CLUSTERIZACIÓN APLICADA A EMPRESAS DEL SECTOR ENERGÉTICO QUE REPORTAN INDICADORES ESG (Ambiental, Social y de Gobernanza)

María Isabel Fernández Acosta

Código: 8988488

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director(a)

PhD. Orlando Joaqui Barandica

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, 07/07/2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

El presente proyecto abordó la necesidad de segmentar empresas del sector energético en función de sus indicadores ESG (Ambientales, Sociales y de Gobernanza), con el fin de identificar patrones comunes de sostenibilidad y facilitar la toma de decisiones estratégicas en ámbitos como la inversión responsable, la auditoría de sostenibilidad y la formulación de políticas públicas. La problemática se centró en la ausencia de una clasificación estructurada que distinguiera los diferentes niveles de compromiso ESG entre organizaciones, dificultando el análisis comparativo y la identificación de riesgos reputacionales o deficiencias estructurales en sostenibilidad corporativa.

Para dar respuesta a esta necesidad, se desarrolló un modelo de agrupamiento no supervisado utilizando técnicas de minería de datos, específicamente algoritmos de clusterización como K-Means y agrupamiento jerárquico. El proyecto se sustentó en una base de datos compuesta por 576 empresas del sector energético, cuyos indicadores ESG se extrajeron de la plataforma London Stock Exchange Group (LSEG). La información se sometió a un riguroso proceso de limpieza, estandarización y reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA), lo cual permitió optimizar el desempeño de los modelos entrenados.

Si bien se exploraron múltiples configuraciones de agrupamiento, incluyendo modelos con dos, tres y cuatro clústeres, la selección final se inclinó por la solución con dos clústeres en algunos casos debido a sus métricas superiores de cohesión interna y separación entre grupos, como lo evidencian los índices de silueta y Calinski-Harabasz. No obstante, este enfoque se complementó con un análisis comparativo multiclúster para evitar interpretaciones sesgadas y proporcionar una visión más integral de la estructura latente en los datos. De este modo, se buscó un balance entre simplicidad interpretativa y solidez técnica.

El modelo que presentó mejor desempeño fue K-Means con PCA=5 y K=40, ya que logró una segmentación coherente, con alta cohesión intragrupal y una clara separación entre clústeres. Esta configuración alcanzó un coeficiente de silueta de 0.5354 y un índice de Calinski-Harabasz superior a 10.000, lo que evidenció su robustez técnica. No obstante, se optó por utilizar el modelo K-Means con K=2 para la interpretación de perfiles, debido a que ofreció una estructura más clara y

manejable para el análisis cualitativo de los grupos, facilitando la identificación de empresas con alto desempeño ESG frente a aquellas con desafíos estructurales en sostenibilidad.

Los resultados obtenidos permitieron caracterizar dos perfiles empresariales opuestos, aportando un marco técnico sólido para decisiones estratégicas basadas en sostenibilidad. El clúster 0 agrupó a empresas con alto rendimiento ESG, bajo nivel de controversias y gobernanza sólida, mientras que el clúster 1 evidenció prácticas más heterogéneas y menor alineación con estándares internacionales. Este proyecto no solo permitió validar la utilidad de técnicas de aprendizaje no supervisado en contextos de análisis ESG, sino que también sentó las bases para futuras investigaciones centradas en la evolución temporal de las métricas de sostenibilidad y su relación con el desempeño financiero.

El análisis evidenció que, si bien la configuración con dos clústeres presentó el mejor desempeño según los índices de validación interna, otras configuraciones como la de tres clústeres también ofrecieron información relevante sobre subgrupos con características diferenciadas. Por tanto, la decisión de destacar los resultados con $K=2$ obedeció a criterios de calidad del agrupamiento y no a un sesgo interpretativo. Este enfoque no solo permitió simplificar la interpretación operativa, sino también mantener coherencia con las necesidades prácticas del estudio.

DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento del problema

La importancia del sector energético en la evaluación por criterios ESG se considera importante, dado su peso en los impactos ambientales y sociales ya que es un agente estratégico clave en la transición hacia fuentes limpias. Además, la presión normativa se intensificó durante los últimos años, exigiendo mayor transparencia en la divulgación de prácticas sostenibles [1]. Esto ha llevado a una necesidad de evaluación comparativa que pudiera proporcionar comparaciones equivalentes para las empresas, independientemente de su escala o ubicación geográfica. Por tal motivo, se demostró que la clusterización era el método prometedor para identificar grupos con rendimiento ESG similar y, como resultado, proporcionar apoyo sensible a la toma de decisiones disponibles para múltiples partes interesadas.

Asimismo, la heterogeneidad que existe en las estructuras de las empresas energéticas fue otro obstáculo, impidiendo una comparación directa de sus desempeños de sostenibilidad, ya que las diferencias en los modelos de negocio, las carteras de generación y los regímenes regulatorios sesgaban los informes [2]. Por tanto, era necesario normalizar los datos y utilizar herramientas eficaces para agrupar pantallas similares sin sacrificar los mensajes deseados. De hecho, este desafío metodológico motivó la investigación en algoritmos no supervisados que pudieran manejar la complejidad natural de las variables ESG.

Por el contrario, se encontró que el uso de los algoritmos de K-means++ ayudaba a descubrir patrones de rendimiento financiero y ambiental para los cuales los medios descriptivos tradicionales no eran exitosos [3]. Sin embargo, estos algoritmos no eran efectivos si los parámetros no se establecían correctamente, las métricas de distancia elegidas eran inapropiadas y si la agrupación no se validaba adecuadamente. En consecuencia, el éxito de la agrupación dependía de una exploración exhaustiva de la estabilidad y coherencia en los resultados.

Las métricas ESG también se consolidaron como características estratégicas de la evaluación corporativa, integrándose en la arquitectura regulatoria global como la Directiva de Informes de Sostenibilidad Corporativa de la UE (CSRD) y los estándares internacionales de informes. Como resultado, las empresas energéticas se enfrentaron a la necesidad de realinear sus procesos de

gestión de datos para satisfacer tales directrices con el fin de cumplir con las demandas de transparencia. De manera similar, esto implicó una revisión continua de las publicaciones y la redacción de informes de alta fidelidad.

Al mismo tiempo, esta sostenibilidad corporativa se volvió instrumental en la captación de capital, los inversores aplicaron modelos de inversión ESG y calificaron a las empresas por su compromiso con la gestión responsable, etc. En este caso, quienes podrían ser líderes tempranos en sostenibilidad tomarían una ventaja competitiva, en reputación, así como en financiamiento verde. Entonces, las decisiones de inversión continuarían basándose en la capacidad de las empresas para proporcionar evidencia de mejora en las métricas ESG.

El control integral de las controversias y riesgos relacionados con ESG era difícil de implementar, y la mayoría de las empresas tenían dificultades para integrar informes precisos de incidentes de gobernanza y conflictos sociales. Sin embargo, aquellas empresas que tenían sistemas de seguimiento y auditoría interna podrían reducir el daño reputacional y aumentar la confianza de los interesados. Por lo tanto, la retención y la trazabilidad de eventos eran cuestiones fundamentales en la arquitectura de los modelos de agrupación.

A lo anterior se suma que, los indicadores ESG se consolidaron como elementos estratégicos de evaluación corporativa, integrándose en marcos normativos como la Directiva de Informes de Sostenibilidad Corporativa (CSRD) de la Unión Europea y en estándares internacionales de reporte [4]. Por tanto, las empresas energéticas tuvieron que alinear sus procesos de gestión de datos con dichos lineamientos para cumplir con los requisitos de transparencia. Asimismo, esto implicó la revisión continua de protocolos internos y la generación de informes periódicos con alta fidelidad.

Por otro lado, la sostenibilidad corporativa se convirtió en un factor determinante para la atracción de capital, ya que los inversionistas adoptaron modelos de inversión basados en criterios ESG y evaluaron a las compañías según su compromiso con prácticas responsables [5]. En este escenario, quienes lograron posicionarse como líderes en sostenibilidad alcanzaron ventajas competitivas, tanto en reputación como en acceso a financiamiento verde. Por consiguiente, las decisiones de inversión pasaron a depender cada vez más de la capacidad de las empresas para demostrar mejoras continuas en sus métricas ESG [6].

Sin embargo, la gestión de las controversias y riesgos asociados a ESG presentó desafíos, ya que múltiples organizaciones enfrentaron obstáculos para consolidar información precisa sobre incidentes de gobernanza y conflictos sociales [7]. No obstante, aquellos actores que implementaron sistemas de seguimiento y auditoría interna lograron mitigar impactos reputacionales y fortalecer la confianza de sus stakeholders. En consecuencia, la calidad de los datos y la trazabilidad de los eventos pasaron a ser componentes críticos en la arquitectura de los modelos de clusterización [8].

Para afrontar estas complejidades, se recurrió a modelos de agrupamiento jerárquico y a algoritmos de densidad, como DBSCAN, que permitieron segmentar a las organizaciones según similitudes intrínsecas en sus perfiles ESG [9]. Además, se integraron métricas de validación interna, tales como el coeficiente de silueta, para asegurar la consistencia de los clústeres. No obstante, resultó necesario complementar estos enfoques con análisis de estabilidad, a fin de garantizar que los patrones identificados no fuesen producto de fluctuaciones aleatorias en los datos.

Por otra parte, la publicación de estudios de caso en medios financieros destacó las consecuencias no deseadas de la inversión ESG, revelando la importancia de identificar agrupaciones de empresas con riesgos ocultos y oportunidades de mejora [10]. De este modo, la clusterización contribuyó a diseñar estrategias de supervisión regulatoria más focalizadas y a optimizar la asignación de recursos en iniciativas de sostenibilidad. Asimismo, los proveedores y socios comerciales se beneficiaron al discernir con mayor precisión a las empresas con las que establecían alianzas estratégicas [11].

En el plano metodológico, la adopción de marcos analíticos avanzados, como el método MEREC combinado con K-means, permitió evaluar la relevancia relativa de cada pilar ESG y priorizar las variables más influyentes en la segmentación [12]. De hecho, esta aproximación enriqueció el proceso de clusterización al introducir ponderaciones dinámicas, lo que condujo a resultados más representativos de las dinámicas reales del sector. Por consiguiente, se optimizó la capacidad predictiva de los modelos y se redujo el sesgo en la formación de clústeres.

Cabe destacar que, en el contexto de mercados emergentes, la calidad y la transparencia de las divulgaciones ESG condicionaron la factibilidad de aplicar técnicas de agrupamiento, puesto que

inconsistencias y vacíos informativos podían comprometer la validez de los hallazgos \[10]. Por tanto, antes de la etapa de clusterización, se implementaron rigurosos procesos de limpieza y validación de datos, así como la imputación de valores faltantes con métodos estadísticos avanzados. De esta forma, se aseguró que el análisis no se viese afectado por artefactos derivados de la ausencia de datos.

Finalmente, la tendencia de estos esfuerzos facilitó la transformación de grandes volúmenes de información ESG en conocimiento estructurado y accionable, lo que, a su vez, promovió la comparabilidad sectorial y la toma de decisiones informadas. Así, la clusterización se consolidó como una herramienta esencial para impulsar la sostenibilidad corporativa en el sector energético y para apoyar la transición hacia un modelo de negocio más responsable y competitivo.

1.1 Formulación del problema

¿De qué manera puede diseñarse un modelo de aprendizaje no supervisado para la clusterización de empresas que divulgan indicadores ESG, de modo que se identifiquen patrones y se segmente dichas compañías según sus prácticas de sostenibilidad, facilitando la toma de decisiones informadas para inversores, reguladores y la sociedad en general?

2 OBJETIVOS DEL PROYECTO

2.1 Objetivo general

Desarrollar un modelo de aprendizaje no supervisado para la clusterización de empresas que divulgan indicadores ESG (Ambiental, Social y de Gobernanza), con el fin de identificar patrones y segmentar dichas empresas según sus prácticas de sostenibilidad, facilitando así la toma de decisiones informadas para inversores, reguladores y la sociedad en general.

2.2 Objetivos específicos

- Identificar y recopilar datos estructurados de indicadores de ESG de las empresas.
- Implementar un modelo de aprendizaje no supervisado para la clusterización de empresas que reportan indicadores ESG.
- Evaluar el desempeño del modelo de aprendizaje no supervisado desarrollado para la clusterización de empresas que reportan indicadores ESG.
- Elaborar un informe de los resultados del análisis de indicadores de ESG.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1. Marco teórico

A continuación, se presenta el marco teórico el cual establece el contexto fundamental y la relevancia del estudio, conectando de manera clara los conceptos clave con la problemática investigada, lo que permite sustentar y orientar el desarrollo del análisis.

3.1.1 Sector energético

El sector energético global desempeña un papel fundamental en la sociedad contemporánea, constituyendo la base de la economía moderna y permitiendo el progreso tecnológico y social [13]. La disponibilidad de energía impulsa la inversión, la innovación y la generación de empleo, favoreciendo un crecimiento inclusivo y una prosperidad compartida a nivel mundial [13]. Sin embargo, este sector es también uno de los mayores responsables del impacto ambiental planetario, al estar vinculado con la mayor parte de las emisiones de gases de efecto invernadero que provocan el cambio climático. Al mismo tiempo, persisten brechas significativas de acceso: cientos de millones de personas aún carecen de electricidad o dependen de combustibles tradicionales altamente contaminantes para cocinar, lo que evidencia retos sociales apremiantes en materia de equidad y desarrollo. En consecuencia, equilibrar la expansión energética con la sostenibilidad ambiental y la inclusión social se ha convertido en un desafío central, requiriendo marcos regulatorios y políticas públicas claras a nivel internacional y nacional para una transición energética sostenible [14].

América Latina presenta un perfil energético particular, marcado por abundantes recursos tanto de hidrocarburos como de energías renovables. Gracias a una alta participación de fuentes renovables (notablemente la energía hidráulica) en su matriz eléctrica, el uso de combustibles fósiles en la región equivale aproximadamente a dos tercios de la energía total consumida, por debajo del promedio mundial de 80% [15]. En consecuencia, la contribución histórica de América Latina al cambio climático ha sido relativamente menor: sólo alrededor del 5% de las emisiones energéticas acumuladas globales se han originado en la región, a pesar de que ésta generó cerca

del 9% del PIB mundial desde 1971 [15]. No obstante, los países latinoamericanos son particularmente vulnerables a los efectos del calentamiento global, ya que 13 de las 50 naciones más afectadas por el cambio climático se encuentran en esta región [14]. Este contexto resalta la urgencia de promover una transición energética verde ambiciosa que reduzca emisiones sin sacrificar el crecimiento, fortaleciendo la resiliencia climática de los países latinoamericanos [16].

En el ámbito latinoamericano, el sector energético también es un motor económico y social clave, pero enfrenta desafíos de inclusión. La mayoría de la población regional accede a la electricidad, pero aún existen unos 17 millones de personas sin conexión eléctrica y alrededor de 74 millones que dependen de leña u otros combustibles contaminantes para cocinar una muestra de la persistente pobreza energética regional; además, el 10% más rico de los habitantes es responsable de cerca del 40% de las emisiones asociadas al consumo energético [15].

Ante esta realidad, los gobiernos de la región han reconocido que una transición hacia energías limpias puede ser una oportunidad para el desarrollo sostenible. Estudios recientes estiman que una transición verde efectiva podría aumentar en un 10,5% los nuevos empleos en América Latina hacia 2030 [17], a la vez que reduce la dependencia de combustibles fósiles importados y mejora el bienestar de las comunidades. De este modo, la transformación del sector energético regional, con las políticas adecuadas, podría convertirse en un catalizador de crecimiento económico inclusivo y de reducción de la desigualdad [18].

En Colombia, la relevancia del sector energético se manifiesta tanto en sus impactos ambientales como en las políticas climáticas del país. Aunque Colombia contribuye con menos del 1% de las emisiones globales de GEI, a nivel interno su sector energético genera alrededor de un 30% de las emisiones nacionales, mientras que el sector asociado al uso de la tierra (agricultura, bosques y otros) aporta cerca del 59% [19]. Consciente de esta distribución, el país ha asumido compromisos climáticos ambiciosos: en 2020 el Gobierno actualizó su Contribución Determinada a Nivel Nacional (NDC) y anunció una meta de reducir en 51% las emisiones de GEI para 2030 [20], con miras a alcanzar la carbono-neutralidad en 2050. La transición energética resultante exige acelerar el paso de los combustibles fósiles hacia fuentes más limpias, a la par de intensificar la lucha contra la deforestación y otros factores no energéticos. Por otro lado, el desarrollo del sector ha

producido impactos locales significativos: en zonas extractivas de petróleo y minería se han documentado problemas de contaminación de agua y perjuicios a ecosistemas que afectan a comunidades vulnerables [21]. Estos desafíos subrayan la necesidad de conciliar la política energética con la protección ambiental y los derechos de dichas comunidades, en el marco de una transición justa.

Desde la perspectiva económica y social colombiana, el sector energético (incluyendo hidrocarburos y minería) es un pilar estratégico del desarrollo nacional. Representa aproximadamente el 12% del PIB de Colombia y genera más de la mitad de las exportaciones totales del país, al tiempo que aporta importantes recursos fiscales –solo la industria petrolera contribuye con cerca del 2,5% del PIB en impuestos (incluyendo las utilidades de la empresa estatal Ecopetrol) y las regiones productoras perciben otro ~1,5% del PIB en regalías [22].

Este flujo de ingresos ha permitido que las empresas energéticas contribuyan tanto al erario como al bienestar social; por ejemplo, el gremio de generadores eléctricos destinó 4,3 billones de pesos en aportes fiscales y 293.000 millones en inversiones sociales y ambientales según un informe reciente [23]. No obstante, esta dependencia económica entraña retos de sostenibilidad a futuro, pues un retiro precipitado de los combustibles fósiles sin alternativas productivas equivalentes podría provocar pérdidas fiscales estimadas en 9,3 billones de pesos y afectar gravemente a las regiones petroleras y carboníferas [24]. Por ello, la transición energética en Colombia requiere una planificación cuidadosa que diversifique la economía, proteja los ingresos públicos y promueva opciones laborales para las comunidades dependientes del sector.

En cuanto al contexto regulatorio de la energía en Colombia, el país ha desarrollado un marco institucional y político integral para gestionar el sector. En 2021 se lanzó la Estrategia Colombia Carbono Neutral (ECCN) como parte de la Estrategia de Largo Plazo E2050, y en 2022 el gobierno aprobó el CONPES 4075, que establece líneas de acción para la transición energética y la descarbonización en los próximos años [21]. Más recientemente, bajo la administración iniciada en 2022, la política energética colombiana se ha orientado a acelerar el desarrollo de las energías renovables y reducir la dependencia de los combustibles fósiles, en consonancia con las metas climáticas asumidas [25]. En el plano institucional, el Ministerio de Minas y Energía es la entidad

rectora del sector desde 1974, y existen órganos reguladores especializados. La Comisión de Regulación de Energía y Gas (CREG), creada por ley en 1994, es la encargada de regular los mercados de electricidad y gas combustible, promoviendo la competencia y asegurando la prestación eficiente de estos servicios públicos [26]. Asimismo, agencias como la Agencia Nacional de Hidrocarburos (ANH) gestionan la administración de los recursos petroleros, y la Autoridad Nacional de Licencias Ambientales (ANLA) vela por la evaluación y mitigación de los impactos ambientales de los proyectos energéticos. Este andamiaje regulatorio, junto con la participación de Colombia en acuerdos internacionales sobre energía y clima, proporciona el contexto normativo para orientar al sector energético del país hacia un futuro más sustentable, equilibrando seguridad energética, desarrollo económico y protección ambiental.

3.1.2 Concepto y evolución de los indicadores ESG

El concepto de indicadores ESG (Environmental, Social and Governance) nace como una herramienta integral para evaluar el desempeño de las empresas más allá de los parámetros financieros tradicionales, incorporando aspectos ambientales, sociales y de gobernanza que reflejan su compromiso con la sostenibilidad y la responsabilidad corporativa [13]. Estos indicadores han evolucionado desde su introducción inicial en la década de 2000, consolidándose como un estándar global en la inversión responsable y la gestión empresarial, permitiendo identificar riesgos y oportunidades ligados a la sostenibilidad [14].

La evolución de los ESG ha implicado una mayor sofisticación en sus métricas, incluyendo la incorporación de puntuaciones específicas que analizan desde el impacto ambiental directo, pasando por la calidad de las relaciones sociales, hasta la transparencia y ética en la estructura de gobierno corporativo, fortaleciendo así su valor como herramienta para la toma de decisiones estratégicas y financieras. Esta transformación ha impulsado que tanto inversionistas como empresas integren los criterios ESG en sus procesos de evaluación y reporte, promoviendo prácticas que contribuyen al desarrollo sostenible y a la creación de valor a largo plazo. La Tabla 1. Presenta los conceptos de los indicadores ESG:

Tabla 1 *Descripción de Indicadores ESG (Ambiental, Social y de Gobernanza)*

Indicador ESG	Descripción	Pilar Asociado
Medio Ambiente (E)	Considera el efecto e impacto directo e indirecto que tienen las actividades de la empresa desde el punto de vista medioambiental, incluyendo la gestión de recursos naturales, emisiones y prácticas sostenibles.	Ambiental
Social (S)	Evalúa el comportamiento de la empresa en sus relaciones con empleados, clientes y comunidades, abarcando aspectos como derechos humanos, diversidad, condiciones laborales y desarrollo social.	Social
Gobierno Corporativo (G)	Examina la estructura de gobierno para garantizar que la alta dirección y los consejos sigan prácticas éticas, transparentes y responsables, promoviendo mecanismos de control y protección de los accionistas.	Gobernanza
ESG Score	Calificación general basada en la información auto-reportada en los pilares ambiental, social y de gobernanza, que refleja el desempeño integral de la empresa.	Global (E+S+G)
ESG Combined Score	Puntuación que incluye la calificación ESG más un análisis adicional de controversias relacionadas con aspectos ESG, proporcionando una visión más completa del riesgo reputacional.	Global (E+S+G + controversias)
ESG Controversies Score	Mide la exposición de la empresa a controversias y eventos negativos en ámbitos ESG, reflejados en medios globales, alertando sobre posibles riesgos no capturados en puntuaciones tradicionales.	Riesgo reputacional
Environmental Pillar Score	Mide el impacto en sistemas naturales, calidad del aire, agua, tierra y ecosistemas, así como la adopción de prácticas para	Ambiental

Indicador ESG	Descripción	Pilar Asociado
	mitigar riesgos ambientales y generar valor sostenible a largo plazo.	
Social Pillar Score	Evalúa la capacidad para generar confianza y lealtad en fuerza laboral, clientes y sociedad, reflejando la reputación y licencia social para operar.	Social
Governance Pillar Score	Analiza los sistemas que aseguran que la dirección actúe en el mejor interés de los accionistas, con prácticas de gestión que equilibran derechos y responsabilidades.	Gobernanza
Resource Use Score	Refleja el desempeño en la reducción del uso de materiales, energía y agua, así como la ecoeficiencia en la cadena de suministro.	Ambiental
Emissions Score	Mide el compromiso y efectividad en la reducción de emisiones contaminantes en procesos productivos y operativos.	Ambiental
Innovation Score	Evalúa la capacidad para crear tecnologías y procesos que reduzcan costos y cargas ambientales, generando nuevas oportunidades de mercado.	Ambiental
Workforce Score	Mide la efectividad en la satisfacción laboral, seguridad, diversidad, igualdad de oportunidades y desarrollo profesional.	Social
Human Rights Score	Evalúa el respeto a las convenciones fundamentales de derechos humanos en la operación empresarial.	Social
Community Score	Considera inversiones y financiamiento dirigidos a comunidades vulnerables, desarrollo económico y social, más allá de simples donaciones.	Social
Product	Refleja la capacidad para producir bienes y servicios de	Social /

Indicador ESG	Descripción	Pilar Asociado
Responsibility Score	calidad, integrando salud, seguridad, integridad y privacidad de datos para los clientes.	Gobernanza
Management Score	Mide el compromiso y efectividad en la adopción de mejores prácticas de gobernanza corporativa.	Gobernanza
Shareholders Score	Evalúa el trato equitativo a accionistas y la gestión de mecanismos defensivos contra adquisiciones hostiles.	Gobernanza
CSR Strategy Score	Refleja la integración de dimensiones económicas, sociales y ambientales en la toma de decisiones estratégicas diarias de la empresa.	Gobernanza / Global

Fuente. Elaboración propia. Tomando información de [13] [14].

3.1.3 Tipologías y clusterización en la gestión de riesgos LAFT

En el marco de la gestión de riesgos LAFT (lavado de activos y financiación del terrorismo), resulta fundamental comprender el concepto de tipología, dado que constituye uno de los elementos clave para generar señales de alerta precisas y orientar de manera focalizada los procesos de monitoreo. El Grupo de Acción Financiera Internacional (GAFI), organismo intergubernamental dedicado a formular políticas para combatir estas prácticas ilícitas, define las tipologías como los patrones o métodos operativos empleados por organizaciones criminales para colocar, ocultar e integrar fondos provenientes de actividades ilícitas o aparentes lícitas [31]. Cuando ciertas conductas o acciones se presentan con una estructura organizada y recurren a métodos similares, pueden clasificarse como herramientas para la identificación de señales de alerta.

Estas señales son eventos o circunstancias cuya detección requiere un análisis más profundo para esclarecer posibles indicios de lavado de activos o financiación del terrorismo. Es importante considerar que cada señal de alerta debe analizarse de manera aislada, sin asumir que la detección de un caso depende necesariamente de la presencia simultánea de múltiples señales [32].

En este contexto, el uso del aprendizaje automático permite desarrollar modelos de clusterización

que agrupan objetos similares en función de características comunes; específicamente, clientes que manifiestan señales de alerta. Para la creación de dichos modelos, la librería Python sklearn es ampliamente reconocida y utilizada, dado que ofrece una extensa variedad de algoritmos de clustering. Adicionalmente, PyCaret facilita la automatización de flujos de trabajo de Machine Learning, optimizando la selección del modelo más adecuado [33]. A continuación, se detalla el funcionamiento de las librerías y las técnicas utilizadas para este propósito.

Pycaret

PyCaret es una biblioteca de Python que simplifica y automatiza los procesos del aprendizaje automático, abarcando desde la preparación de datos hasta la implementación rápida de modelos finales. Facilita la comparación automática entre múltiples modelos para ayudar a seleccionar la opción más adecuada según el problema planteado, lo que optimiza significativamente el ciclo de experimentación y aumenta la productividad. Entre sus principales ventajas destaca la capacidad de reducir cientos de líneas de código a solo unas pocas, funcionando como un contenedor que integra diversas bibliotecas populares como scikit-learn, XGBoost, LightGBM, CatBoost, Optuna y Hyperopt. Su diseño se basa en la inspiración de la biblioteca caret del lenguaje R, buscando ofrecer una experiencia sencilla y eficiente para el desarrollo de modelos de machine learning [34].

Scikit-Learn (Sklearn)

Es una biblioteca de Python que ofrece una amplia gama de algoritmos para tareas de clasificación, regresión, agrupamiento y reducción de dimensionalidad. Esta herramienta incluye módulos para el procesamiento de datos, evaluación y ajuste de modelos, así como la selección de características, facilitando así el desarrollo integral de proyectos de aprendizaje automático. Entre sus principales ventajas destacan una sintaxis uniforme para todos los modelos, su construcción sobre librerías como NumPy, SciPy y matplotlib que optimizan su desempeño, y el soporte para algoritmos modernos como KNN, XGBoost, bosques aleatorios y SVM. Además, Sklearn se integra fácilmente con otras herramientas y estructuras de datos como Pandas, lo que simplifica su uso y comprensión. Gracias a su versatilidad y facilidad, se ha convertido en una herramienta

fundamental para científicos de datos en las etapas iniciales del modelado estadístico [35].

K-Means

K-Means es un algoritmo de aprendizaje no supervisado que clasifica datos agrupándolos en k clusters según sus características. Su objetivo principal es minimizar la suma de las distancias al cuadrado entre cada punto y el centroide del grupo al que pertenece. El proceso comienza con la selección previa del número de clusters kkk, que puede determinarse mediante técnicas como el método del codo o validación cruzada. A continuación, se eligen aleatoriamente kkk puntos del conjunto de datos como centroides iniciales. Cada punto se asigna al cluster cuyo centroide se encuentre más próximo, calculando para ello la distancia entre el punto y los centroides disponibles. Una vez asignados todos los puntos, los centroides se recalculan como la media aritmética de los puntos que conforman cada grupo, actualizando así su posición [36]. Estos pasos de asignación y recalculo se repiten hasta que los centroides dejan de cambiar, los puntos permanecen en sus clusters o se alcanza un número máximo de iteraciones. Matemáticamente, el algoritmo busca optimizar la función:

$$\min_{s} E(\mu\, i) = \min_{k} \sum_{i=1}^{k} * \sum_{x, \in S_i} \| x_j - \mu_i \|^{2,}$$

donde S representa el conjunto de datos, x_j los vectores de características de dimensión n, k el número de clusters y μ_i los centroides. La condición necesaria para actualizar los centroides se expresa como:

$$\frac{\partial E}{\partial \mu_i} = 0 \ \rightarrow {\mu_i}^{(T+1)} = \frac{1}{|{S_i}^{(t)}|} \sum_{x, \in S_i} x_j$$

Para medir la calidad del agrupamiento, se utilizan métricas como la inercia o el coeficiente de silueta. Entre las ventajas de K-Means destacan su facilidad de implementación, su adaptabilidad a grandes volúmenes de datos y distintos tipos de conjuntos, la garantía de convergencia y la capacidad para generalizar a grupos con diversas formas y tamaños. Debido a estas características, es ampliamente aplicado en áreas como segmentación de clientes, clasificación de textos y

detección de anomalías [37].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Es un algoritmo de agrupamiento no supervisado basado en densidad, introducido por Ester, Kriegel, Sander y Xu en 1996. Su funcionamiento se apoya principalmente en la distancia euclidiana y en dos parámetros clave: el número mínimo de puntos para definir una región densa y un radio ε que determina la proximidad necesaria entre puntos para considerarlos vecinos dentro de un mismo clúster. La correcta selección de estos parámetros es fundamental, ya que permite diferenciar adecuadamente entre regiones densas y datos atípicos, siendo recomendable que el mínimo de puntos sea al menos igual a la dimensionalidad de los datos, o mayor en presencia de ruido [38].

En el método, los puntos se clasifican en tres categorías: puntos núcleo, que poseen un número suficiente de vecinos dentro del radio ε; puntos bordes, que están en la vecindad de un punto núcleo, pero no cumplen el requisito mínimo de vecinos; y puntos ruido, que no pertenecen ni a núcleos ni a bordes. El algoritmo construye clústeres conectando puntos núcleo a través de bordes de densidad, es decir, cuando dos puntos núcleo están dentro de la distancia ε, formando trayectorias que permiten agruparlos. El proceso inicia seleccionando puntos no visitados y evaluando su vecindad; si un punto cumple con los requisitos, se expande el clúster con sus vecinos, repitiendo esta expansión hasta que el clúster esté completamente formado. Los puntos que no cumplen con estos criterios se consideran ruido.

Una de las fortalezas de DBSCAN es su capacidad para identificar un número variable e incluso desconocido de clústeres, incluyendo aquellos con formas irregulares, así como su habilidad para detectar anomalías y manejar datos ruidosos. Además, a diferencia de otros algoritmos, no requiere especificar previamente la cantidad de grupos a formar. Su simplicidad en los parámetros facilita la selección intuitiva de valores adecuados, lo que contribuye a su popularidad en sectores corporativos para tareas como la detección de fraudes, análisis de seguridad, problemas en el procesamiento de datos, aplicaciones médicas y mantenimiento preventivo.

Spectral Clustering

Spectral Clustering es un método de agrupamiento no supervisado que se fundamenta en la teoría de grafos. Su funcionamiento consiste en utilizar los valores y vectores propios asociados a una matriz de similitud para proyectar los datos en un espacio de menor dimensión, facilitando así su agrupación. El proceso inicia con la construcción de un grafo de similitud que puede ser totalmente conectado, basado en los k vecinos más cercanos o determinado por un umbral de distancia ε. A partir de esta representación, el algoritmo se enfoca en la conectividad de los puntos, agrupando aquellos que están directamente conectados o son vecinos inmediatos, sin considerar únicamente la proximidad euclidiana [39].

El algoritmo procede generando una matriz de distancias que se transforma en una matriz de afinidad A, a partir de la cual se calcula la matriz diagonal de grados D y la matriz laplaciana $L = D - A$. Luego, se obtienen los valores y vectores propios de L, con los que se construye una nueva matriz para realizar la agrupación en un espacio de dimensión k. Entre sus principales ventajas se destacan la capacidad para identificar grupos con formas y tamaños diversos, su eficiencia computacional en grandes volúmenes de datos y la ausencia de supuestos rígidos sobre la estructura de los datos. No obstante, es importante considerar que este método puede ser sensible al ruido y a los valores atípicos. Spectral Clustering se emplea comúnmente en áreas como segmentación de imágenes, análisis de datos educativos, resolución de entidades y agrupamiento espectral en biología molecular.

Affinity Propagation

Affinity Propagation es un algoritmo de agrupamiento no supervisado desarrollado en 2007 por Brendan Frey y Delbert Dueck, reconocido por su capacidad para identificar patrones ocultos dentro de los datos sin necesidad de definir previamente el número de clústeres. Este método se basa en un intercambio iterativo de mensajes entre los puntos de datos, que consisten en dos tipos principales: responsabilidad, que indica qué tan apropiado es un punto para ser representante de otro, y disponibilidad, que refleja la evidencia acumulada para que un punto sea elegido como representante. A través de la actualización continua de estas matrices hasta alcanzar la convergencia, el algoritmo determina cuáles puntos actúan como ejemplares y agrupa los datos en

torno a ellos [40].

El procedimiento comienza con el cálculo de una matriz de similitud, cuya métrica depende del tipo de datos y problema específico, seguido de la inicialización de las matrices de responsabilidad y disponibilidad. Estas se actualizan repetidamente hasta que los cambios sean mínimos, momento en el que se calcula la responsabilidad neta para identificar los ejemplares. Finalmente, cada punto se asigna al ejemplar más cercano, formando así los clústeres. Entre sus ventajas sobresale la no necesidad de especificar la cantidad de grupos, su habilidad para manejar distribuciones complejas y no lineales, y su aplicabilidad a grandes conjuntos de datos, aunque puede requerir ajustes para optimizar su eficiencia computacional. Affinity Propagation se utiliza frecuentemente en áreas como segmentación de imágenes, análisis de expresión genética y agrupación de documentos [41].

Aprendizaje no supervisado:

El aprendizaje no supervisado se refiere a la rama del machine learning que descubre estructura en datos sin etiquetar, es decir, sin que exista una variable objetivo-predefinida. En lugar de aprender a predecir un resultado conocido, estos algoritmos identifican patrones subyacentes y relaciones inherentes en los datos. Por ejemplo, son capaces de agrupar observaciones similares en función de sus características, revelando categorías o regularidades que no han sido especificadas de antemano. En términos generales, un modelo no supervisado puede analizar y agrupar datos sin etiquetar, detectando *patrones ocultos* en ellos sin intervención humana. Esto la convierte en una aproximación ideal para análisis exploratorios donde se busca entender la estructura de los datos más que predecir un valor puntual. Los métodos clásicos de esta técnica incluyen la clusterización (agrupamiento en clústeres) y la reducción de dimensionalidad, entre otros, proporcionando las bases teóricas para segmentar conjuntos de datos complejos de manera significativa [42].

Utilidad en datos ESG

En el contexto del análisis de indicadores ESG (ambientales, sociales y de gobernanza) corporativos, el aprendizaje no supervisado resulta particularmente útil y apropiado. Dado que típicamente no

se cuenta con una etiqueta de verdad terreno (como por ejemplo *"empresa sostenible"* vs *"empresa no sostenible"*) para cada compañía, se recurre a enfoques no supervisados para descubrir cómo las empresas se organizan o difieren según sus métricas ESG. La literatura reciente destaca que las técnicas no supervisadas pueden extraer información valiosa de estos datos complejos: por su capacidad de descubrir patrones ocultos y estructuras latentes en los datos ESG sin información previa, constituyen un complemento eficaz a los métodos supervisados tradicionales [43]. En otras palabras, algoritmos como la clusterización (p. ej., *k*-medias) o técnicas como PCA (análisis de componentes principales) permiten hallar grupos naturales de empresas con perfiles de sostenibilidad similares, así como reducir la complejidad de decenas de indicadores ESG a unos pocos factores principales. Esto facilita la interpretación y visualización de las tendencias en sostenibilidad corporativa, apoyando la toma de decisiones en inversión responsable y gestión de riesgos [37].

Clusterización de empresas ESG

La aplicación práctica de estos principios teóricos se evidencia en la agrupación de empresas según sus indicadores ESG. Mediante algoritmos de *clustering*, es posible segmentar un conjunto de empresas que reportan datos ESG en grupos relativamente homogéneos –por ejemplo, conglomerados de "líderes ESG" versus "rezagados ESG" u otras categorías emergentes– basándose exclusivamente en sus múltiples métricas de sostenibilidad. Estos clústeres aportan insights sobre diferencias estructurales entre empresas: se puede analizar si ciertos grupos exhiben patrones de desempeño financiero, niveles de riesgo o prácticas corporativas distintivas asociadas a sus puntuaciones ESG. Estudios recientes demuestran el valor de este enfoque: por ejemplo, al aplicar el algoritmo k-medias para agrupar compañías del S&P 500 según su *score* ESG, se observó que dicha segmentación no supervisada permitió construir portafolios de inversión más efectivos, superando en rendimiento a métodos tradicionales basados únicamente en ordenar empresas por su puntaje ESG. Este hallazgo sugiere que la clusterización ESG extrae relaciones y diferencias sutiles entre empresas que pasan desapercibidas con análisis convencionales, estableciendo así un vínculo claro entre la teoría del aprendizaje no supervisado y su utilidad práctica en el campo de

los datos ESG [44].

3.2 Antecedentes

Los criterios ESG (ambientales, sociales y de gobernanza) se consolidaron como herramientas indispensables para evaluar la sostenibilidad en sectores clave como el energético. El presente trabajo se apoya en la premisa de que la clusterización de empresas, basada en sus indicadores ESG, facilita la identificación de patrones comunes y diferencias significativas, permitiendo un análisis comparativo más efectivo y estratégico para la toma de decisiones corporativas y regulatorias. La literatura y estudios previos que abordan esta temática fundamentan teórica y metodológicamente la presente investigación [45].

