluster, separación inter-cluster y compacidad. En consecuencia, la selección de una segmentación basada en una única métrica puede introducir sesgos metodológicos y favorecer determinados paradigmas de clustering.

Particularmente, algunas métricas como la Inercia y el índice de Calinski-Harabasz presentan una dependencia estructural hacia modelos basados en centroides, mientras que otras, como el índice de Silhouette, ofrecen una evaluación más neutral respecto a la forma geométrica de los clusters.

Por este motivo, se adoptó un enfoque de **evaluación multicriterio ponderada**, donde cada métrica contribuye al resultado final de acuerdo con su relevancia teórica y su grado de neutralidad respecto al algoritmo de clustering utilizado.

4.6.2 Ponderación de las métricas de evaluación

Se definió una ponderación equilibrada considerando los siguientes criterios:

- Neutralidad respecto al paradigma de clustering.
- Capacidad de evaluar cohesión y separación simultáneamente.
- Robustez geométrica.
- Reconocimiento en la literatura científica.

La Tabla 104 presenta la ponderación asignada a cada métrica.

Tabla 104

Ponderación de métricas para la evaluación multicriterio del clustering

Métrica	Peso	Justificación
Silhouette	0,40	Métrica más neutral, evalúa cohesión y separación sin asumir forma de clusters
Calinski-Harabasz	0,25	Basada en varianza global, mide separación entre clusters
Davies-Bouldin	0,20	Penaliza clusters solapados y dispersos
Inercia	0,15	Mide compacidad, aunque dependiente de centroides

4.6.3 Comparación técnica de las métricas utilizadas

Para fundamentar la selección de las métricas y su ponderación, se presenta en la Tabla 105 una comparación técnica de sus propiedades principales.

Tabla 105

Comparación técnica de métricas de validación de clustering

Métrica	Cohesión	Separación	Dependencia de centroides	Aplicable a DBSCAN
Inercia	Sí	No	Alta	No
Davies-Bouldin	Sí	Sí	Media	Parcial
Silhouette	Sí	Sí	Baja	Sí
Calinski-Harabasz	Sí	Sí	Alta	Parcial

4.6.4 Definición del score global

A partir de la ponderación establecida, se definió un **score global** para cada configuración de clustering, con el objetivo de integrar las distintas métricas en un único criterio de decisión.

$$Score_{global} = 0,40 \cdot S + 0,25 \cdot CH + 0,20 \cdot DBI + 0,15 \cdot Inercia \tag{21}$$

donde:

- S corresponde al índice de Silhouette.
- CH corresponde al índice de Calinski-Harabasz.
- DBI corresponde al índice de Davies-Bouldin (ajustado para que valores mayores representen mejor calidad).
- $Inercia$ corresponde a la suma de errores cuadráticos intra-cluster (normalizada).

Previamente al cálculo del score global, todas las métricas fueron normalizadas para garantizar su comparabilidad en una misma escala.

4.6.5 Criterio de selección de la segmentación

Finalmente, la segmentación más adecuada se definió como aquella que maximiza el $Score_{global}$, garantizando simultáneamente:

- Alta cohesión intra-cluster.
- Alta separación inter-cluster.
- Evaluación equilibrada entre diferentes paradigmas de clustering.

A continuación, se presentan los resultados obtenidos del cálculo del score global para cada configuración evaluada, permitiendo identificar la segmentación óptima.

4.6.6 Análisis del ranking de segmentaciones

Tabla 106

Ranking de las 30 mejores segmentaciones

Algoritmo	Ejecución	Número de Clusters	Semilla	Métrica				Ranking				Score Global
				Silhouette	Davies-Bouldin	Calinski-Harabasz	Inercia	Silhouette	Calinski-Harabasz	Davies-Bouldin	Inercia	
Kmeans	386723	5	86723	0,417	0,871	258,655	19,962	3	1	15	1	4,6
Kmeans	323613	5	23613	0,417	0,873	258,619	19,964	2	5	19	5	6,6
Kmeans	318569	5	18569	0,416	0,872	258,625	19,964	7	3	17	3	7,4
Kmeans	341472	5	41472	0,417	0,874	258,538	19,969	1	9	20	9	8
Kmeans	381810	5	81810	0,417	0,875	258,625	19,964	4	4	25	4	8,2
Kmeans	315583	5	15583	0,417	0,874	258,590	19,966	5	7	21	7	9
Kmeans	396018	5	96018	0,416	0,871	258,512	19,970	6	10	14	10	9,2
Kmeans	333087	5	33087	0,416	0,873	258,655	19,962	13	2	18	2	9,6
Kmeans	362464	5	62464	0,416	0,874	258,619	19,964	8	6	24	6	10,4
Kmeans	331576	5	31576	0,416	0,874	258,587	19,966	9	8	22	8	11,2
Kmeans	396770	5	96770	0,416	0,916	256,971	20,057	10	14	26	13	14,65
Kmeans	332792	5	32792	0,415	0,872	257,394	20,033	18	12	16	11	15,05
Kmeans	380393	5	80393	0,416	0,917	256,948	20,058	12	15	27	14	16,05
Kmeans	364926	5	64926	0,414	0,874	257,383	20,033	19	13	23	12	17,25
Kmeans	332633	5	32633	0,415	0,918	256,920	20,059	17	16	28	15	18,65
BisectingKMeans	326851	5	26851	0,416	0,925	254,119	20,219	14	18	31	16	18,7
BisectingKMeans	345256	5	45256	0,416	0,925	254,119	20,220	14	18	31	19	19,15
BisectingKMeans	328841	5	28841	0,416	0,925	254,119	20,228	14	18	31	20	19,3
BisectingKMeans	378598	5	78598	0,416	0,928	253,738	20,248	11	21	32	22	19,35
BisectingKMeans	356666	5	56666	0,416	0,925	254,119	20,236	14	18	31	21	19,45
BisectingKMeans	382768	5	82768	0,416	0,923	254,108	20,219	16	19	29	17	19,5
BisectingKMeans	362276	5	62276	0,416	0,924	254,104	20,220	15	20	30	18	19,7
BisectingKMeans	375362	5	75362	0,411	0,946	251,227	20,386	22	22	44	23	26,55
BisectingKMeans	310337	5	10337	0,411	0,946	251,227	20,387	22	22	44	24	26,7
BisectingKMeans	314335	5	14335	0,411	0,946	251,227	20,388	22	22	44	25	26,85
BisectingKMeans	367470	5	67470	0,411	0,946	251,227	20,388	22	22	44	26	27
BisectingKMeans	332716	5	32716	0,411	0,946	251,227	20,395	22	22	44	28	27,3
BisectingKMeans	383326	5	83326	0,411	0,943	251,161	20,390	23	23	42	27	27,4
BisectingKMeans	310823	5	10823	0,411	0,947	251,072	20,402	21	24	45	29	27,75

La Tabla 106 presenta las 30 configuraciones de clustering mejor posicionadas según el marco de evaluación multicriterio propuesto. Es importante señalar que, con el fin de garantizar la utilidad del análisis, se eliminaron segmentaciones duplicadas (entendidas como aquellas que producían particiones equivalentes del espacio de datos), priorizando así la diversidad estructural de las soluciones evaluadas.

A partir de los resultados obtenidos, se observa una clara predominancia de configuraciones generadas mediante el algoritmo KMeans en las primeras posiciones del ranking. En particular, las mejores segmentaciones corresponden consistentemente a configuraciones con cinco clusters, lo cual sugiere que dicho nivel de granularidad logra un equilibrio adecuado entre cohesión intra-cluster y separación inter-cluster.

En términos de métricas, los valores de Silhouette se mantienen relativamente estables en el rango de 0,416 a 0,417 para las mejores soluciones, lo que indica una estructura de clusters razonablemente bien definida. Asimismo, los valores del índice

de Calinski-Harabasz alcanzan sus máximos en estas configuraciones, evidenciando una adecuada separación global entre los grupos.