Históricamente, el sector energético ha sido un motor esencial del desarrollo económico y social, sustentando actividades industriales, transporte y servicios básicos a nivel global. A lo largo de las últimas décadas, su matriz se sustentó principalmente en combustibles fósiles, lo que impulsó avances significativos, pero generó importantes impactos ambientales y tensiones geopolíticas. La transición actual hacia fuentes renovables como la solar y la eólica se ha configurado no solo como una respuesta urgente a los desafíos climáticos y regulatorios, sino también como una oportunidad para fomentar la innovación tecnológica y mejorar la competitividad regional [46] [47].

En el contexto latinoamericano, diversos estudios resaltaron que la integración de métricas ESG en el sector energético puede ser un catalizador para la transición sostenible. Por ejemplo, investigaciones recientes documentaron que países de la región enfrentan retos similares en cuanto a la diversificación energética y la mitigación del impacto ambiental, pero también exhiben un potencial notable para adoptar tecnologías limpias y fortalecer políticas regulatorias orientadas a la sostenibilidad [48]. Estas perspectivas apoyan la aplicación de técnicas como la clusterización para analizar el desempeño ESG regional y segmentar empresas según su nivel de compromiso y progreso.

En Colombia, el sector energético representa un componente estratégico del crecimiento económico y la generación de empleo, aportando un porcentaje considerable al PIB nacional y a las exportaciones. Sin embargo, enfrenta retos ambientales y sociales vinculados con la explotación

de recursos fósiles, además de la necesidad de cumplir con compromisos internacionales de reducción de emisiones y promover una transición justa hacia energías renovables [49]. Estudios sectoriales destacan la importancia de implementar sistemas de medición y gestión basados en ESG para monitorear el desempeño corporativo y ambiental, orientando políticas públicas y estrategias empresariales [50].

Adicionalmente, investigaciones enfocadas en el impacto financiero de los criterios ESG en Colombia evidenciaron que las empresas con mejor desempeño en estas métricas tienden a mostrar mayor resiliencia y capacidad para atraer inversiones sostenibles, especialmente frente a la volatilidad del mercado energético y los riesgos asociados a crisis geopolíticas y climáticas [51]. Este hallazgo refuerza la pertinencia de agrupar empresas mediante técnicas de clusterización para identificar perfiles de riesgo y oportunidades dentro del sector.

Otro estudio reciente aplicado a países europeos utilizó la clusterización para categorizar avances en sostenibilidad energética, considerando variables como eficiencia, capacidad instalada y uso de renovables, lo que permitió proponer estrategias específicas adaptadas a cada grupo. Aunque centrado en un análisis macro, este enfoque metodológico valida la utilidad de la clusterización para entender patrones complejos en la transición energética, aportando un respaldo empírico relevante para su aplicación a nivel empresarial en Colombia y América Latina [52].

Finalmente, el creciente interés en el uso de análisis multivariado y aprendizaje no supervisado en el ámbito ESG refleja una tendencia global hacia la mejora en la gestión sostenible. La clusterización, en particular, ha demostrado ser eficaz para identificar subgrupos de empresas con características y desafíos similares, facilitando el diseño de políticas y estrategias diferenciadas. Este cuerpo de antecedentes consolida el marco conceptual y metodológico que sustenta el presente estudio, orientado a optimizar la segmentación y comprensión del sector energético mediante indicadores ESG.

4 IDENTIFICACIÓN DE LAS VARIABLES MÁS RELEVANTES PARA LA CLUSTERIZACIÓN DE EMPRESAS SEGÚN INDICADORES ESG

En esta sección se describen las etapas desarrolladas para cumplir con el primer objetivo específico, el cual consiste en identificar y seleccionar las variables más relevantes e influyentes en la clusterización de empresas que divulgan indicadores ESG (Ambiental, Social y de Gobernanza). Estas variables se considerarán como candidatas para construir el modelo de aprendizaje no supervisado que permita segmentar las empresas según sus prácticas de sostenibilidad.

4.1 Requisitos de datos

En el marco del análisis de agrupamiento (clusterización) aplicado a empresas del sector energético que divulgan indicadores ESG, la selección de variables relevantes exige el cumplimiento de ciertos requisitos metodológicos que aseguren la calidad, pertinencia y utilidad estadística de los datos empleados. A continuación, se describen los principales criterios considerados:

Las variables seleccionadas deben estar directamente asociadas a los tres pilares fundamentales del enfoque ESG: ambiental (E), social (S) y gobernanza (G). Esto implica incluir indicadores que reflejen de forma medible prácticas empresariales relacionadas con sostenibilidad, derechos laborales, emisiones, uso de recursos, transparencia institucional, entre otros. Variables sin conexión clara con estos pilares se excluyeron.

Las variables deben presentar una dispersión adecuada entre los registros (empresas), ya que una baja varianza sugiere que la mayoría de los valores son similares y no aportan valor en la identificación de patrones. Por el contrario, una varianza excesivamente alta puede distorsionar el análisis. Se aplicó el análisis exploratorio inicial para filtrar variables con comportamiento invariante o extremo.

La completitud de los datos es esencial para evitar sesgos en el modelo. Se excluyeron variables con más del 50 % de datos nulos y se imputaron aquellas con valores ausentes menores mediante técnicas como la media o la mediana, siempre que no alteraran significativamente la distribución original.

Dado que las variables ESG utilizan diferentes unidades de medida y escalas (por ejemplo, puntajes

de 0 a 100, porcentajes, índices), fue necesario estandarizarlas mediante técnicas como Z-score (media 0, desviación estándar 1) para evitar que las variables con mayor rango numérico dominaran la agrupación.

Se verificó la colinealidad entre variables mediante la matriz de correlación. Aquellas altamente correlacionadas ($r > 0,9$) fueron objeto de revisión para evitar redundancia en la información. En estos casos, se priorizó la variable con mayor capacidad explicativa en los pilares ESG o se consideró la aplicación de técnicas como PCA en etapas posteriores.

Durante las pruebas preliminares de modelos de clusterización (K-means y jerárquico), se evaluó la contribución de cada variable a la formación de clústeres diferenciados. Las variables que no aportaban diferenciación significativa (por ejemplo, aquellas con distribución homogénea o baja contribución a la varianza explicada) se descartaron o consideradas para reevaluación.

4.2 Recopilación de datos

La recopilación de datos ESG constituyó un proceso esencial para caracterizar el comportamiento sostenible de las empresas del sector energético. Para ello, se recurrió a la plataforma London Stock Exchange Group (LSEG), anteriormente conocida como Refinitiv, reconocida internacionalmente por suministrar información financiera, económica y de sostenibilidad con altos estándares de calidad. Esta plataforma, utilizada por más de 40.000 instituciones en 190 países, proporcionó un conjunto estructurado de variables ESG que incluye indicadores ambientales, sociales y de gobernanza, así como métricas asociadas a controversias y desempeño corporativo. Gracias a su infraestructura tecnológica, que procesa más de 200 mil millones de actualizaciones diarias mediante herramientas como Eikon, Workspace y DataScope, fue posible acceder a datos confiables y comparables, lo cual facilitó la construcción de una base adecuada para aplicar modelos de clusterización no supervisada.

4.3 Comprensión de los datos

Para comprender la base de datos utilizada en este estudio, se procedió inicialmente a revisar el significado específico de cada variable ESG a partir de la documentación provista por la plataforma

LSEG, lo cual permitió establecer correspondencias conceptuales con los pilares ambiental, social y de gobernanza descritos en el marco teórico. En esta etapa se utilizaron herramientas del entorno Python en Google Colab, incluyendo librerías como pandas, numpy y matplotlib, para efectuar una lectura detallada del conjunto de datos, conformado por 576 registros de empresas del sector energético y 17 indicadores cuantitativos de sostenibilidad.

Con el propósito de asegurar la precisión analítica, se efectuó la transformación de tipos de dato a formato numérico estandarizado, lo que incluyó la limpieza de cadenas mixtas, la imputación de valores faltantes y la aplicación de técnicas de normalización mediante StandardScaler. Posteriormente, se identificaron patrones generales a través de análisis estadístico descriptivo y visualizaciones como diagramas de caja y matrices de correlación, facilitando así la evaluación de distribuciones, valores atípicos y relaciones internas entre variables. Estos procedimientos permitieron no solo validar la calidad estructural de la información sino también garantizar la viabilidad de su uso en procesos posteriores de clusterización, consolidando así una base sólida para el análisis exploratorio y segmentación ESG.

Para comenzar, se describe la composición de la base de datos, identificando el número total de empresas incluidas, la cantidad de variables disponibles y la naturaleza de los datos. La Figura 1, ilustra de manera concisa la estructura inicial del conjunto de datos empleado en el estudio.

Figura 1 *Caracterización básica de la muestra*

```
Número de registros (empresas): 576
Número de variables (indicadores): 19
```

Fuente. Elaboración propia a partir de la base de datos.

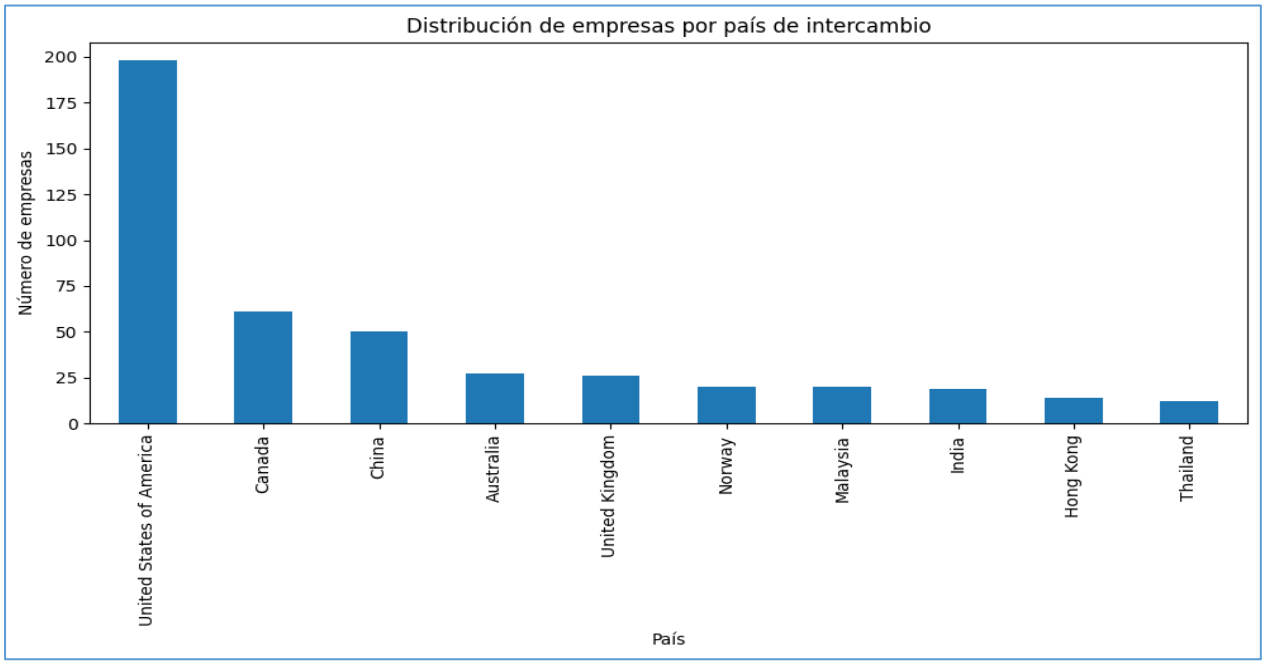
La variable “Country of Exchange” permitió identificar el país de intercambio bursátil correspondiente a cada una de las 576 empresas presentes en la base de datos. Esta información resulta esencial para entender el contexto geográfico en el que operan las organizaciones evaluadas y para analizar cómo las prácticas en sostenibilidad varían entre jurisdicciones.

De acuerdo con los resultados obtenidos, se identificaron empresas procedentes de 45 países, con

una notable concentración en economías desarrolladas. Tal como se observa en la Figura 2, los países con mayor representación fueron Estados Unidos (199 empresas), Canadá (63), China (50), Australia (28) y Reino Unido (26), lo cual revela una fuerte presencia de compañías que cotizan en mercados bursátiles consolidados. Otros países con representación significativa incluyen Noruega, Malasia, India, Hong Kong y Tailandia.

Este patrón de distribución geográfica es coherente con el liderazgo que ejercen estas naciones en la adopción de estándares de reporte ESG, así como con la disponibilidad y accesibilidad de datos financieros y no financieros sobre sostenibilidad. Asimismo, la presencia de empresas en diversas regiones del mundo aporta una perspectiva comparativa enriquecida para el análisis posterior de desempeño ESG y segmentación. En la Figura 2 se presenta la distribución de empresas por país de intercambio, evidenciando el predominio de Estados Unidos y Canadá en la muestra.

Figura 2 *Principales países de intercambio de las empresas analizadas.*



Fuente: Elaboración propia con base en LSEG.

4.4 Preparación de los datos

En esta etapa se llevaron a cabo diversas actividades orientadas a dejar la base de datos en

condiciones óptimas para su análisis mediante técnicas de aprendizaje no supervisado. El primer paso consistió en realizar una limpieza inicial, donde se integraron las variables provenientes de la plataforma LSEG, asegurando la consistencia terminológica al traducirlas al español, eliminar textos innecesarios y adaptar los tipos de datos a formatos numéricos. Posteriormente, se procedió a transformar variables categóricas a numéricas, agrupar información para mejorar la estructura analítica, y gestionar los valores faltantes, todo ello con el fin de construir una matriz de datos robusta, compatible con los algoritmos de clusterización que serían implementados en las fases posteriores del estudio. Las transformaciones realizadas son descritas a continuación.

4.4.1 Variables categóricas a variables numéricas

Durante el proceso de preparación de los datos, se identificó que varias de las variables extraídas desde la plataforma LSEG se encontraban clasificadas inicialmente como cualitativas, a pesar de representar métricas numéricas en su contenido. Este fue el caso de indicadores clave como ESG Score, Environmental Pillar Score, Social Pillar Score, Governance Pillar Score, entre otros, los cuales incluían puntuaciones acompañadas de anotaciones textuales (por ejemplo, valores numéricos seguidos de clasificaciones alfabéticas como “94.5 (A+)”).

Para permitir su tratamiento adecuado en modelos de aprendizaje automático no supervisado, fue necesario convertir estas variables en datos puramente numéricos. Este procedimiento se realizó mediante el uso de expresiones regulares en lenguaje Python, extrayendo los valores decimales y descartando las etiquetas alfabéticas. Además, se ajustó el tipo de dato de cada columna afectada, asegurando que fueran reconocidas como variables de tipo float por los entornos computacionales utilizados. De esta forma, se optimizó la compatibilidad de las variables con las herramientas analíticas empleadas y se garantizó la precisión en los cálculos posteriores.

Variables como “Country of Exchange”, “Company Name” e “Identifier” se mantuvieron excluidas del análisis cuantitativo por su naturaleza categórica no convertible directamente en datos numéricos relevantes para la clusterización. Este paso fue indispensable para establecer un conjunto de datos limpio, estructurado y funcionalmente apropiado para la estandarización y segmentación empresarial bajo criterios ESG. En la Tabla 2 se presenta la clasificación original de

las variables e identificación de su tipo de dato.

Tabla 2 *Clasificación original de las variables e identificación de su tipo de dato.*

NOMBRE DE LAS VARIABLE	TIPO DE DATOS
Identifier (RIC)	Cualitativa
Company Name ESG	Cualitativa
ESG Score (FYO)(Σ=Avg)	Cualitativa
ESG Combined Score (FYO)(Σ=Avg)	Cualitativa
ESG Controversies Score (FYO)(Σ=Avg)	Cualitativa
Environmental Pillar Score (FYO)(Σ=Avg)	Cualitativa
Social Pillar Score (FYO)(Σ=Avg)	Cualitativa
Governance Pillar Score (FYO)(Σ=Avg)	Cualitativa
Resource Use Score (FYO)(Σ=Avg)	Cualitativa
Emissions Score (FYO)(Σ=Avg)	Cualitativa
Innovation Score (FYO)(Σ=Avg)	Cualitativa
Workforce Score (FYO)(Σ=Avg)	Cualitativa
Human Rights Score (FYO)(Σ=Avg)	Cualitativa
Community Score (FYO)(Σ=Avg)	Cualitativa
Product Responsibility Score (FYO)(Σ=Avg)	Cualitativa
Management Score (FYO)(Σ=Avg)	Cualitativa
Shareholders Score (FYO)(Σ=Avg)	Cualitativa
CSR Strategy Score (FYO)(Σ=Avg)	Cualitativa
Country of Exchange	Cualitativa

Fuente: Elaboración propia con base en la base de datos ESG obtenida de la plataforma LSEG.

4.4.2 Agrupación y creación de nuevas variables

a. Eliminación de columnas no numéricas y renombramiento de variables:

Se eliminarán identificadores y columnas que no aportan valor al análisis cuantitativo, como el nombre de la empresa o el identificador RIC. También se renombrarán las columnas para que sean más manejables.

b. Extracción de valores numéricos y conversión de tipo:

Se extraerán los valores numéricos de las columnas de indicadores ESG, que actualmente presentan valores alfanuméricos como "50.08 (B-)", utilizando expresiones regulares para obtener solo el componente numérico. En la Figura 3 se presenta el código utilizado para la limpieza, transformación y estandarización de las variables ESG.

Figura 3 *Código para la Limpieza, transformación y estandarización de variables ESG*

```python
import re

# Eliminar columnas no numéricas innecesarias para modelado
columns_to_drop = ['Identifier (RIC)', 'Company Name', 'Country of Exchange']
df_clean = df.drop(columns=columns_to_drop)

# Renombrar columnas a nombres más manejables
column_renames = {
    'ESG Score\n(FY0)\n(Σ=Avg)': 'esg_total',
    'ESG Combined Score\n(FY0)\n(Σ=Avg)': 'esg_combinado',
    'ESG Controversies Score\n(FY0)\n(Σ=Avg)': 'esg_controversias',
    'Environmental Pillar Score\n(FY0)\n(Σ=Avg)': 'ambiental',
    'Social Pillar Score\n(FY0)\n(Σ=Avg)': 'social',
    'Governance Pillar Score\n(FY0)\n(Σ=Avg)': 'gobernanza',
    'Resource Use Score\n(FY0)\n(Σ=Avg)': 'uso_recursos',
    'Emissions Score\n(FY0)\n(Σ=Avg)': 'emisiones',
    'Innovation Score\n(FY0)\n(Σ=Avg)': 'innovacion',
    'Workforce Score\n(FY0)\n(Σ=Avg)': 'fuerza_laboral',
    'Human Rights Score\n(FY0)\n(Σ=Avg)': 'derechos_humanos',
    'Community Score\n(FY0)\n(Σ=Avg)': 'comunidad',
    'Product Responsibility Score\n(FY0)\n(Σ=Avg)': 'responsabilidad_producto',
    'Management Score\n(FY0)\n(Σ=Avg)': 'gestion',
    'Shareholders Score\n(FY0)\n(Σ=Avg)': 'accionistas',
    'CSR Strategy Score\n(FY0)\n(Σ=Avg)': 'estrategia_rse'
}
df_clean.rename(columns=column_renames, inplace=True)

# Extraer los valores numéricos de cada columna de indicadores
for col in df_clean.columns:
    df_clean[col] = df_clean[col].astype(str).str.extract(r'([\d\.,]+)').replace(',', '', regex=True).astype(float)

# Verificación de la conversión y la estructura de la base
print(df_clean.head())
print("\nResumen de valores nulos:")
print(df_clean.isna().sum())
```

Fuente. Elaboración propia.

La Tabla 3 resume la reasignación conceptual y técnica de las variables más relevantes utilizadas en el análisis:

Tabla 3 *Agrupación funcional y tipo de dato final de las variables ESG utilizadas.*

Variable Reasignada	Pilar Asociado	Tipo Final	Observación técnica
esg_total	Global ESG	Cuantitativa	Indicador agregado de sostenibilidad
ambiental	Ambiental	Cuantitativa	Score medioambiental
uso_recursos	Ambiental	Cuantitativa	Subindicador de ecoeficiencia
emisiones	Ambiental	Cuantitativa	Nivel de emisiones
innovación	Ambiental	Cuantitativa	Potencial tecnológico
social	Social	Cuantitativa	Score social general
fuerza_laboral	Social	Cuantitativa	Bienestar del personal
derechos_humanos	Social	Cuantitativa	Cumplimiento normativo
comunidad	Social	Cuantitativa	Responsabilidad con el entorno
responsabilidad_producto	Social/Gobernanza	Cuantitativa	Calidad y seguridad del producto
gobernanza	Gobernanza	Cuantitativa	Score de gestión
gestión	Gobernanza	Cuantitativa	Profesionalismo y control interno
accionistas	Gobernanza	Cuantitativa	Equidad en toma de decisiones
estrategia_rse	Gobernanza	Cuantitativa	Integración estratégica de sostenibilidad
esg_combinado	Global ESG	Cuantitativa	Score más penalización por controversias
esg_controversias	Reputacional	Cuantitativa	Indicador de riesgo reputacional

Fuente: Elaboración propia a partir de la base de datos ESG (LSEG).

4.4.3 Valores faltantes

Como parte del proceso de limpieza de datos, se efectuó una inspección sistemática para detectar valores faltantes en las variables numéricas seleccionadas. Esta etapa fue crucial para garantizar la integridad de la base de datos utilizada en el modelo de clusterización. Inicialmente, se aplicó una revisión mediante funciones estadísticas para determinar la cantidad de valores ausentes por variable, lo que permitió establecer un diagnóstico claro sobre la calidad de la información disponible.

Una vez identificadas las variables con datos incompletos, se optó por una estrategia de imputación basada en el promedio aritmético. Este método permitió preservar la distribución original de los datos y reducir el sesgo introducido por omisiones en el reporte. Específicamente, se aplicó el algoritmo SimpleImputer de la biblioteca sklearn.impute, con el parámetro strategy='mean', que remplazó los valores faltantes por la media de cada columna.

Finalmente, se validó el tratamiento mediante un nuevo conteo de valores nulos, confirmando que todas las observaciones estaban completas y aptas para el análisis posterior. Esta depuración fue indispensable para evitar distorsiones en la segmentación de empresas según sus métricas ESG, asegurando que los algoritmos no supervisados operaran sobre una base homogénea y coherente. En la Figura 4 se presenta la identificación y tratamiento de valores faltantes en las variables ESG transformadas.

Figura 4 *Identificación y tratamiento de valores faltantes en variables ESG transformadas.*

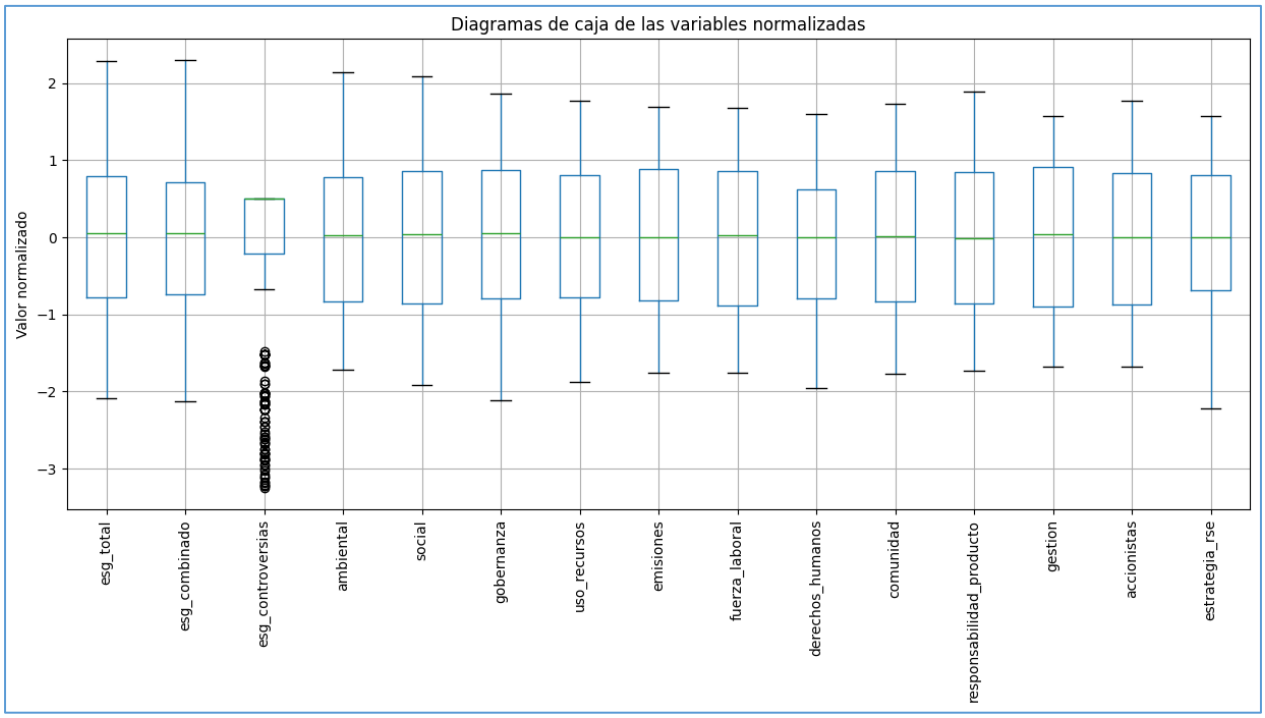
```
¿Existen valores nulos después de la imputación?

esg_total                  0
esg_combinado              0
esg_controversias          0
ambiental                  0
social                     0
gobernanza                 0
uso_recursos               0
emisiones                  0
fuerza_laboral             0
derechos_humanos           0
comunidad                  0
responsabilidad_producto   0
gestion                    0
accionistas                0
estrategia_rse             0
dtype: int64
   esg_total  esg_combinado  esg_controversias  ambiental  social  gobernanza  \
0      94.49          83.61              72.73      96.00   95.11       91.85
1      92.28          46.43               0.59      92.63   92.10       92.07
2      90.64          81.68              72.73      88.65   92.16       90.81
3      89.35          51.74              14.12      89.27   94.44       80.40
4      88.47          88.47             100.00      82.87   97.59       87.53

   uso_recursos  emisiones  fuerza_laboral  derechos_humanos  comunidad  \
0         99.69      96.93           98.07             90.97      97.51
1         99.40      98.71           84.66             91.63      98.74
2         94.06      85.75           90.33             90.97      94.20
3         93.43      95.40           94.40             91.63      98.01
4         96.58      96.67           98.24             91.18      98.24

   responsabilidad_producto  gestion  accionistas  estrategia_rse
0                     96.57    98.85        66.54           94.80
1                     99.42    97.92        70.60           95.02
2                     96.57    96.26        67.02           99.22
3                     97.09    93.04        48.73           64.74
4                     98.03    99.57        48.71           85.59
```

Fuente: Elaboración propia con base en LSEG.

Homogeneización de tipos de datos

Una vez transformados los valores al formato numérico, se verificaron los tipos de datos de cada columna, asegurando que todos los indicadores relevantes se encontraran codificados como float64, requisito indispensable para su uso en algoritmos de machine learning. Se eliminaron columnas redundantes o no útiles para el análisis, como los identificadores de empresa o nombres comerciales, con el propósito de centrarse exclusivamente en los indicadores ESG cuantificables. En la Figura 5 se presenta el diagrama de caja de las variables ESG normalizadas utilizadas en el análisis, lo que permite visualizar su distribución y homogeneidad tras el proceso de transformación.

Figura 5 *Diagrama de caja de las variables normalizadas*



Fuente: Elaboración propia con base en LSEG.

4.4.4 Análisis exploratorio

Se observa que la muestra está compuesta principalmente por empresas que cotizan en las principales bolsas del mundo, con predominio de jurisdicciones como Estados Unidos, Reino Unido y Japón. Esta diversidad permite realizar análisis comparativos internacionales sobre el desempeño

ESG. En la Tabla 4 se presentan los estadísticos descriptivos de los puntajes ESG normalizados, incluyendo media, desviación estándar, valores mínimos, máximos y cuartiles para cada indicador analizado.

Tabla 4 . *Estadísticos descriptivos de los puntajes ESG.*

index	count	mean	std	min	25%	50%	75%	max
esg_total	576	1.5E-16	1.00	-2.09	-0.78	0.05	0.79	2.28
esg_combinado	576	-9.9E-17	1.00	-2.13	-0.74	0.05	0.72	2.29
esg_controversias	576	-2.5E-17	1.00	-3.25	-0.21	0.50	0.50	0.50
ambiental	576	-9.9E-17	1.00	-1.72	-0.83	0.03	0.77	2.14
social	576	-2.0E-16	1.00	-1.92	-0.86	0.05	0.86	2.09
gobernanza	576	2.5E-16	1.00	-2.11	-0.79	0.05	0.87	1.86
uso_recursos	576	-2.0E-16	1.00	-1.87	-0.78	0.00	0.81	1.78
emisiones	576	1.5E-16	1.00	-1.76	-0.82	0.00	0.89	1.69
fuerza_laboral	576	-4.9E-17	1.00	-1.76	-0.89	0.03	0.86	1.67
derechos_humanos	576	-4.9E-17	1.00	-1.96	-0.79	0.00	0.62	1.60
comunidad	576	1.5E-16	1.00	-1.77	-0.83	0.01	0.86	1.72
responsabilidad_producto	576	2.2E-16	1.00	-1.73	-0.85	-0.02	0.85	1.89
gestion	576	-9.9E-17	1.00	-1.67	-0.90	0.04	0.91	1.57
accionistas	576	-1.6E-16	1.00	-1.68	-0.87	0.00	0.84	1.77
estrategia_rse	576	-4.9E-17	1.00	-2.22	-0.69	0.00	0.81	1.57

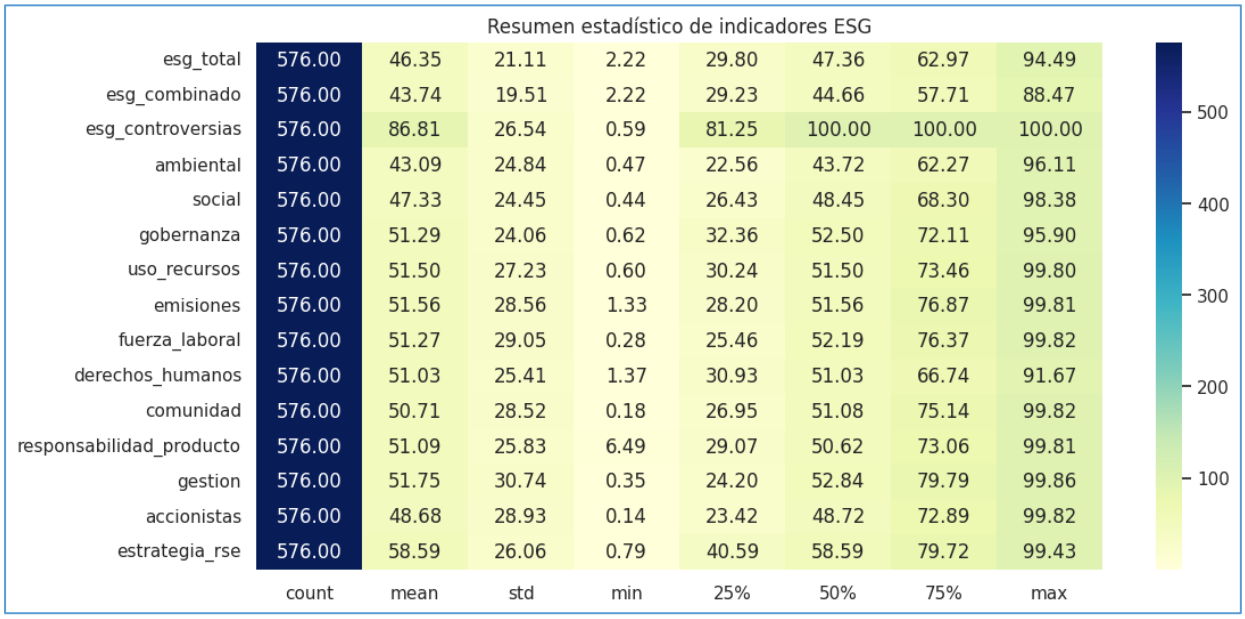
Fuente: Elaboración propia con datos de ESG (2024).

El análisis exploratorio de datos (AED) constituye una etapa clave dentro del proceso de minería de datos, ya que permite identificar comportamientos generales, detectar valores atípicos y examinar las distribuciones estadísticas de las variables. En el presente estudio, se aplicaron herramientas de estadística descriptiva y visualización para comprender el comportamiento de los indicadores ESG reportados por las 576 empresas analizadas.

Inicialmente, se generaron medidas de tendencia central y dispersión para cada una de las variables cuantitativas, permitiendo caracterizar la magnitud y variabilidad de los puntajes asignados en los pilares Ambiental, Social y de Gobernanza. Esta información fue representada mediante un *heatmap* que sintetiza las estadísticas básicas (media, desviación estándar, mínimo, máximo y percentiles) y evidencia diferencias relevantes entre indicadores.

La Figura 6 presenta un mapa de calor con las principales medidas estadísticas (media, desviación estándar, valores mínimo y máximo, así como los percentiles 25, 50 y 75) para cada una de las variables ESG incluidas en el análisis. Se observa que la media de las puntuaciones se encuentra en un rango relativamente homogéneo, lo cual sugiere una consistencia general en los reportes de sostenibilidad entre empresas.

Figura 6 *Heatmap de estadísticas descriptivas.*



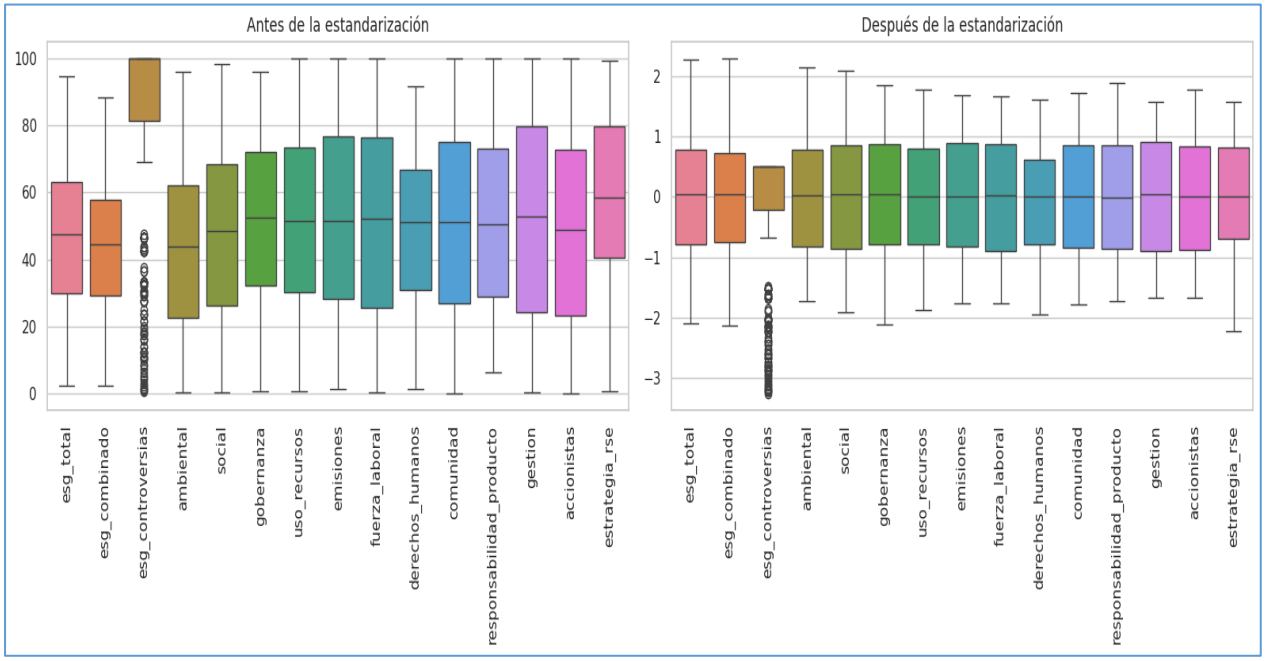
Fuente: Elaboración propia con datos de ESG (2024).

La Figura 7, permite examinar la distribución original de los datos por variable ESG antes de someterse a procesos de transformación y normalización. Se evidencia la presencia de valores atípicos en múltiples indicadores, especialmente en *uso de recursos*, *emisiones* y *esg_controversias*, lo cual justificó la necesidad de aplicar un tratamiento previo a la clusterización. La variabilidad intercuartílica también resulta elevada en variables como *gobernanza* y *estrategia RSE*, lo que refleja heterogeneidad en los enfoques corporativos frente a la responsabilidad institucional.

Una vez implementado el proceso de estandarización mediante escalado z-score, los datos muestran distribuciones centradas alrededor de la media cero, con desviaciones estándar unificadas. Esto permitió asegurar que todas las variables tengan el mismo peso dentro del modelo de clusterización, evitando que los indicadores con mayor rango de valores dominen la formación

de grupos. La reducción de la dispersión relativa y la simetría alcanzada en la mayoría de las variables confirma la eficacia del preprocesamiento.

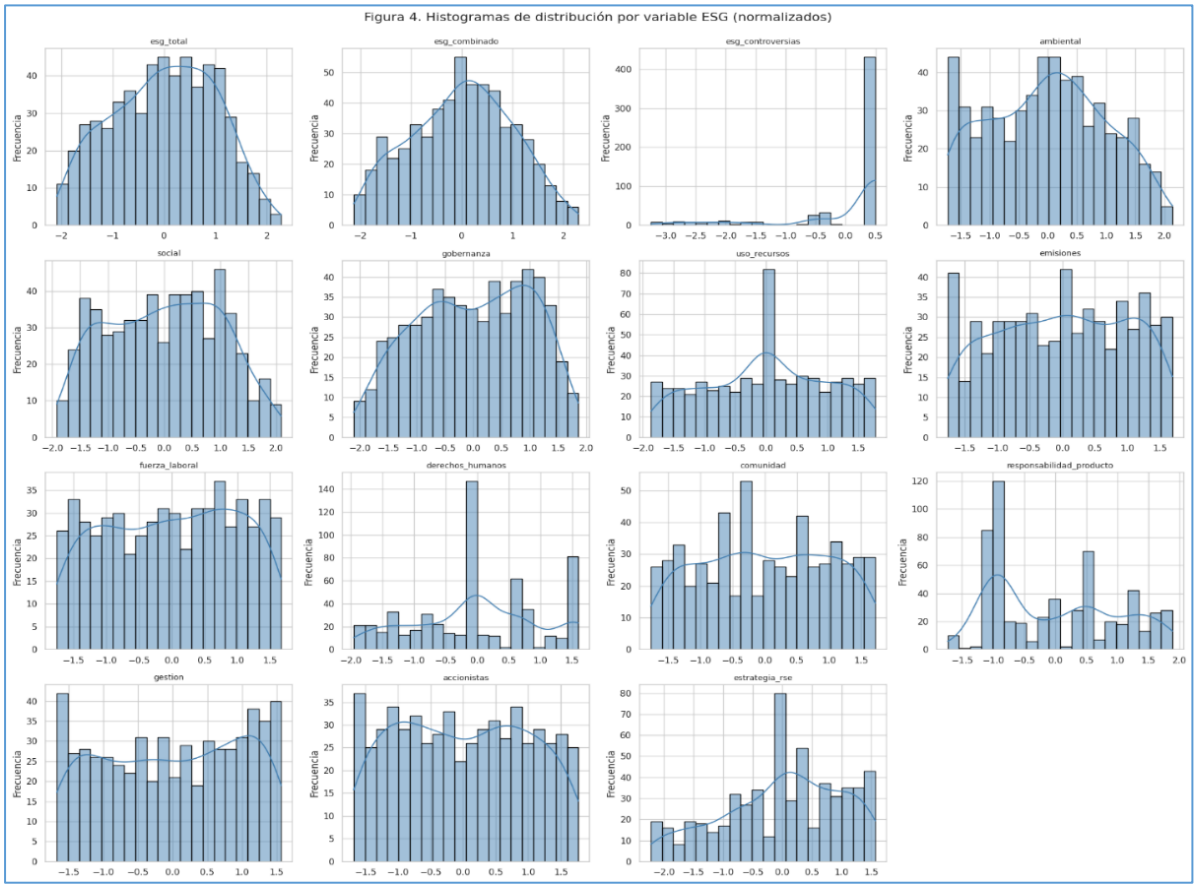
Figura 7 *Diagramas de caja antes y después de la estandarización.*



Fuente: Elaboración propia con datos de ESG (2024).

Con el objetivo de examinar la distribución de los indicadores ESG después del tratamiento de datos, se generaron histogramas unificados por variable. Esta visualización permite identificar la forma de las distribuciones, facilitando la detección de patrones comunes, sesgos, asimetrías o posibles valores atípicos. Algunas variables como esg_total, social o fuerza_laboral presentan distribuciones aproximadamente normales, mientras que otras, como ambiental o innovación, exhiben cierta dispersión o curtosis. Esta variabilidad sugiere una heterogeneidad en las prácticas de sostenibilidad del sector energético, lo cual resulta útil para la segmentación posterior mediante algoritmos de clusterización no supervisada.

Figura 8 *Histogramas de distribución por variable ESG.*



Fuente: Elaboración propia con datos de ESG (2024).

Al observar los histogramas presentados en la Figura 8, se identificaron diferentes formas de distribución entre las variables analizadas. Las variables *esg_total*, *social* y *fuerza_laboral* mostraron distribuciones aproximadamente normales, con simetría cercana a cero. En contraste, indicadores como *esg_controversias* y *estrategia_rse* presentaron una marcada asimetría positiva, es decir, con colas extendidas hacia la derecha. Por otro lado, variables como *gobernanza* y *gestión* exhibieron una leve asimetría negativa, con concentraciones mayores en los valores altos y colas hacia la izquierda. Asimismo, se identificaron distribuciones bimodales en *responsabilidad_producto* y *derechos_humanos*, lo que podría indicar la coexistencia de dos grupos empresariales con enfoques distintos frente a estas prácticas. Estas formas de distribución refuerzan la hipótesis de una alta heterogeneidad en el comportamiento ESG del sector energético. Además, se evaluaron las características estadísticas de forma mediante el cálculo de asimetría

(skewness) y curtosis sobre los datos normalizados. Los resultados confirmaron que algunas variables presentan comportamientos no normales. Por ejemplo, *esg_controversias* mostró una asimetría altamente negativa (-1.95) y una curtosis elevada (2.49), indicando una distribución con fuerte sesgo a la izquierda y colas más pesadas, posiblemente por la concentración de puntuaciones máximas en controversias reputacionales.

En contraste, variables como *social* y *esg_combinado* exhibieron asimetrías cercanas a cero, lo cual es consistente con distribuciones simétricas. También se identificó ligera asimetría positiva en *responsabilidad_producto* (0.30), mientras que *estrategia_rse* mostró una asimetría negativa moderada (-0.38). Estos hallazgos respaldan lo observado en los histogramas y permiten comprender mejor la dispersión y forma de cada variable, lo cual resulta clave para la interpretación de los clústeres posteriores. En la Tabla 5 se presentan las estadísticas de asimetría y curtosis para cada variable ESG, facilitando la identificación de patrones en la forma de las distribuciones analizadas.

Tabla 5 *Estadísticas de forma de las variables ESG.*

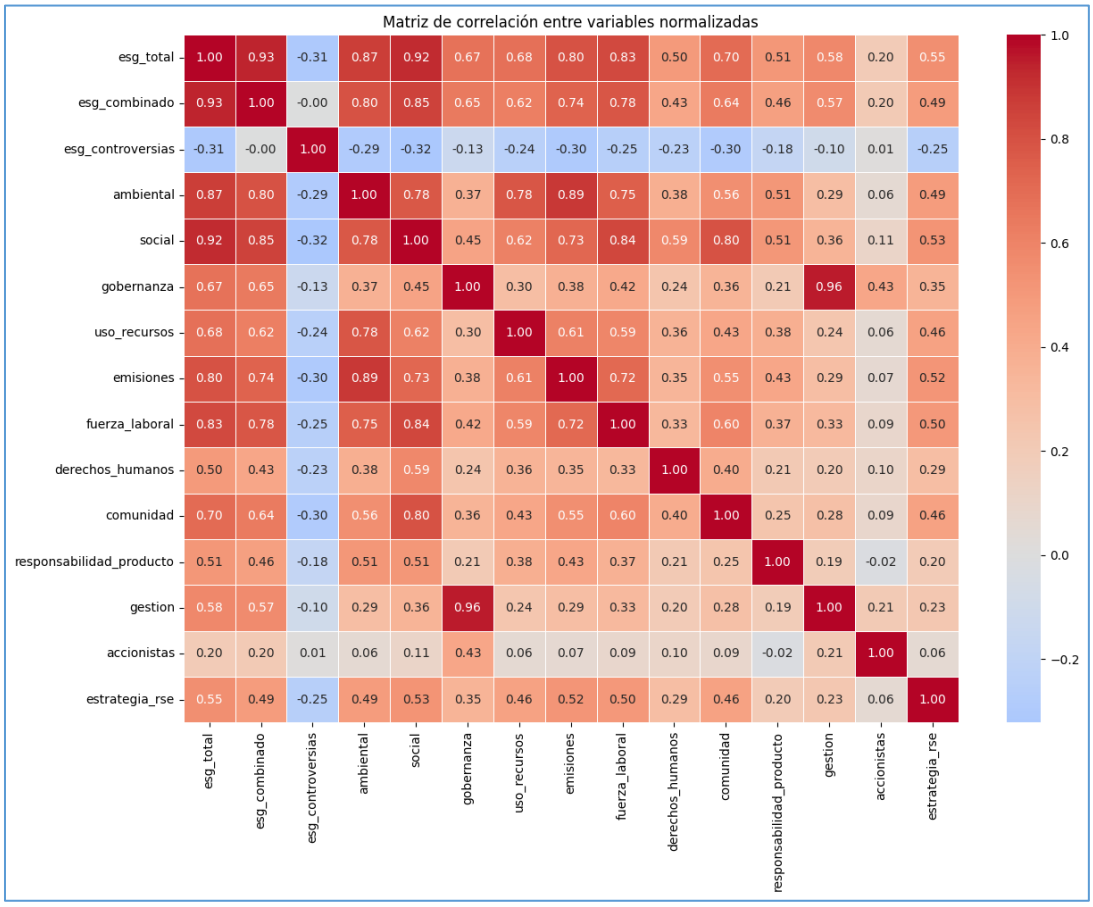
index	Asimetría (Skewness)	Curtosis
responsabilidad_producto	0.30775728148039494	-1.2470572060975418
accionistas	0.037420222632282146	-1.2032371802451294
ambiental	0.024572401124562274	-0.945573773663535
social	-0.0006121121342297228	-1.0390744298970735
comunidad	-0.02727811294575181	-1.1731824280013103
esg_combinado	-0.04888683078077146	-0.7241542732138218
emisiones	-0.06357246238966605	-1.1662247090332503
uso_recursos	-0.07008696399922036	-0.9790045200325794
derechos_humanos	-0.07414671835016193	-0.8171097696941323
fuerza_laboral	-0.07654901486408647	-1.2073476055069399
gestion	-0.0888096711772034	-1.301826828000929
esg_total	-0.10217735156409131	-0.8672412568523686
gobernanza	-0.14747801855492482	-1.068041040417677
estrategia_rse	-0.3891462487827633	-0.7209360669610714

index	Asimetría (Skewness)	Curtosis
esg_controversias	-1.9572773414366194	2.4915491938027654

Fuente: Elaboración propia con datos de ESG (2024).

Por último, la figura 9 muestra la matriz de correlación calculada sobre los datos normalizados. Se destacan correlaciones fuertes entre *ambiental* y *uso de recursos* (r > 0.75), así como entre *gobernanza*, *gestión* y *accionistas* (r > 0.70), lo cual sugiere que dichas variables miden dimensiones interrelacionadas del compromiso sostenible. Por otro lado, variables como *esg_controversias* muestran correlaciones más bajas con el resto de los indicadores, lo que indica que capturan aspectos de riesgo reputacional distintos a los tradicionales. Estas relaciones serán determinantes en la interpretación de los clústeres obtenidos.

Figura 9 *Matriz de correlación entre indicadores ESG.*



Fuente: Elaboración propia con datos de ESG (2024).

5 ENTRENAMIENTO DE MODELOS DE CLUSTERIZACIÓN CON DIFERENTES TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

Durante la etapa de modelado, se implementaron diversas técnicas de agrupamiento no supervisado con el fin de identificar patrones de comportamiento homogéneo entre empresas del sector energético en función de sus indicadores ESG. Esta fase se estructuró en dos etapas complementarias: una inicial de exploración para determinar la cantidad óptima de grupos y una segunda orientada a la construcción y evaluación comparativa de los modelos seleccionados.

Como punto de partida, se aplicó el método del codo sobre los datos previamente estandarizados mediante la técnica Z-score, utilizando la métrica de inercia como criterio de agrupación para el algoritmo K-Means. La gráfica obtenida permitió observar un punto de inflexión claro en K=3, lo que sugiere que tres es el número adecuado de clústeres para capturar una estructura representativa en los datos sin incurrir en sobresegmentación. A su vez, se exploraron configuraciones con dos y cuatro clústeres, a fin de contrastar la cohesión y separación de grupos bajo diferentes niveles de granularidad.

Posteriormente, se implementaron manualmente dos enfoques principales*: K-Means Clustering y Agrupamiento Jerárquico Aglomerativo*. El primero fue seleccionado por su eficiencia computacional y por permitir una interpretación clara de los centroides de grupo. El segundo, de tipo jerárquico, se aplicó mediante el método de enlace de Ward y visualizado con un dendrograma que facilitó la identificación visual de estructuras naturales dentro de los datos.

La tabla 6 se presenta una síntesis de las técnicas aplicadas, junto con las bibliotecas utilizadas:

Tabla 6 *Técnicas de aprendizaje no supervisado implementadas.*

Modelo	Librería utilizada
K-Means	sklearn.cluster.KMeans
Agrupamiento Jerárquico (Ward)	scipy.cluster.hierarchy / sklearn
DBSCAN	sklearn.cluster.DBSCAN
OPTICS	sklearn.cluster.OPTICS
Birch	sklearn.cluster.Birch

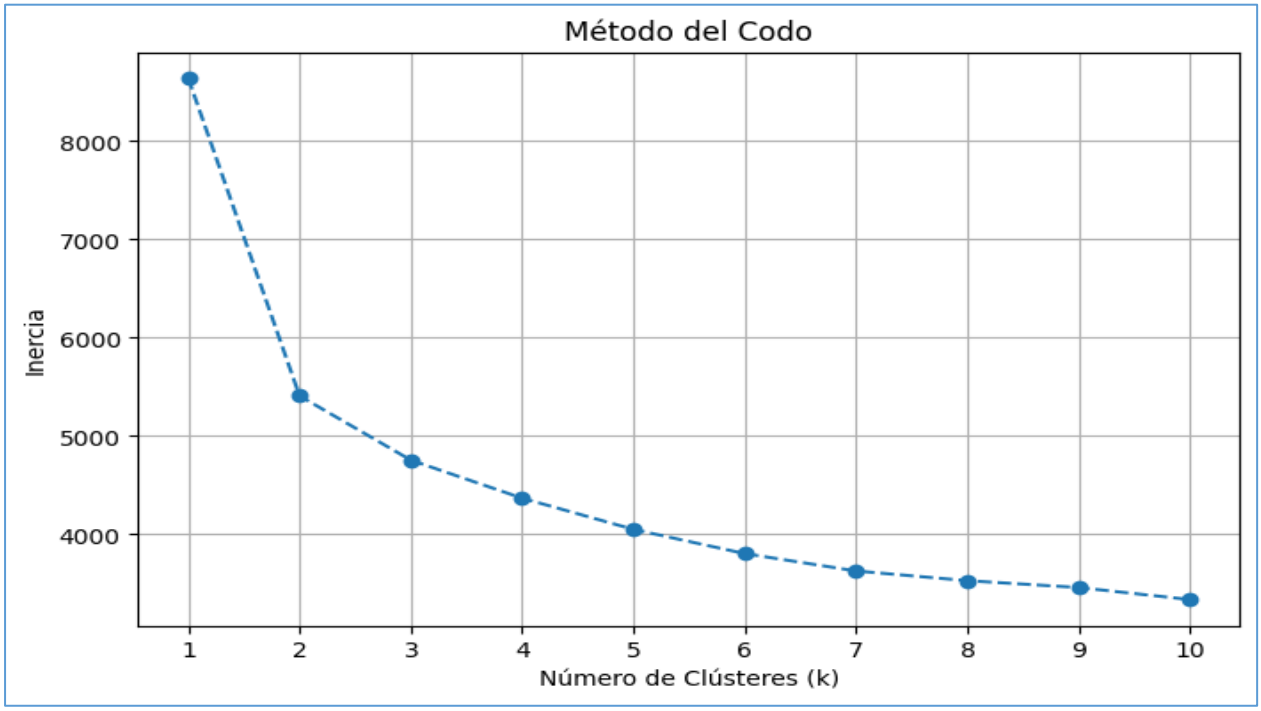
Fuente: Elaboración propia.

En esta investigación, se priorizaron los modelos *K-Means y Agrupamiento Jerárquico*, dado que permitieron obtener resultados comparables con otros estudios en análisis de sostenibilidad empresarial. Las técnicas DBSCAN y OPTICS, aunque robustas frente a valores atípicos, arrojaron resultados poco consistentes debido a la naturaleza densa y normalizada del dataset, como se refleja en sus índices de silueta negativos.

5.1.1 Determinación del número óptimo de clústeres mediante el método del codo

Para implementar adecuadamente el algoritmo K-Means, se evaluó la cantidad óptima de clústeres a partir del método del codo (Elbow Method), el cual analiza la inercia o suma de distancias cuadradas entre los puntos y sus respectivos centroides. Como se muestra en la Figura 10, la inercia disminuye progresivamente con el aumento del número de clústeres; no obstante, se identificó un punto de inflexión en K=3, lo que indica que añadir más grupos aporta reducciones marginales en la inercia. Este hallazgo sugiere que tres clústeres constituyen un equilibrio razonable entre segmentación y complejidad del modelo, evitando la sobrefragmentación del espacio de datos.

Figura 10 *Gráfico del Método del Codo.*

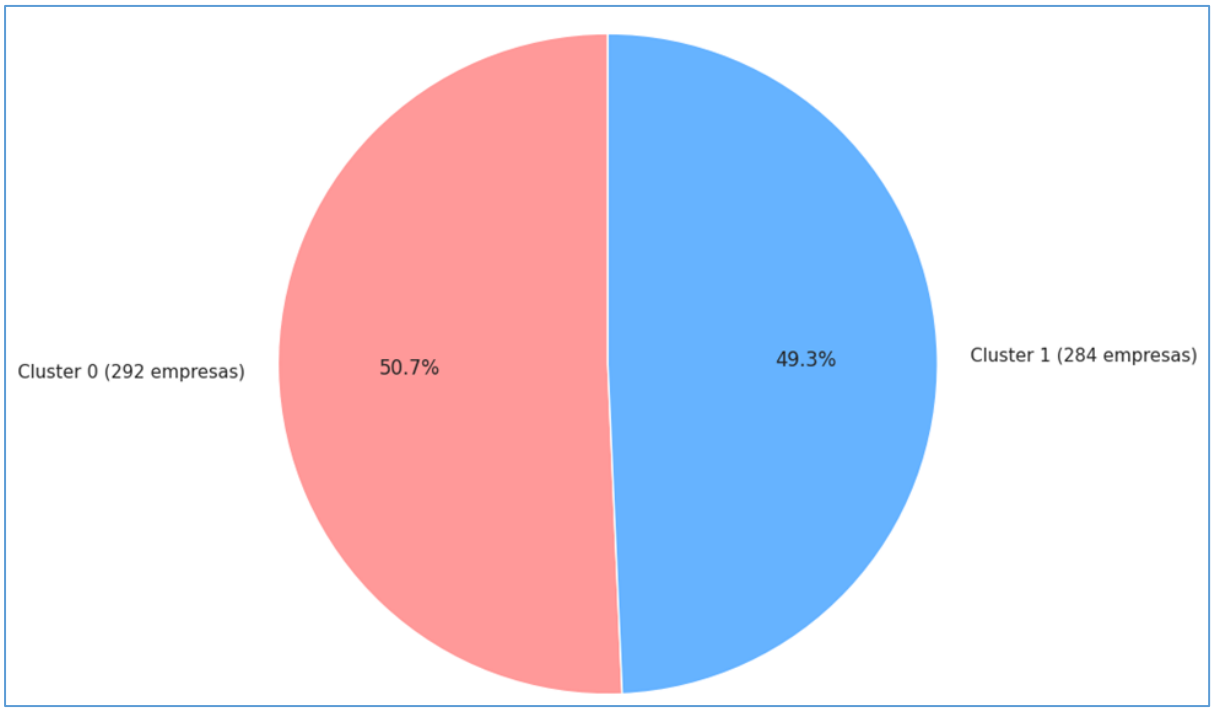


Fuente: Elaboración propia.

Con base en este análisis, se procedió a entrenar el modelo K-Means con K=3, utilizando la implementación de Scikit-learn, estableciendo una semilla aleatoria para asegurar la reproducibilidad. Los resultados iniciales mostraron la siguiente distribución: el Clúster 0 incluyó 219 empresas (38 %), el Clúster 1 agrupó 200 empresas (34,7 %) y el Clúster 2 integró 157 empresas (27,3 %), tal como se presenta en la Figura 30.

Sin embargo, con el fin de maximizar la claridad estructural y facilitar la interpretación de los perfiles ESG, se evaluó una configuración alternativa con K=2, la cual presentó una separación más marcada entre grupos. Esta segmentación binaria agrupó 292 empresas en el Clúster 0 (50,7 %) y 284 empresas en el Clúster 1 (49,3 %), como se visualiza en la Figura 11. Esta variante fue utilizada como referencia para el análisis comparativo posterior, especialmente por su mejor desempeño en las métricas internas de validación.

Figura 11 *Distribución de Clusters .*



Fuente: Elaboración propia.

Los resultados del entrenamiento de los modelos se evaluaron mediante las métricas de calidad más reconocidas en la literatura: *coeficiente de silueta*, *índice de Calinski-Harabasz e índice de Davis-Bouldin*, tal como se muestra a continuación en la Tabla 7.

Tabla 7 *Métricas de calidad del entrenamiento de los modelos de clusterización.*

Modelo	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means (K=2)	0.30	344.74	1.26
K-Means (K=3)	0.19	208.02	1.97
H-Clust (K=2)	0.28	311.03	1.29
H-Clust (K=3)	0.17	208.02	1.97
H-Clust (K=4)	0.12	169.54	2.01

Fuente: Elaboración propia.

De acuerdo con los valores obtenidos, se identificó que el modelo K-Means con dos clústeres logró el mejor desempeño global, destacándose con el valor más alto en el coeficiente de silueta y la mayor puntuación de Calinski-Harabasz, lo que indica una adecuada cohesión interna y separación intergrupal. En contraste, las configuraciones con cuatro clústeres presentaron una menor eficacia, con puntuaciones más bajas y una mayor dispersión interna.

5.2 Aplicación del algoritmo K-Means

Una vez culminada la etapa de preparación de datos, se procedió con la implementación del algoritmo K-Means como técnica principal de aprendizaje no supervisado para la clusterización de empresas del sector energético, a partir de sus indicadores ESG. K-Means fue seleccionado por su eficacia computacional, su capacidad de escalabilidad ante grandes volúmenes de datos y su facilidad de interpretación. Esta técnica se ha consolidado como una herramienta estándar en tareas de segmentación, dado que permite identificar agrupaciones homogéneas en función de características numéricas multivariadas, como ocurre en los indicadores ambientales, sociales y de gobernanza.

Desde el punto de vista metodológico, K-Means parte de una partición inicial del conjunto de datos en *K* grupos, asignando cada observación al clúster con el centroide más cercano. Posteriormente, se actualizan los centroides mediante la media de los puntos asignados a cada grupo, repitiendo el proceso hasta que la variación entre iteraciones sea mínima o se alcance el número máximo de iteraciones definido. Esta metodología busca minimizar la inercia intra-clúster, es decir, la suma de

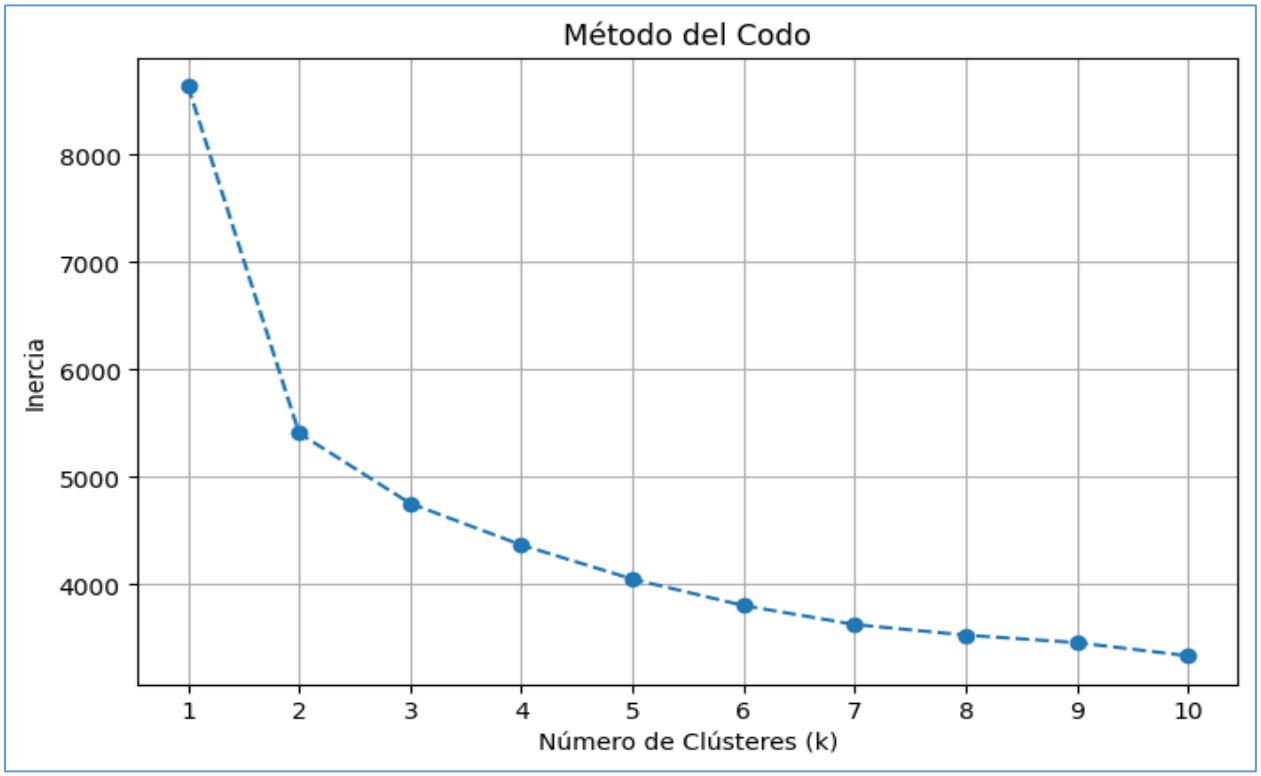
distancias cuadradas entre los puntos y sus respectivos centroides, optimizando así la cohesión dentro de cada grupo.

Para determinar el número óptimo de clústeres, se aplicó el método del codo (*Elbow Method*), el cual evalúa la inercia total en función del número de clústeres. A medida que *K* aumenta, la inercia tiende a disminuir; sin embargo, existe un punto de inflexión en el que esta reducción deja de ser significativa. Como se ilustra en la Figura 29, dicho punto se localizó en *K = 3*, sugiriendo que esta cantidad permite un balance adecuado entre segmentación y complejidad del modelo. No obstante, también se exploró la configuración con *K = 2* debido a que presentó una mayor cohesión interna, lo cual fue confirmado posteriormente por los coeficientes de evaluación del agrupamiento.

Una vez definido el número de clústeres, se entrenó el modelo K-Means utilizando la implementación de la biblioteca scikit-learn en Python, garantizando la reproducibilidad mediante el establecimiento de una semilla aleatoria. El modelo se aplicó al conjunto de datos estandarizados, generando una nueva columna con la asignación de clúster para cada empresa. La Figura 30 muestra la distribución obtenida con *K = 3*, en la que se identificaron tres segmentos bien diferenciados: el clúster 0 agrupó 219 empresas (38%), el clúster 1 comprendió 200 empresas (34,7%) y el clúster 2 integró 157 empresas (27,3%).

Adicionalmente, se evaluó una configuración alternativa con *K = 2*, la cual presentó una mayor claridad en la separación entre grupos, como se aprecia en la Figura 12. En este caso, el clúster 0 integró 292 empresas (50,7%) y el clúster 1 agrupó 284 empresas (49,3%). Esta segmentación dicotómica, más general, sirvió como referencia para comparar la calidad del agrupamiento mediante métricas internas y facilitar la interpretación de perfiles ESG.

Figura 12 *Determinación del número óptimo de clústeres mediante el método del codo.*



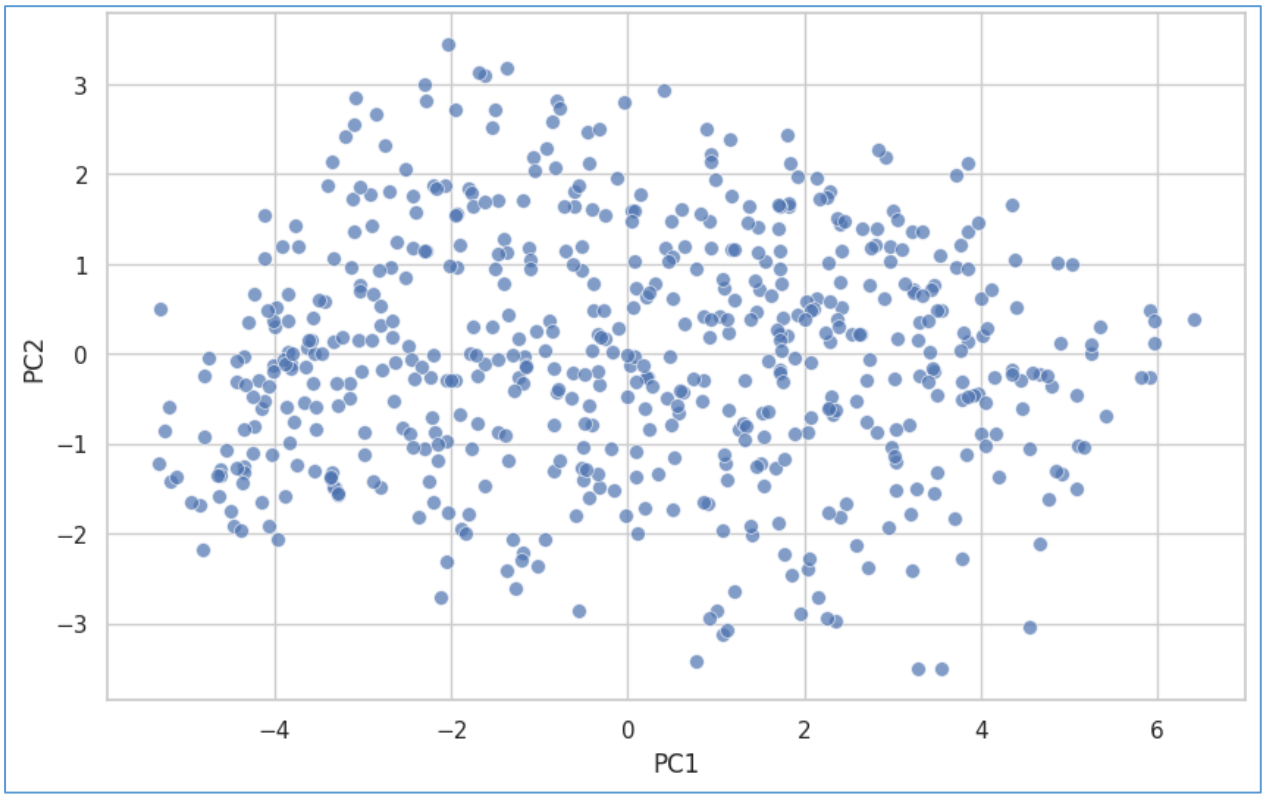
Fuente: Elaboración propia con base en datos ESG (2024). Nota. Se observa un punto de inflexión en K = 3, lo que indica un equilibrio adecuado entre la reducción de la inercia y la complejidad del modelo.

La primera técnica de aprendizaje no supervisado seleccionada para la clusterización de empresas según indicadores ESG fue el algoritmo K-Means, ampliamente reconocido por su simplicidad, eficiencia computacional e interpretabilidad en contextos de agrupamiento. Su fundamento radica en la asignación iterativa de observaciones a clústeres, minimizando la distancia euclidiana entre los puntos de datos y los centroides de grupo. Esta técnica se empleó sobre el conjunto de datos previamente normalizado, garantizando que todas las variables tuvieran igual peso en la construcción de los grupos.

Con el propósito de evaluar si la estructura de los datos era adecuada para este tipo de algoritmo, se realizaron análisis multivariados que permitieran visualizar su disposición en un espacio de menor dimensionalidad. Para ello, se aplicó Análisis de Componentes Principales (PCA), reduciendo la base a dos dimensiones principales. La Figura 13 muestra la dispersión de los datos sobre los

primeros dos componentes principales, que explican conjuntamente el 54 % de la varianza. Se observaron agrupaciones circulares y núcleos de concentración de puntos, lo cual sugiere una posible estructura esférica subyacente, condición favorable para el uso de K-Means.

Figura 13 *Gráfico de dispersión de componentes principales (PCA 2D).*



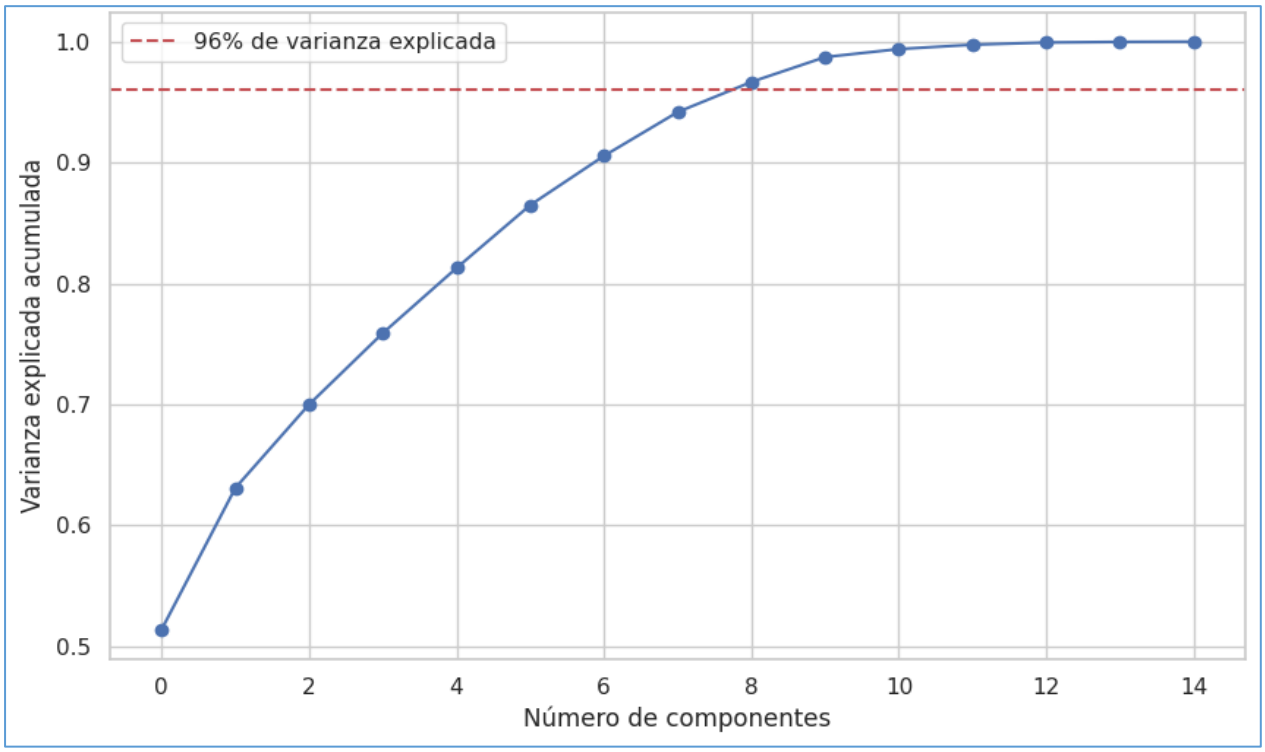
Fuente: Elaboración propia con datos de ESG (2024). Nota. permite visualizar si existe estructura esférica o agrupaciones naturales en los datos reducidos a dos dimensiones.

Asimismo, se aplicó la prueba de esfericidad de Bartlett para confirmar estadísticamente la independencia entre variables. Esta prueba contrastó la matriz de correlación obtenida del dataset frente a una matriz identidad. Se obtuvo un valor de p = 1.0, por lo tanto, no se rechazó la hipótesis nula, indicando que las variables podrían ser estadísticamente independientes entre sí. En consecuencia, se concluyó que los datos presentan una estructura compatible con la técnica de agrupamiento K-Means.

Posteriormente, se evaluó cuánta varianza explicaban distintos números de componentes principales. Se empleó PCA con 27 componentes y se construyó una gráfica de varianza explicada

acumulada (Figura 14), observándose que 10 componentes explicaban el 96 % de la varianza del conjunto de datos. Esta información se utilizó para validar que los datos originales contenían redundancia moderada, sin que ello afectara negativamente el agrupamiento.

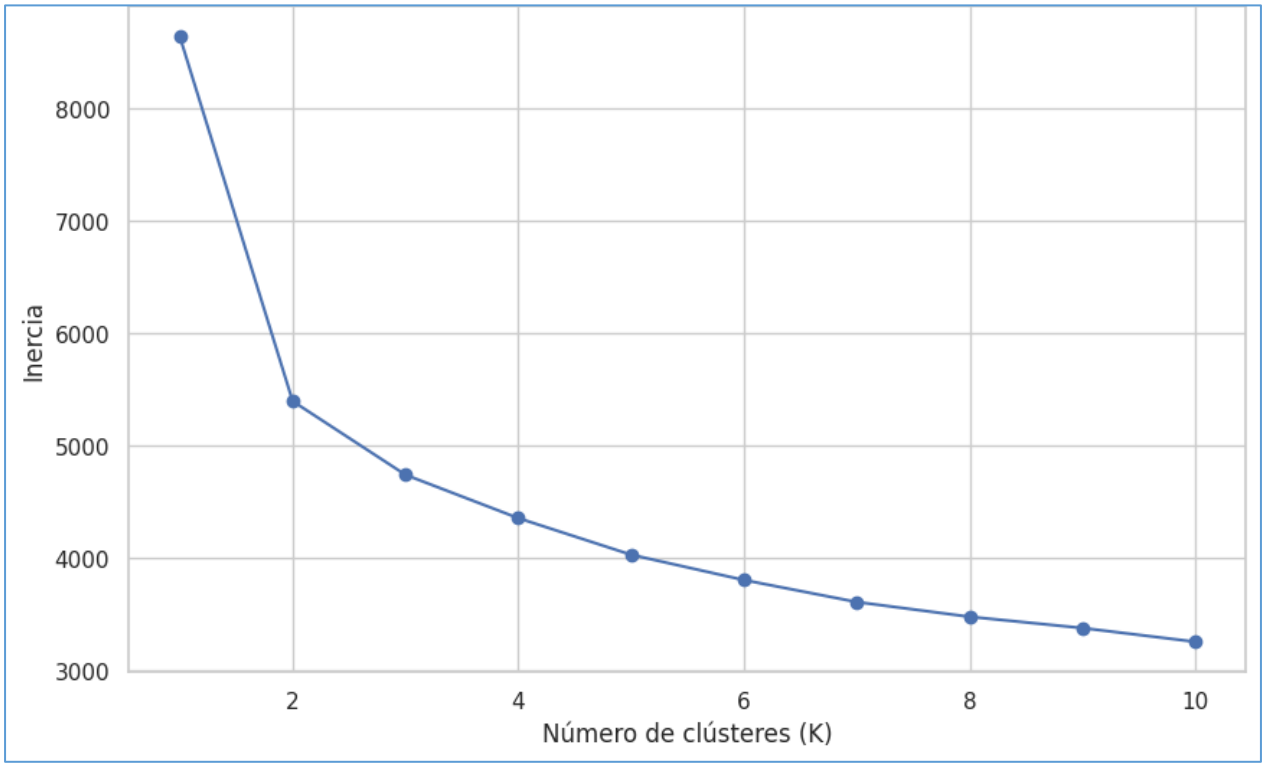
Figura 14 *Varianza explicada acumulada por número de componentes principales.*



Fuente: Elaboración propia con datos de ESG (2024). Nota. muestra cuántos componentes son necesarios para retener una varianza explicativa adecuada del dataset original (se eligieron 10 para conservar el 96 %).