Por otro lado, el índice de Davies-Bouldin presenta valores moderados (alrededor de 0,87), lo que sugiere que, si bien existe cierta proximidad entre algunos clusters, la relación entre dispersión interna y separación externa se mantiene dentro de rangos aceptables. En cuanto a la Inercia, los valores más bajos corresponden también a las primeras posiciones del ranking, reflejando una alta compacidad de los clusters generados por KMeans.

En contraste, las configuraciones obtenidas mediante BisectingKMeans, aunque presentes en el ranking, aparecen en posiciones inferiores. Estas segmentaciones muestran valores ligeramente menores en el índice de Calinski-Harabasz y mayores en Inercia, lo que indica una menor separación global y una mayor dispersión interna en comparación con KMeans. Asimismo, los valores del índice de Davies-Bouldin son superiores, evidenciando una mayor superposición entre clusters.

Otro aspecto relevante es la estabilidad de las soluciones encontradas. Se observa que múltiples ejecuciones de KMeans con diferentes semillas producen resultados muy similares en términos de métricas y ranking, lo cual sugiere una convergencia hacia estructuras de clustering consistentes. Esto constituye un indicio favorable respecto a la robustez de la segmentación.

Adicionalmente, es importante destacar el comportamiento del algoritmo DBSCAN dentro del ranking global. A pesar de haber sido evaluado bajo múltiples configuraciones de parámetros, las segmentaciones generadas por este algoritmo no lograron posicionarse dentro de las primeras posiciones. En particular, la mejor configuración de DBSCAN alcanzó un $Score_{global}$ de 1249,95, ubicándose recién en la posición N° 1571 del ranking general.

Este resultado evidencia una clara desventaja comparativa de DBSCAN frente a los algoritmos basados en centroides en el contexto específico del conjunto de datos analizado. Desde una perspectiva geométrica, esto sugiere que la estructura de los datos no presenta patrones de densidad claramente definidos que puedan ser explotados por algoritmos basados en densidad. En cambio, los resultados indican que los clusters tienden

a organizarse en estructuras más compactas y aproximadamente convexas, lo cual favorece el desempeño de algoritmos como KMeans.

Asimismo, desde el punto de vista metodológico, este hallazgo refuerza la importancia de utilizar un marco de evaluación multicriterio, ya que permite evidenciar de manera objetiva las limitaciones de ciertos paradigmas de clustering en función de la naturaleza de los datos. En este caso, DBSCAN no logra un equilibrio adecuado entre cohesión y separación, lo que se traduce en un desempeño significativamente inferior en términos del score global definido.

En consecuencia, se descarta el uso de DBSCAN como candidato para la segmentación final, priorizando aquellos algoritmos que demuestran un comportamiento consistente y superior en las métricas evaluadas.

En conjunto, los resultados indican que las configuraciones basadas en KMeans no solo optimizan el score global definido, sino que también presentan una estructura geométrica más favorable en términos de cohesión, separación y compacidad. Por tanto, estas segmentaciones constituyen candidatas sólidas para su posterior análisis en términos de interpretabilidad y aplicabilidad empresarial.

Finalmente, en la siguiente subsección se procederá a analizar en detalle las configuraciones mejor posicionadas, con el objetivo de seleccionar la segmentación óptima que será utilizada en la etapa de interpretación mediante árboles de decisión.

Selección de la segmentación óptima

Con el objetivo de determinar la segmentación final a utilizar en la etapa de interpretación, se consideraron las configuraciones mejor posicionadas en el ranking global obtenido mediante el marco de evaluación multicriterio.

Dado que el $Score_{global}$ integra de manera ponderada las métricas de cohesión, separación y compacidad, su maximización permite identificar la segmentación que presenta el mejor equilibrio entre estos criterios. En este sentido, la primera posición del ranking corresponde a la configuración que optimiza simultáneamente todas las métricas

consideradas bajo el esquema de ponderación definido.

Asimismo, el análisis de las primeras posiciones del ranking evidencia que las diferencias entre las configuraciones mejor evaluadas son relativamente pequeñas, lo cual sugiere la existencia de múltiples soluciones de alta calidad estructural. Sin embargo, la configuración ubicada en la primera posición presenta consistentemente los mejores valores agregados, lo que justifica su selección como la segmentación óptima.

En consecuencia, se seleccionó como segmentación final la configuración correspondiente al algoritmo **KMeans**, con **5 clusters** y semilla **86723**, la cual obtuvo el menor valor de ranking agregado y el mejor $Score_{global}$ dentro del conjunto de configuraciones evaluadas.

4.6.7 Descripción y visualización de la segmentación

Descripción de la segmentación

La Tabla 107 presenta los estadísticos descriptivos de las variables edad, monto de operaciones y número de visitas para cada uno de los clusters identificados. Este análisis permite caracterizar los distintos segmentos de clientes en función de sus patrones de comportamiento.

En primer lugar, respecto a la variable edad, se observa que el Cluster 0 agrupa a los clientes de mayor edad, con un promedio de 55,56 años, mientras que los Clusters 1 y 3 concentran a los clientes más jóvenes, con promedios de 27,56 y 25,23 años respectivamente. El Cluster 4 presenta una edad intermedia, mientras que el Cluster 2 agrupa clientes con edades relativamente homogéneas, evidenciado por su baja desviación estándar.

En cuanto al monto de operaciones, los Clusters 2 y 4 destacan por presentar los valores promedio más elevados, alcanzando 86 100 y 89 603,17 respectivamente, lo que indica la presencia de clientes de alto valor económico en estos segmentos. Por el contrario, el Cluster 3 presenta el menor monto promedio (25 727,27), caracterizando a un grupo de clientes con menor volumen de operaciones. Los Clusters 0 y 1 muestran valores

Tabla 107

Estadísticos descriptivos por cluster

Variable	Estadístico	Número de Clusters				
		0	**1**	**2**	**3**	**4**
Edad	count	119	94	80	44	63
	mean	55,56	27,56	32,94	25,23	41,24
	std	8,22	7,12	3,86	5,24	9,79
	min	42	17	26	18	19
	median	53	27	32	23	42
	max	71	40	41	36	60
Monto de Operaciones	count	119	94	80	44	63
	mean	49126,05	50010,64	86100	25727,27	89603,17
	std	14057,12	16786,13	16212,67	7564,8	16466,52
	min	19000	15000	68000	15000	71000
	median	49000	54000	79000	24000	86000
	max	79000	76000	137000	40000	137000
Número de Visitas	count	119	94	80	44	63
	mean	123,55	130,73	244,48	237,84	50,52
	std	47,72	41,16	29,8	30,92	29,34
	min	9	18	174	183	1
	median	138	141	249	231	48
	max	182	183	292	297	119

intermedios, con comportamientos relativamente similares en términos de gasto.

En relación con el número de visitas, los Clusters 2 y 3 presentan los promedios más altos, con 244,48 y 237,84 visitas respectivamente, lo que sugiere un alto nivel de interacción con la empresa. En contraste, el Cluster 4 presenta el menor promedio (50,52), indicando clientes con baja frecuencia de visitas. Los Clusters 0 y 1 presentan niveles intermedios de interacción.