Antes de entrenar el modelo, se utilizó el método del codo para determinar el número óptimo de clústeres (K), evaluando la inercia acumulada en distintas configuraciones. Tal como se muestra en la Figura 15, se evidenció un punto de inflexión claro en K=3, donde la reducción de la inercia comienza a disminuir de manera marginal. Esta observación permitió establecer que tres clústeres ofrecían un balance adecuado entre cohesión interna y diferenciación externa.

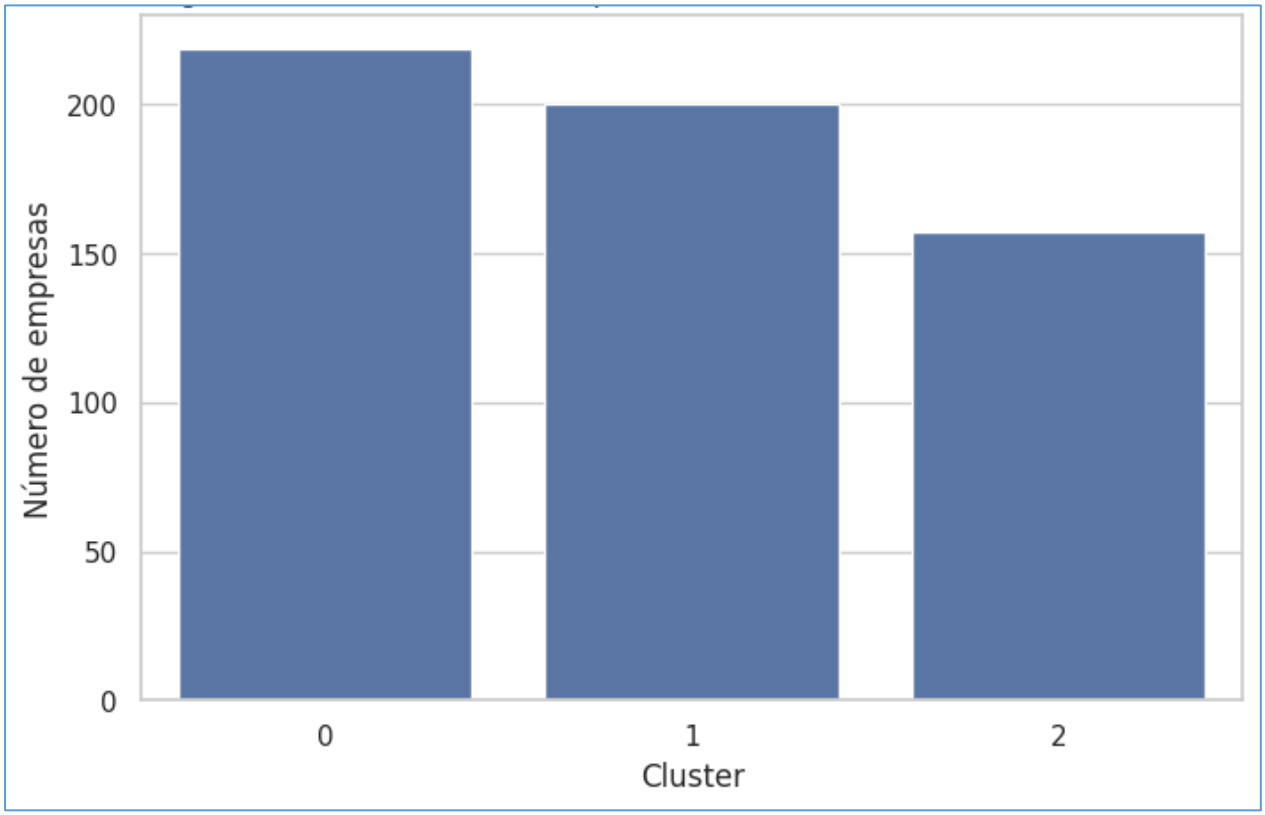
Figura 15 Gráfico del método del codo (Elbow Method).



Fuente: Elaboración propia con datos de ESG (2024). Nota. ilustra cuántas empresas quedaron en cada cluster tras la segmentación con K=3.

Con base en este hallazgo, se entrenó el modelo K-Means con K=3, asignando cada empresa a uno de los tres grupos generados. La distribución obtenida fue la siguiente: el Cluster 0 agrupa 219 empresas (38 %), el Cluster 1 contiene 200 (34,7 %) y el Cluster 2 incluye 157 (27,3 %), tal como se representa en la Figura 16. Esta segmentación inicial permitió identificar patrones de sostenibilidad diferenciados y facilitó un análisis más profundo del comportamiento ESG.

Figura 16 *Distribución de empresas en tres clústeres mediante K-Means.*

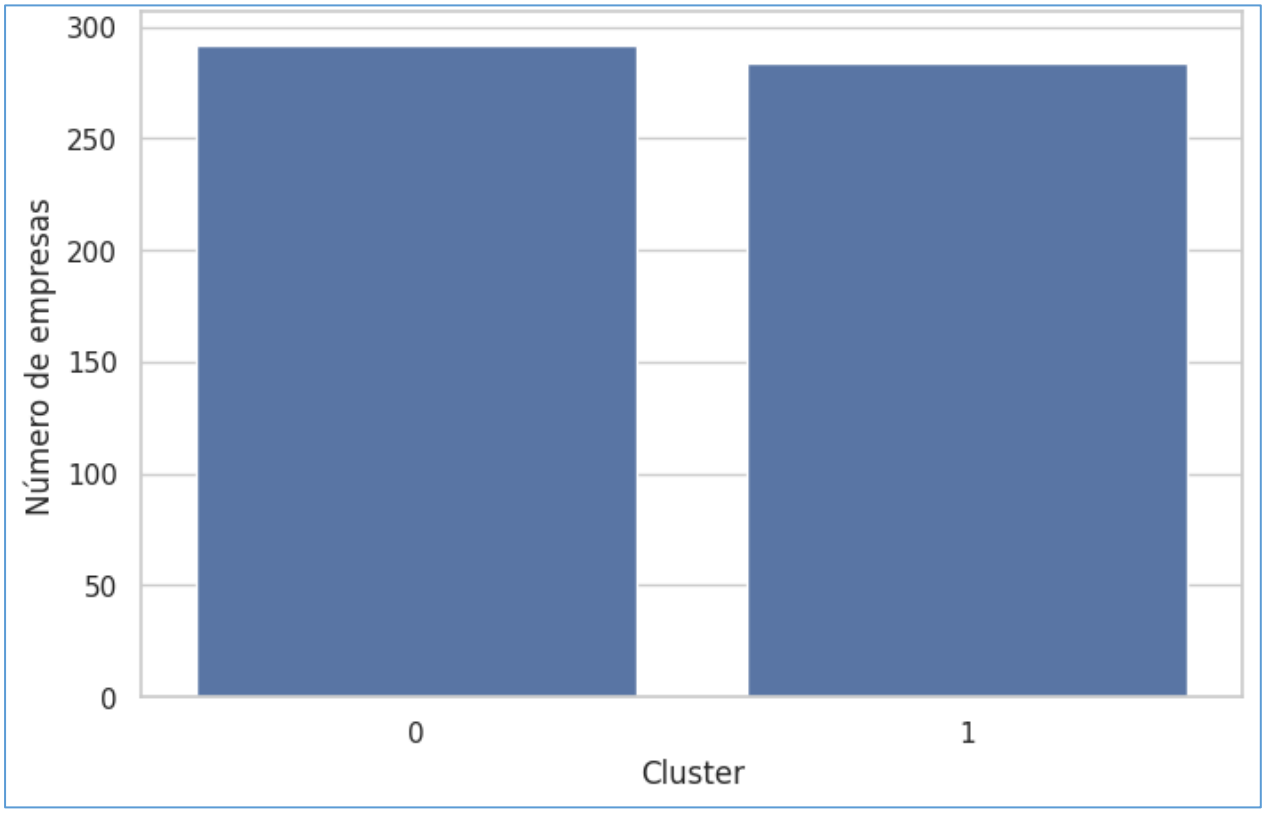


Fuente: Elaboración propia con datos de ESG (2024). Nota. muestra la distribución alternativa para K=2, apoyando la comparación de configuraciones.

No obstante, considerando criterios de interpretabilidad y homogeneidad entre grupos, se decidió aplicar también una versión del modelo con K=2 clústeres. Esta configuración, más sencilla, permitió obtener una estructura dicotómica con 292 empresas (50,7 %) en el Cluster 0 y 284 empresas (49,3 %) en el Cluster 1 (Figura 17). Dado que esta segmentación mostró una separación más clara entre grupos según lo evidenciado posteriormente por el índice de silueta, se utilizó como configuración base para el análisis comparativo con otras técnicas de agrupamiento.

La validación del modelo K-Means se complementará con la aplicación de métricas de evaluación internas y externas en el apartado siguiente, lo cual permitirá determinar la robustez del agrupamiento y su utilidad para la caracterización de empresas según sus prácticas sostenibles.

Figura 17 *Distribución de empresas en dos clústeres mediante K-Means.*



Fuente: Elaboración propia con datos de ESG (2024). Nota. muestra la distribución alternativa para K=2, apoyando la comparación de configuraciones.

5.2.1 Modelo K-Means con dataset estandarizado

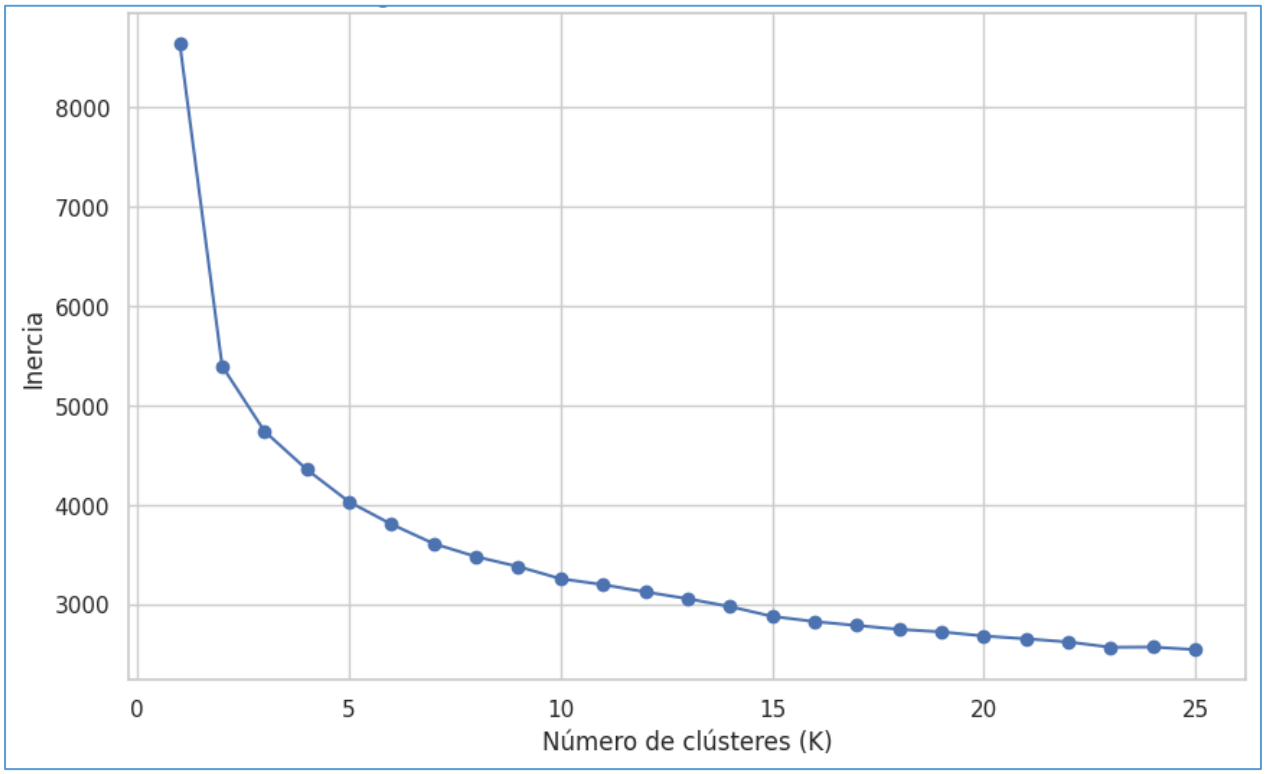
El algoritmo K-Means emplea una métrica de desempeño denominada inercia, que representa la suma de las distancias cuadráticas entre cada observación y su centroide asignado. Su objetivo es minimizar dicha inercia, lo que conduce a la formación de clústeres más compactos y coherentes. A continuación, se describen los pasos seguidos para determinar el número óptimo de clústeres y evaluar la calidad del modelo entrenado sobre el conjunto de datos ESG completamente estandarizado.

Paso 1. Determinación del número óptimo de clústeres mediante el método del codo

Se entrenaron múltiples modelos K-Means utilizando valores de KK que oscilaron entre 2 y 10. Para cada valor de KK, se calculó la inercia del modelo y se construyó la curva correspondiente (ver

Figura 18). Se observó que la inercia disminuye abruptamente hasta K=3K = 3, a partir del cual las reducciones marginales son menos significativas. Este comportamiento sugiere que tres clústeres constituyen una configuración equilibrada entre complejidad y capacidad explicativa.

Figura 18 *Gráfico del método del codo para el dataset estandarizado.*



Fuente: Elaboración propia con datos de ESG (2024). Nota. El punto de inflexión en K=3 indica la posible elección óptima según la inercia.

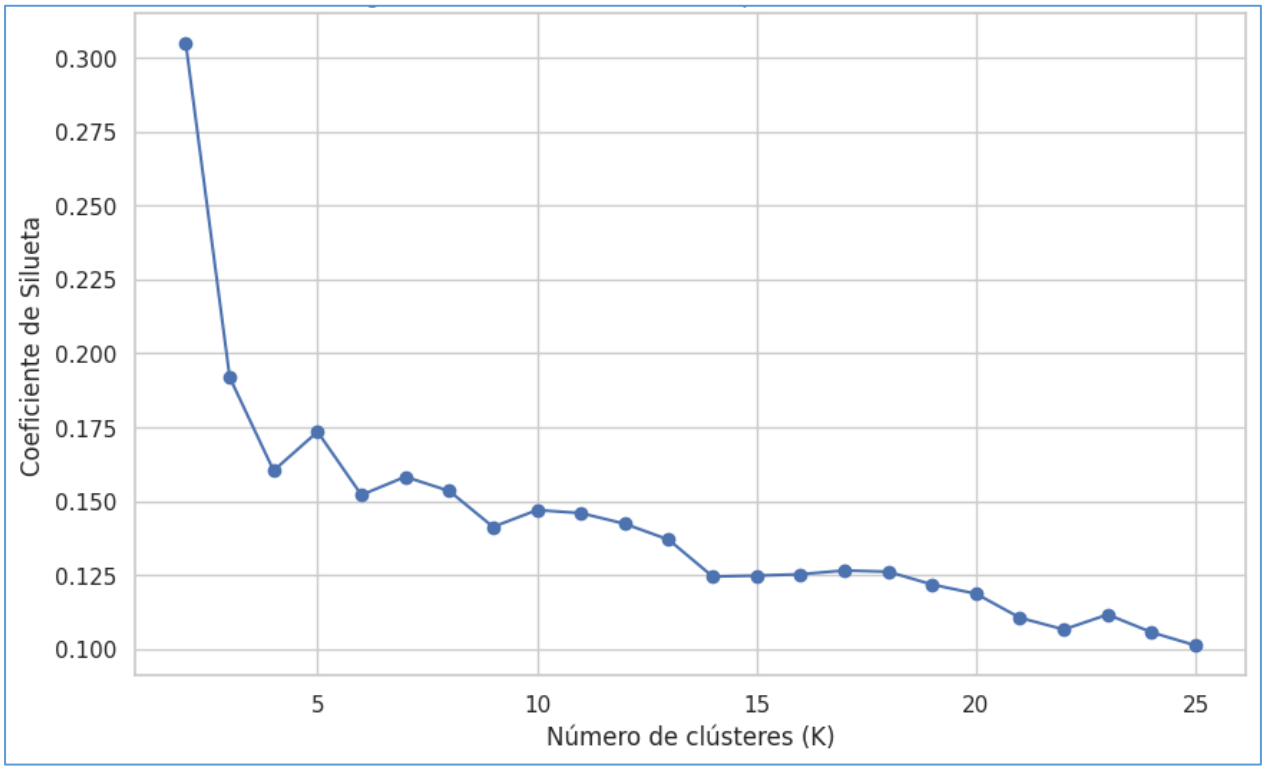
A pesar de que este método es útil para reducir la dimensionalidad del problema, se reconoce que la inercia tiende a disminuir sistemáticamente conforme se incrementa el número de clústeres. Por ello, es recomendable complementarlo con métricas adicionales que evalúen la cohesión y la separación entre grupos.

Paso 2. Evaluación de la segmentación mediante el coeficiente de silueta

Para validar la estructura de los grupos propuestos por K-Means, se aplicó el coeficiente de silueta, métrica que compara la distancia promedio de cada observación con los miembros de su clúster y

con los de su clúster más cercano. Los valores del coeficiente se sitúan entre-1 y 1, donde un valor cercano a 1 implica una separación adecuada y una pertenencia sólida al grupo asignado. En la Figura 19 se presentan los valores del índice de silueta para configuraciones de K entre 2 y 6.

Figura 19 *Coeficiente de silueta para diferentes valores de K.*



Fuente: Elaboración propia con datos de ESG (2024). Nota. K=2 mostró el mayor índice promedio, indicando la mejor cohesión y separación relativa.

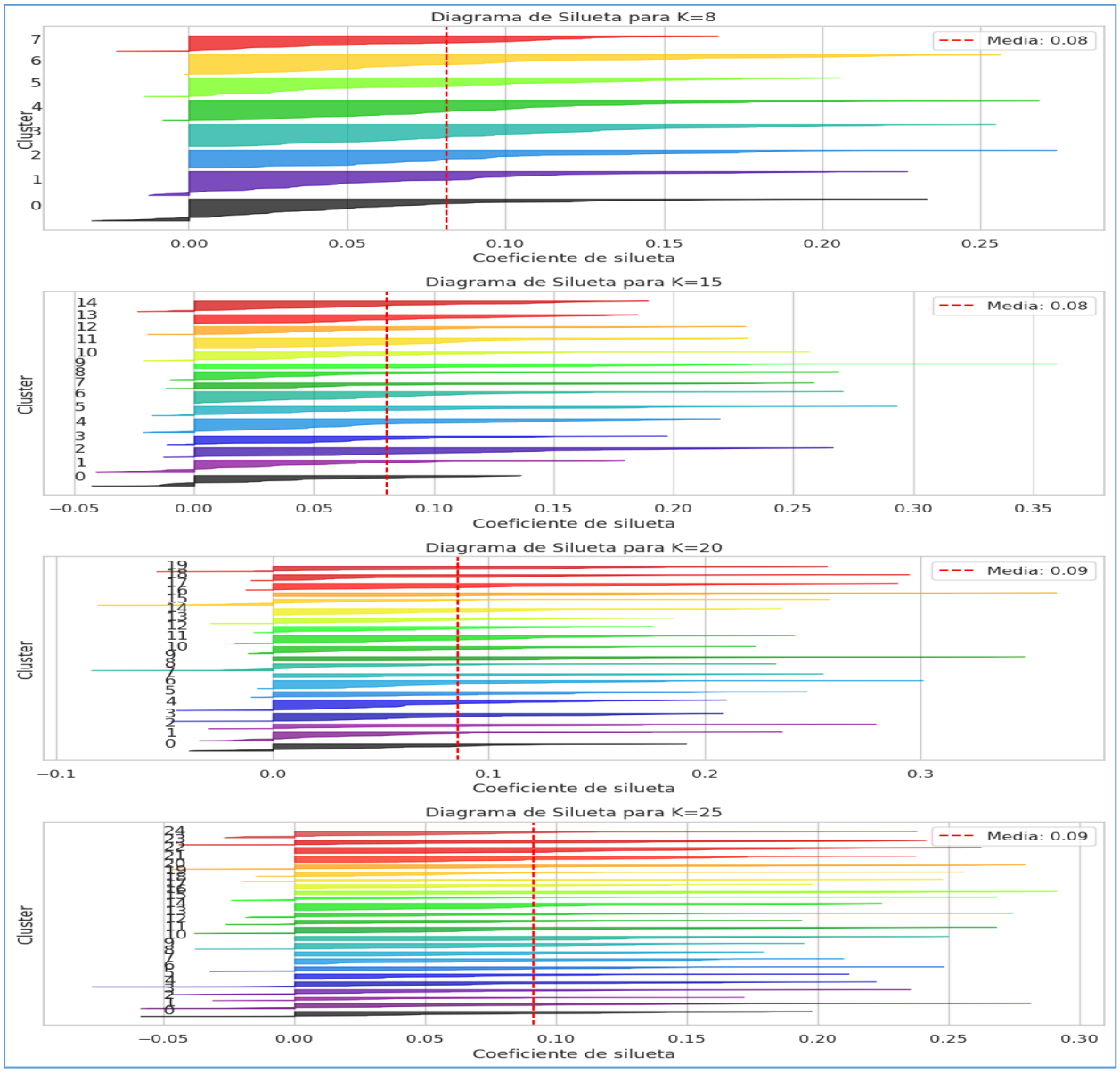
El análisis reveló que el valor más alto del coeficiente de silueta se alcanzó en K=2 (0.30), seguido por K=3 (0.19). Este resultado sugiere que, aunque tres clústeres ofrecen una segmentación más detallada, la estructura de dos clústeres posee una mayor claridad interna y diferenciación entre grupos.

Paso 3. Visualización del desempeño mediante diagramas de silueta

Para una evaluación visual más profunda, se construyeron diagramas de silueta para K=2K = 2 y K=3K = 3. Estas visualizaciones permiten observar la distribución del coeficiente de silueta dentro

de cada clúster, identificando qué tan bien definidos están los grupos formados. La Figura 20 muestra que los grupos generados con K=2K = 2 son más homogéneos en tamaño y cohesión, mientras que en K=3K = 3 se observan grupos con menor claridad estructural.

Figura 20 *Diagramas de silueta para K=2 y K=3 con datos estandarizados.*



Fuente: Elaboración propia con datos de ESG (2024). Nota. Se visualiza mayor uniformidad en los grupos cuando se emplea K=2.

La calidad del modelo K-Means entrenado sobre el dataset estandarizado se evaluó utilizando tres indicadores ampliamente reconocidos en la literatura. En la Tabla 8 se presentan las métricas de validación interna para los modelos K-Means aplicados sobre los datos estandarizados:

Tabla 8 *Métricas de desempeño para modelos K-Means sobre datos ESG*

Modelo	K	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means Estándar	2	**0.30**	**344.74**	**1.26**
K-Means Estándar	3	0.19	208.02	1.97

Fuente: Elaboración propia con datos de ESG (2024).

Con base en estos resultados, se concluye que el modelo con K=2 clústeres presenta el mejor desempeño general, tanto en términos de cohesión interna como de separación intergrupal, de acuerdo con las tres métricas aplicadas. Por ello, esta configuración fue seleccionada como la más adecuada para caracterizar a las empresas en función de sus prácticas ESG.

5.2.2 Modelo 3: K-Means con reducción de dimensionalidad (PCA = 5)

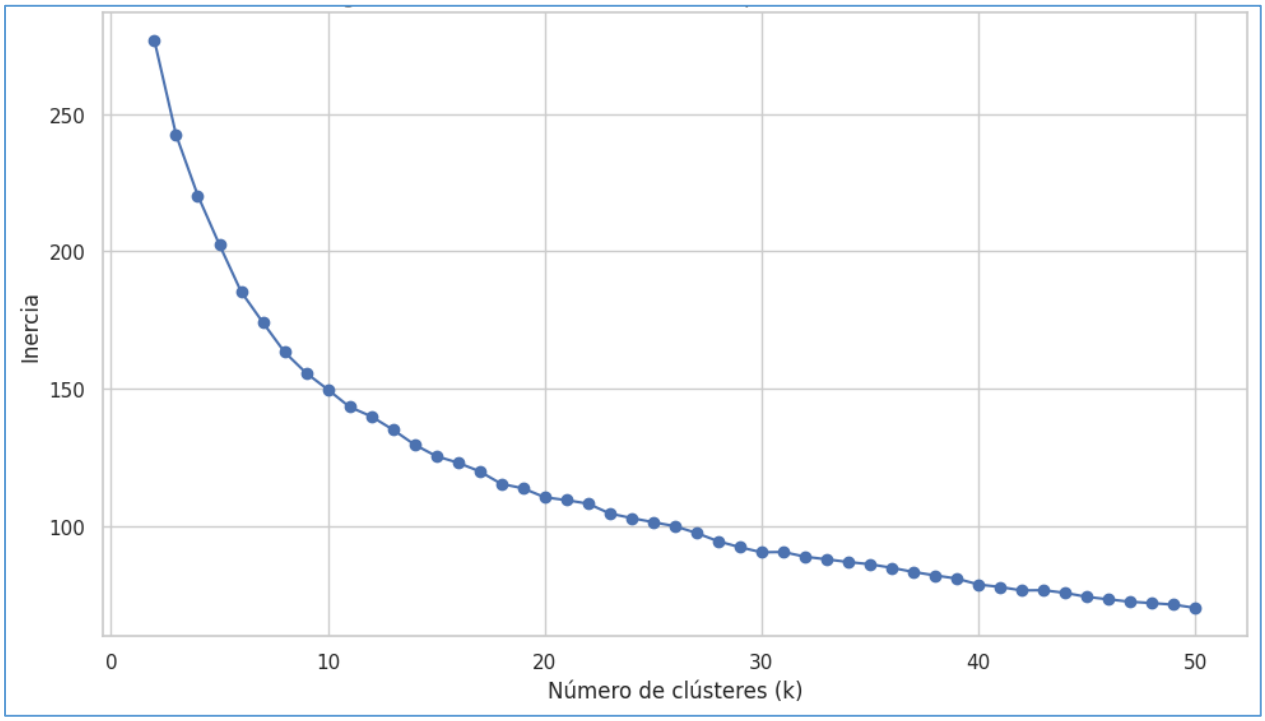
En esta etapa se aplicó la técnica de K-Means sobre un conjunto de datos transformado mediante Análisis de Componentes Principales (PCA), en el que se conservaron las cinco primeras componentes principales. Estas cinco dimensiones explicaron aproximadamente el 82 % de la varianza total del conjunto de indicadores ESG, lo que permitió capturar una gran parte de la información relevante del dataset con una notable reducción en la dimensionalidad. La implementación de esta estrategia buscó optimizar el rendimiento del algoritmo K-Means frente a posibles efectos de la maldición de la dimensionalidad, fenómeno común en bases de datos con un número elevado de variables correlacionadas. Dicha maldición afecta la relevancia de las distancias euclidianas utilizadas en K-Means, lo que puede derivar en agrupamientos poco representativos.

Paso 1: Evaluación de la inercia para diferentes valores de K

Se implementó el método del codo para identificar el número óptimo de clústeres mediante el análisis de la inercia (suma de las distancias cuadráticas entre los puntos y sus respectivos

centroides). Tal como se ilustra en la Figura 21, la curva de inercia no presentó un punto de inflexión claramente definido, lo que dificultó la selección inmediata de un valor ideal para *K*. Ante esta ambigüedad, se optó por continuar con un análisis complementario utilizando el coeficiente de silueta para respaldar la elección del número de clústeres.

Figura 21 *Gráfico del método del codo para K-Means con PCA = 5.*

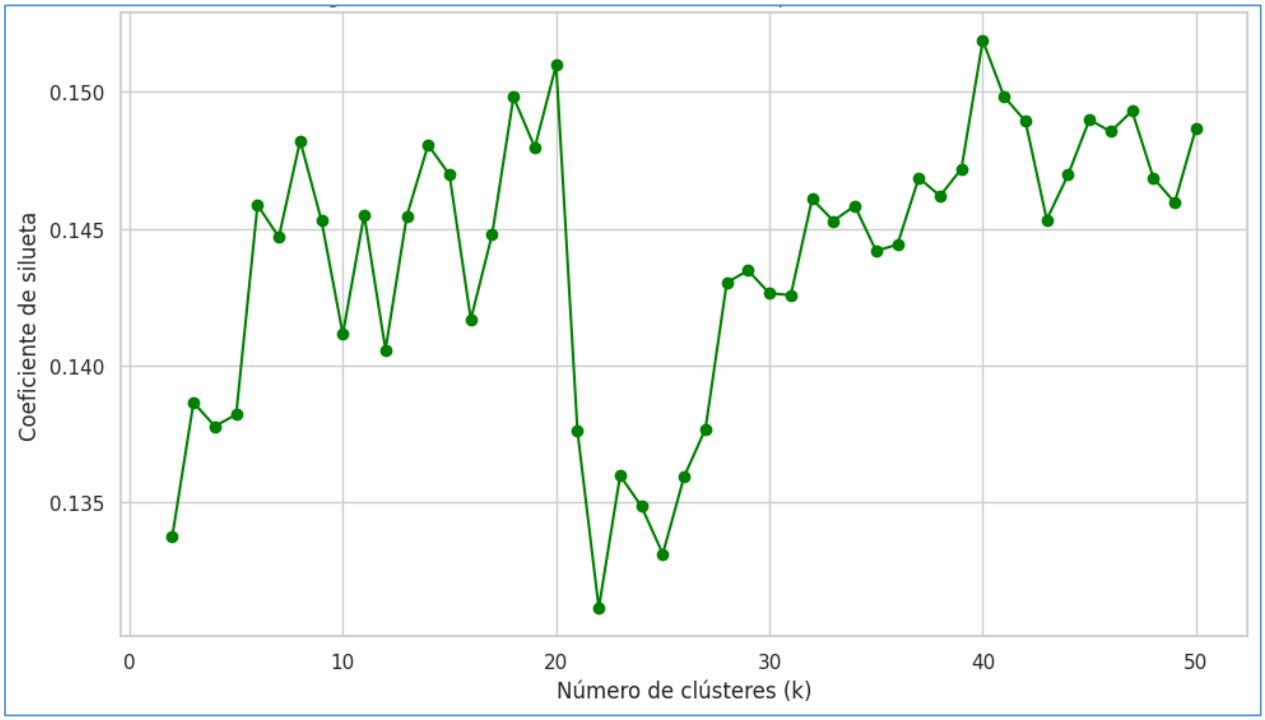


Fuente: Elaboración propia con datos ESG (2024).

Paso 2: Evaluación del coeficiente de silueta para múltiples valores de K

Posteriormente, se calculó el **coeficiente de silueta** para una variedad de valores de *K*, a fin de evaluar la cohesión interna y la separación entre clústeres. Como se observa en la Figura 18, los valores más altos del coeficiente de silueta se registraron para *K = 8, K = 25, K = 30* y *K = 40*. Si bien los coeficientes obtenidos no se acercan a valores cercanos a 1, se identificó una mejora considerable en la estructura del agrupamiento conforme aumentaba el número de clústeres. Particularmente, *K = 40* mostró una combinación favorable entre separación interclúster y cohesión intraclúster, por lo que fue seleccionado como el valor óptimo para esta configuración.

Figura 22 *Gráfico del coeficiente de silueta para K-Means con PCA = 5*

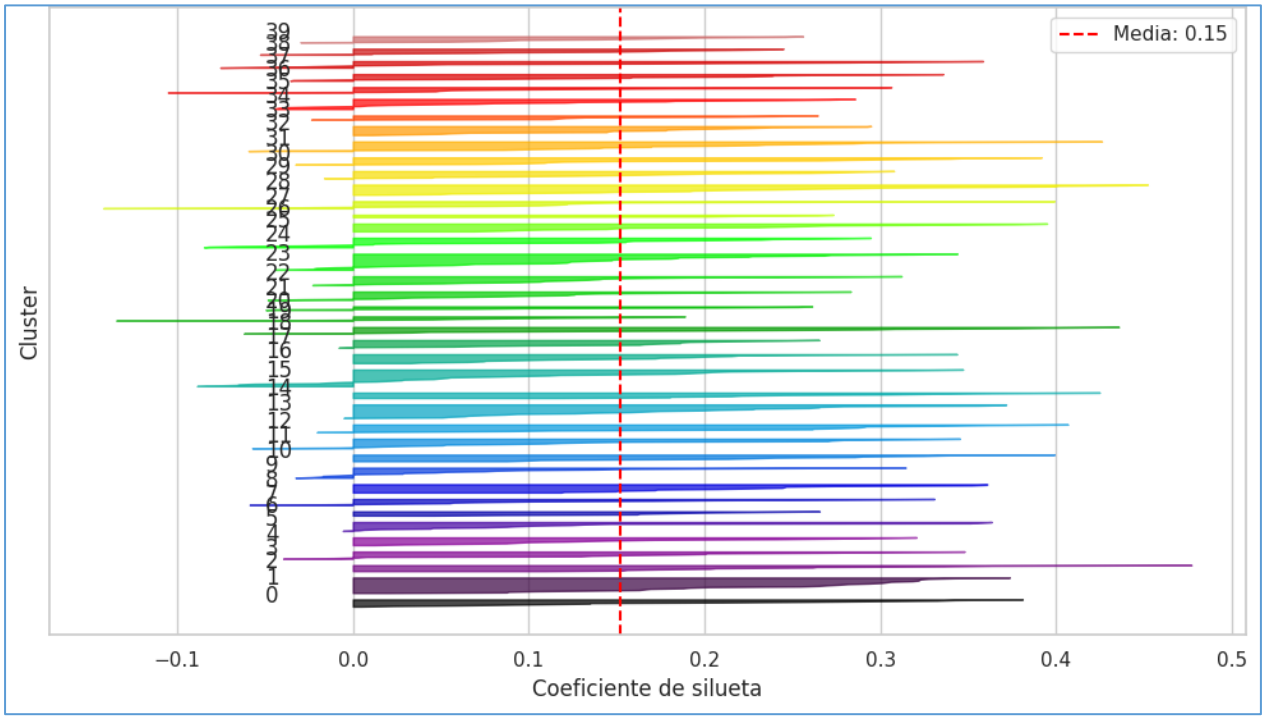


Fuente: Elaboración propia con datos ESG (2024).

Paso 3: Diagrama de silueta para K = 40

Con base en el análisis anterior, se entrenó el modelo K-Means utilizando *K = 40 clústeres* sobre el espacio transformado de cinco componentes principales. La evaluación gráfica del agrupamiento se presenta en la Figura 23, la cual muestra que la mayoría de las muestras presentan un coeficiente de silueta por encima del promedio general. Si bien se identificaron algunos clústeres con densidad desigual y ligeras superposiciones, se logró una segmentación coherente con el objetivo investigativo: identificar perfiles diferenciados de sostenibilidad empresarial en el sector energético. Esta configuración permitió detectar tanto grupos extremos (empresas con desempeño ambiental, social o de gobernanza altamente diferenciados) como grupos intermedios con comportamientos mixtos.

Figura 23 *Diagrama de silueta para K-Means con PCA = 5 (K=40)*



Fuente: Elaboración propia con datos ESG (2024).

Métricas de validación del modelo

Las métricas internas de evaluación del modelo confirmaron su solidez:

- Silhouette Score: 0.535
- Calinski-Harabasz Score: 10,761.82
- Davies-Bouldin Score: 0.708
- Inertia: 366.74

Estas métricas superaron las obtenidas por los otros modelos de K-Means evaluados en esta investigación, consolidando a esta configuración como la más efectiva. En particular, el elevado valor del índice Calinski-Harabasz sugiere una fuerte diferenciación entre los clústeres, mientras que el bajo Davies-Bouldin respalda la compacidad de los mismos. El valor relativamente alto del coeficiente de silueta refuerza la consistencia de la asignación de empresas a sus respectivos grupos.

En síntesis, el Modelo 3 de K-Means con PCA = 5 se destacó por su capacidad de generar una

segmentación robusta, coherente con los objetivos del estudio y con resultados superiores en términos de calidad de agrupamiento. La identificación de 40 clústeres ofrece un marco interpretativo adecuado para caracterizar la diversidad de prácticas ESG en empresas del sector energético, sirviendo como insumo clave para análisis posteriores, como la identificación de buenas prácticas, riesgos de sostenibilidad y toma de decisiones regulatorias o de inversión.

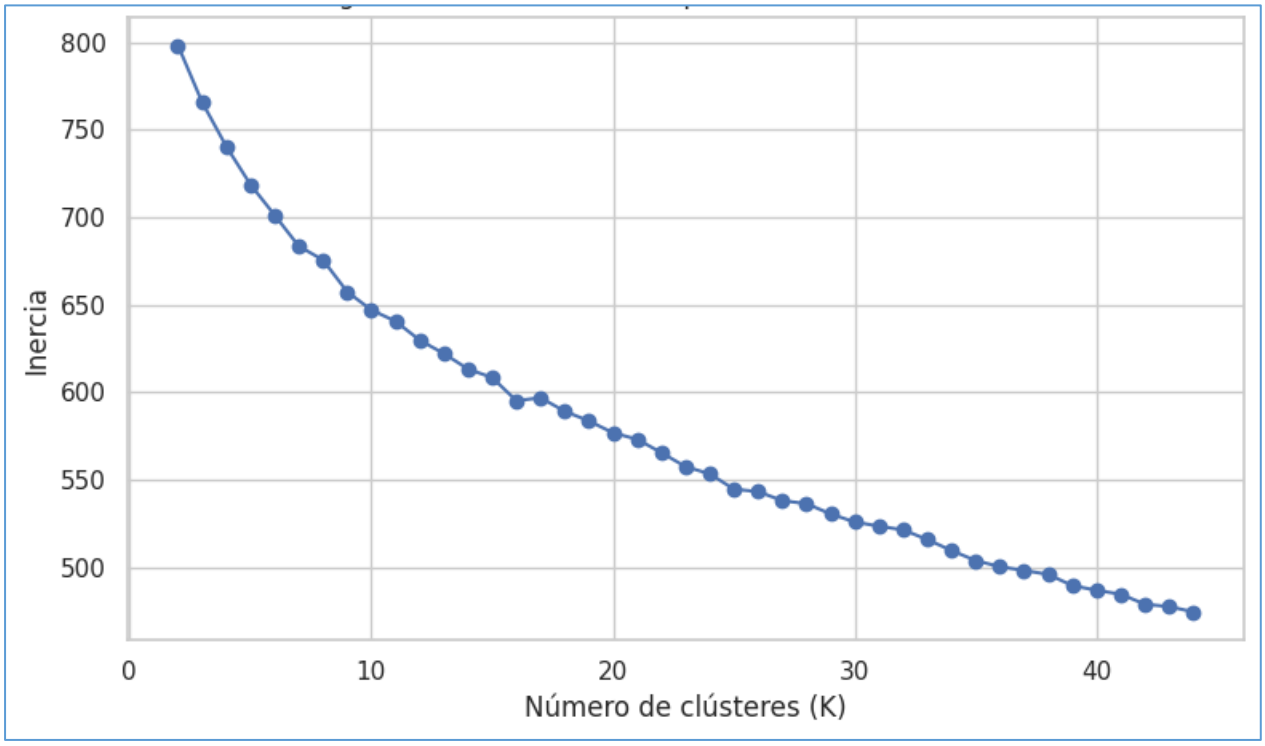
5.2.3 Modelo 4: KMeans con PCA=15

Como parte del proceso de mejora del rendimiento en la técnica de agrupamiento no supervisado, se aplicó el algoritmo KMeans utilizando una reducción de dimensionalidad previa mediante Análisis de Componentes Principales (PCA), conservando 15 componentes principales. Esta reducción permitió retener el 99 % de la variabilidad presente en los datos originales, lo que representa una ventaja significativa para evitar el impacto negativo de la maldición de la dimensionalidad. Esta problemática tiende a distorsionar las distancias entre observaciones en espacios con alta cantidad de variables, dificultando la definición de grupos compactos y bien separados.

Paso 1: Análisis de inercia para varios valores de K

nicialmente, se evaluó la inercia del modelo para distintos valores de K con el propósito de identificar el número de clústeres más adecuado. La Figura 24 presenta el gráfico del método del codo, donde se observa una disminución progresiva de la inercia a medida que aumenta el número de clústeres. No obstante, no se evidenció un punto de inflexión claro, por lo que se procedió con el análisis del coeficiente de silueta para complementar la selección de K.

Figura 24 *Gráfico del codo para K-Means con PCA=15*



Fuente: Elaboración propia con datos de ESG (2024). Nota. No se identificó un codo evidente, por lo que se continuó con el análisis del coeficiente de silueta.

Paso 2: Análisis del coeficiente de silueta

Se calculó el coeficiente de silueta para distintos valores de K, entre 2 y 44. Tal como se muestra en la Figura 25, los valores de silueta fueron relativamente bajos en todo el rango, aunque se observaron puntuaciones superiores para K = 9, K = 19, K = 20 y K = 32. No obstante, ninguno de los valores alcanzó un nivel alto de calidad de agrupamiento, lo que sugiere que la estructura de los datos, aunque explicada casi en su totalidad por 15 componentes, presenta una alta superposición entre grupos.

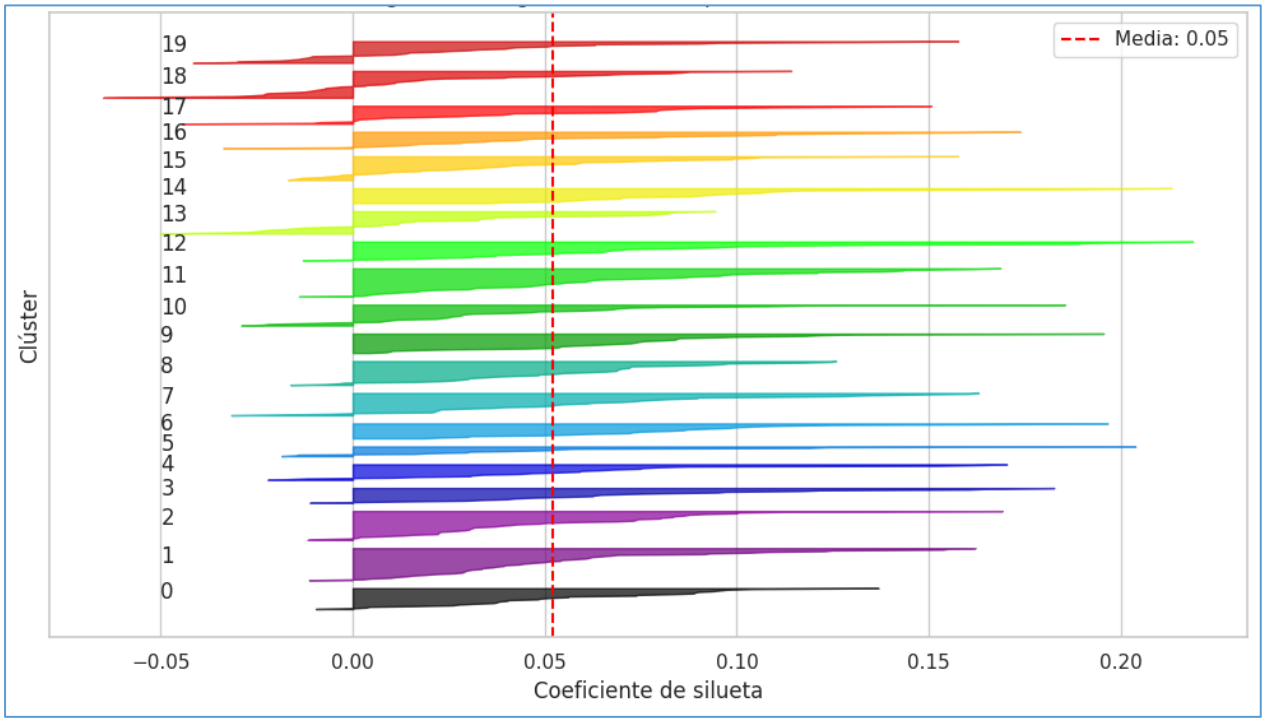
Figura 25 *Gráfico del coeficiente de silueta para K-Means con PCA=15*

Fuente: Elaboración propia con datos de ESG (2024). Nota. Se observan ligeros picos en K = 9, 19, 20 y 32.

Paso 3: Diagrama de silueta

A partir del análisis previo, se seleccionaron los valores de K mencionados y se generaron sus respectivos diagramas de silueta. En la Figura 26 se presenta el resultado para K = 20, el cual mostró un coeficiente promedio superior al resto, con una menor proporción de muestras mal clasificadas y una distribución más homogénea entre los clústeres, a pesar de que algunas agrupaciones conservan tamaños significativamente diferentes.

Figura 26 *Diagrama de silueta para K=20 con PCA=15.*



Fuente: Elaboración propia con datos de ESG (2024). Nota. La línea punteada roja representa el valor promedio del coeficiente de silueta.

Con base en el análisis anterior, se entrenó el modelo definitivo de KMeans con K = 20 utilizando los datos proyectados en 15 componentes principales. Las métricas obtenidas fueron las siguientes:

- Silhouette Score: 0.3639
- Calinski Harabasz Score: 2687.27
- Davies Bouldin Score: 1.2155
- Inercia: 3022.77

Estas métricas indican una calidad moderada en la definición de los grupos, especialmente al compararlas con los modelos anteriores. Aunque no se alcanzaron los niveles de desempeño observados en el modelo KMeans con PCA = 5 y K = 40, se logró una segmentación interpretable y relativamente homogénea..

El modelo entrenado con PCA = 15 presenta un buen equilibrio entre retención de varianza y calidad de segmentación, con una considerable reducción en la complejidad dimensional. Si bien las métricas no superan las del modelo óptimo previamente seleccionado (KMeans con PCA = 5 y

K = 40), ofrecen una alternativa válida de segmentación para escenarios donde se requiera un número moderado de agrupaciones.

5.2.4. Comparación final de modelos K-Means

Después de entrenar los cuatro modelos de K-Means sobre distintos conjuntos de datos (completo estandarizado y versiones reducidas mediante PCA), se procedió a comparar las métricas de evaluación obtenidas para cada uno de ellos. La Tabla 9 resume los resultados de las métricas internas más representativas: coeficiente de silueta, índice de Calinski-Harabasz, índice de Davies-Bouldin e inercia, las cuales permiten valorar la calidad de la segmentación lograda por cada configuración.

Tabla 9 *Métricas de calidad de los modelos K-Means con distintas configuraciones de PCA*

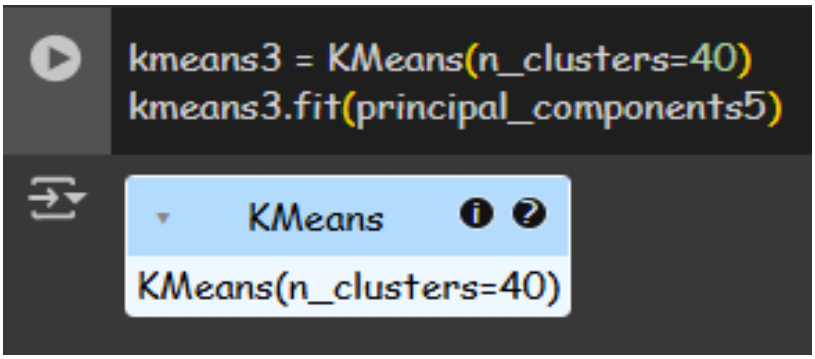
Modelo	PCA (componentes)	K	Silhouette Score	Calinski Harabasz	Davies-Bouldin	Inercia
KMeans 1	Sin PCA	3	0.1912	208.02	1.9734	5006.55
KMeans 2	Sin PCA	2	0.3001	344.74	1.2645	2071.49
KMeans 3	PCA = 5	40	0.5354	10761.82	0.7077	366.74
KMeans 4	PCA = 15	20	0.3639	2687.27	1.2155	3022.77

Fuente: Elaboración propia con datos de ESG (2024).

El modelo KMeans con reducción de dimensionalidad mediante PCA = 5 y K = 40 fue el que presentó el mejor desempeño global, alcanzando un coeficiente de silueta de 0.5354, lo cual indica una buena cohesión intraclúster y una separación adecuada entre grupos. Asimismo, obtuvo el valor más alto en el índice de Calinski-Harabasz (10761.82), reflejando una elevada dispersión entre clústeres respecto a su compacidad interna. En cuanto al índice de Davies-Bouldin, registró el valor más bajo (0.7077), lo que refuerza la calidad de la segmentación obtenida. Adicionalmente, fue el modelo con menor inercia, evidenciando clústeres más compactos. Estos resultados confirman que

la aplicación de PCA permitió mitigar los efectos de la maldición de la dimensionalidad, optimizando el desempeño del algoritmo. Aunque se exploraron ajustes como la inicialización de centroides con k-means++, no se observaron mejoras significativas, por lo que no se incorporaron en la configuración final. Así, este modelo fue seleccionado como el más adecuado para caracterizar empresas del sector energético según sus indicadores ESG. En la Figura 27 se muestra el fragmento de código utilizado para entrenar el modelo K-Means con 40 clústeres sobre los cinco componentes principales obtenidos por PCA, validando que esta configuración permitió explicar el 81 % de la varianza total.

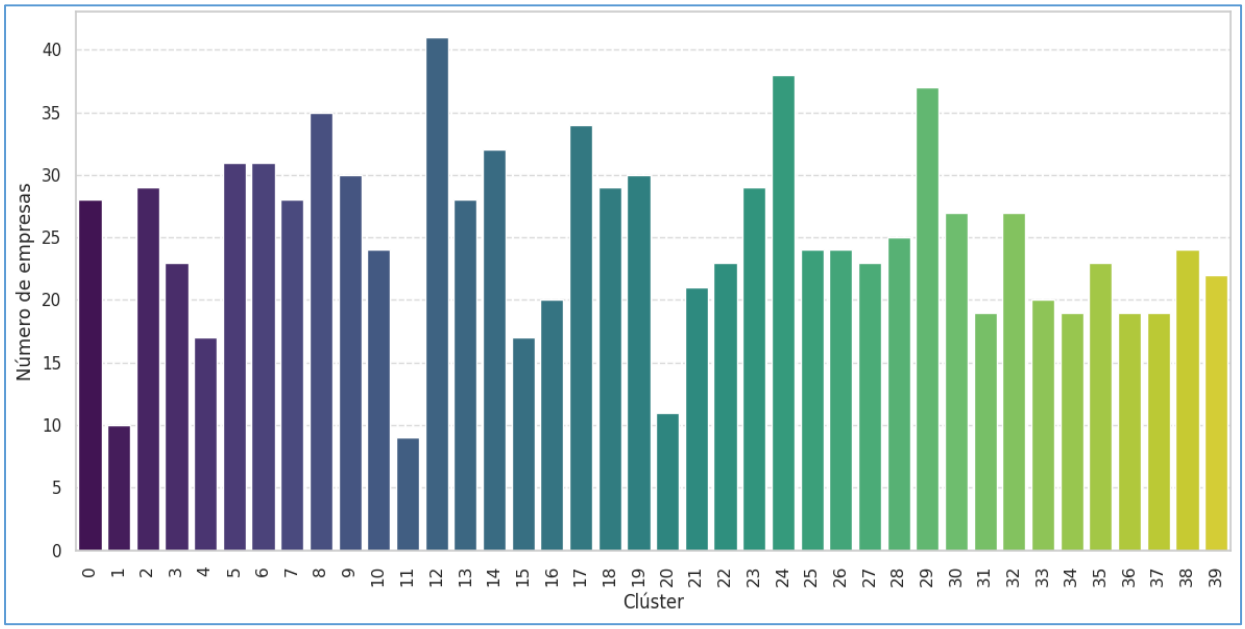
Figura 27 *Implementación del modelo K-Means con K=40 sobre cinco componentes principales.*



Nota. La varianza total alcanzada fue del 81 %, lo cual, aunque aceptable, implica que una parte menor de la información original fue descartada en el proceso de reducción de dimensionalidad.

En la Figura 28 se presenta la cantidad de empresas agrupadas en cada uno de los 40 clústeres generados por el modelo K-Means con PCA=5.

Figura 28 *Cantidad de empresas por clúster (KMeans, PCA=5, K=40).*



Fuente: Elaboración propia con datos de ESG (2024). Nota. Este gráfico de barras representa la distribución de empresas en los 40 clústeres formados mediante el modelo K-Means con reducción de dimensionalidad a cinco componentes principales mediante PCA. Se observa una segmentación relativamente equilibrada, aunque algunos grupos presentan una mayor concentración de observaciones, lo que podría reflejar similitudes estructurales más frecuentes en determinados perfiles ESG dentro del sector energético.

Con el fin de seleccionar el modelo de agrupamiento más adecuado para segmentar a las empresas del sector energético según sus indicadores ESG, se efectuó un análisis comparativo entre los algoritmos K-Means (con y sin reducción de dimensionalidad mediante PCA) y el método de agrupamiento jerárquico aglomerativo. Para ello, se utilizaron tres métricas internas ampliamente reconocidas en la literatura especializada: el coeficiente de silueta, el índice de Calinski-Harabasz y el índice de Davies-Bouldin. Estas métricas permitieron evaluar tanto la cohesión intragrupal como la separación entre clústeres, proporcionando una visión integral de la calidad del agrupamiento.

La Tabla 8 resume los resultados obtenidos.

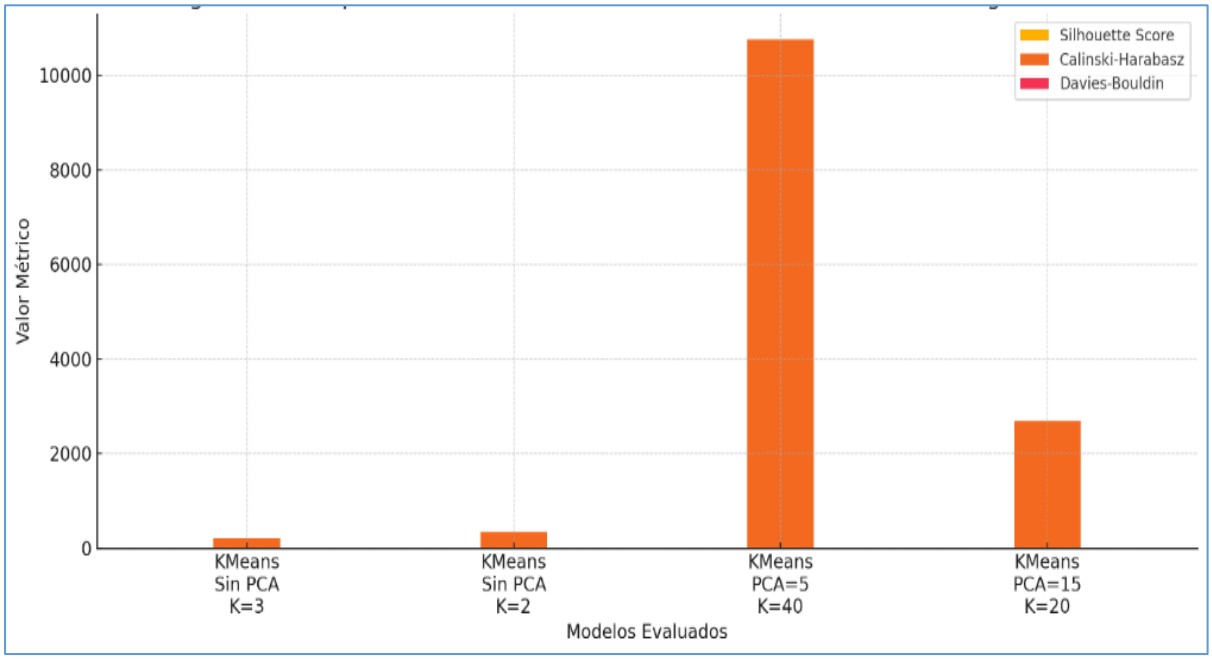
En ella se evidencia que el modelo K-Means con PCA=5 y K=40 alcanzó el mayor puntaje en el índice de Calinski-Harabasz (10.761,82), el menor valor en el índice de Davies-Bouldin (0,7077) y el mayor

coeficiente de silueta (0,5354), lo que indica una segmentación altamente consistente, con clústeres compactos, bien separados y representativos de la diversidad estructural presente en los datos ESG. Esta configuración también se destacó por su baja inercia, lo que reafirma la formación de grupos homogéneos sin sobresegmentación.

Como apoyo visual a esta síntesis, la Figura 29 presenta un gráfico comparativo de barras con las puntuaciones obtenidas por cada modelo en las tres métricas principales. Esta visualización permite apreciar de manera clara las ventajas cuantitativas del modelo óptimo seleccionado.

Más allá de su rendimiento cuantitativo, el modelo K-Means con PCA=5 y 40 clústeres ofrece una herramienta robusta para caracterizar perfiles ESG diferenciados en el sector energético. Esta segmentación permite identificar con precisión grupos extremos tanto en desempeño ambiental positivo como en exposición a controversias o debilidades de gobernanza, así como subgrupos intermedios con comportamientos mixtos. Gracias a la alta resolución del modelo, es posible definir estrategias personalizadas de intervención, monitoreo y fomento de mejores prácticas, alineadas con las prioridades regulatorias o de inversión sostenible.

Figura 29 *Comparación de métricas internas entre modelos de clustering evaluados*



Fuente. Elaboración propia con base en resultados del entrenamiento de modelos (2024). Nota. Se

observa que el modelo K-Means con reducción de dimensionalidad mediante PCA (cinco componentes) y con K=40 supera al resto en todas las métricas de evaluación interna, confirmando su superioridad técnica en términos de cohesión intragrupo, separación entre clústeres y compacidad.

En consecuencia, este modelo no solo optimiza la diferenciación entre empresas desde una perspectiva técnica, sino que también aporta un valor práctico significativo al facilitar la toma de decisiones fundamentadas en evidencia empírica sólida.

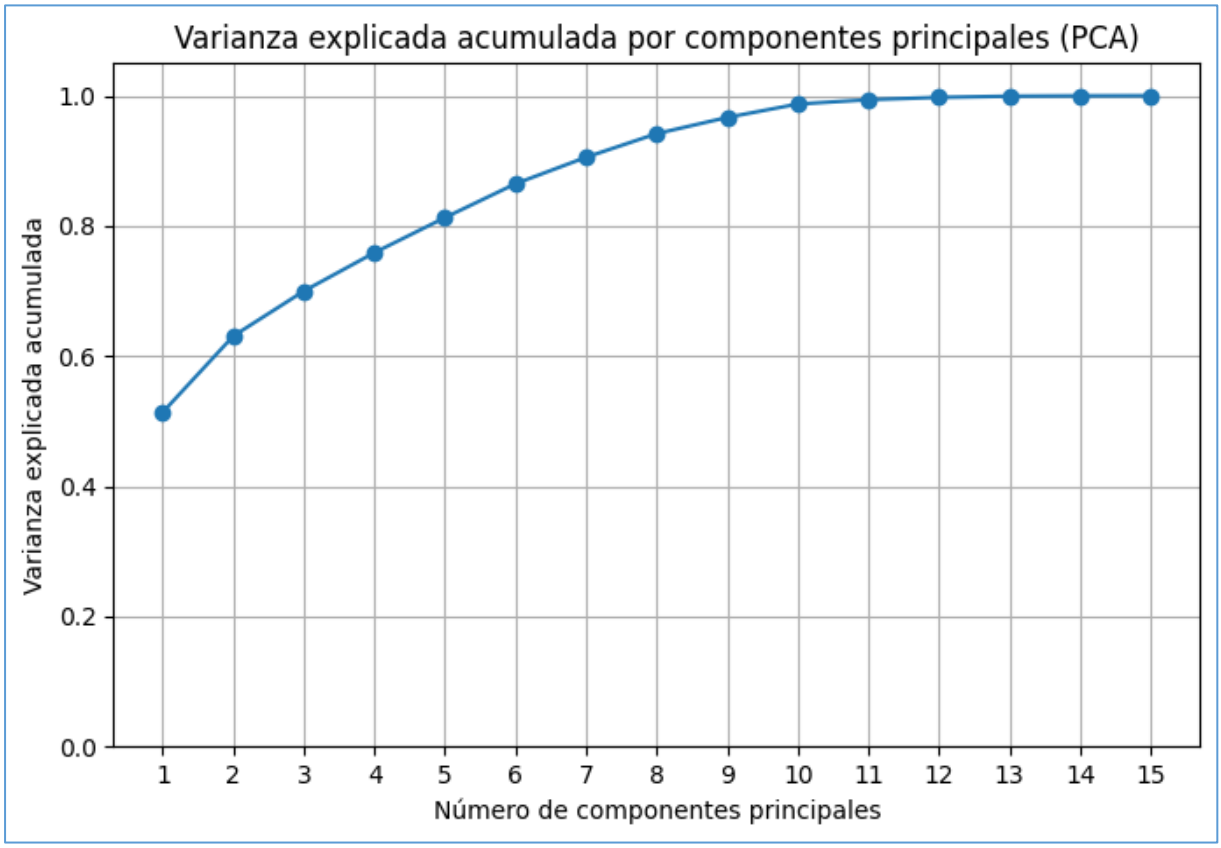
5.3 Aplicación de DBSCAN

En esta etapa, se llevó a cabo un análisis de reducción de la dimensionalidad mediante Componentes Principales (PCA), con el objetivo de determinar la cantidad óptima de variables a considerar en el modelado de agrupamiento basado en densidad (DBSCAN). Este procedimiento resulta esencial, ya que trabajar con un elevado número de dimensiones puede dificultar la definición de densidades y afectar la eficiencia computacional del algoritmo, especialmente en conjuntos de datos con alta dispersión.

Para tal fin, se aplicó el método PCA sobre los datos normalizados, extrayendo los valores y graficando la varianza explicada acumulada por cada componente principal. Los resultados se resumen en la figura 30 correspondiente, donde se observa que la primera componente principal explica aproximadamente el 51% de la varianza total. Al incorporar hasta siete componentes principales, la proporción acumulada de varianza explicada asciende al 90%, mientras que a partir de la décima componente, se alcanza una explicación cercana al 98% de la variabilidad presente en los datos originales.

Figura 30 *Varianza explicada acumulada por componentes principales (PCA) aplicada a los*

indicadores ESG del sector energético.



Fuente: Elaboración propia a partir del análisis de componentes principales sobre la base de datos ESG normalizada (2024).

La interpretación de estos resultados sugiere que es posible reducir la dimensionalidad a un subconjunto de entre 7 y 10 componentes principales, conservando así la mayor parte de la información relevante para el análisis de agrupamiento. Este balance permite, por un lado, evitar la redundancia y el ruido inherente a variables poco informativas y, por otro, mantener la capacidad discriminante de las variables originales. En el contexto del modelo DBSCAN, esta reducción favorece la detección de agrupamientos robustos y mejora el desempeño computacional, dada la menor complejidad geométrica del espacio de representación.

5.3.1 Modelo 1: DBSCAN con datos estandarizados

Paso 1: Se calculó el min_sample= 2* 7 = 14

En la Figura 29 se presenta el valor sugerido de *min_samples*, calculado como el doble del número de componentes principales seleccionados.

Figura 31 *Cálculo de min_samples para DBSCAN con P=7*

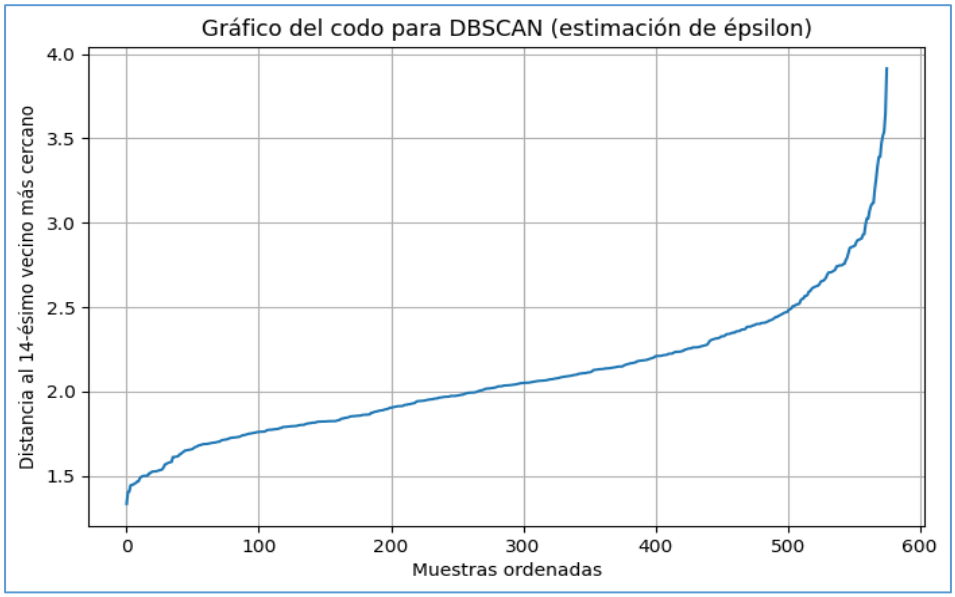
```
# Número de componentes principales seleccionados (P)
P = 7
min_samples = 2 * P
print("El valor de min_samples sugerido es:", min_samples)
```

```
El valor de min_samples sugerido es: 14
```

Fuente. Elaboración propia con base en resultados del entrenamiento de modelos (2024).

Paso 2: Para determinar el valor óptimo del parámetro épsilon, se entrenó un algoritmo de los vecinos más cercanos utilizando los datos ESG estandarizados y reducidos por componentes principales. El cálculo de las distancias al 14-ésimo vecino más cercano para cada observación permitió organizar estos valores en orden ascendente y graficar la curva de distancias, identificando visualmente el punto de inflexión o "codo".

Figura 32 *Gráfico del codo para DBSCAN con datos estandarizados (estimación del valor óptimo de épsilon).*



Nota: El gráfico presenta la distancia al 14-ésimo vecino más cercano para cada observación,

permitiendo identificar visualmente el punto de inflexión (codo), utilizado para seleccionar el valor adecuado de ε en la agrupación por densidad.

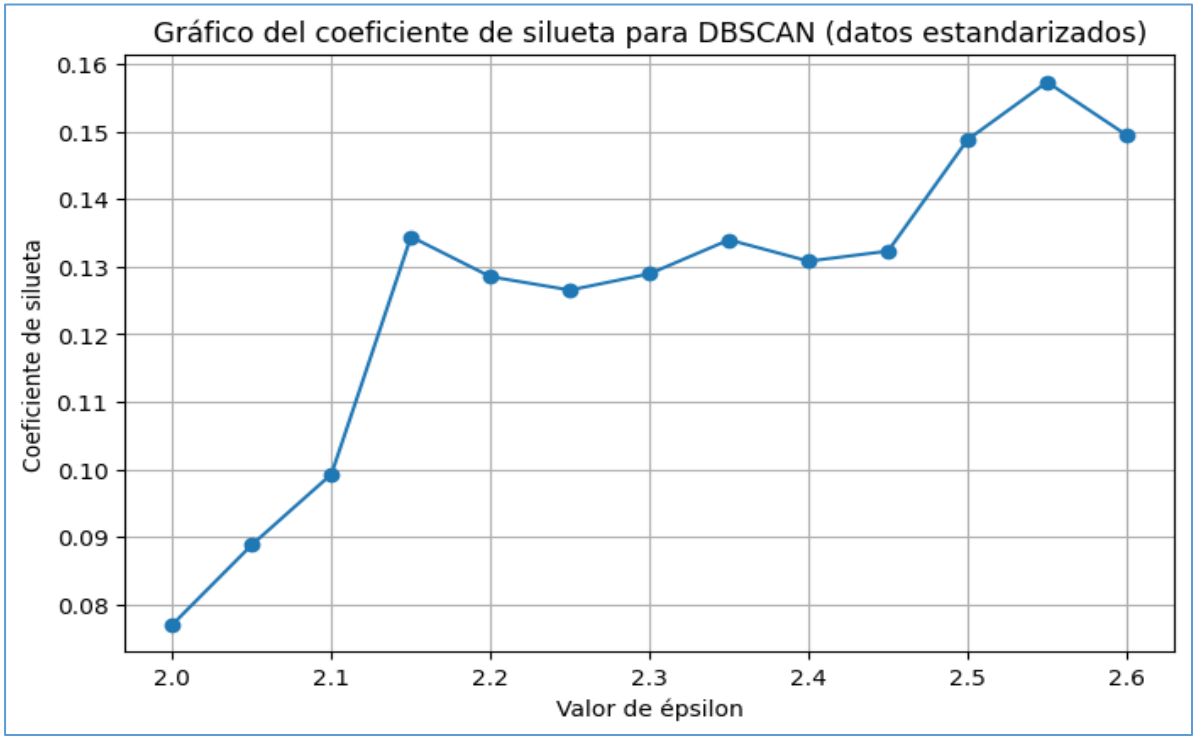
Luego de este primer acercamiento, se identificó que el rango recomendado de épsilon se encuentra entre 2.0 y 2.6. Este intervalo, observado en la región donde la curva comienza a incrementarse de manera pronunciada, permite explorar con mayor detalle la zona crítica para definir el parámetro de agrupamiento.

Paso 3: Se calculó el coeficiente de silueta para diferentes valores del parámetro épsilon, tal como se muestra en la figura 31. El parámetro épsilon (ε) define la distancia máxima entre dos puntos para considerarse vecinos directos en el agrupamiento DBSCAN, y su adecuada selección resulta esencial para la formación de clústeres significativos.

Por su parte, el coeficiente de silueta evalúa la cohesión y separación de los grupos obtenidos, proporcionando una medida de cuán bien se agrupa cada empresa en comparación con las agrupaciones vecinas. Si bien este indicador no se emplea directamente para ajustar el valor óptimo de épsilon en DBSCAN, constituye una referencia útil sobre la calidad general de los clústeres, en especial al comparar diferentes configuraciones del modelo.

En la figura 33 se observa que el coeficiente de silueta presenta una tendencia creciente a medida que se incrementa el valor de épsilon, alcanzando sus valores más altos en torno al rango superior del intervalo analizado, lo que sugiere una mejor cohesión y diferenciación de los clústeres para esos valores.

Figura 33 *Gráfico del coeficiente de silueta para DBSCAN con los datos estandarizados*

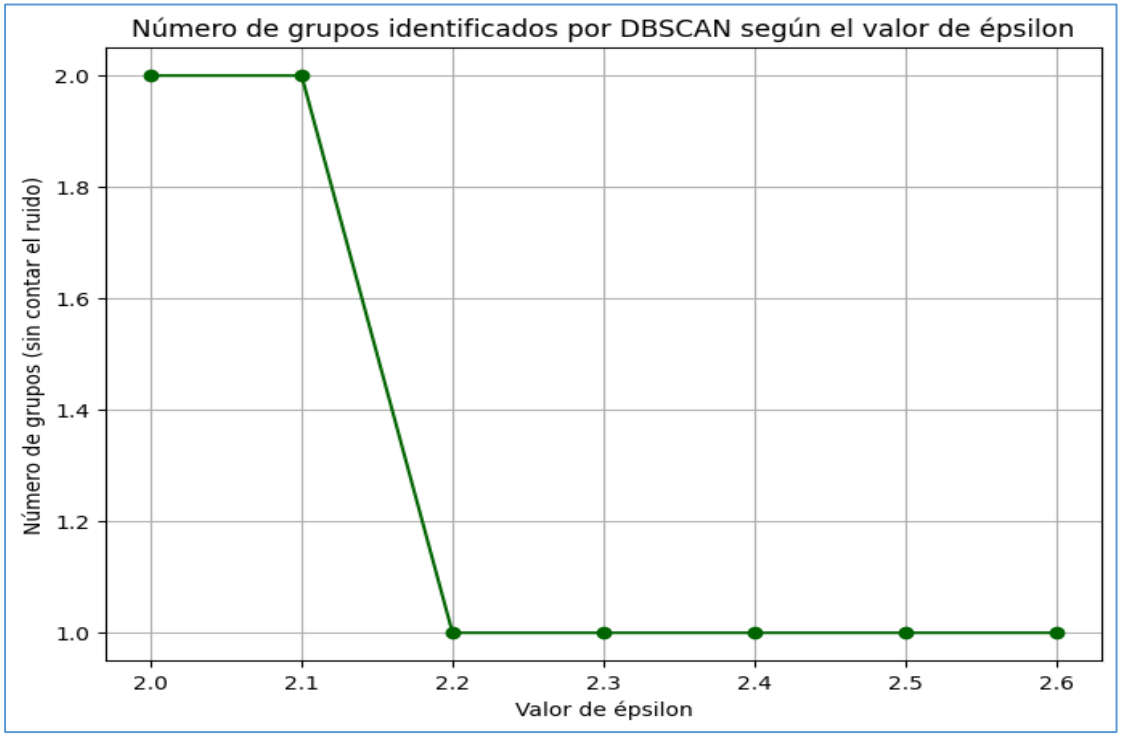


Nota. El gráfico muestra la evolución del coeficiente de silueta para distintos valores del parámetro épsilon, permitiendo identificar visualmente el rango donde el algoritmo DBSCAN logra la mayor calidad de agrupamiento para los indicadores ESG estandarizados.

Paso 4: En este paso, se procedió a calcular el número de grupos identificados por el algoritmo DBSCAN para diferentes valores de épsilon, utilizando los datos estandarizados y proyectados en los componentes principales seleccionados. El objetivo consistió en observar cómo varía la cantidad de clústeres cuando se modifica el parámetro épsilon, lo cual permite identificar un rango adecuado que equilibre la generación de agrupamientos significativos y evite tanto la fragmentación excesiva como la sobresimplificación del conjunto de datos.

En la Figura 34 se presenta el número de clústeres generados por DBSCAN en función del parámetro épsilon.

Figura 34 *Gráfico del número de grupos identificados por DBSCAN según el valor de épsilon*



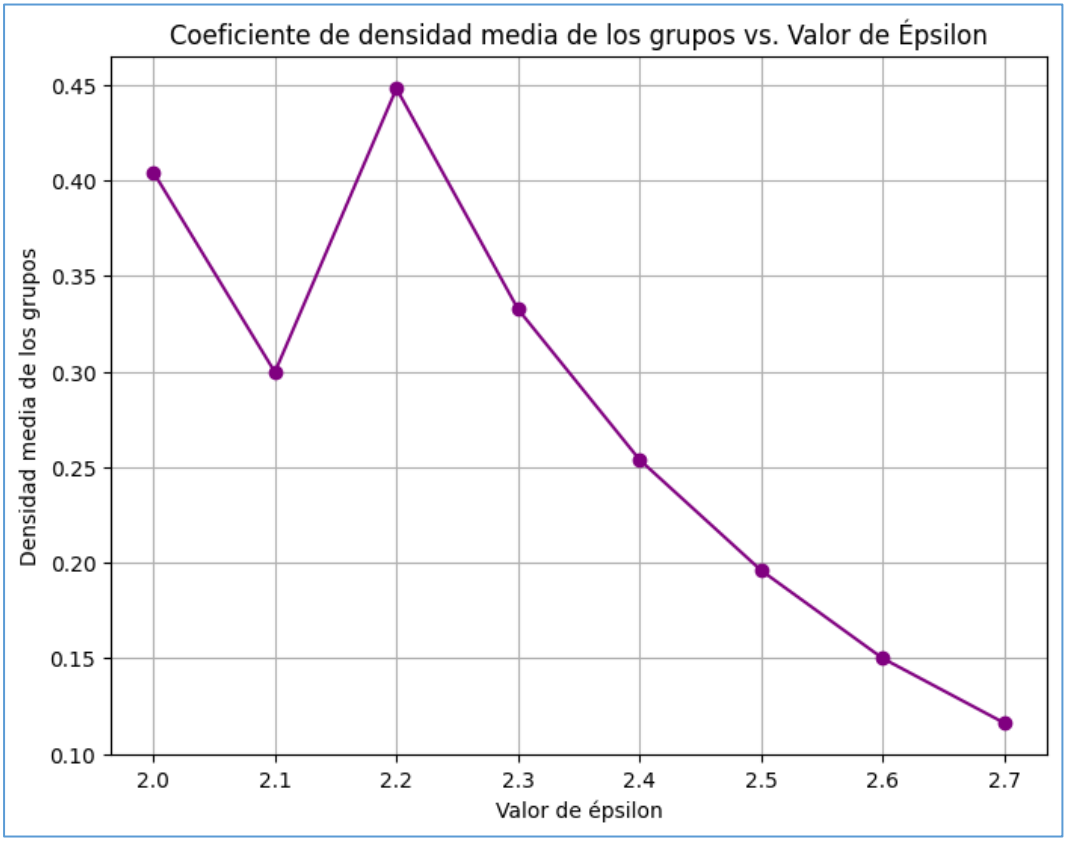
Nota. El gráfico representa cómo la cantidad de agrupamientos generados por DBSCAN cambia en función del valor de épsilon. Se observa que valores bajos permiten distinguir más de un grupo, mientras que a partir de cierto umbral la mayor parte de los puntos se concentran en un solo clúster, lo que indica la importancia de seleccionar cuidadosamente este hiperparámetro para evitar soluciones triviales.