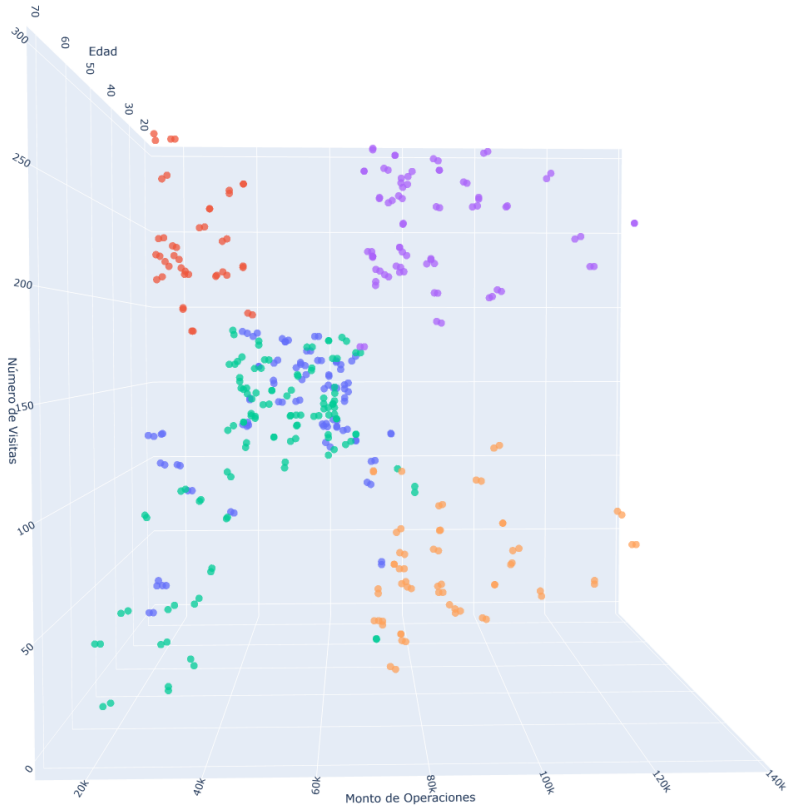
En conjunto, estos resultados evidencian la existencia de perfiles diferenciados de clientes. En particular, se identifican segmentos de alto valor y alta frecuencia (Clusters 2 y 3), segmentos de bajo valor y baja interacción (Cluster 4), así como segmentos intermedios con características mixtas (Clusters 0 y 1). Esta caracterización resulta fundamental para el diseño de estrategias diferenciadas orientadas a cada grupo de clientes.

Visualización de la segmentación

Con el objetivo de complementar el análisis cuantitativo de la segmentación obtenida, se presentan visualizaciones tridimensionales que permiten observar la distribución de los clientes en función de las variables consideradas: monto de operaciones, número de visitas y edad. Estas representaciones facilitan la interpretación geométrica de los clusters y permiten identificar patrones estructurales en los datos.

Figura 29

Vista monto de operaciones y número de visitas

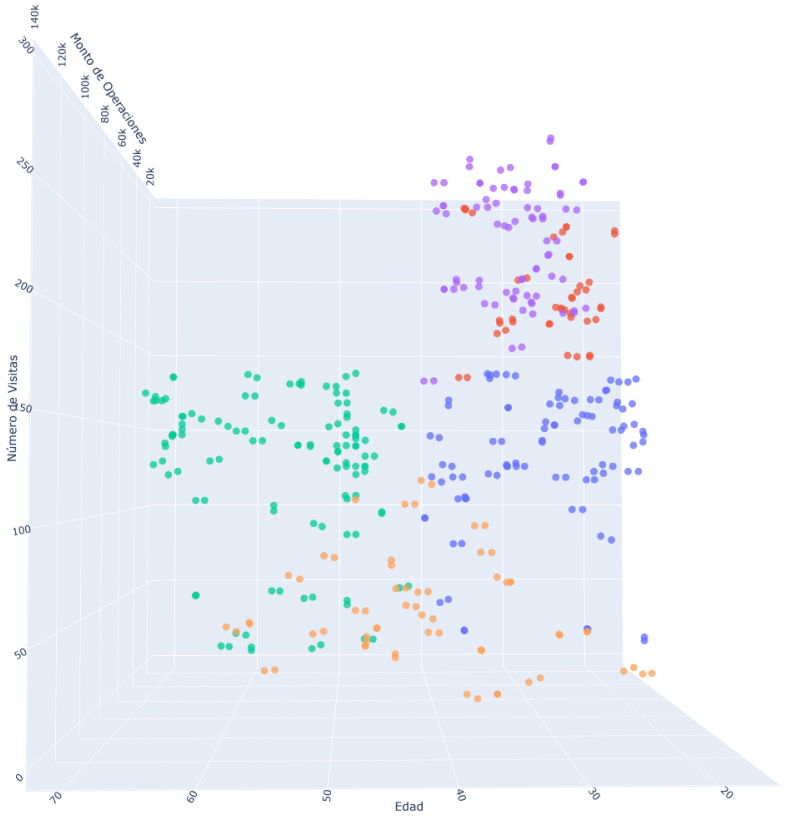


En la Figura 29 se observa la relación entre el monto de operaciones y el número de visitas. En esta proyección, los clusters muestran una diferenciación clara en términos de intensidad de interacción con la empresa. Se identifican grupos de clientes con alto número de visitas y montos elevados, así como segmentos con menor frecuencia y menor volumen de operaciones. Esta separación sugiere que ambas variables constituyen dimensiones clave para la segmentación, permitiendo distinguir perfiles de clientes intensivos frente

a clientes ocasionales.

Figura 30

Vista edad y número de visitas



La Figura 30 presenta la relación entre la edad de los clientes y su número de visitas. En esta visualización se evidencia que ciertos segmentos se agrupan en rangos etarios específicos con comportamientos diferenciados en términos de frecuencia. Por ejemplo, se observan clusters de clientes más jóvenes con patrones de visitas distintos a los de clientes de mayor edad, lo que sugiere la existencia de dinámicas de comportamiento asociadas a la edad. Sin embargo, también se aprecia cierto solapamiento entre clusters en algunos rangos, lo cual indica que la edad, si bien relevante, no es el único factor determinante de la frecuencia de interacción.

Figura 31

Vista edad y monto de operaciones

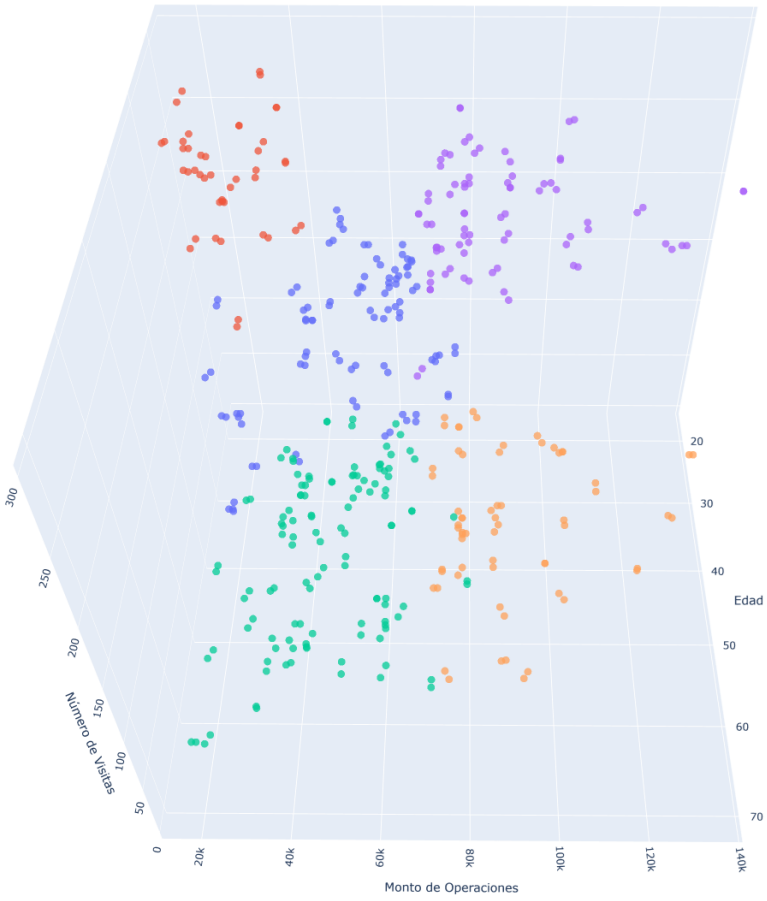
Finalmente, en la Figura 31 se analiza la relación entre la edad y el monto de operaciones. En esta proyección se observa una diferenciación más clara entre segmentos, donde ciertos grupos presentan montos significativamente mayores en rangos etarios específicos. Asimismo, se identifican clusters con montos bajos distribuidos en diferentes edades, lo que sugiere la coexistencia de perfiles heterogéneos dentro de ciertos rangos demográficos.