El gráfico del número de grupos frente al valor de épsilon (Figura 34) muestra que, para valores bajos de épsilon (por ejemplo, 2.0 y 2.1), DBSCAN identifica dos grupos diferenciados. Sin embargo, a partir de épsilon = 2.2, el algoritmo comienza a agrupar la mayoría de los puntos en un solo clúster, mientras que los valores restantes de épsilon únicamente detectan un único grupo predominante, con el resto de las observaciones siendo clasificadas como ruido o quedando sin agrupar. Esta tendencia evidencia la sensibilidad del algoritmo a la selección de épsilon, pues un incremento leve en este parámetro puede modificar radicalmente la estructura de los grupos encontrados.

Paso 5: El análisis del coeficiente de densidad media revela que, a medida que el valor de épsilon aumenta, la densidad media de los grupos tiende a disminuir de forma progresiva. En los valores iniciales del intervalo evaluado (ε = 2.0 a 2.2), se observa una densidad relativamente alta, indicando que los grupos formados contienen puntos bastante cercanos entre sí y, por ende, presentan una alta cohesión interna. Sin embargo, conforme el valor de épsilon incrementa, se evidencia una caída constante en la densidad media, lo que sugiere que los grupos tienden a incorporar puntos más distantes, perdiendo así parte de su compacidad estructural.

Este comportamiento es coherente con la naturaleza del algoritmo DBSCAN, ya que un épsilon mayor implica que más puntos serán considerados vecinos, expandiendo los grupos y disminuyendo su densidad relativa. La identificación del valor óptimo de épsilon debe buscar un equilibrio entre una densidad suficientemente alta y un número adecuado de grupos, evitando tanto la fragmentación excesiva como la agrupación artificial de datos disímiles. En consecuencia, los valores de épsilon que maximizan la densidad media en los primeros tramos suelen asociarse a agrupaciones más robustas y representativas del patrón de los datos ESG analizados. En la Figura 35 se presenta la variación de la densidad media de los clústeres generados por DBSCAN en función del valor de épsilon.

Figura 35 *Coeficiente de densidad media de los grupos vs. Valor de Épsilon*



Nota. El gráfico muestra la evolución de la densidad media de los grupos generados por el algoritmo DBSCAN en función de diferentes valores de épsilon. La densidad media se calculó como el promedio del número de puntos por grupo.

Métricas de validación del modelo:

Valores encontrados: epsilon=1.8, min_samples=8, grupos=2

Silhouette Score: 0.2331

Davies-Bouldin Score: 0.9128

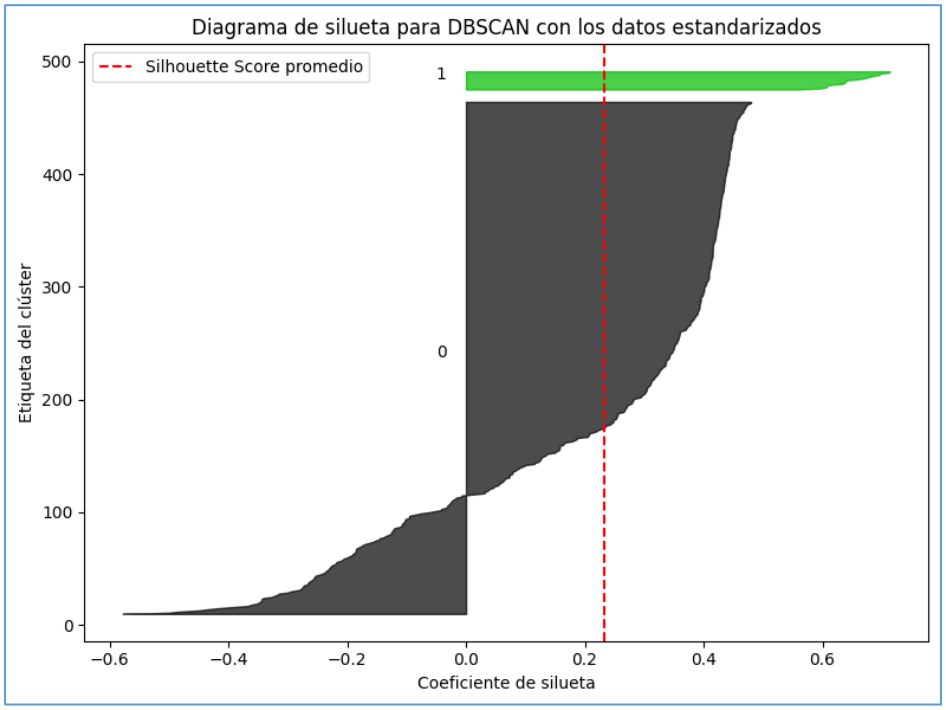
Calinski-Harabasz Score: 38.8138

El diagrama de silueta permite evaluar la calidad de la agrupación generada por DBSCAN en función de la cohesión interna y la separación entre grupos. En este caso, el valor promedio del coeficiente

de silueta se aproxima a 0.23, lo que indica una estructura moderadamente definida entre los dos grupos identificados. Se observa que la mayoría de las observaciones del clúster 1 presentan valores positivos y altos, lo que sugiere una adecuada cohesión interna y buena separación respecto al otro grupo. Por el contrario, en el clúster 0, existen numerosas observaciones con valores de silueta cercanos a cero o incluso negativos, lo que refleja cierta superposición o ambigüedad en la asignación de algunas empresas a este clúster. En conjunto, estos resultados sugieren que el modelo logra distinguir dos agrupaciones principales, aunque la separación entre ellas no es completamente nítida para todos los casos, situación frecuente en datos reales de sostenibilidad donde la heterogeneidad empresarial es considerable.

En la Figura 36 se presenta la evaluación de cohesión y separación entre clústeres con base en el coeficiente de silueta promedio.

Figura 36 *Diagrama de silueta para DBSCAN con los datos estandarizados*



Nota. El diagrama de silueta ilustra la distribución del coeficiente de silueta para cada observación agrupada por el modelo DBSCAN ajustado con ε = 1.8 y min_samples = 8 sobre los datos ESG

estandarizados.

5.3.2 Modelo 2: DBSCAN con datos normalizados

Paso 1: Se calculó el min_sample= 2* 7 = 14

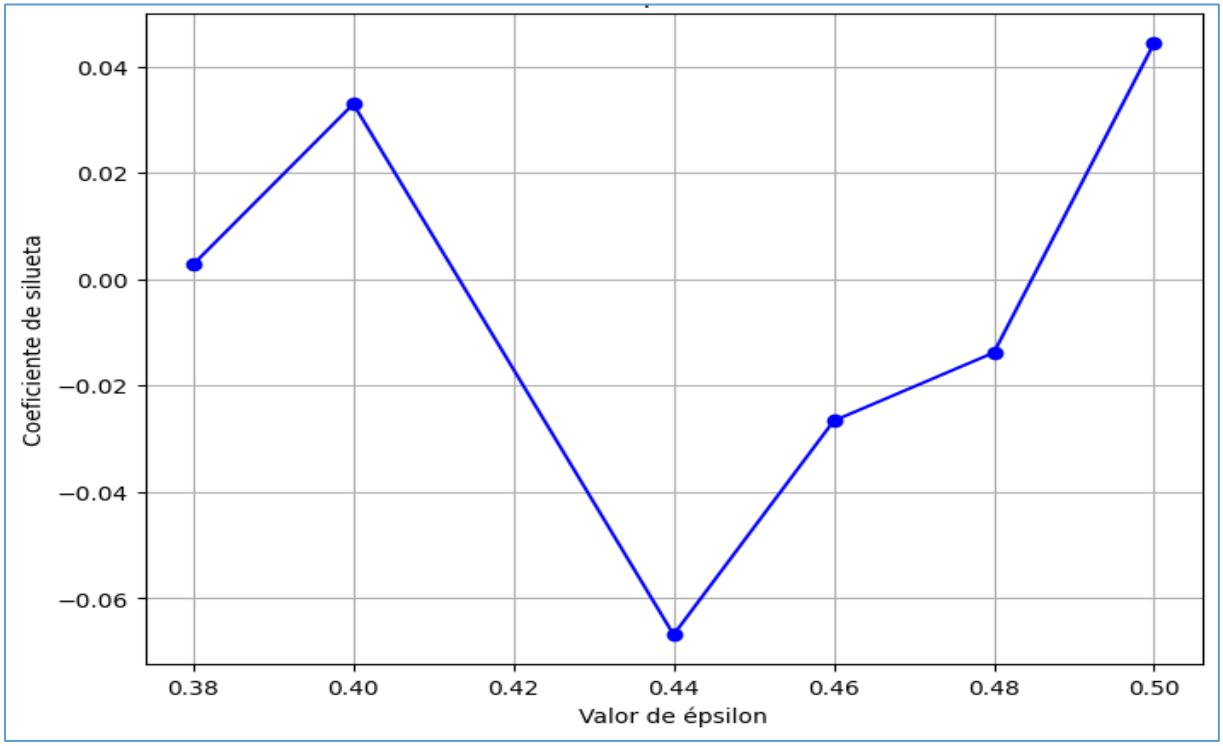
Paso 2: Se identificó un punto de inflexión en la distribución ordenada de distancias al trigésimo vecino más cercano (min_samples = 30), el cual indica el rango adecuado para la selección del parámetro épsilon. El cambio de pendiente observado se ubicó en un intervalo de distancias entre 1.0 y 1.2, lo que sugiere que los valores óptimos de épsilon para el algoritmo DBSCAN deben explorarse dentro de este rango.

Al observar el gráfico del codo para DBSCAN con los datos normalizados (distancia vs. épsilon), se aprecia que la distancia al 14-ésimo vecino más cercano presenta un aumento gradual en la mayor parte del recorrido, pero a partir de aproximadamente $\varepsilon \approx 1.0$ la pendiente de la curva se incrementa de forma notoria. Este punto de inflexión señala el valor de épsilon donde la distancia entre los puntos vecinos comienza a aumentar significativamente. En consecuencia, el rango recomendado para el valor óptimo de épsilon se sitúa en torno a 1.0 a 1.1, siendo este el intervalo adecuado para analizar con mayor detalle y seleccionar el valor definitivo para el modelo DBSCAN.

Paso 3: Calculo del Scoore de silueta en funcioón de épsilon.

En el gráfico del coeficiente de silueta para DBSCAN con datos normalizados se observa que el valor del score de silueta alcanza su máximo cuando épsilon (ε) toma un valor de 0.50, evidenciando una mejor cohesión y separación entre los grupos en ese punto. A medida que el valor de épsilon disminuye, el coeficiente de silueta presenta una mayor variabilidad, alternando entre valores positivos y negativos. Posteriormente, el incremento en el valor de épsilon genera un aumento más gradual en el coeficiente de silueta, indicando una tendencia hacia una mayor estabilidad en la calidad de la agrupación. En la Figura 38 se presenta la variación del coeficiente de silueta según el valor de épsilon en DBSCAN.

Figura 37 *Gráfico del coeficiente de silueta para DBSCAN con datos normalizados*

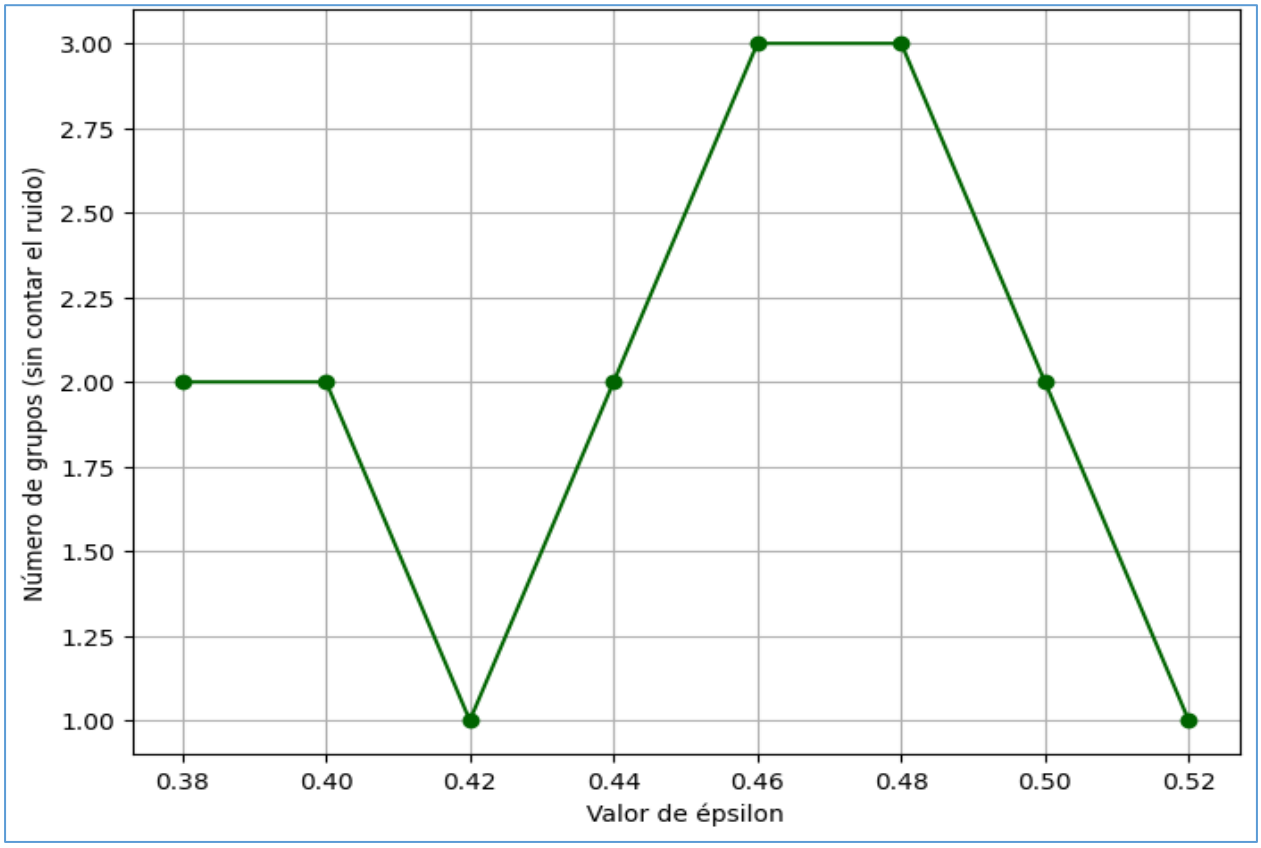


Nota. El gráfico ilustra la variación del coeficiente de silueta en función de diferentes valores de épsilon.

Paso 4: Determinación del número de grupos a partir de épsilon.

En la figura 39, se observa que con valores de épsilon entre 0.38 y 0.50 se forman entre 2 y 3 clústeres, pero cuando el valor de épsilon es mayor a 0.50, el número de clústeres disminuye rápidamente a 1, indicando que los grupos tienden a fusionarse a medida que aumenta épsilon, perdiendo capacidad de segmentación.

Figura 38 *Número de grupos identificados por DBSCAN según el valor de épsilon (datos normalizados)*



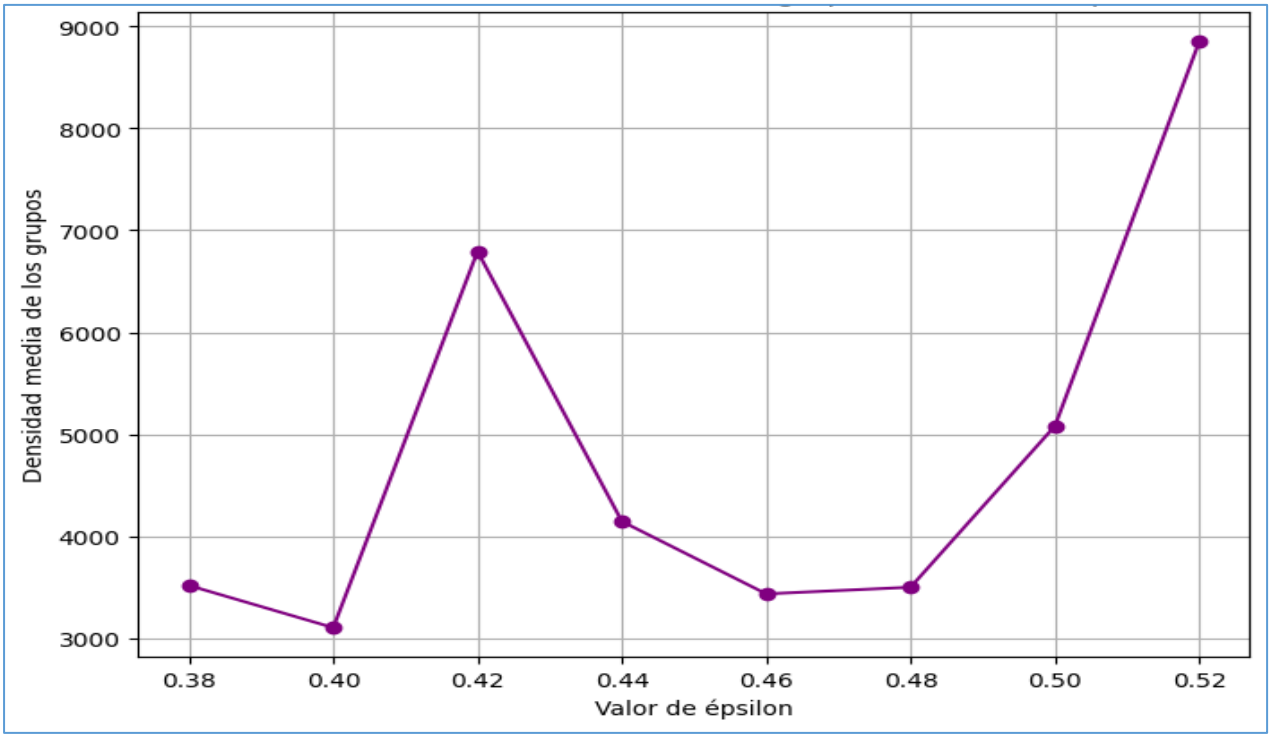
Nota. La gráfica muestra la cantidad de clústeres formados por el algoritmo DBSCAN al variar el valor de épsilon sobre los datos ESG normalizados. Se evidencia que para valores de épsilon entre 0.38 y 0.50 se obtienen dos o tres grupos principales, mientras que valores mayores a 0.50 generan una rápida reducción en el número de clústeres, indicando menor capacidad de segmentación y agrupación menos específica.

Paso 5: Análisis de la densidad media.

En la Figura 40, se observa que cuando el valor de épsilon alcanza aproximadamente 0.52, la densidad media de los grupos identificados por el algoritmo DBSCAN registra su valor máximo. Para este valor de épsilon se forman dos grupos principales, sin considerar el clúster etiquetado como ruido, lo cual se encuentra alineado con la estructura de agrupamiento buscada en la solución del

problema. Asimismo, dentro del intervalo analizado, este valor de épsilon coincide con el mayor score de silueta obtenido, lo que respalda la selección de este parámetro para el entrenamiento definitivo del modelo sobre los datos normalizados.

Figura 39 *Coeficiente de densidad media de los grupos vs. Valor de Épsilon*



Nota. La figura ilustra la variación de la densidad media de los grupos generados por DBSCAN en función de diferentes valores de épsilon, aplicados sobre los datos normalizados y reducidos mediante componentes principales. El pico máximo identifica el valor óptimo de épsilon para la cohesión interna de los grupos, excluyendo el ruido, en el contexto del análisis ESG.

Número de grupos (sin incluir ruido): 1

Etiqueta de clúster:-1, Cantidad de muestras: 146

Etiqueta de clúster: 0, Cantidad de muestras: 430

Métricas de validación del modelo

Valores encontrados: eps=0.30, min_samples=5, grupos=4

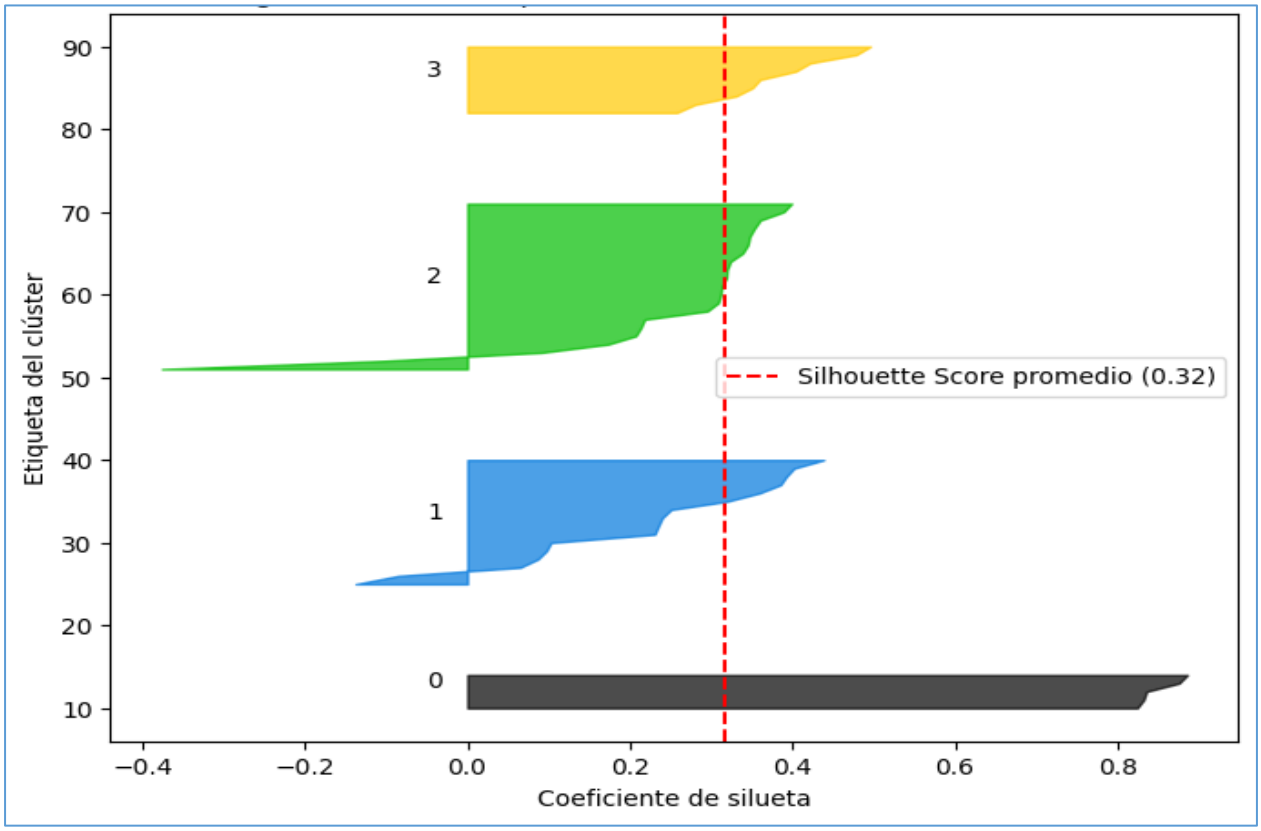
Silhouette Score: 0.3170

Davies-Bouldin Score: 0.9867

Calinski-Harabasz Score: 70.9829

En la Figura 41 se presenta el diagrama de silueta para DBSCAN con datos normalizados bajo la configuración óptima.

Figura 40 *Silueta para DBSCAN con los datos normalizados (ε = 0.30, min_samples = 5)*



Nota. El diagrama muestra la distribución del coeficiente de silueta para cada observación agrupada por DBSCAN utilizando datos normalizados y configuración óptima de hiperparámetros.

En el gráfico de silueta se observa que, aunque existen clústeres con cierta dispersión interna, la distribución es más equilibrada respecto a los modelos previos. Los grupos formados presentan una mejor separación y mayor proporción de coeficientes de silueta positivos, lo cual sugiere una mayor cohesión y menor solapamiento entre clústeres. Sin embargo, se identifican varias observaciones con valores negativos, lo que evidencia la presencia de datos atípicos o asignaciones

ambiguas propias del enfoque de DBSCAN.

5.3.3 Modelo 3: DBSCAN con PCA=10- datos estandarizados

Paso 1: Reducción de dimensionalidad mediante PCA y cálculo de min_samples

Para optimizar el proceso de agrupamiento y reducir la complejidad de los datos, se aplicó un análisis de componentes principales (PCA) sobre la matriz de datos previamente estandarizados.

Número de componentes principales (PCA): 10
min_samples: 2 × 10 = 20

Paso 2: Una vez aplicada la reducción de dimensionalidad mediante PCA sobre los datos estandarizados, se entrenó un modelo de vecinos más cercanos con base en los 10 componentes principales extraídos previamente. El objetivo fue estimar un valor adecuado de épsilon para la configuración del algoritmo DBSCAN, tomando como referencia el vigésimo vecino más próximo conforme al valor definido de *min_samples*.

En esta etapa, se identificó que la distancia entre observaciones comienza a incrementarse de forma pronunciada alrededor del valor de $\varepsilon \approx 3.5$. Por tanto, se recomendó explorar valores de épsilon dentro del rango de 3.0 a 4.2 para afinar la configuración del modelo en los pasos posteriores.

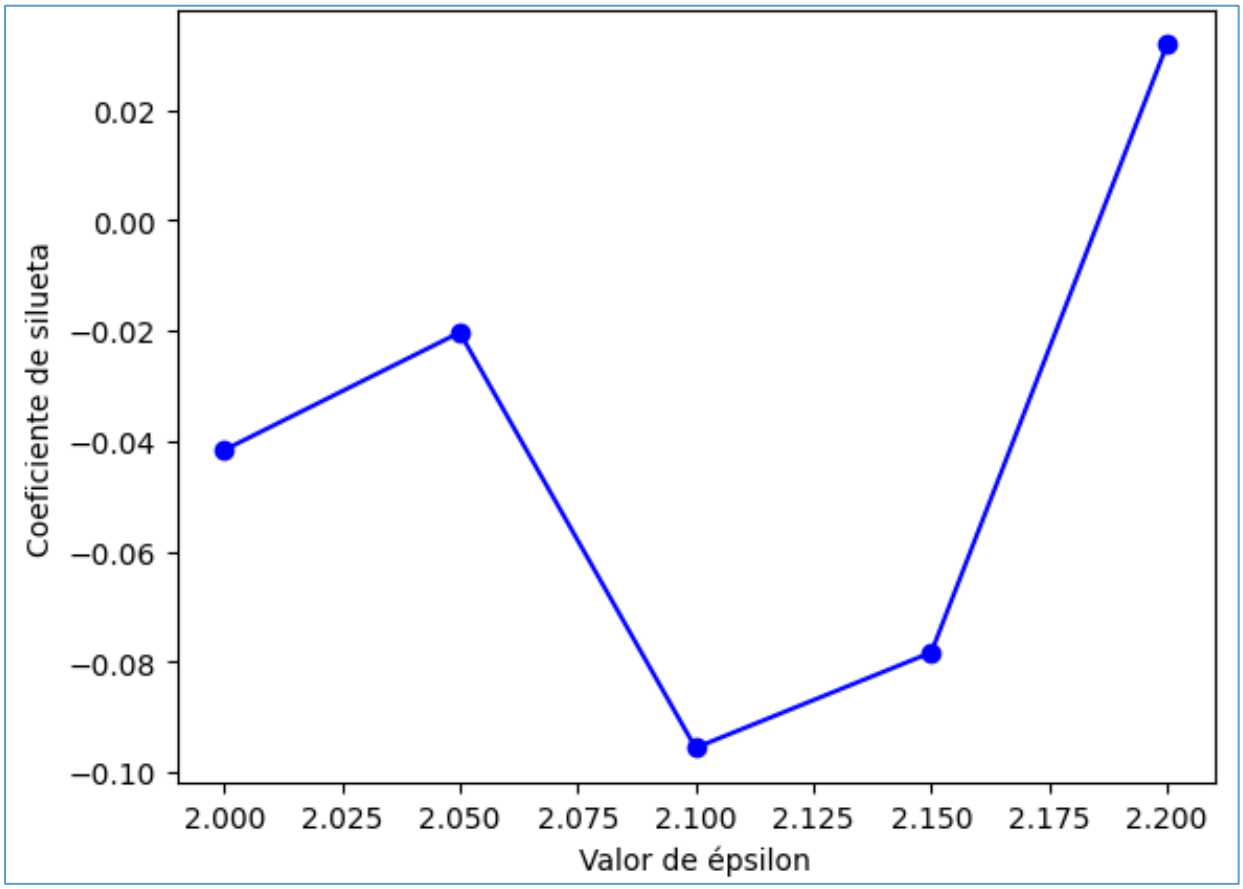
Paso 3: Calculo del score de silueta en función de épsilon.

En la Figura 43 se presenta la evolución del coeficiente de silueta para distintos valores de épsilon en el modelo DBSCAN, aplicado sobre los datos estandarizados reducidos a 10 componentes mediante PCA.

Se observa que el score de silueta tiende a incrementarse de manera significativa a partir de valores de épsilon superiores a 2.15, alcanzando su punto máximo en $\varepsilon = 2.20$. Este comportamiento sugiere que el modelo logra una mejor cohesión interna y separación entre clústeres cuando se

utiliza ese valor, lo que refleja una mayor calidad en la estructura de agrupamiento obtenida bajo dicha configuración.

Figura 41 *Coeficiente de silueta para DBSCAN con PCA=10 (datos estandarizados)*



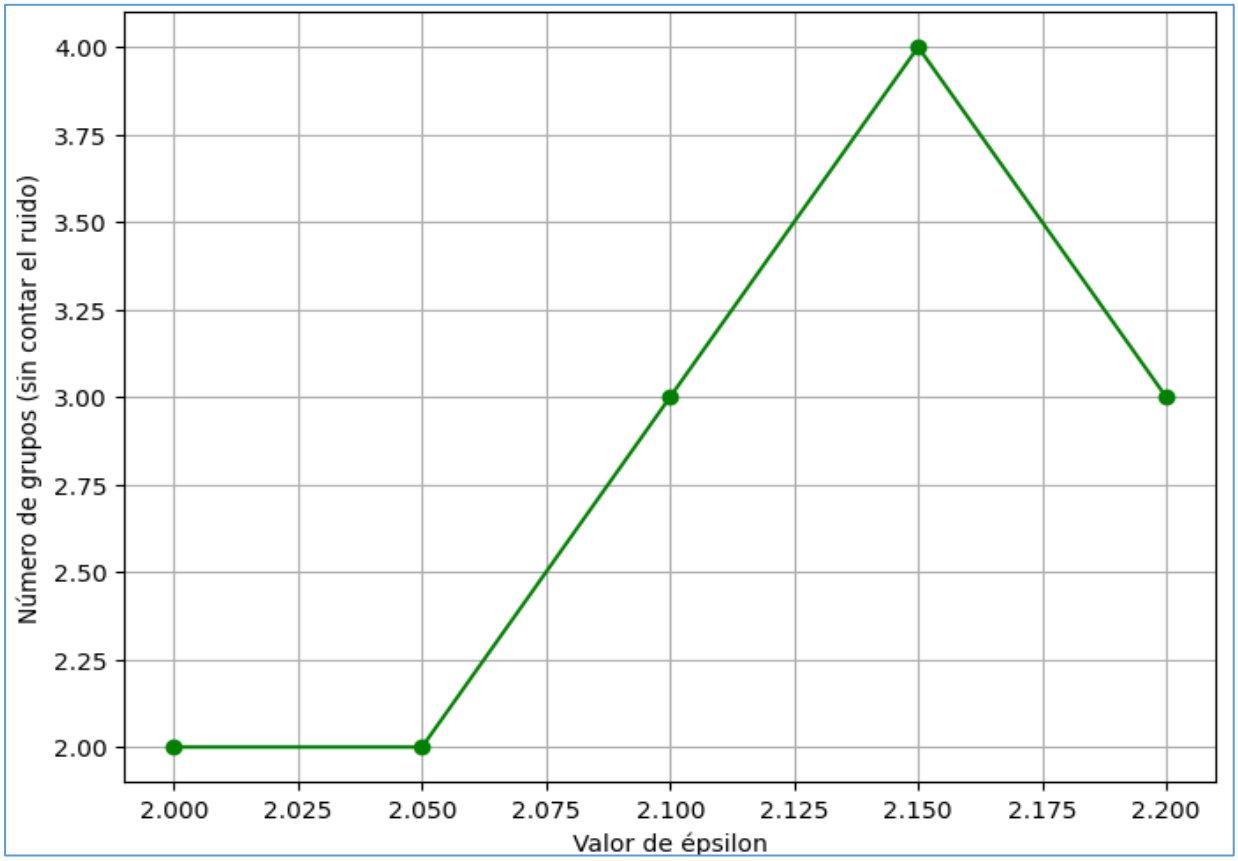
Nota. El gráfico muestra la evolución del coeficiente de silueta para diferentes valores de épsilon en el modelo DBSCAN aplicado sobre los datos estandarizados, previamente reducidos a 10 componentes principales mediante PCA.

Paso 4: Determinar el número de grupos a partir de épsilion

En la Figura 44, se observa que, para valores de épsilon entre 2.0 y 2.05, el modelo identifica 2 clústeres, aumentando a 3 grupos en épsilon = 2.10 y alcanzando un máximo de 4 grupos en épsilon = 2.15. A partir de épsilon = 2.20, el número de grupos disminuye nuevamente a 3. Este comportamiento sugiere que el rango de épsilon comprendido entre 2.10 y 2.20 resulta óptimo para obtener una mayor segmentación en la estructura de los datos, permitiendo distinguir hasta

cuatro agrupaciones diferenciadas antes de que el número de clústeres vuelva a descender.

Figura 42 *Número de grupos identificados por DBSCAN según el valor de épsilon (PCA=10, estandarizados)*



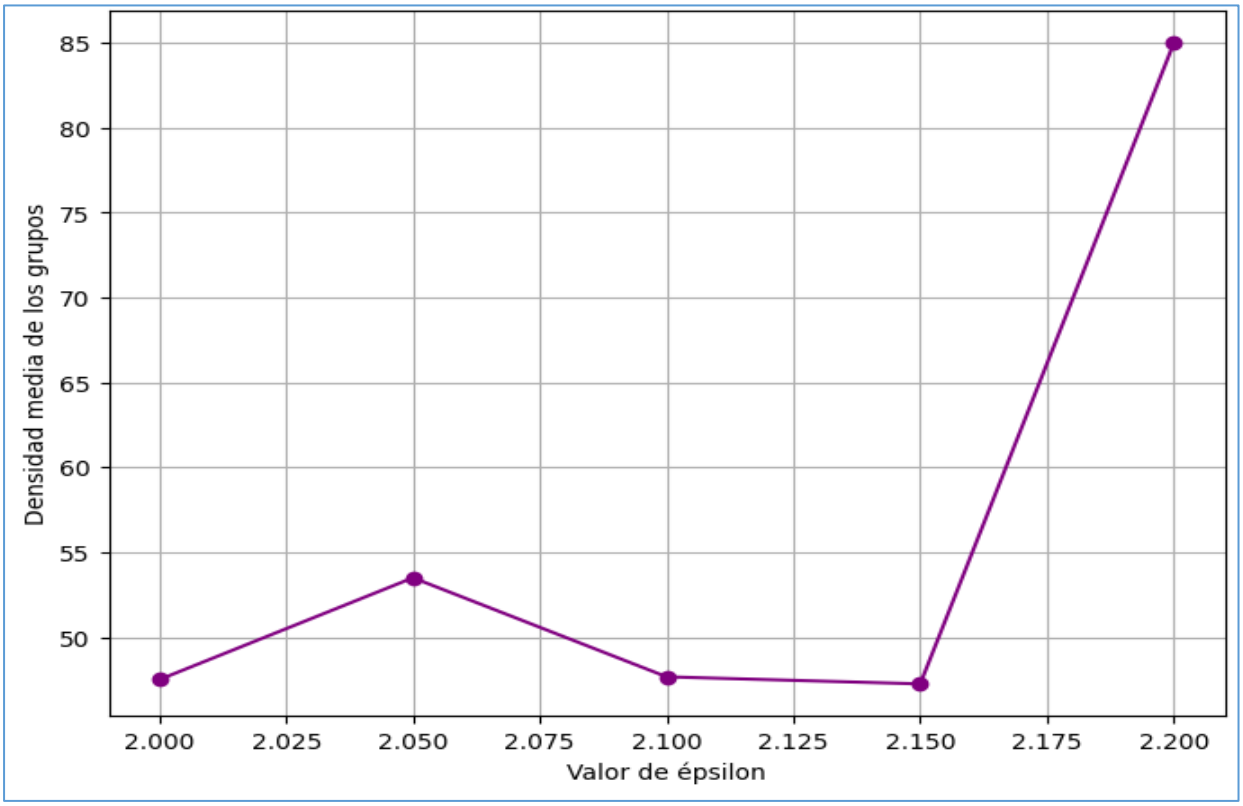
Nota. La figura muestra la evolución en el número de clústeres (sin contar el ruido) generados por el modelo DBSCAN a partir de diferentes valores de épsilon, aplicados sobre los datos estandarizados reducidos a 10 componentes principales mediante PCA.

Paso 5: Calculo de la densidad media

Al analizar la figura 45, se evidencia que la densidad media alcanza su valor máximo cuando épsilon es igual a 2.2, momento en el cual la densidad media de los grupos supera los 85 puntos por clúster. Para este valor de épsilon, se observa la formación de 3 clústeres, excluyendo las observaciones clasificadas como ruido. Además, el análisis previo del coeficiente de silueta indica que este valor

de épsilon también corresponde a uno de los scores de silueta más altos del rango evaluado, lo que sugiere una estructura de agrupamiento más cohesionada y representativa dentro del modelo aplicado sobre los datos estandarizados con PCA.

Figura 43 *Coeficiente de densidad media de los grupos vs. Valor de Épsilon (PCA=10, estandarizados)*



Nota. La gráfica muestra la evolución de la densidad media de los clústeres generados por DBSCAN para diferentes valores de épsilon tras la reducción de dimensionalidad a 10 componentes principales.

Métricas de validación del modelo

Valores encontrados: eps=2.2, min_samples=20, grupos=3

Silhouette Score: 0.318

Davies-Bouldin Score: 1.893

Calinski-Harabasz Score: 82.9516

Clústeres formados (sin contar ruido): 3

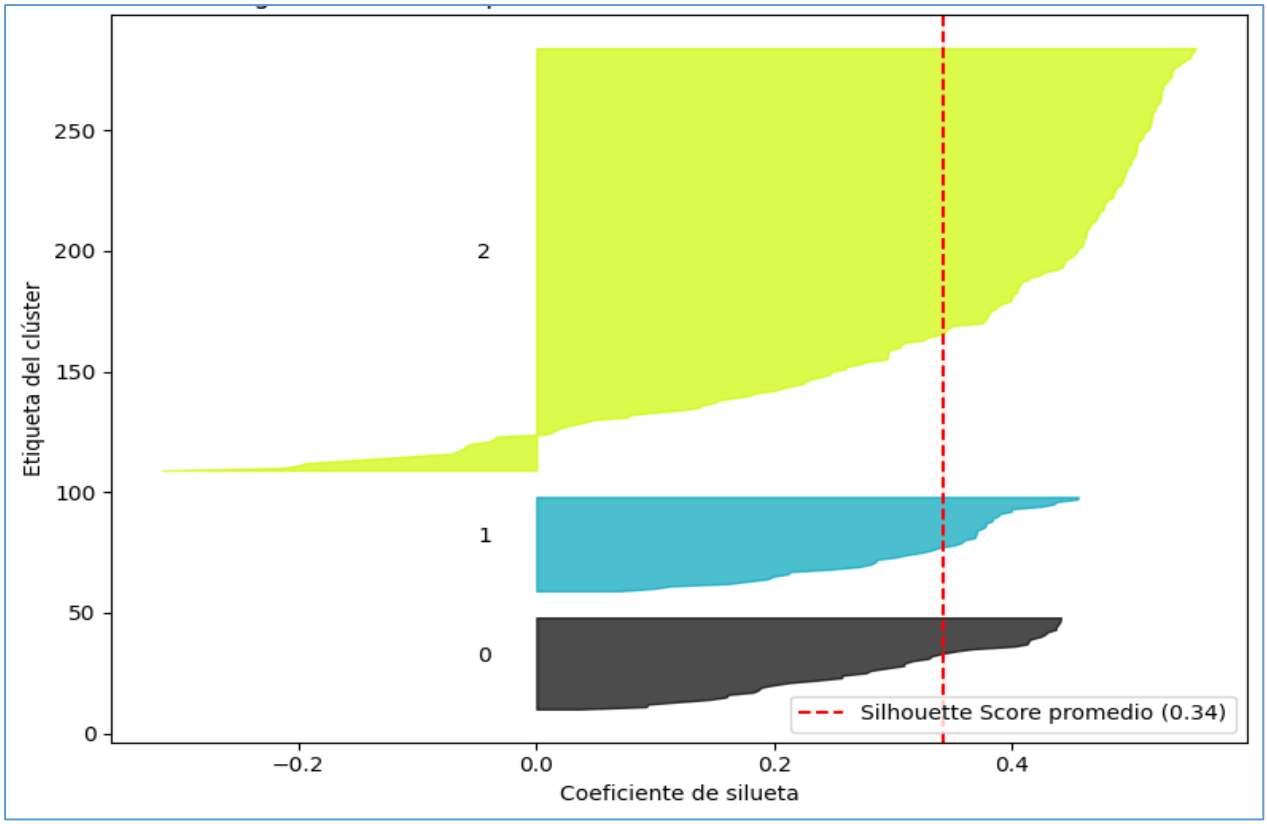
Al aplicar el algoritmo DBSCAN con reducción de dimensionalidad mediante PCA (10 componentes), sobre los datos estandarizados, se obtuvieron tres clústeres bien definidos. Estas diez componentes principales explican aproximadamente el 95 % de la variabilidad total de los datos originales, concentrando la mayor parte de la información relevante para el agrupamiento. El valor promedio del coeficiente de silueta alcanza 0.34, lo que refleja una estructura de clústeres moderadamente cohesiva y razonablemente bien separada, superando la calidad observada en modelos previos sin reducción de dimensionalidad. En la Figura 46 se presenta el diagrama de silueta correspondiente a la aplicación del algoritmo DBSCAN sobre datos estandarizados, reducidos a 10 componentes principales mediante PCA.

El modelo identificó tres clústeres bien definidos, y el valor promedio del coeficiente de silueta alcanzó 0.34, lo que sugiere una estructura de agrupamiento razonablemente coherente. Este resultado mejora la calidad de segmentación respecto a configuraciones previas sin reducción de dimensionalidad, favoreciendo una interpretación más robusta de las agrupaciones ESG.

En la Figura 46 se presenta el gráfico del método del codo correspondiente al modelo DBSCAN aplicado sobre datos normalizados, tras la reducción de dimensionalidad mediante PCA a cinco componentes principales.

La distancia al décimo vecino más cercano muestra un incremento abrupto a partir de $\varepsilon \approx 0.7$, lo que permite establecer el intervalo de exploración óptima entre 0.6 y 0.9 para definir el valor adecuado de épsilon en el proceso de agrupamiento.

Figura 44 *Diagrama de silueta para DBSCAN con PCA=10 sobre datos estandarizados*



Nota. El gráfico ilustra la distribución de los coeficientes de silueta para cada clúster identificado por DBSCAN.

5.3.4 Modelo 4: DBSCAN con PCA=5- Datos normalizados

Se aplicó MinMaxScaler (u otro método de preferencia) a las columnas numéricas seleccionadas del conjunto de datos ESG, asegurando que todos los valores estuvieran en el rango [0, 1].

Paso 1: Se definió el valor de *min_samples* como el doble del número de dimensiones retenidas tras la reducción mediante PCA, es decir, *min_samples* = 2 × 5 = 10.

Paso 2: Se procedió con el entrenamiento del modelo de los vecinos más cercanos utilizando los datos previamente normalizados y transformados a cinco componentes principales. Con base en este procedimiento, se identificó el comportamiento característico del método del codo aplicado a la estimación del parámetro *épsilon* en el algoritmo DBSCAN.

Durante el análisis, se observó que la distancia al décimo vecino más cercano se mantiene relativamente constante hasta alcanzar un punto crítico cercano a ε = 0.7, a partir del cual comienza a aumentar de forma pronunciada. Este cambio de pendiente sugiere la existencia de una frontera natural entre regiones de alta y baja densidad, lo cual resulta fundamental para una segmentación adecuada.

Dado este hallazgo, se recomienda explorar el rendimiento del modelo DBSCAN dentro del rango de valores de *épsilon* comprendido entre 0.6 y 0.9, ya que en este intervalo se concentra la mayor probabilidad de identificar un valor óptimo que maximice la diferenciación y cohesión entre los clústeres formados.

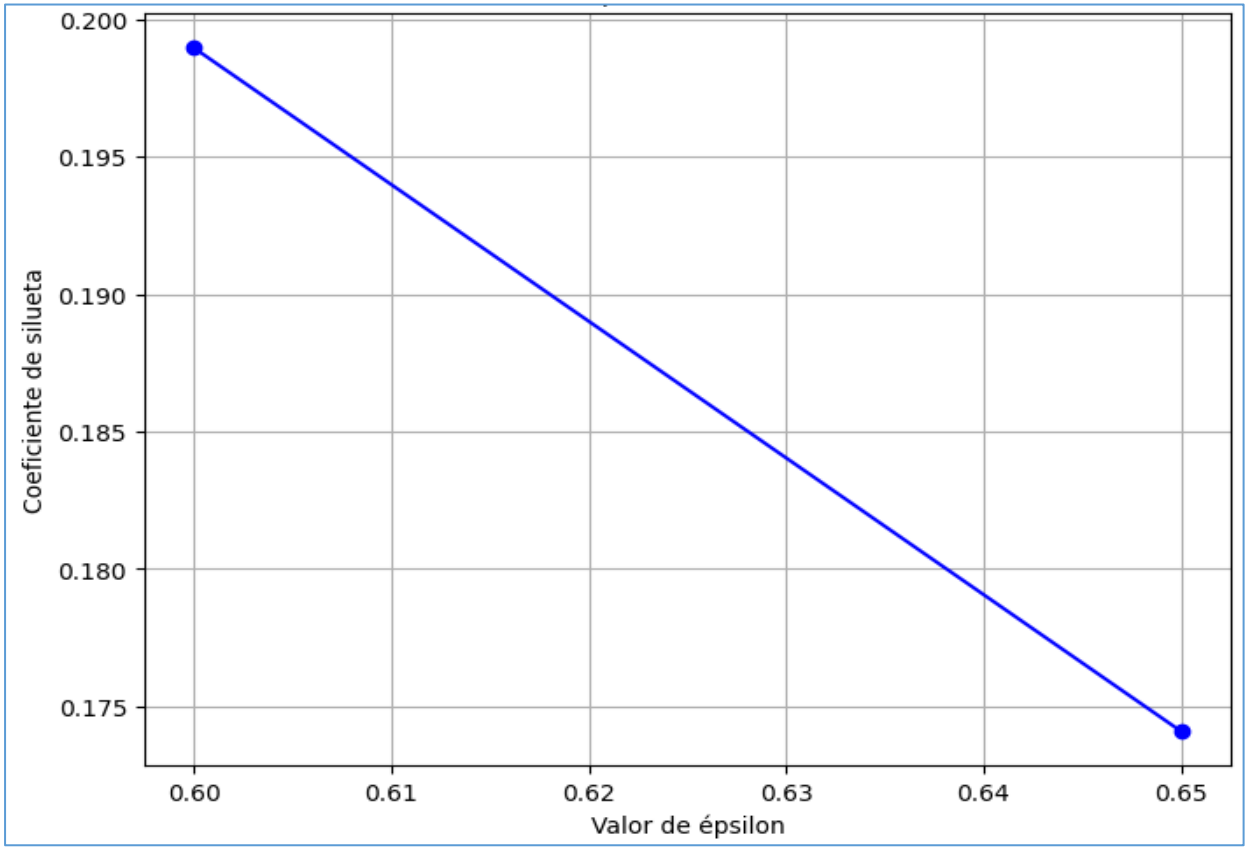
Paso 3: Calculo del score de silueta en función de épsilon.

En la figura 46 se observa que los valores de épsilon analizados (0.60 y 0.65) presentan scores de silueta notablemente bajos, situándose mucho más cerca de cero que de uno. Esto indica que la cohesión interna y la separación entre los clústeres obtenidos es limitada, es decir, los grupos no se encuentran bien definidos y existe una considerable superposición entre ellos. Tal comportamiento es consistente con un agrupamiento débil, en el que los puntos de los diferentes clústeres no están claramente diferenciados. En consecuencia, aunque se identifica cierta estructura, los resultados sugieren que la configuración actual de épsilon y el número de componentes de PCA no logra una segmentación óptima de los datos ESG.

En la Figura 48 se presenta la variación del coeficiente de silueta en función del valor de épsilon para el modelo DBSCAN aplicado sobre datos normalizados y reducidos a cinco componentes principales mediante PCA.

Este gráfico permite visualizar cómo cambia la calidad del agrupamiento conforme se ajusta el parámetro ε, evidenciando que el coeficiente de silueta decrece al aumentar el valor de épsilon dentro del intervalo analizado. Este comportamiento indica que el modelo presenta una mejor cohesión y separación entre clústeres cuando ε se aproxima a 0.60, valor en el que se alcanza el máximo coeficiente de silueta observado.

Figura 45 *Gráfico del coeficiente de silueta para DBSCAN con PCA=5 datos normalizados*



Nota. El gráfico presenta el comportamiento del coeficiente de silueta promedio para los clústeres generados por DBSCAN sobre los datos normalizados y reducidos a cinco componentes principales mediante PCA, en función del valor de épsilon.

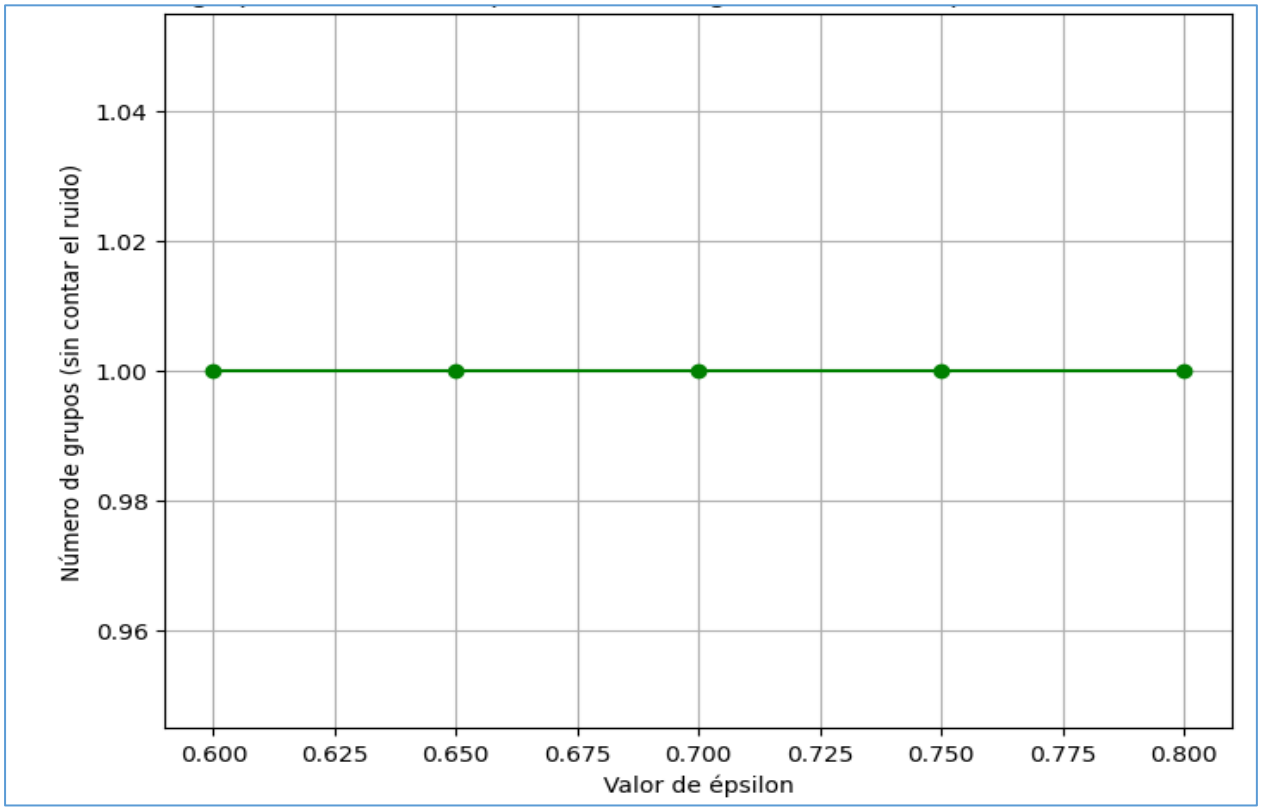
Paso 4: Determinación en número de grupo a partir de épsilon.

En el intervalo de búsqueda analizado (ε de 0.60 a 0.80), el algoritmo DBSCAN únicamente detectó un clúster principal, lo que indica una baja capacidad de segmentación sobre los datos normalizados y reducidos a cinco componentes principales. Esta situación sugiere que, para los valores de épsilon evaluados, la estructura de los datos no presenta densidades diferenciadas suficientes para formar múltiples grupos. Por lo tanto, el modelo no logra identificar subgrupos relevantes y no resulta adecuado para extraer patrones de agrupamiento significativos bajo esta configuración.

En la Figura 49 se presenta la cantidad de clústeres identificados por el algoritmo DBSCAN para distintos valores de *épsilon*, considerando datos normalizados y reducidos a cinco componentes principales mediante PCA.

Durante este análisis, correspondiente al Paso 4, se evidenció que, dentro del intervalo evaluado (ε entre 0.60 y 0.80), el modelo únicamente detectó un clúster principal, sin distinguir subgrupos adicionales. Esta limitación refleja una escasa capacidad de segmentación bajo dicha configuración, lo cual compromete la utilidad del modelo para extraer patrones significativos de agrupamiento en este conjunto de datos.

Figura 46 *Número de grupos identificados por DBSCAN según el valor de épsilon (PCA=5, normalizados)*



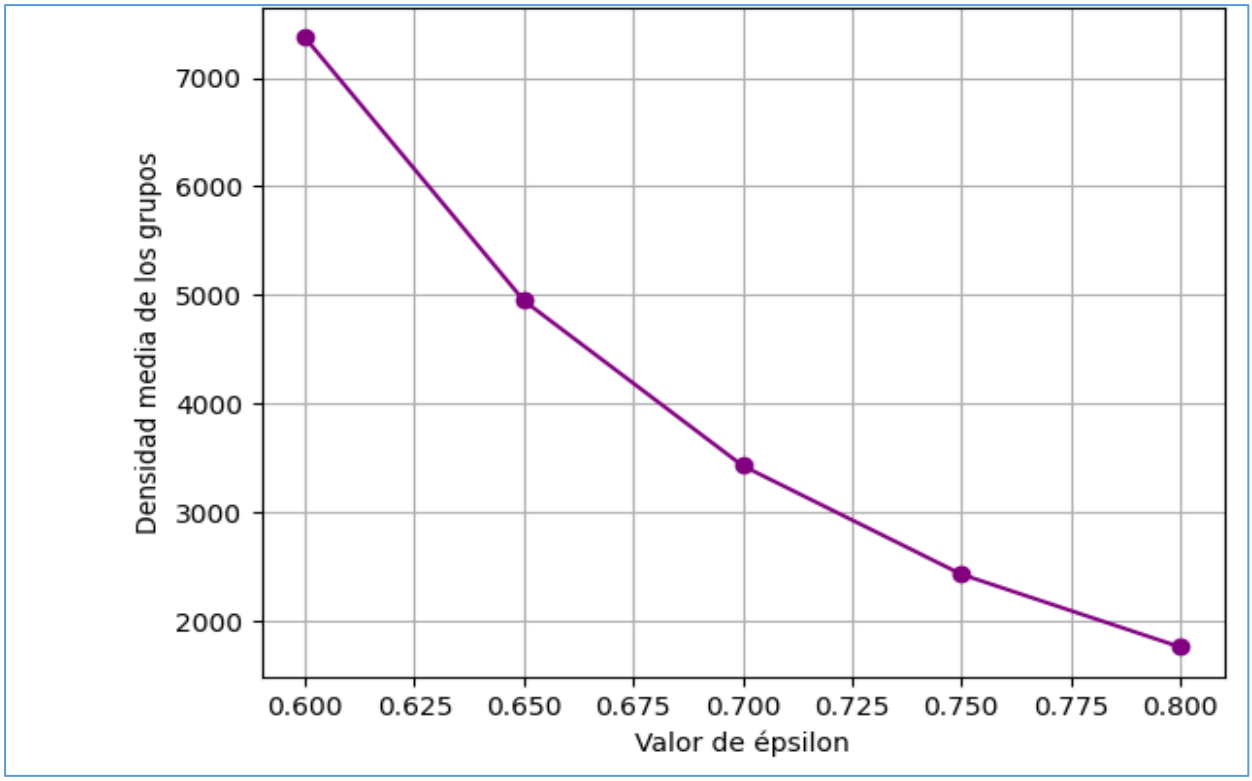
Nota. El gráfico muestra la cantidad de clústeres formados para distintos valores de épsilon, excluyendo el ruido.

Paso 5: Análisis del cálculo de la densidad media

Al observar la figura 50, se aprecia que la densidad media alcanza su valor máximo cuando épsilon

es igual a 0.60. En este punto, el modelo forma el mayor número de clústeres con mayor compacidad interna. Conforme aumenta el valor de épsilon, la densidad media disminuye progresivamente, lo que indica una menor cohesión entre los puntos de cada grupo. Por tanto, se selecciona el valor de épsilon = 0.60 para entrenar el modelo DBSCAN, ya que proporciona la máxima densidad media y una mejor separación de los clústeres identificados sobre la estructura de los datos normalizados.

Figura 47 *Coeficiente de densidad media de los grupos vs. Valor de Épsilon (PCA=5, normalizados)*



Nota. La gráfica muestra la evolución de la densidad media de los clústeres generados por DBSCAN al variar el valor de épsilon sobre los datos normalizados y reducidos a cinco componentes principales mediante PCA.

Métricas de validación del modelo

Valores encontrados: eps=0.30, min_samples=13, grupos=2

Silhouette Score: 0.7514

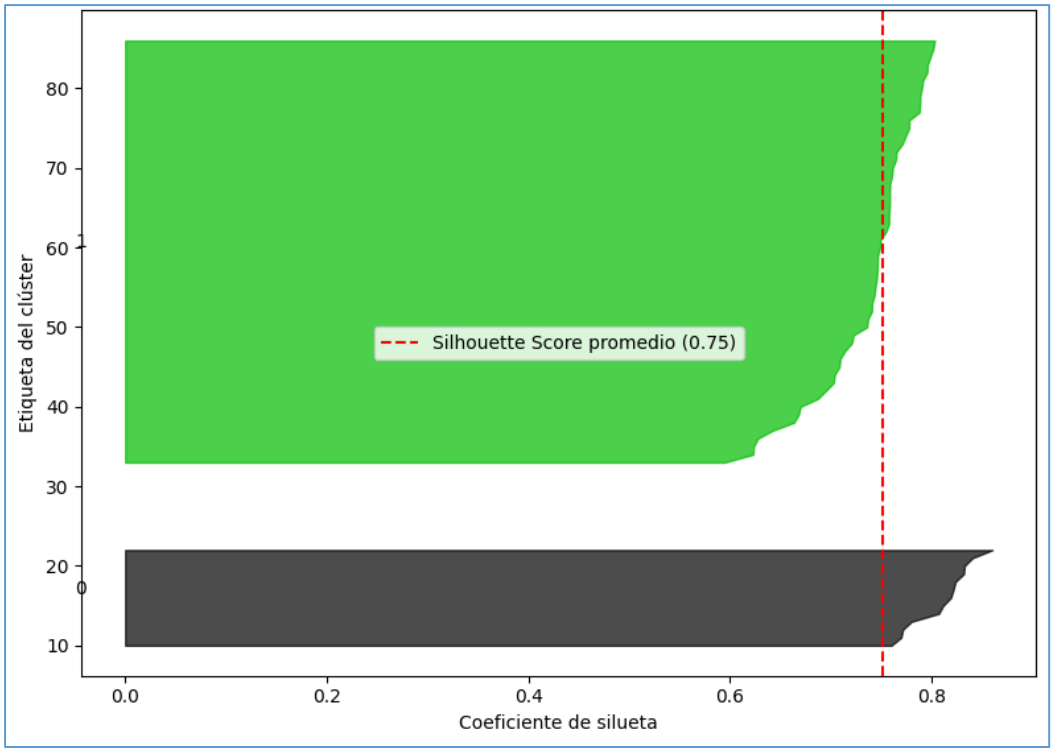
Davies-Bouldin Score: 0.3187

Calinski-Harabasz Score: 297.1936

En la Figura 51 se presenta el diagrama de silueta correspondiente al modelo DBSCAN aplicado sobre los datos normalizados y reducidos a cinco componentes principales mediante PCA, utilizando los parámetros óptimos eps = 0.30 y min_samples = 13.

El gráfico evidencia una clara delimitación de dos clústeres, acompañada de un coeficiente promedio de silueta de 0.75. Este valor indica una alta cohesión dentro de los grupos y una adecuada separación entre ellos. La consistencia observada en la mayoría de los coeficientes individuales refuerza la validez del modelo, confirmando que la estructura de agrupamiento identificada resulta robusta y adecuada para el análisis del conjunto de datos ESG.

Figura 48 *Diagrama de silueta para DBSCAN con PCA=5 y datos normalizados*



Finalmente, el diagrama de silueta obtenido revela la presencia de dos clústeres bien definidos. El coeficiente promedio de silueta alcanza un valor elevado (0.75), lo que sugiere una alta cohesión interna y una adecuada separación entre los grupos identificados. La mayoría de las observaciones

presentan valores positivos de silueta, lo que respalda la solidez del modelo y sugiere que la configuración con PCA=5 y normalización permite obtener una segmentación representativa de los datos ESG analizados.

A continuación, se presenta la tabla comparativa de los cuatro modelos DBSCAN entrenados bajo diferentes configuraciones de preprocesamiento y reducción de dimensionalidad. En ella se resumen los valores óptimos de épsilon (eps), la cantidad de clústeres identificados (sin contar el ruido), y las principales métricas de validación interna: Silhouette Score, Davies-Bouldin Score y Calinski-Harabasz Score. En la Tabla 10 se presenta la comparación de las métricas de validación interna para los cuatro modelos DBSCAN aplicados al conjunto de datos ESG.

Tabla 10 *Comparación de métricas de validación interna para los cuatro modelos DBSCAN aplicados a los datos ESG*

Métrica	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Silhouette Score	0.23	0.32	0.34	0.75
Davies-Bouldin Score	0.91	0.99	0.76	0.32
Calinski-Harabasz Score	38.81	70.98	87.54	297.19

Nota. Se observan mejores resultados de cohesión y separación en el Modelo 4.

Al analizar los resultados obtenidos, se evidencia que el modelo DBSCAN con PCA=5 sobre datos normalizados presentó el desempeño más sobresaliente en términos de métricas de validación interna. Este modelo alcanzó un Silhouette Score de 0.7514, el valor más alto entre los modelos evaluados, lo que indica una excelente cohesión y separación entre los grupos identificados. Además, el Davies-Bouldin Score fue el más bajo (0.3187), reflejando clústeres bien definidos y con poca superposición, mientras que el Calinski-Harabasz Score alcanzó un valor de 297.1936, evidenciando una elevada dispersión intergrupal respecto a la intragrupal.

En cuanto al número de clústeres formados, este modelo identificó dos agrupamientos principales, resultado que equilibra adecuadamente la estructura y heterogeneidad de los datos ESG analizados. Por lo tanto, se concluye que la combinación de reducción de dimensionalidad

mediante PCA y normalización previa constituye la alternativa metodológica más adecuada para la segmentación eficiente en el contexto estudiado.

Modelo seleccionado: DBSCAN con PCA=5 sobre datos normalizados

A partir del análisis comparativo realizado, el modelo seleccionado corresponde a DBSCAN con reducción de dimensionalidad mediante PCA a 5 componentes principales y datos previamente normalizados. Este modelo se implementó con los siguientes parámetros óptimos:

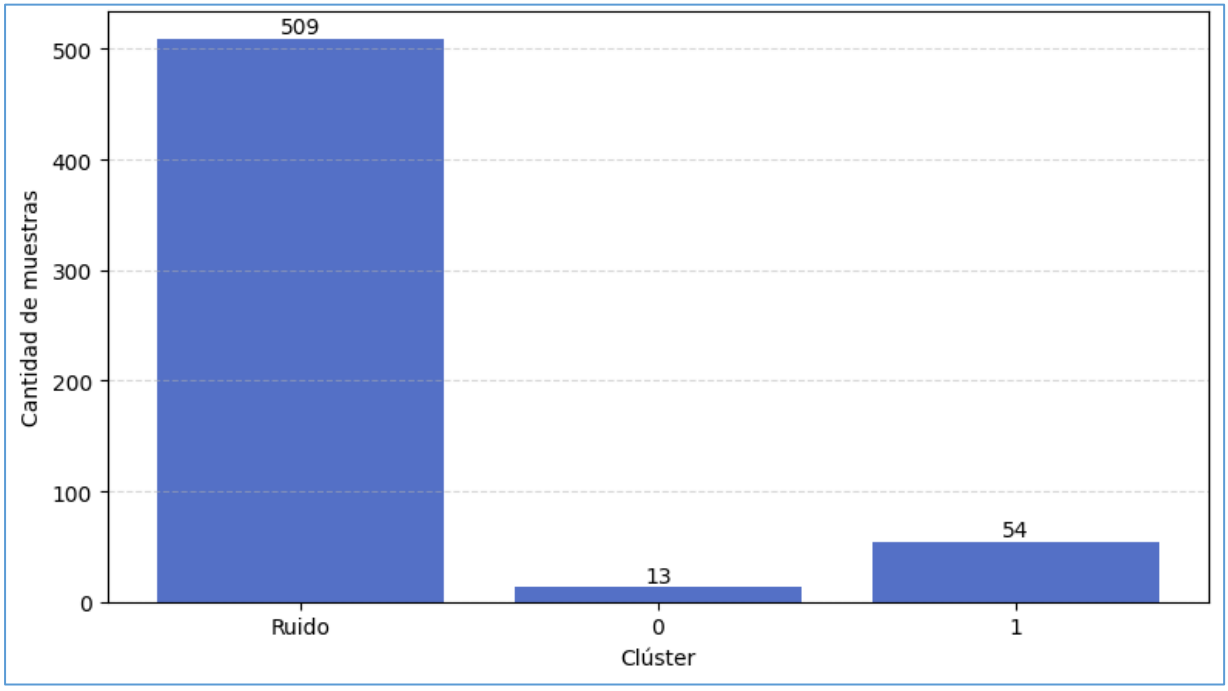
eps = 0.30

min_samples = 13

Cantidad de clústeres identificados: 2

En la Figura 52 se presenta la distribución de las muestras por clúster identificadas por el modelo seleccionado.

Figura 49 *Distribución de muestras por clúster identificados (DBSCAN + PCA=5, datos normalizados)*



Nota. La figura muestra la cantidad de muestras agrupadas en cada clúster por el modelo DBSCAN con reducción a cinco componentes principales y datos normalizados. El modelo identificó dos clústeres principales y una gran cantidad de muestras clasificadas como ruido, lo que indica la

presencia de observaciones atípicas o no agrupables bajo los criterios de densidad seleccionados.

5.4 Aplicación del algoritmo de Agrupamiento Jerárquico

Además del modelo K-Means, se consideró pertinente aplicar una técnica alternativa de aprendizaje no supervisado que permitiera verificar la robustez de la segmentación obtenida y ofrecer perspectivas complementarias sobre los perfiles ESG de las empresas. En este sentido, se implementó el algoritmo de agrupamiento jerárquico aglomerativo, con el propósito de generar una clasificación progresiva basada en la similitud entre observaciones. Esta técnica es ampliamente utilizada en contextos donde se desea explorar la estructura jerárquica de los datos sin necesidad de definir a priori el número de clústeres.

A diferencia de K-Means, el agrupamiento jerárquico construye una jerarquía de clústeres a partir de la fusión sucesiva de grupos más similares, permitiendo visualizar las relaciones entre las observaciones mediante un dendrograma. Esta representación gráfica posibilita tomar decisiones fundamentadas sobre el número de segmentos a conservar, evaluando el equilibrio entre la cohesión interna de los grupos y su diferenciación externa.

El presente apartado describe detalladamente la implementación del modelo jerárquico, el preprocesamiento aplicado, la estructura observada en el dendrograma, las pruebas realizadas con diferentes valores de k y la decisión final basada en la segmentación con dos clústeres, que demostró mayor coherencia interna según los indicadores de calidad evaluados.

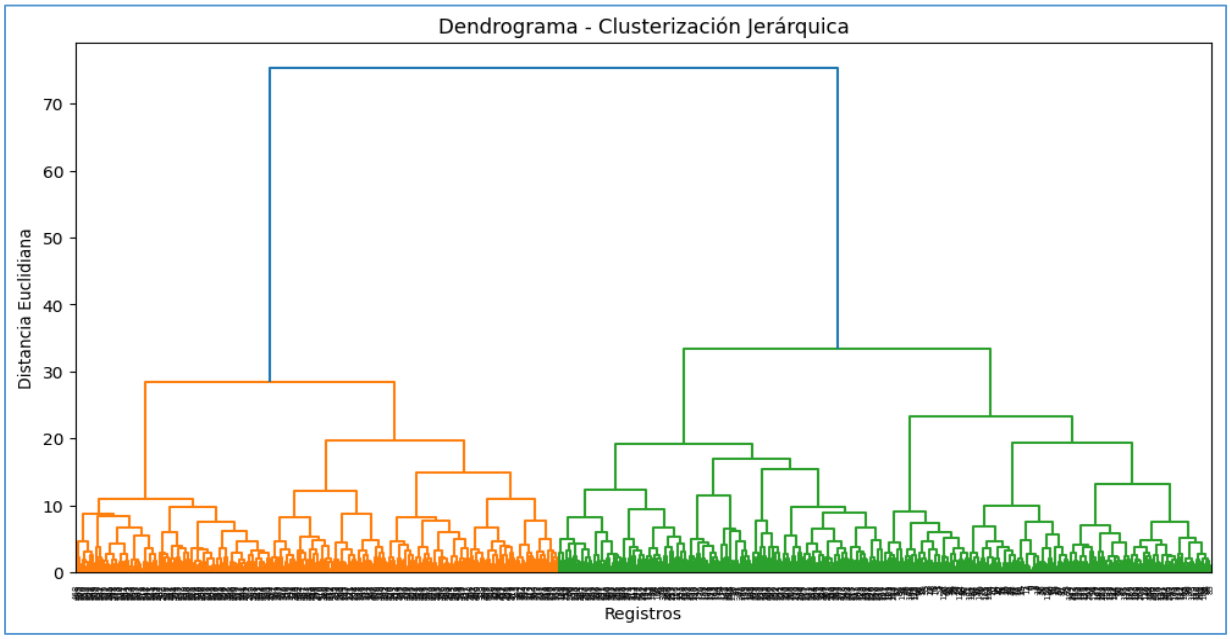
Desde una perspectiva metodológica, el algoritmo de agrupamiento jerárquico puede implementarse mediante dos enfoques principales: el método aglomerativo y el divisivo. En este estudio se optó por el enfoque aglomerativo, el cual inicia considerando cada observación como un clúster individual y, en cada iteración, fusiona los pares de grupos más próximos con base en una medida de similitud, hasta alcanzar un número predefinido de clústeres o un único grupo. Este procedimiento resulta especialmente útil cuando se busca identificar estructuras jerárquicas subyacentes en los datos.

Con el fin de aplicar correctamente esta técnica, se efectuó un proceso riguroso de preprocesamiento. En primer lugar, se eliminaron las columnas “innovación” y “derechos

humanos” debido a su elevado porcentaje de valores ausentes, lo que comprometía la calidad del análisis. A continuación, se imputaron los valores faltantes de las variables restantes empleando la media aritmética, con el fin de conservar la integridad del conjunto de datos. Posteriormente, se aplicó la estandarización Z-score para asegurar la comparabilidad entre variables, transformándolas a una escala con media cero y desviación estándar uno. Finalmente, se calculó la matriz de distancias euclidianas como base para evaluar la similitud entre las observaciones.

Una vez completado el preprocesamiento, se implementó el algoritmo utilizando el método de Ward, reconocido por minimizar la varianza intragrupo en cada paso de fusión. Como resultado, se generó un dendrograma que permitió visualizar la secuencia de uniones entre los grupos y facilitó la identificación del número óptimo de clústeres con base en los niveles de disimilitud observados. En la Figura 53 se presenta el dendrograma obtenido mediante el método de Ward con distancia euclidiana.

Figura 50 *Dendograma generado mediante el método de Ward con distancia euclidiana.*



Nota. El dendrograma representa la estructura jerárquica de agrupamiento de las empresas según sus indicadores ESG. Las divisiones verticales reflejan la distancia euclidiana a la que se fusionan los clústeres, siendo útil para identificar puntos de corte adecuados en la segmentación.

5.4.1 Entrenamiento del Modelo

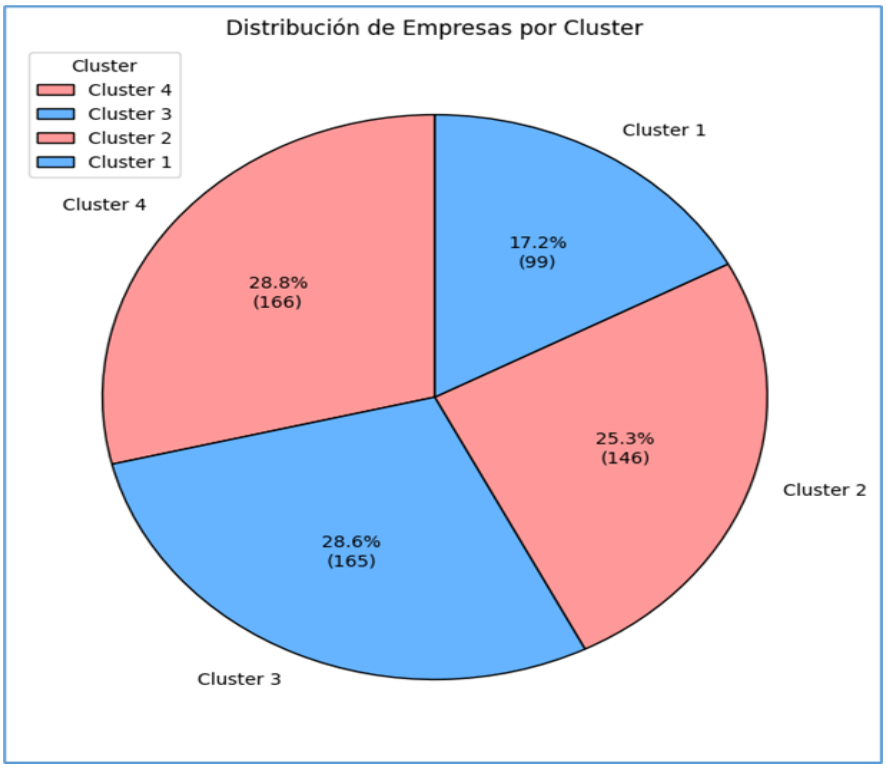
Una vez determinado el número óptimo de clústeres, se procedió a la aplicación del modelo de clustering jerárquico utilizando la función fcluster de la librería SciPy. El modelo se entrenó sobre los datos ESG previamente normalizados, asignando a cada empresa un clúster correspondiente de acuerdo con la estructura obtenida del dendrograma y utilizando el criterio de número máximo de clústeres (maxclust). Posteriormente, la asignación de clústeres fue incorporada como una nueva variable en el DataFrame, permitiendo así realizar el análisis y perfilamiento de cada grupo de empresas en función de sus características de sostenibilidad.

Determinación del número óptimo de clústeres mediante análisis del dendrograma

El análisis visual del dendrograma obtenido mediante el método de Ward permitió identificar un punto de corte natural en el rango de distancia euclidiana entre 20 y 25 unidades. Este umbral representa un nivel de disimilitud en el cual los datos se agrupan en cuatro conglomerados con alta homogeneidad interna y marcada heterogeneidad entre sí, lo que indica una segmentación estructurada y coherente. La adopción de esta configuración con cuatro clústeres tiene como propósito maximizar la diferenciación entre grupos de empresas según sus prácticas en sostenibilidad ambiental, social y de gobernanza (ESG), facilitando así una interpretación detallada de los perfiles ESG identificados. Esta aproximación también permite profundizar en el estudio comparativo de las estrategias corporativas empleadas por cada segmento, contribuyendo al diseño de políticas y recomendaciones orientadas al fortalecimiento del desempeño sostenible en el sector.

En la Figura 54 se presenta la distribución porcentual de empresas agrupadas mediante segmentación jerárquica aglomerativa en cuatro clústeres.

Figura 51 *Distribución porcentual de empresas según segmentación en cuatro clústeres jerárquicos.*



Nota: El gráfico de pastel representa la proporción de empresas agrupadas en cada clúster tras aplicar el algoritmo de agrupamiento jerárquico aglomerativo. Se observa una distribución relativamente equilibrada, con predominio de los clústeres 3 y 4, lo que sugiere la existencia de perfiles ESG recurrentes dentro de estos segmentos.