La Figura 32 presenta una visualización tridimensional que integra simultáneamente las tres variables analizadas: monto de operaciones, número de visitas y edad. A diferencia de las proyecciones bidimensionales anteriores, esta representación permite observar de manera más completa la estructura espacial de los clusters.

En esta vista se evidencia una separación más clara entre los distintos segmentos, confirmando que la segmentación obtenida no es únicamente un artefacto de proyecciones parciales, sino que responde a una estructura consistente en el espacio tridimensional.

Figura 32

Vista tridimensional de la segmentación de clientes



Los grupos se distribuyen en regiones diferenciadas, lo que indica que las variables consideradas permiten discriminar adecuadamente entre perfiles de clientes.

Asimismo, esta visualización refuerza la coherencia geométrica de la segmentación seleccionada, mostrando que los clusters presentan fronteras relativamente bien definidas y con bajo grado de solapamiento. Esto respalda la validez de la segmentación obtenida y fortalece su idoneidad para la siguiente etapa de interpretación mediante modelos explicativos.

En conjunto, las cuatro visualizaciones permiten confirmar que la segmentación obtenida presenta una estructura coherente en el espacio de variables, con una separación razonable entre clusters y patrones diferenciados en términos de comportamiento de los clientes.

4.7 Interpretar los segmentos obtenidos de los clientes de la empresa D&C Comunicaciones

Una vez seleccionada la segmentación óptima y caracterizados los distintos grupos de clientes, se procede a la etapa de interpretación de los segmentos. Esta fase tiene como objetivo traducir la estructura obtenida mediante técnicas de clustering en reglas claras y operativamente aplicables dentro del contexto empresarial.

Dado que los algoritmos de clustering son, por naturaleza, algoritmos no supervisados que no generan reglas explícitas de clasificación, se empleó un enfoque de interpretación basado en modelos supervisados. En particular, se utilizó un árbol de decisión como modelo sustituto (*surrogate model*), donde la variable objetivo corresponde a la etiqueta de cluster previamente asignada.

Este enfoque permite aproximar la segmentación mediante un conjunto de reglas jerárquicas basadas en las variables originales (edad, monto de operaciones y número de visitas), facilitando la comprensión de los criterios que definen cada segmento.

Para este propósito, se construyó un modelo de clasificación utilizando el algoritmo de inteligencia artificial supervisado *Decision Tree*, considerando como variables predictoras las características originales de los clientes y como variable dependiente el cluster asignado. Con el fin de garantizar la interpretabilidad del modelo, se restringió la complejidad del árbol mediante los siguientes hiperparámetros:

- Profundidad máxima del árbol igual a 4 niveles (*max_depth = 4*).
- Mínimo de 40 observaciones por nodo terminal (*min_samples_leaf = 40*).

Estas restricciones permiten la obtención de reglas simples, comprensibles y fácilmente implementables en el entorno empresarial.

El modelo de árbol de decisión fue entrenado utilizando la totalidad de los datos disponibles, evaluando su capacidad para reproducir la segmentación original mediante la métrica de exactitud (*accuracy*). Este indicador mide el porcentaje de observaciones correctamente clasificadas en su respectivo cluster.

El resultado obtenido muestra un alto nivel de exactitud, lo cual indica que la estructura de clustering puede ser representada adecuadamente mediante reglas de decisión relativamente simples. Esto sugiere que los segmentos identificados presentan una separación suficientemente clara en el espacio de variables, permitiendo su interpretación mediante un modelo supervisado de baja complejidad.

A partir del árbol de decisión entrenado, es posible extraer reglas que describen las condiciones bajo las cuales un cliente es asignado a un determinado segmento. Estas reglas se basan en umbrales definidos sobre las variables analizadas, tales como el monto de operaciones, la edad y el número de visitas.

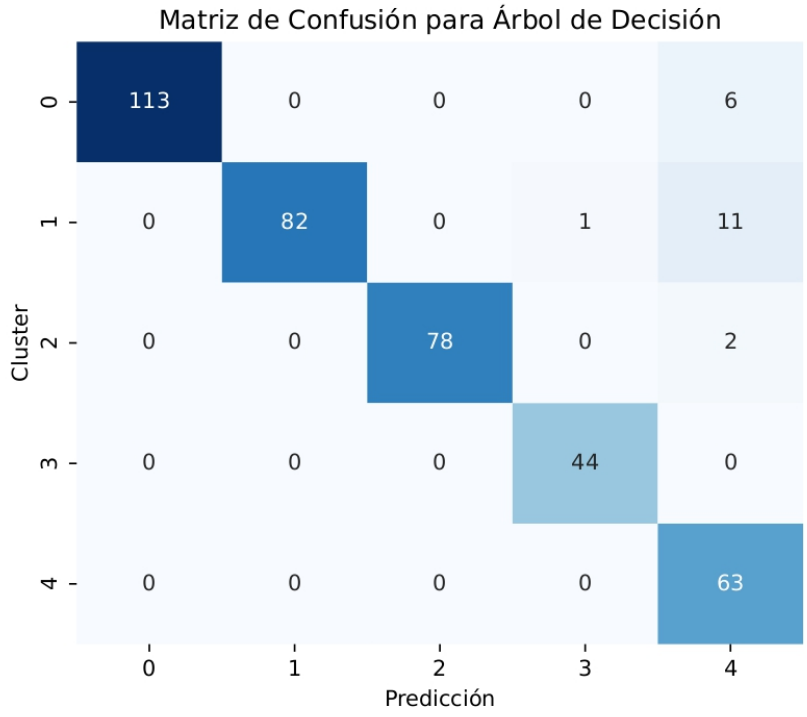
Este enfoque permite transformar la segmentación obtenida en un conjunto de criterios explícitos, facilitando su implementación en sistemas de gestión empresarial, tales como estrategias de marketing segmentado, programas de fidelización o políticas de atención diferenciada.

En las siguientes subsecciones se presentan los resultados obtenidos a partir del árbol de decisión, incluyendo su matriz de confusión, el reporte de clasificación y la estructura del modelo utilizada para interpretar los segmentos.

4.7.1 Matriz de confusión

Figura 33

Matriz de confusión del modelo Árbol de Decisión para la reproducción de los segmentos de clientes.



La Figura 33 presenta la matriz de confusión obtenida para el modelo Árbol de Decisión utilizado como modelo sustituto para interpretar la segmentación de clientes. Se observa una concentración predominante de valores en la diagonal principal, lo que indica que la mayor parte de las observaciones fueron clasificadas correctamente en su respectivo cluster. Este comportamiento evidencia que el árbol de decisión logra reproducir adecuadamente la estructura generada previamente por el algoritmo de clustering.

De manera específica, se aprecia que los clusters 3 y 4 presentan una clasificación completamente correcta en los datos evaluados, mientras que los clusters 0, 1 y 2 también muestran un desempeño favorable, aunque con algunas confusiones puntuales. En el

cluster 0 se identifican 113 clasificaciones correctas y 6 asignaciones erróneas hacia el cluster 4. En el cluster 1 se registran 82 clasificaciones correctas, además de 1 caso asignado al cluster 3 y 11 casos asignados al cluster 4. En el cluster 2 se observan 78 clasificaciones correctas y 2 casos clasificados como cluster 4. Por su parte, el cluster 3 presenta 44 clasificaciones correctas sin errores, y el cluster 4 registra 63 clasificaciones correctas sin confusiones hacia otros grupos.

En términos globales, la matriz evidencia un elevado nivel de concordancia entre la segmentación original y la clasificación generada por el árbol de decisión. La suma de los valores de la diagonal principal asciende a 380 observaciones correctamente clasificadas de un total de 400 registros evaluados, lo que equivale aproximadamente a una exactitud del 95 %. Este resultado confirma que la segmentación obtenida puede ser representada mediante reglas de decisión relativamente simples, favoreciendo así la interpretación operativa de los segmentos dentro del contexto empresarial.

Reporte de clasificación

Tabla 108

Reporte de clasificación del modelo Árbol de Decisión

Cluster	Precision	Recall	F1-score	Support
0	1,00	0,95	0,97	119
1	1,00	0,87	0,93	94
2	1,00	0,97	0,99	80
3	0,98	1,00	0,99	44
4	0,77	1,00	0,87	63
Accuracy			0,95	400
Macro avg	0,95	0,96	0,95	400
Weighted avg	0,96	0,95	0,95	400

La Tabla 108 presenta el reporte de clasificación del modelo Árbol de Decisión empleado como modelo sustituto para reproducir la segmentación obtenida mediante clustering. Los resultados evidencian un desempeño global alto, con una exactitud (*accuracy*) de 0,95 sobre un total de 400 observaciones evaluadas. Este valor indica que el

modelo logra clasificar correctamente el 95 % de los casos en el cluster correspondiente, lo cual confirma una adecuada capacidad para representar la estructura de segmentación original.

A nivel de clases individuales, se observa que los clusters 0, 2 y 3 presentan métricas particularmente favorables. En los clusters 0, 1 y 2 la precisión es igual a 1,00, lo que indica que, cuando el modelo asigna una observación a estos grupos, lo hace sin generar prácticamente falsos positivos. Asimismo, el cluster 3 presenta una precisión de 0,98 y un recall de 1,00, lo cual refleja un desempeño igualmente sólido. En el caso del cluster 2, el modelo alcanza un F1-score de 0,99, evidenciando un equilibrio muy alto entre precisión y sensibilidad.

Por otro lado, el cluster 1 presenta un recall de 0,87, inferior al de los demás grupos, lo que sugiere que una proporción de sus observaciones fue asignada a otros clusters. Esta situación es consistente con lo observado en la matriz de confusión, donde dicho cluster exhibe algunas discrepancias respecto de la clasificación ideal. En contraste, el cluster 4 alcanza un recall de 1,00, lo que significa que todas las observaciones reales de este grupo fueron identificadas por el modelo. Sin embargo, su precisión desciende a 0,77, indicando que este cluster recibe observaciones pertenecientes a otros grupos, lo cual reduce su capacidad discriminativa específica.

En términos agregados, tanto el promedio macro como el promedio ponderado presentan valores altos y equilibrados. El *macro average* alcanza 0,95 en precisión, 0,96 en recall y 0,95 en F1-score, mientras que el *weighted average* registra 0,96, 0,95 y 0,95, respectivamente. Estos resultados refuerzan la evidencia de que el árbol de decisión reproduce de manera satisfactoria la segmentación original, permitiendo traducir los clusters obtenidos en reglas interpretables y útiles para la caracterización operativa de los perfiles de clientes.

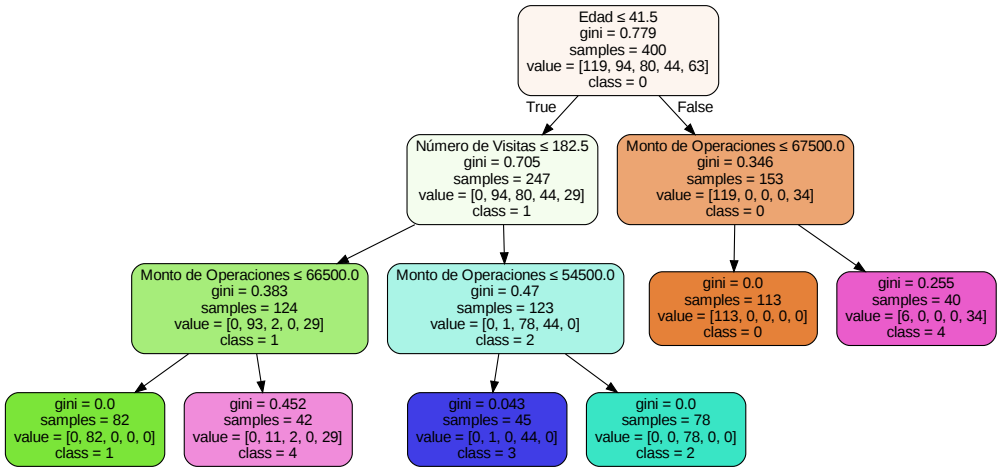
En este sentido, los resultados obtenidos no solo validan la capacidad del modelo para aproximar la segmentación original, sino que también evidencian que los segmentos identificados presentan fronteras relativamente claras en el espacio de variables. Esto es especialmente relevante en el contexto de aplicación empresarial, ya que implica que

los perfiles de clientes pueden ser diferenciados mediante criterios simples y medibles, facilitando su utilización en procesos operativos.

4.7.2 Árbol de decisión

Figura 34

Árbol de decisión para la interpretación de los clusters de clientes.



La Figura 34 presenta la estructura del árbol de decisión generado para aproximar la segmentación de clientes obtenida mediante el algoritmo de clustering. A través de este modelo, es posible identificar reglas jerárquicas construidas a partir de variables como el monto de operaciones, la edad y el número de visitas, las cuales permiten explicar la asignación de los clientes a cada segmento.

La principal ventaja de esta representación radica en que transforma la segmentación, originalmente obtenida por un algoritmo no supervisado, en un conjunto de reglas explícitas y comprensibles. De este modo, cada nodo interno del árbol representa una condición de partición sobre una variable específica, mientras que cada nodo terminal agrupa observaciones con características similares y asociadas predominantemente a un cluster determinado.

Adicionalmente, la estructura del árbol permite identificar la importancia relativa de

las variables en la segmentación. Se observa que variables como el monto de operaciones y el número de visitas aparecen en niveles superiores del árbol, lo que indica su mayor capacidad discriminativa en comparación con la edad. Este hallazgo es consistente con el análisis previo, donde dichas variables mostraron una mayor influencia en la diferenciación de los segmentos.

4.7.3 Reglas de clasificación de los segmentos

A partir del árbol de decisión entrenado, se extrajeron las reglas de clasificación que permiten asignar a cada cliente a un segmento específico en función de sus características. Estas reglas se basan en umbrales definidos sobre las variables edad, monto de operaciones y número de visitas, proporcionando una representación interpretable de la segmentación obtenida.

La Tabla 109 presenta las principales reglas de decisión derivadas del modelo.

Tabla 109

Reglas de clasificación de los segmentos de clientes

Condición	Cluster
Edad >41,5 y Monto de Operaciones $\leq$ 67500	Cluster 0
Edad >41,5 y Monto de Operaciones >67500	Cluster 4
Edad $\leq$ 41,5 y Número de Visitas $\leq$ 182,5 y Monto de Operaciones $\leq$ 66500	Cluster 1
Edad $\leq$ 41,5 y Número de Visitas $\leq$ 182,5 y Monto de Operaciones >66500	Cluster 4
Edad $\leq$ 41,5 y Número de Visitas $>$ 182,5 y Monto de Operaciones $\leq$ 54500	Cluster 3
Edad $\leq$ 41,5 y Número de Visitas $>$ 182,5 y Monto de Operaciones >54500	Cluster 2

Es importante destacar que las reglas presentadas constituyen una aproximación simplificada de la estructura de segmentación, capturando los patrones más representativos identificados por el modelo. Si bien pueden existir casos particulares que no se ajusten estrictamente a estas condiciones, las reglas permiten describir de manera general el comportamiento predominante de cada grupo de clientes.

En términos prácticos, este modelo permite identificar perfiles de clientes a partir de umbrales concretos, facilitando la interpretación de los segmentos y su traducción a decisiones empresariales. Así, por ejemplo, los grupos pueden diferenciarse según

combinaciones de alto o bajo monto de operaciones, mayor o menor frecuencia de visitas y rangos etarios específicos. Esta estructura aporta un valor aplicado importante, ya que hace posible incorporar la segmentación en procesos de negocio como campañas de marketing dirigidas, estrategias de fidelización, priorización comercial y diseño de ofertas diferenciadas.

En consecuencia, el árbol de decisión no solo confirma que la segmentación obtenida posee coherencia interna, sino que además proporciona una herramienta interpretable para comprender cómo se distinguen los distintos perfiles de clientes dentro de la empresa D&C Comunicaciones.

En síntesis, la utilización de un árbol de decisión como modelo sustituto permitió transformar la segmentación obtenida mediante técnicas no supervisadas en un conjunto de reglas explícitas, comprensibles y aplicables en el entorno empresarial. Este proceso no solo facilita la interpretación de los resultados, sino que también contribuye a la integración de la segmentación en la toma de decisiones estratégicas.

De este modo, la investigación logra no solo identificar patrones en los datos, sino también traducirlos en conocimiento accionable, alineando el análisis de ciencia de datos con las necesidades operativas de la empresa D&C Comunicaciones. Esto refuerza el valor práctico del modelo propuesto, permitiendo su implementación en escenarios reales de gestión y planificación comercial.

DISCUSIONES

Los resultados obtenidos en la presente investigación evidencian que la segmentación de clientes mediante algoritmos de inteligencia artificial no supervisada constituye una estrategia efectiva para identificar patrones de comportamiento en los clientes de la empresa D&C Comunicaciones. En particular, el análisis comparativo entre K-Means, Bisecting K-Means y DBSCAN permitió identificar diferencias significativas en el desempeño de los algoritmos, tanto desde una perspectiva descriptiva como inferencial, respaldada por pruebas estadísticas y análisis post-hoc.

En primer lugar, los resultados muestran que los algoritmos basados en centroides, especialmente K-Means y su variante Bisecting K-Means, presentan un desempeño consistente en términos de métricas como el índice de Silhouette y Calinski-Harabasz, lo cual indica una adecuada cohesión intra-cluster y separación inter-cluster. Este hallazgo es coherente con lo reportado por Reddy et al. (2023) y Priyadarshni et al. (2023), quienes destacan la eficiencia de K-Means para segmentar datos estructurados en contextos comerciales. En este sentido, la estructura relativamente simple de los datos utilizados en esta investigación (edad, monto de operaciones y número de visitas) favorece el desempeño de este tipo de algoritmos.

Sin embargo, los resultados también evidencian que DBSCAN presenta un comportamiento diferenciado, especialmente en escenarios donde los datos pueden contener ruido o distribuciones no homogéneas. Aunque en algunos casos sus métricas no superan a K-Means, su capacidad para identificar clusters de forma arbitraria sin necesidad de definir previamente el número de grupos representa una ventaja conceptual importante. Este comportamiento es consistente con lo señalado por Joga et al. (2022) y Jadwal et al. (2022), quienes destacan la robustez de DBSCAN frente a estructuras no lineales. No obstante, tal como indican Ganesh et al., 2024, su rendimiento depende en gran medida de la adecuada selección de parámetros, lo cual se refleja en la variabilidad observada en los resultados experimentales.

Desde una perspectiva estadística, la incorporación de pruebas de normalidad, homogeneidad de varianzas y comparaciones post-hoc permitió validar rigurosamente las

diferencias entre algoritmos. Los resultados de las pruebas globales sugieren la existencia de diferencias estadísticamente significativas en varias de las métricas evaluadas, lo que respalda la hipótesis alternativa planteada en la investigación. Este enfoque fortalece el rigor metodológico del estudio, superando limitaciones identificadas en investigaciones previas como Buleje Calderón (2022) y Salazar Gebol (2015), donde la comparación entre algoritmos no se desarrolló con suficiente profundidad inferencial.

Asimismo, el uso de un enfoque multicriterio para la selección de la mejor segmentación constituye un aporte relevante, ya que permite integrar distintas métricas de evaluación en un único score global. Este enfoque resulta consistente con la necesidad, señalada por Kumar et al. (2025), de considerar múltiples dimensiones de calidad al evaluar algoritmos de clustering, en lugar de depender de una sola métrica. En este sentido, la selección final de la segmentación no solo responde a criterios técnicos, sino también a la interpretabilidad y aplicabilidad de los resultados en el contexto empresarial.

En relación con los antecedentes nacionales, los resultados obtenidos coinciden con lo reportado por Alikhan Trujillo et al. (2023) y Díaz Pulido (2023), quienes señalan que la elección del algoritmo de clustering depende del tipo de datos y del objetivo de la segmentación. En particular, el hecho de que K-Means proporcione segmentaciones más estructuradas mientras que DBSCAN capture patrones más complejos refuerza la idea de que no existe un algoritmo universalmente superior, sino que su idoneidad es contextual.

Por otro lado, la interpretación de los segmentos mediante un modelo supervisado (árbol de decisión) permitió traducir los clusters en reglas comprensibles, lo cual constituye un elemento clave para la aplicabilidad práctica de la segmentación. Este resultado es consistente con lo planteado por Manchego Melendez y Rimay Pelaez (2023), donde se resalta la importancia de generar conocimiento accionable a partir de los modelos de segmentación. La capacidad de explicar los segmentos facilita su adopción en la toma de decisiones empresariales, cerrando la brecha entre análisis técnico y aplicación estratégica.

Finalmente, es importante destacar que los resultados obtenidos deben interpretarse considerando las limitaciones del estudio, especialmente en relación con el tamaño

del dataset y la naturaleza de las variables utilizadas. Tal como señalan Othayoth y Muthalagu (2022) y Paramasivan et al. (2024), la incorporación de más variables y técnicas complementarias, como reducción de dimensionalidad u optimización de hiperparámetros, podría mejorar aún más la calidad de la segmentación.

En síntesis, la presente investigación confirma que los algoritmos de clustering son herramientas efectivas para la segmentación de clientes, y que su desempeño depende tanto de las características de los datos como del enfoque metodológico utilizado. La combinación de análisis estadístico riguroso, evaluación multicriterio e interpretación supervisada permite obtener segmentaciones no solo precisas, sino también útiles para la toma de decisiones empresariales.

CONCLUSIONES

1. Se describieron las características de los clientes de la empresa D&C Comunicaciones mediante análisis estadísticos descriptivos, univariados, bivariados y correlacionales (incluyendo coeficientes de correlación de Pearson y Spearman), evidenciándose que las variables monto de operaciones y número de visitas presentan patrones diferenciados entre clientes, así como relaciones positivas moderadas que permiten distinguir comportamientos de consumo relevantes para la segmentación.

2. Se seleccionaron y preprocesaron adecuadamente las características relevantes de los clientes, aplicando técnicas de limpieza de datos, tratamiento de valores atípicos y normalización (escalamiento Min-Max), que permitieron estandarizar las variables edad, monto de operaciones y número de visitas, reduciendo la dispersión y asegurando que ninguna variable domine el proceso de clustering, lo que se reflejó en una mayor estabilidad en las métricas de evaluación.

3. Se aplicaron algoritmos de inteligencia artificial no supervisados, específicamente K-Means, Bisecting K-Means y DBSCAN, generándose múltiples configuraciones de segmentación mediante variaciones en el número de clusters (k) y parámetros como ϵ y min_samples, donde los algoritmos basados en centroides produjeron agrupaciones más homogéneas, mientras que DBSCAN mostró mayor variabilidad dependiendo de los parámetros utilizados.

4. Se compararon los algoritmos de inteligencia artificial no supervisados mediante métricas como Silhouette, Davies-Bouldin, Calinski-Harabasz e Inercia, evidenciándose que K-Means y Bisecting K-Means alcanzaron consistentemente mejores valores promedio y menor dispersión en Silhouette y Calinski-Harabasz, lo que indica una mejor cohesión intra-cluster y separación inter-cluster en comparación con DBSCAN.

5. Se evaluó la calidad de las segmentaciones obtenidas mediante pruebas estadísticas, incluyendo pruebas de normalidad (Shapiro-Wilk), homogeneidad

de varianzas (Levene) y pruebas inferenciales no paramétricas (Kruskal-Wallis) con comparaciones post-hoc (Dunn), evidenciándose diferencias estadísticamente significativas entre los algoritmos ($p < 0,05$), lo que permitió rechazar la hipótesis nula y confirmar que el desempeño varía según el método de clustering utilizado.

6. Se seleccionó la segmentación más adecuada mediante un método de evaluación multicriterio basado en la ponderación de métricas de calidad (normalizadas y agregadas en un score global), identificándose una configuración óptima dentro del ranking de segmentaciones que presentó el mejor equilibrio entre cohesión, separación y estabilidad, constituyéndose como la solución más robusta para el problema planteado.

7. Se interpretaron los segmentos obtenidos mediante un modelo de inteligencia artificial supervisado basado en árboles de decisión (con parámetros como profundidad máxima y tamaño mínimo de hoja), obteniéndose reglas de clasificación con alta capacidad explicativa que evidencian que variables como el monto de operaciones y la frecuencia de visitas son determinantes en la diferenciación de clientes, permitiendo identificar perfiles claramente definidos y aplicables en estrategias comerciales.

RECOMENDACIONES

1. Se recomienda que futuras investigaciones que aborden problemas de segmentación de clientes adopten un enfoque experimental basado en múltiples ejecuciones y análisis de estabilidad, en lugar de depender de una única corrida de los algoritmos, ya que se evidenció que la variabilidad en la inicialización y parametrización puede influir significativamente en los resultados.

2. Se sugiere incorporar de manera sistemática el uso de métricas de validación interna múltiples (Silhouette, Davies-Bouldin, Calinski-Harabasz e Inercia), evitando la dependencia de una sola métrica, dado que cada una captura distintas propiedades de los clusters (cohesión, separación y compactación), lo que permite una evaluación más integral de la calidad de la segmentación.

3. Se recomienda complementar la evaluación de modelos no supervisados con pruebas estadísticas inferenciales, especialmente en contextos donde se comparan múltiples algoritmos, utilizando enfoques no paramétricos cuando no se cumplan supuestos de normalidad, con el fin de sustentar las conclusiones con evidencia estadística robusta.

4. Se sugiere que los estudios de segmentación adopten metodologías de evaluación multicriterio para la selección de la mejor solución, integrando métricas normalizadas en un score global, lo cual permite tomar decisiones más objetivas y reproducibles frente a escenarios donde las métricas individuales presentan resultados contradictorios.

5. Se recomienda documentar explícitamente el proceso de preprocesamiento de datos, incluyendo técnicas de normalización, tratamiento de valores atípicos y selección de variables, dado que estas etapas tienen un impacto directo en el desempeño de los algoritmos de clustering y son fundamentales para la replicabilidad del estudio.

6. Se sugiere que futuras investigaciones consideren la exploración sistemática del espacio de hiperparámetros (por ejemplo, número de clusters en K-Means o parámetros ϵ y min_samples en DBSCAN), mediante estrategias automatizadas

o experimentales extensivas, con el fin de identificar configuraciones óptimas y comprender la sensibilidad de los algoritmos.

7. Se recomienda complementar los modelos de segmentación con técnicas supervisadas de interpretación, como árboles de decisión, ya que permiten traducir los resultados del clustering en reglas comprensibles y transferibles a contextos prácticos, mejorando la utilidad del modelo más allá de la agrupación numérica.

8. Se sugiere que los estudios en segmentación de clientes utilicen datasets estructurados y variables relevantes que combinen dimensiones demográficas, transaccionales y conductuales, ya que la integración de múltiples perspectivas mejora la capacidad de los algoritmos para identificar patrones significativos.

9. Se recomienda que futuras investigaciones evalúen la robustez de los resultados mediante análisis de consistencia entre diferentes particiones de datos o subconjuntos muestrales, con el objetivo de garantizar que los patrones identificados no sean producto de particularidades del conjunto de datos original.

10. Se sugiere promover la reproducibilidad de los estudios mediante la publicación del código, configuraciones experimentales y criterios de evaluación utilizados, facilitando la validación de los resultados y su aplicación en otros contextos empresariales o académicos.

REFERENCIAS BIBLIOGRÁFICAS

Alikhan Trujillo, K. F., Aspiazu Neyra, L. E., Auccapiña Guillen, J. A., Ayna Benegas, I., & Cardenas Pijo, M. C. (2023). Propuesta de segmentación de clientes aplicando técnicas de Machine Learning para mejorar la estrategia de ventas de productos de bebidas en el departamento de Ica.

Arias, F. G. (2012). *El proyecto de investigación. Introducción a la metodología científica. 6ta*. Fidias G. Arias Odón.

Bartels, C. (2022). Cluster analysis for customer segmentation with open banking data. *2022 3rd Asia Service Sciences and Software Engineering Conference*, 87-94.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning* (Vol. 1). MIT press Cambridge, MA, USA.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Buleje Calderón, W. W. (2022). Segmentación de la cartera de clientes de una empresa textil del Perú en el año 2022.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1-27. https : / / doi . org / 10 . 1080 / 03610927408827101

CEPAL, N. (2013). Economía digital para el cambio estructural y la igualdad.

Davenport, T. H., & Dyché, J. (2013). Big data in big companies. *International Institute for Analytics*, *3*(1-31).

Díaz Pulido, J. A. (2023). Aplicación de la minería de datos espaciales basada en técnicas de agrupamiento al congestionamiento del tráfico vehicular en la Ciudad de Trujillo, Perú.

Dresch, A., Lacerda, D. P., & Antunes Jr, J. A. V. (2014). Design science research. En *Design science research: A method for science and technology advancement* (pp. 67-102). Springer.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*(34), 226-231.

Ganesh, C., Hermalin, C. S., Boomikha, S., Sneha, M., & Shanmugasundaram, R. (2024). Customer Segmentation with Hyperparameter Tuning Using Machine Learning. *2024 International Conference on Smart Systems for Electrical, Electronics, Communication and Computer Engineering (ICSSEECC)*, 279-283. https://doi.org/https://doi.org/10.1109/ICSSEECC61126.2024.10649417

Gartner. (2022). *Market Trends: Personalization in Customer Segmentation* [Gartner Reports].

Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning: data mining, inference, and prediction.

Hernández-Sampieri, R., & Mendoza, C. (2018). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta.

Jadwal, P. K., Pathak, S., & Jain, S. (2022). Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis. *Microsystem Technologies*, *28*(12), 2715-2721. https://doi.org/https://doi.org/10.1504/IJBIDM.2022.123218

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Joga, P., Harshini, B., & Sahay, R. (2022). Comparative Analysis of Machine Learning Models for Customer Segmentation. *International Conference on Intelligent Systems Design and Applications*, 49-61. https://doi.org/https://doi.org/10.1007/978-3-031-35510-3_6

Kotler, P., & Keller, K. L. (2016). *Marketing Management*. Pearson.

Kumar, S., Rani, R., Pippal, S. K., & Agrawal, R. (2025). Customer segmentation in e-commerce: K-means vs hierarchical clustering. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *23*(1), 119-128. https://doi.org/https://doi.org/10.12928/TELKOMNIKA.v23i1.26384

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(14), 281-297.

Manchego Melendez, E. M., & Rimay Pelaez, C. M. (2023). Sistema de información para la segmentación de clientes por medio del algoritmo de K-Means para el Área de Marketing del sector de Medios de Comunicación.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., Barton, D., et al. (2012). Big data. *The management revolution. Harvard Bus Rev*, *90*(10), 61-67.

Michael Chui, M. M., James Manyika. (2018). AI adoption advances, but foundational barriers remain. *McKinsey Quarterly*. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/ai-adoption-advances-but-foundational-barriers-remain

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Narayana, V. L., Sirisha, S., Divya, G., Pooja, N. L. S., & Nouf, S. A. (2022). Mall customer segmentation using machine learning. *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 1280-1288. https://doi.org/https://doi.org/10.1109/ICEARS53579.2022.9752447

Othayoth, S. P., & Muthalagu, R. (2022). Customer segmentation using various machine learning techniques. *International Journal of Business Intelligence and Data Mining*, *20*(4), 480-496.

Paramasivan, C., Dhinakaran, D. P., Selvam, S. S. P., Mukherjee, S., Juliet, A. P., & Devi, S. R. (2024). Comparing Supervised and Unsupervised Learning Technologies for Customer Segmentation in Marketing. *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, 1-6. https://doi.org/https://doi.org/10.1109/ICONSTEM60960.2024.10568702

Peña Escobar, S., Ramírez Reyes, G. S., & Osorio Gómez, J. C. (2015). Evaluación de una estrategia de fidelización de clientes con dinámica de sistemas. *Revista Ingenierías Universidad de Medellín*, *14*(26), 87-104.

Preethi, P., Sathvika, S. V. D., Vyshnavi, D., & Kumar, P. V. J. (2023). Enhancement of shopping experience for customers at malls. En *Artificial Intelligence, Blockchain, Computing and Security Volume 2* (pp. 224-229). CRC Press. https://doi.org/https://doi.org/10.1201/9781032684994-34

Priyadarshni, S., Fathima, R., Urolagin, S., Bongale, A. M., & Dharrao, D. S. (2023). Unveiling Customer Segmentation Patterns in Credit Card Data using K-Means Clustering: A Machine Learning Approach. *2023 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom)*, 362-366. https://doi.org/https://doi.org/10.1109/MoSICom59118.2023.10458783

Rahmet, S., & Gürkan, F. (2024). Integration of Pre-processing and Machine Learning Methods for Improved Customer Segmentation. *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1-6. https://doi.org/https://doi.org/10.1109/IDAP64064.2024.10710767

Reddy, B. S. V., Rishikeshan, C., Dagumati, V., Prasad, A., & Singh, B. (2023). Customer segmentation analysis using clustering algorithms. *International Conference on Machine Learning, IoT and Big Data*, 353-368. https://doi.org/https://doi.org/10.1007/978-981-99-3932-9_31

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Pearson Education.

Salazar Gebol, J. S. (2015). Segmentación de clientes en base a su comportamiento de consumo a través del modelo de segmentación K-MEANS en una entidad Bancaria.

Schiffman, L. G., Wisenblit, J., & Kumar, S. R. (2019). *Consumer Behavior | By Pearson*. Pearson Education India.

Seidor. (2023). Clusterización y su papel en la captación y fidelización de clientes. *Seidor Blog*. https://www.seidor.com/es-es/blog/clusterizacion-papel-captacion-fidelizacion-clientes

Slupu. (2023). Clustering en marketing digital: cómo segmentar tu público. *Slupu Blog*. https://slupu.com/blog/clustering-en-marketing-digital-segmentacion/

Solomon, M., Russell-Bennett, R., & Previte, J. (2012). *Consumer behaviour*. Pearson Higher Education AU.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.

Stone, M., & Jacobs, N. (2008). *Successful Direct Marketing Methods*. McGraw-Hill.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining.

Wieringa, R. (2014). *Design science methodology for information systems and software engineering*. Springer.

ANEXOS

ANEXO 1: MATRIZ DE CONSISTENCIA

SEGMENTACIÓN DE CLIENTES UTILIZANDOALGORITMOS DE INTELIGENCIA ARTIFICIALEN LA EMPRESA D&C COMUNICACIONES

<table>
<tr><th>PROBLEMA</th><th>OBJETIVO</th><th>HIPÓTESIS</th><th colspan="2">VARIABLES E INDICADORES</th><th>METODOLOGÍA</th></tr>
<tr>
<td>PRINCIPAL:
¿Cómo segmentar los clientes de la empresa D&C Comunicaciones utilizando algoritmos de inteligencia artificial de manera que se obtengan segmentos interpretables y útiles para la toma de decisiones en la empresa?
SUBPROBLEMAS
• ¿Cuáles son las características de los clientes de la empresa D&C Comunicaciones?
• ¿Cómo seleccionar y preprocesar las características relevantes de los clientes de la empresa D&C Comunicaciones para su utilización en algoritmos de inteligencia artificial?
• ¿Cómo generar segmentaciones de clientes de la empresa D&C Comunicaciones mediante algoritmos de inteligencia artificial no supervisados?
• ¿Qué algoritmo de inteligencia artificial no supervisado genera la mejor segmentación de clientes de la empresa D&C Comunicaciones?
• ¿Cómo evaluar la calidad de las segmentaciones obtenidas de los clientes de la empresa D&C Comunicaciones mediante métricas de clusterización?
• ¿Cómo seleccionar la segmentación más adecuada de los clientes de la empresa D&C Comunicaciones mediante un método de evaluación multicriterio basado en métricas de calidad de segmentación?
• ¿Cómo interpretar los segmentos obtenidos de los clientes de la empresa D&C Comunicaciones mediante un modelo de inteligencia artificial supervisado que permita generar reglas de clasificación comprensibles para la empresa?</td>
<td>GENERAL:
Segmentar los clientes de la empresa D&C Comunicaciones utilizando algoritmos de inteligencia artificial, con el fin de obtener segmentos interpretables y útiles para la toma de decisiones en la empresa.
ESPECÍFICOS:
• Describir las características de los clientes de la empresa D&C Comunicaciones.
• Seleccionar y preprocesar las características relevantes de los clientes de la empresa D&C Comunicaciones para su utilización en algoritmos de inteligencia artificial.
• Aplicar algoritmos de inteligencia artificial no supervisados para generar segmentaciones de clientes de la empresa D&C Comunicaciones.
• Comparar los algoritmos de inteligencia artificial no supervisados a través de sus métricas en la generación de segmentaciones de clientes de la empresa D&C Comunicaciones.
• Evaluar la calidad de las segmentaciones obtenidas de los clientes de la empresa D&C Comunicaciones mediante métricas de clusterización.
• Seleccionar la segmentación más adecuada de los clientes de la empresa D&C Comunicaciones mediante un método de evaluación multicriterio basado en métricas de calidad de segmentación.
• Interpretar los segmentos obtenidos de los clientes de la empresa D&C Comunicaciones mediante un modelo de inteligencia artificial supervisado</td>
<td>GENERAL:
Es posible segmentar los clientes de la empresa D&C Comunicaciones utilizando algoritmos de inteligencia artificial, con el fin de obtener segmentos interpretables y útiles para la toma de decisiones en la empresa.
ESPECÍFICAS :
• Es posible describir las características de los clientes de la empresa D&C Comunicaciones.
• Es posible seleccionar y preprocesar las características relevantes de los clientes de la empresa D&C Comunicaciones para su utilización en algoritmos de inteligencia artificial.
• Es posible aplicar algoritmos de inteligencia artificial no supervisados para generar segmentaciones de clientes de la empresa D&C Comunicaciones.
• Existen diferencias estadísticamente significativas en al menos una de las métricas de calidad de segmentación entre los algoritmos de inteligencia artificial no supervisados evaluados.
• Es posible evaluar la calidad de las segmentaciones obtenidas de los clientes de la empresa D&C Comunicaciones mediante métricas de clusterización.
• Es posible seleccionar la segmentación más adecuada de los clientes de la empresa D&C Comunicaciones mediante un método de evaluación multicriterio basado en métricas de calidad de segmentación.
• Es posible interpretar los segmentos obtenidos de los clientes de la empresa D&C Comunicaciones mediante un modelo de inteligencia artificial supervisado.</td>
<td colspan="2">Variable 1:
Algoritmos inteligencia artificial

Variable 2:
Segmentación de clientes

Dimensiones | Indicadores

Métricas
• Índice de Silhouette.
• Índice de David-Bouldin.
• Índice de Calinski-Harabasz.
• Inercia</td>
<td>Tipo de investigación:
Aplicada

Diseño de investigación:
Experimental

Población:
400 clientes
Muestra:
400 clientes

Técnicas de recolección de datos:
Observación Estruturada

Instrumento de recolección de datos:
Ficha digital de observación</td>
</tr>
</table>