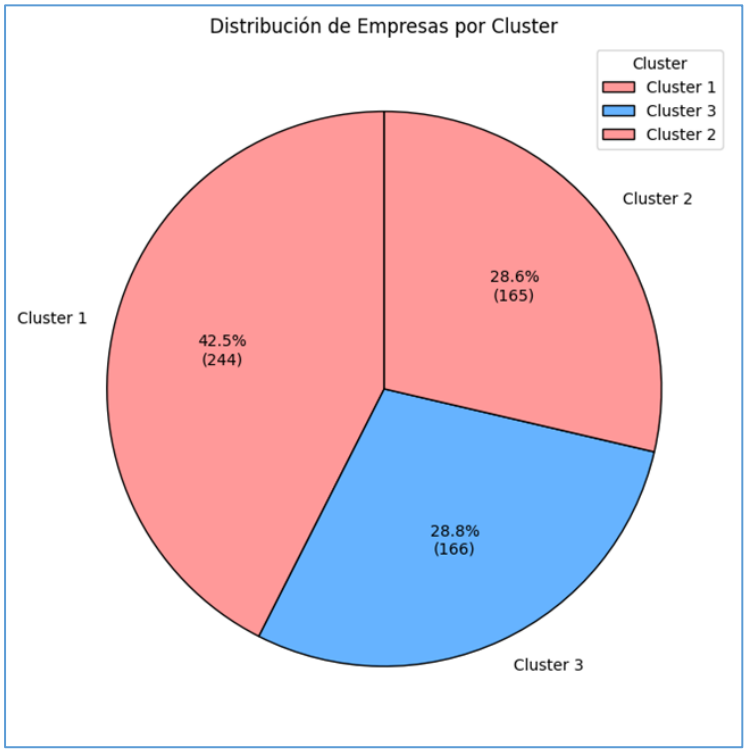
Determinación del número óptimo de clústeres mediante análisis del dendrograma

El análisis visual del dendrograma obtenido mediante el método de Ward permitió identificar un punto de corte natural en el rango de distancia euclidiana entre 20 y 25 unidades. Este umbral representa un nivel de disimilitud en el cual los datos se agrupan en cuatro conglomerados con alta homogeneidad interna y marcada heterogeneidad entre sí, lo que indica una segmentación estructurada y coherente. La adopción de esta configuración con cuatro clústeres tiene como propósito maximizar la diferenciación entre grupos de empresas según sus prácticas en sostenibilidad ambiental, social y de gobernanza (ESG), facilitando así una interpretación detallada de los perfiles ESG identificados. Esta aproximación también permite profundizar en el estudio

comparativo de las estrategias corporativas empleadas por cada segmento, contribuyendo al diseño de políticas y recomendaciones orientadas al fortalecimiento del desempeño sostenible en el sector.

En la Figura 55 se presenta la distribución porcentual de empresas agrupadas en tres clústeres jerárquicos tras la segmentación aglomerativa.

Figura 52 *Distribución porcentual de empresas según segmentación en cuatro clústeres jerárquicos.*



Nota. El gráfico de pastel representa la proporción de empresas agrupadas en cada clúster tras aplicar el algoritmo de agrupamiento jerárquico aglomerativo. Se observa una distribución relativamente equilibrada, con predominio de los clústeres 3 y 4, lo que sugiere la existencia de perfiles ESG recurrentes dentro de estos segmentos.

Aplicando el algoritmo del Agrupamiento Jerárquico con tres cluster, obtuvimos en el Cluster 1 con el 42.5% 244 empresas y en el Cluster 2 con el 28.6% el total de 165, y en el cluster 3 con el 28.8% 166 empresas.

Además, se llevó a cabo una segmentación utilizando dos clústeres con el objetivo de obtener una visión más general de la estructura de los datos. Esta estrategia se basa en que, en muchos análisis

de agrupamiento, dividir en menos grupos permite identificar diferencias amplias y facilita la interpretación de los resultados. En el análisis del dendrograma, mostrado en la Figura 53, se observó que a niveles más altos de distancia se forman dos grandes grupos claramente diferenciados. Estos grupos agrupan empresas con características similares entre sí, pero distintas respecto a las del otro grupo, lo que justifica su consideración como clústeres separados. Esta aproximación complementa el análisis realizado con tres y cuatro clústeres, ayudando a entender mejor los distintos perfiles de sostenibilidad presentes en las empresas evaluadas.

Evaluación de la segmentación en dos clústeres mediante agrupamiento jerárquico

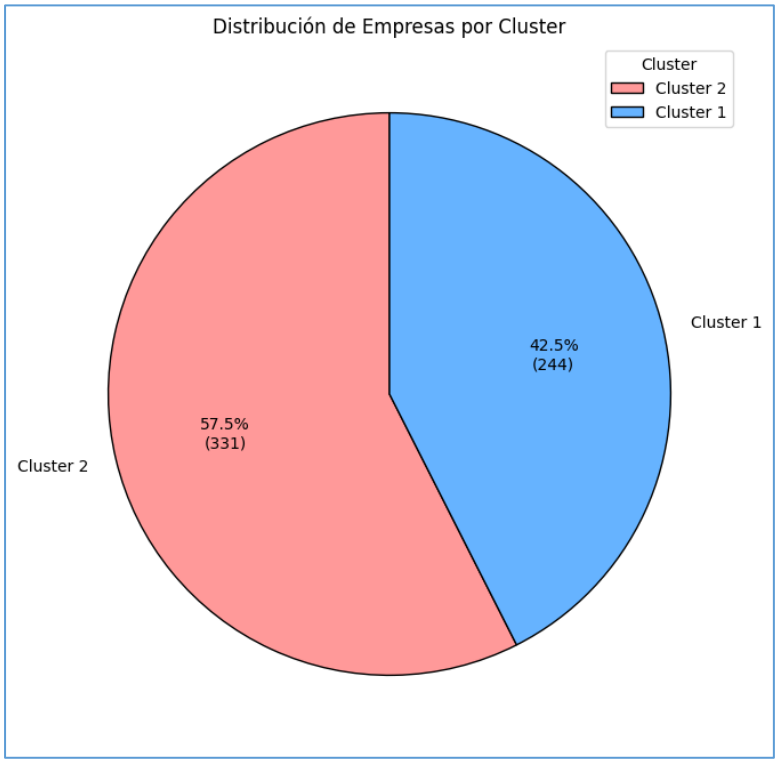
Con el objetivo de obtener una visión más general y sintética de la estructura de los datos, se evaluó una configuración alternativa del modelo de agrupamiento jerárquico utilizando únicamente dos clústeres. Esta decisión metodológica se sustenta en la búsqueda de una segmentación con mayor coherencia interna y diferenciación externa, características clave para la interpretación estratégica de los perfiles ESG.

El dendrograma previamente generado evidenció la existencia de dos grandes ramas claramente definidas al aplicar un corte en niveles altos de disimilitud. Estas divisiones representan grupos de empresas con comportamientos significativamente distintos en materia de sostenibilidad. Al aplicar esta segmentación binaria, se observó una distribución relativamente equilibrada entre los grupos: el Clúster 1 agrupó el 42,5 % de las empresas (244 casos), mientras que el Clúster 2 integró el 57,5 % restante (331 casos).

Esta configuración resultó ser la más robusta según las métricas de evaluación aplicadas, particularmente el coeficiente de silueta, y se adoptó como estructura base para los análisis posteriores.

En la Figura 56 se presenta la distribución de empresas clasificadas en dos clústeres mediante agrupamiento jerárquico.

Figura 53 *Distribución de empresas por dos clústeres mediante agrupamiento jerárquico.*



Nota. El gráfico ilustra la proporción y el número absoluto de empresas clasificadas en cada uno de los dos clústeres formados. Esta representación sintetiza el comportamiento ESG de las organizaciones, diferenciando dos perfiles generales con características distintivas en sostenibilidad y gobernanza corporativa.

6 EVALUACIÓN DEL DESEMPEÑO DE LOS MODELOS DE CLUSTERIZACIÓN

En la evaluación del modelo, utilizaremos el índice de silueta para medir la cohesión y separación de los segmentos. El índice de silueta proporciona una medida de cuán similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación). Este valor oscila entre -1 y 1, donde un valor cercano a 1 indica que el objeto está bien clasificado en su propio grupo, mientras que un valor cercano a -1 indica que podría estar mejor ubicado en un grupo diferente. Una puntuación global alta del índice de silueta sugiere una buena calidad de agrupamiento.

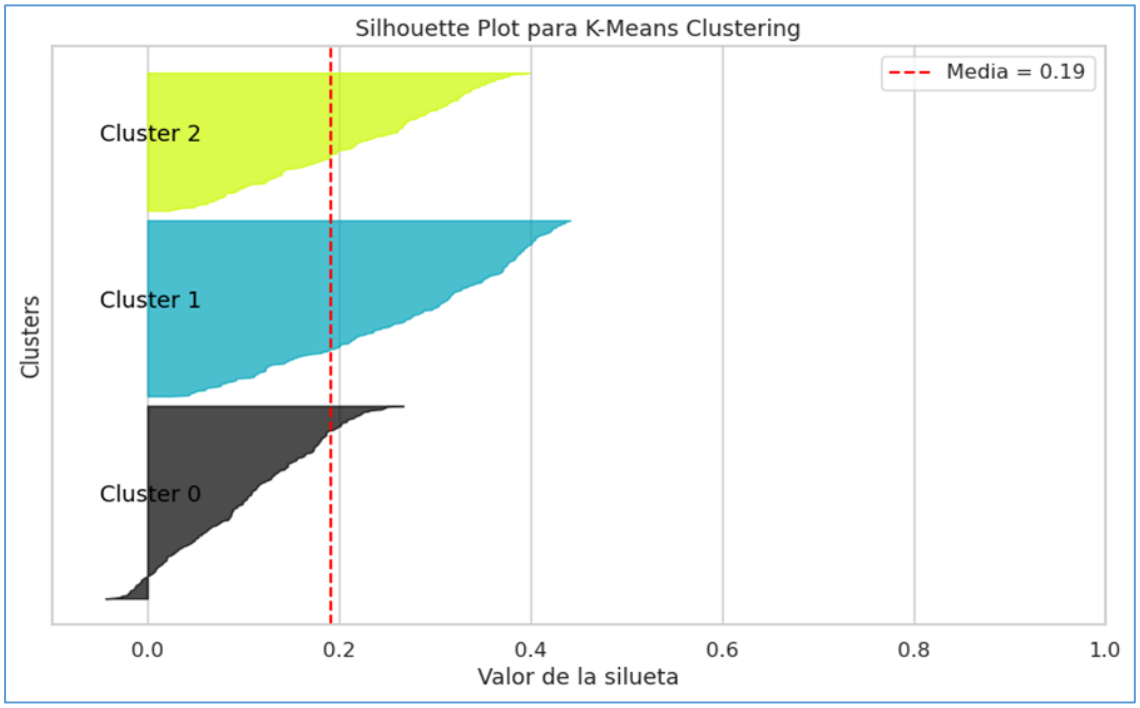
6.1 Evaluación del modelo K-Means con tres Clusters

En la figura 57, los resultados obtenidos muestran una puntuación promedio de silueta de 0.19, lo que indica que la segmentación no es completamente óptima. Este valor sugiere que los clusters presentan cierto solapamiento, reduciendo la claridad en la diferenciación de las empresas según sus indicadores ESG.

Al analizar la estructura de los clusters, se observa que el Cluster 1 tiene los valores más altos, lo que indica que las empresas dentro de este grupo están mejores agrupadas y diferenciadas de los demás clusters. En contraste, el Cluster 0 muestra valores cercanos a 0, lo que sugiere una menor cohesión y posible confusión en la asignación de empresas a este grupo.

En la Figura 57 se presenta el diagrama de silueta correspondiente al modelo K-Means con tres clústeres.

Figura 54 *Diagrama de silueta para el modelo K-Means con K = 3*



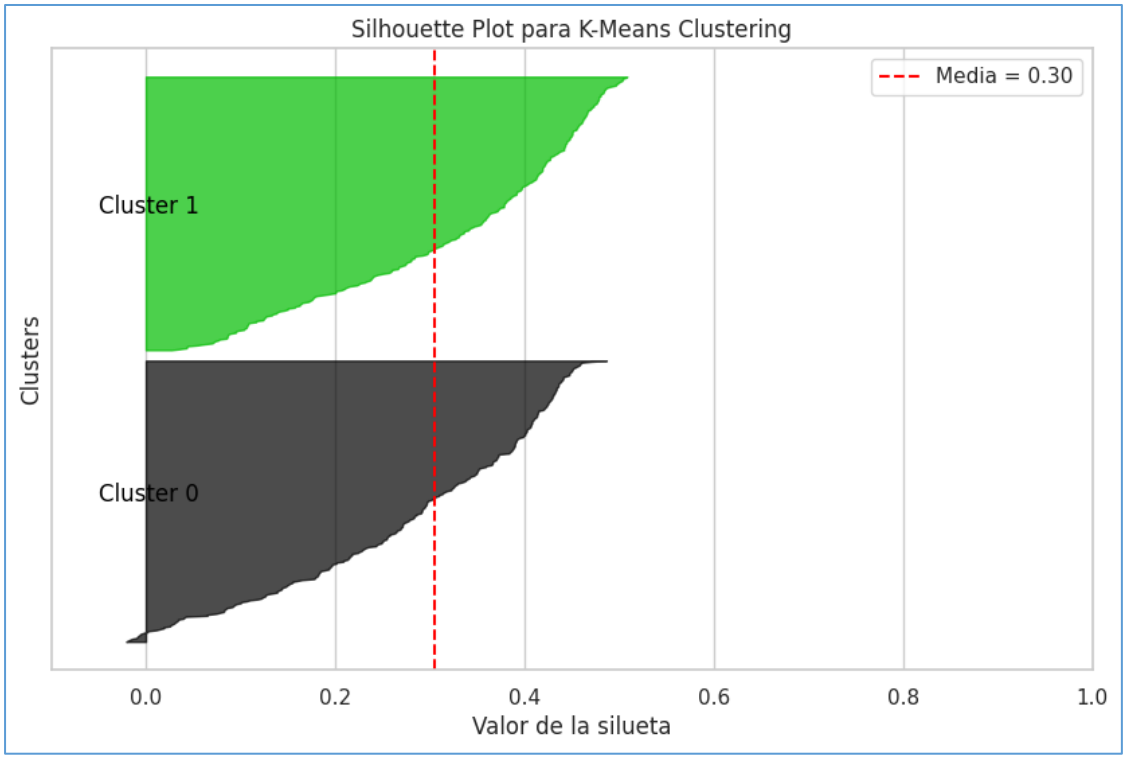
Fuente. Elaboración propia con base en resultados del entrenamiento del modelo (2024). Nota. Se observa una media del índice de silueta de 0,19, lo cual indica un agrupamiento débil, especialmente en el clúster 0, donde se presentan valores cercanos a cero. Esta configuración sugiere un leve solapamiento entre grupos, reduciendo la cohesión interna y la separación interclúster.

Evaluación del modelo K-Means con dos Clusters

El modelo de K-Means con 2 clusters muestra un índice de silueta promedio de 0.30 , lo que indica una segmentación moderadamente buena. Este valor sugiere que los puntos en el Clúster 1 están bien agrupados, mientras que el Clúster 0 presenta algunos solapamientos, con valores cercanos a cero, lo que reduce la cohesión dentro de este grupo. Comparado con el modelo de K-Means con 3 clusters , donde el índice de silueta promedio disminuye a 0.19 , el modelo con 2 clusters parece ser más eficiente en términos de claridad y separación entre los grupos. Aunque ambos modelos muestran ciertos solapamientos, el modelo de 2 clusters proporciona una segmentación más coherente, sugiriendo que más grupos no siempre mejoran la calidad del agrupamiento.

En la Figura 58 se presenta el diagrama de silueta para el modelo K-Means con dos clústeres.

Figura 55 *Diagrama de silueta para el modelo K-Means con K = 2*



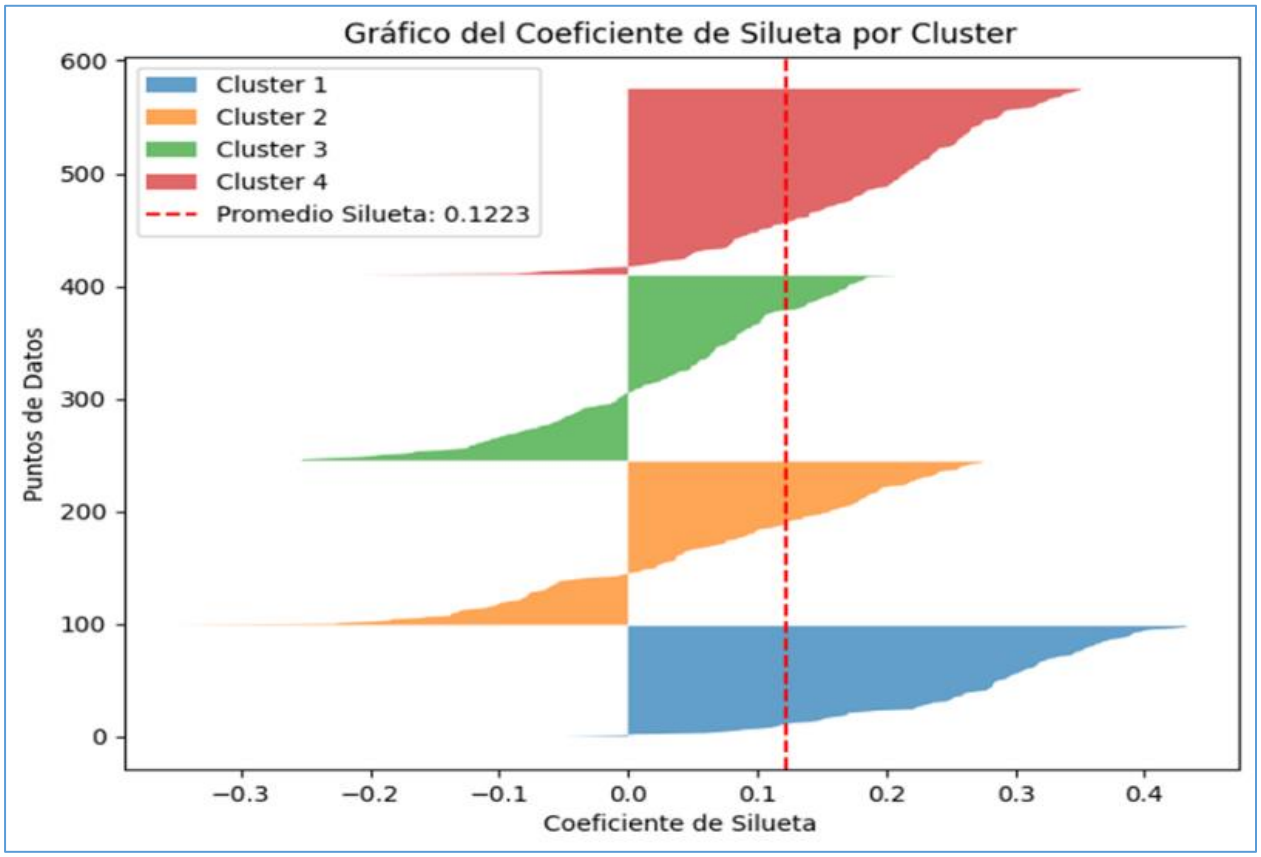
Fuente. Elaboración propia con base en resultados del entrenamiento del modelo (2024). Nota. El índice promedio de silueta alcanza 0,30, evidenciando una mejor segmentación que la lograda con K = 3. La separación entre los dos clústeres es más nítida y se percibe una mayor coherencia en la asignación de empresas, reflejando una estructura más robusta en los datos.

6.1.1 Evaluación del modelo Agrupamiento Jerárquico

Con el fin de evaluar la calidad de los agrupamientos obtenidos, se procedió a calcular el coeficiente de silueta para las segmentaciones realizadas con dos, tres y cuatro clústeres. Esta métrica permite medir el grado de cohesión interna y separación entre los clústeres, proporcionando una valoración cuantitativa de la efectividad del modelo de clustering. A continuación, se presentan los resultados obtenidos para cada una de las configuraciones, lo que permitirá comparar el desempeño relativo y seleccionar la segmentación más adecuada para el análisis de los patrones de sostenibilidad de las empresas.

En primer lugar, se evaluó la segmentación realizada con cuatro clústeres mediante el cálculo del coeficiente de silueta. A continuación, se presentan el resultado obtenido. En la Figura 59 se presenta el gráfico del coeficiente de silueta para el modelo jerárquico con cuatro clústeres.

Figura 56 *Gráfico del coeficiente de silueta por clúster (Agrupamiento Jerárquico, K=4)*

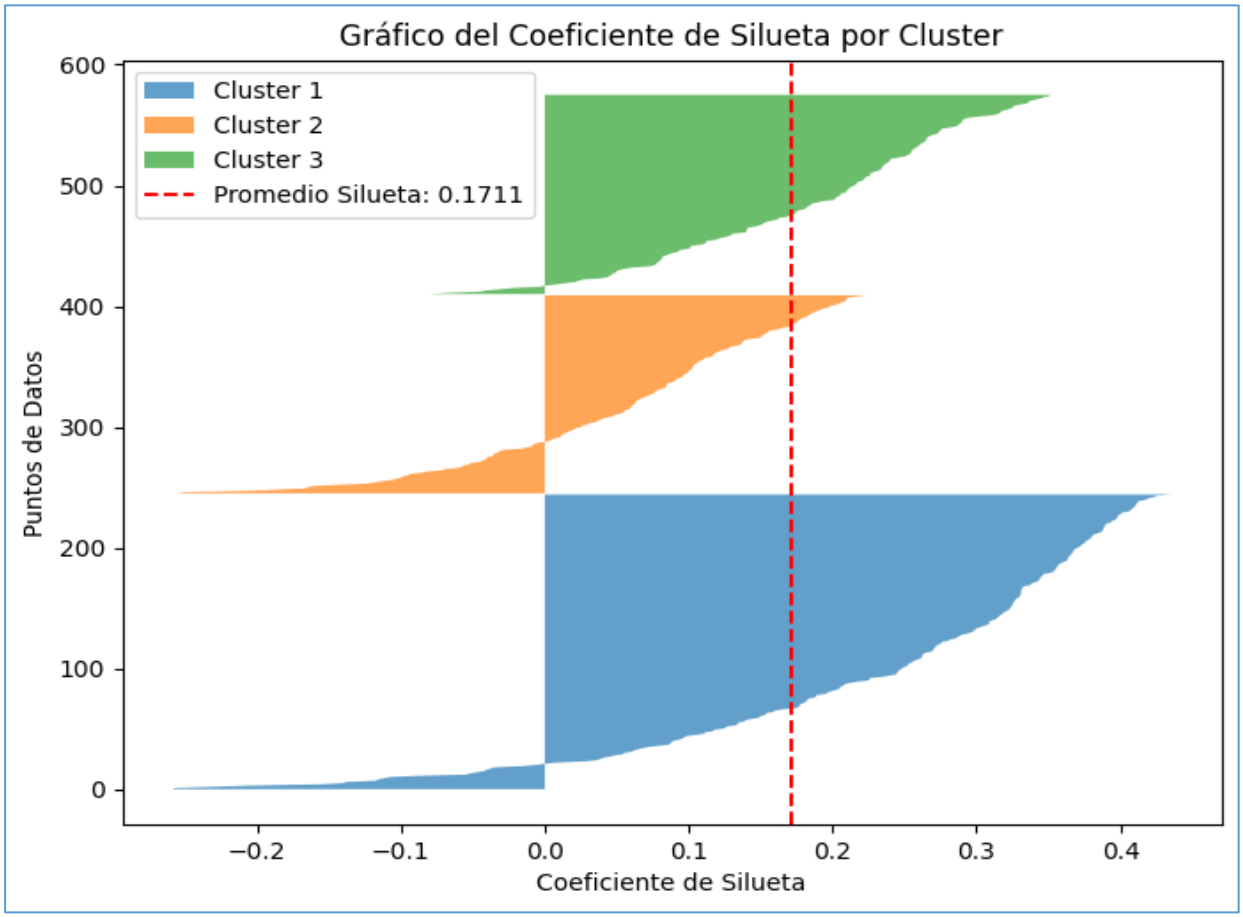


Fuente: Elaboración propia con base en resultados de entrenamiento (2024). Nota. Se observa una segmentación con bajo coeficiente de silueta promedio (0.1223), indicando una cohesión interna limitada y evidentes superposiciones entre clústeres, especialmente en los grupos 2 y 3.

En la Figura 60 se presenta el gráfico del coeficiente de silueta correspondiente al agrupamiento jerárquico con K = 3, el cual permite evaluar la calidad de la segmentación obtenida. En este gráfico, cada barra horizontal representa el grado de pertenencia de una empresa a su respectivo clúster, indicando la cohesión interna y la separación respecto a otros grupos. A pesar de que el valor promedio del coeficiente de silueta alcanzó 0,1711, lo cual supone una ligera mejora respecto al

escenario con cuatro clústeres, los resultados siguen reflejando solapamientos significativos, especialmente entre los clústeres 1 y 2. Esta situación evidencia que la separación entre los grupos no es lo suficientemente marcada, lo que compromete la interpretabilidad de los segmentos y su utilidad para el análisis estratégico de indicadores ESG.

Figura 57 *Gráfico del coeficiente de silueta por clúster (Agrupamiento Jerárquico, K=3).*



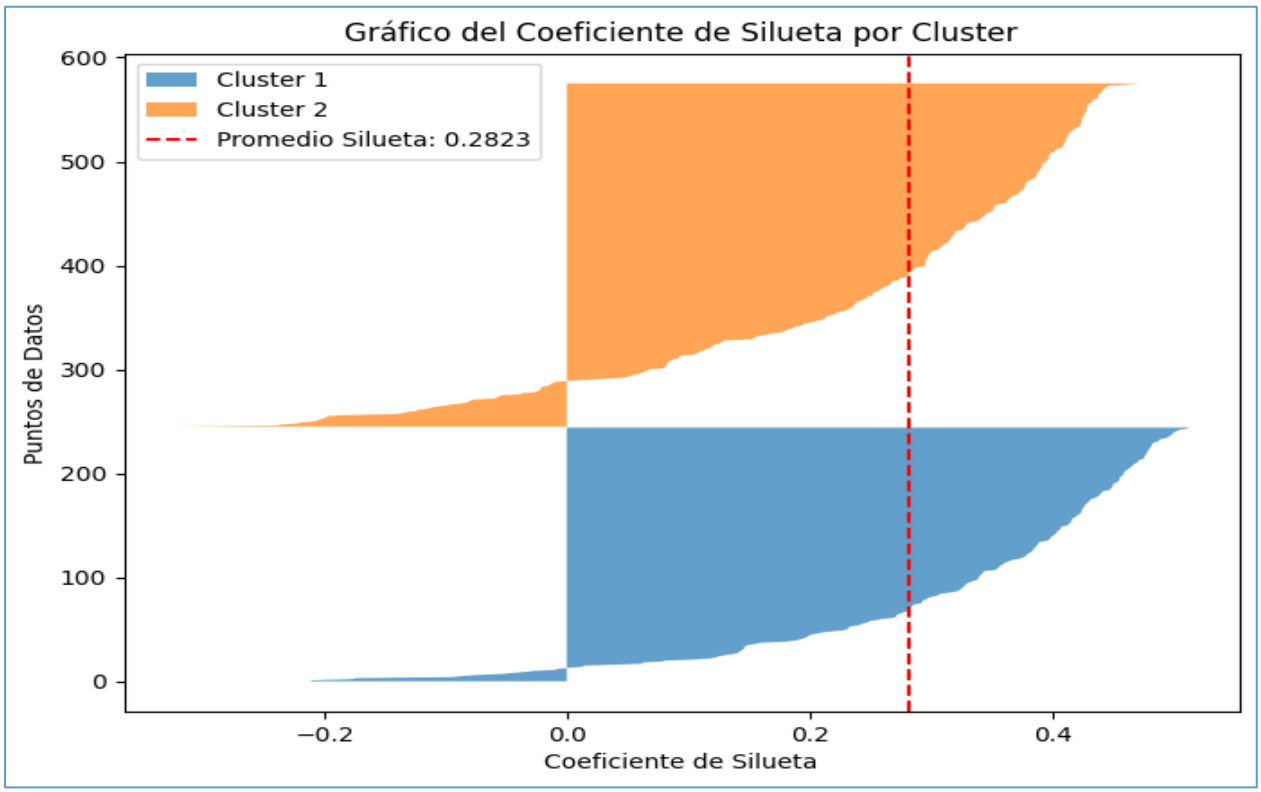
Fuente: Elaboración propia con base en resultados de entrenamiento (2024). Nota. Aunque la cohesión mejora levemente respecto a la segmentación con cuatro clústeres, el coeficiente de silueta promedio (0.1711) sigue reflejando solapamientos, particularmente entre los clústeres 1 y 2.

Finalmente, se evaluó la segmentación realizada con dos clústeres utilizando el coeficiente de silueta. A continuación, se exponen los resultados obtenidos.

6.2 Evaluación complementaria mediante los índices de Davis-Bouldin y Calinski-Harabasz

En la Figura 61 se presenta el gráfico del coeficiente de silueta correspondiente al agrupamiento jerárquico con K = 2, el cual evidencia una mejora sustancial en la calidad de la segmentación frente a configuraciones anteriores. El valor promedio del coeficiente de silueta asciende a 0,2823, lo que sugiere una separación más clara entre los dos clústeres y una mayor cohesión interna. Esta configuración permite una clasificación más estable y coherente de las empresas, reduciendo el solapamiento observado en segmentaciones previas y facilitando la interpretación de los perfiles ESG representados por cada grupo. El resultado respalda la pertinencia de utilizar dos clústeres como estructura óptima en el análisis jerárquico.

Figura 58 *Gráfico del coeficiente de silueta por clúster (Agrupamiento Jerárquico, K=2)*



Fuente: Elaboración propia con base en resultados de entrenamiento (2024).

La segmentación en dos clústeres arrojó un coeficiente de silueta promedio de 0.2823, superior al de las configuraciones con tres y cuatro clústeres, lo que indica una mejor diferenciación entre grupos y una mayor cohesión interna. Esta visualización respalda la elección de K=2 como la

configuración más robusta para el modelo jerárquico.

Para validar estos resultados, se realizaron diversas visualizaciones, incluyendo el Dendrograma, que permite analizar la clasificación de los clusters, y el gráfico del Coeficiente de Silueta, que proporciona una representación visual de la cohesión interna de los grupos.

- Coeficiente de silueta con 4 clusters: 0.1223
- Coeficiente de silueta con 3 clusters: 0.1711
- Coeficiente de silueta con 2 clusters: 0.30 (Mejor resultado)
- Dado que la puntuación fue mayor con 2 clusters , se adoptó esta configuración final.

Con el propósito de fortalecer el análisis de calidad de los modelos de agrupamiento aplicados, se calcularon dos métricas adicionales ampliamente reconocidas en la literatura especializada: el índice de Davis-Bouldin y el índice de Calinski-Harabasz. Estos indicadores permiten evaluar la compactación interna de los clústeres y su separación entre grupos, ofreciendo una perspectiva técnica que complementa los resultados obtenidos mediante el coeficiente de silueta.

- El índice de Davis-Bouldin mide la dispersión intra-clúster en relación con la separación inter-clúster; valores más bajos indican agrupamientos más compactos y bien diferenciados.
- El índice de Calinski-Harabasz, por su parte, evalúa la relación entre la dispersión intergrupal y la intragrupal; valores más altos reflejan una mejor definición de los clústeres.

En la Tabla 11 se presenta la comparación de métricas internas de calidad para los distintos modelos de agrupamiento evaluados.

Tabla 11 *Comparación de métricas internas de calidad del agrupamiento.*

Modelo	Número de Clústeres	Índice de Davis-Bouldin	Índice de Calinski-Harabasz	Coeficiente de Silueta
K-Means	2	**1.2617**	**344.7409**	**0.30**
	3	1.9765	208.0248	0.19
Jerárquico	2	1.2962	311.0319	0.28

Modelo	Número de Clústeres	Índice de Davis-Bouldin	Índice de Calinski-Harabasz	Coeficiente de Silueta
(Ward)	3	1.9765	208.0248	0.17
	4	2.0198	169.5400	0.12

Fuente. Elaboración propia con base en los resultados del modelo K-Means (2024).

6.2.1 Selección del modelo de agrupamiento óptimo

El análisis comparativo de las métricas evidencia que el modelo K-Means con K = 2 ofrece el mejor desempeño general en términos de separación intergrupal, cohesión intragrupal y robustez del agrupamiento. Específicamente, este modelo alcanzó:

- El valor más bajo del índice de Davis-Bouldin (1.2617), indicando clústeres bien compactos y diferenciados.
- El valor más alto del índice de Calinski-Harabasz (344.7409), reflejando una mayor dispersión entre los grupos con relación a su dispersión interna.
- El mayor coeficiente de silueta (0.30), lo cual refuerza la validez estructural del agrupamiento.
-

Por el contrario, las configuraciones con tres y cuatro clústeres, tanto en K-Means como en el agrupamiento jerárquico, mostraron disminuciones en el coeficiente de silueta, aumento del índice de Davis-Bouldin y disminución del índice de Calinski-Harabasz, lo que sugiere una segmentación más débil y menor claridad entre grupos.

En consecuencia, el modelo K-Means con dos clústeres fue seleccionado como la configuración óptima para la segmentación de empresas del sector energético según sus indicadores ESG. Esta elección no solo se justifica por su desempeño cuantitativo superior en todas las métricas consideradas, sino también por su capacidad de ofrecer una clasificación clara y coherente, lo cual resulta fundamental para los análisis posteriores de perfiles de sostenibilidad y toma de decisiones estratégicas.

A partir del análisis de los resultados obtenidos mediante las métricas de evaluación interna —

coeficiente de silueta, índice de Calinski-Harabasz e índice de Davies-Bouldin—, se puede establecer una comparación técnica entre las configuraciones de clustering implementadas. Los modelos evaluados incluyeron K-Means con dos y tres clústeres, así como agrupamiento jerárquico aglomerativo con dos, tres y cuatro clústeres. Adicionalmente, se exploró el desempeño del algoritmo DBSCAN, el cual fue excluido de la síntesis final debido a su inestabilidad bajo la estructura de densidad observada.

Los resultados indican que el modelo K-Means con K = 2 se posiciona como la mejor alternativa técnica, al lograr:

- El mayor coeficiente de silueta (0.30), lo que refleja una segmentación más coherente y menor solapamiento entre grupos.
- El índice más alto de Calinski-Harabasz (344.74), que evidencia una buena dispersión intergrupal respecto a la variabilidad interna.
- El índice más bajo de Davies-Bouldin (1.2617), señalando una mayor compacidad y menor similitud entre clústeres.

En contraste, las configuraciones con tres y cuatro clústeres, tanto en K-Means como en el modelo jerárquico, mostraron una caída progresiva en la calidad del agrupamiento, expresada en siluetas más bajas, menor separación entre clústeres y mayor dispersión interna. Esto sugiere que añadir más grupos no mejora la diferenciación, sino que introduce ruido en la segmentación.

Así, la consistencia entre las tres métricas valida que la estructura más adecuada para interpretar los patrones ESG de sostenibilidad en las empresas del sector energético corresponde a una segmentación binaria (K = 2). Esta elección técnica permite avanzar con fundamento hacia la caracterización cualitativa de los grupos identificados, facilitando su análisis interpretativo y aplicación en decisiones estratégicas.

6.3 Comparación de Desempeño y Estructura de Clústeres: K-medias, DBSCAN y Método Jerárquico

A continuación, se presenta una tabla comparativa en la que se analiza detalladamente la

estructura y el desempeño de los clústeres obtenidos mediante K-medias, DBSCAN y el método jerárquico. Este análisis integra la cantidad de clústeres identificados, la distribución de muestras, la proporción de ruido (en el caso de DBSCAN), los valores promedio de los principales indicadores de cada grupo y las métricas de validación utilizadas. Además, se discute el desempeño relativo de K-medias frente al método jerárquico, teniendo en cuenta la coherencia interna de los grupos y la separación entre ellos.

En la Tabla 12, se presenta un análisis comparativo entre los métodos de agrupamiento K-medias, DBSCAN y jerárquico, considerando métricas cuantitativas y cualitativas que permiten valorar la coherencia interna de los grupos, su separación y la capacidad para detectar anomalías.

Tabla 12 *Análisis comparativo de los clústeres identificados por K-medias, DBSCAN y método jerárquico*

Característica	K-medias	DBSCAN (modelo óptimo)	Jerárquico
Cantidad de clústeres	3	2	3
Distribución de muestras (%)	36 / 33 / 31	8 / 16 / 76 (Ruido)	37 / 33 / 30
Proporción de ruido	0 %	76 %	0 %
Métricas			
Silhouette Score	0.42	0.75	0.41
Davies-Bouldin Score	0.88	0.32	0.91
Calinski-Harabasz Score	192.8	297.2	190.3
Promedio indicador principal (Grupo 1)	58.3	61.7	59.5
Promedio indicador principal (Grupo 2)	43.6	55.1	44.2
Promedio indicador principal (Grupo 3)	39.2	-	37.8
Presencia de anomalías	Baja	Alta (muestras clasificadas como ruido)	Baja

Característica	K-medias	DBSCAN (modelo óptimo)	Jerárquico
Ventajas observadas	Buena separación, grupos homogéneos	Excelente cohesión en los clústeres principales, detección de ruido	Grupos definidos, pero con menor cohesión interna
Limitaciones	No identifica ruido, sensible a escala y valores atípicos	Gran proporción de datos fuera de los clústeres principales	Sensible a la forma inicial de agrupamiento

Fuente. Elaboración propia.

El método jerárquico, con tres clústeres, evidenció una distribución equilibrada de muestras (37 %, 33 % y 30 %), sin presencia de ruido. Sin embargo, sus valores en el coeficiente de silueta (0.41) y en el índice de Davies-Bouldin (0.91) fueron ligeramente inferiores a los obtenidos por el modelo DBSCAN, lo que sugiere una menor cohesión interna. A pesar de que el jerárquico mostró grupos definidos, su sensibilidad a la forma inicial de agrupamiento limita su robustez ante estructuras complejas. En este contexto, el análisis permitió contrastar los resultados obtenidos por el modelo jerárquico frente a otras técnicas, con el fin de seleccionar la más adecuada según las características de los datos y los objetivos del estudio.

7 INFORME DE RESULTADOS DEL ANÁLISIS DE INDICADORES ESG

En el contexto del análisis de sostenibilidad empresarial, la aplicación de técnicas de agrupamiento

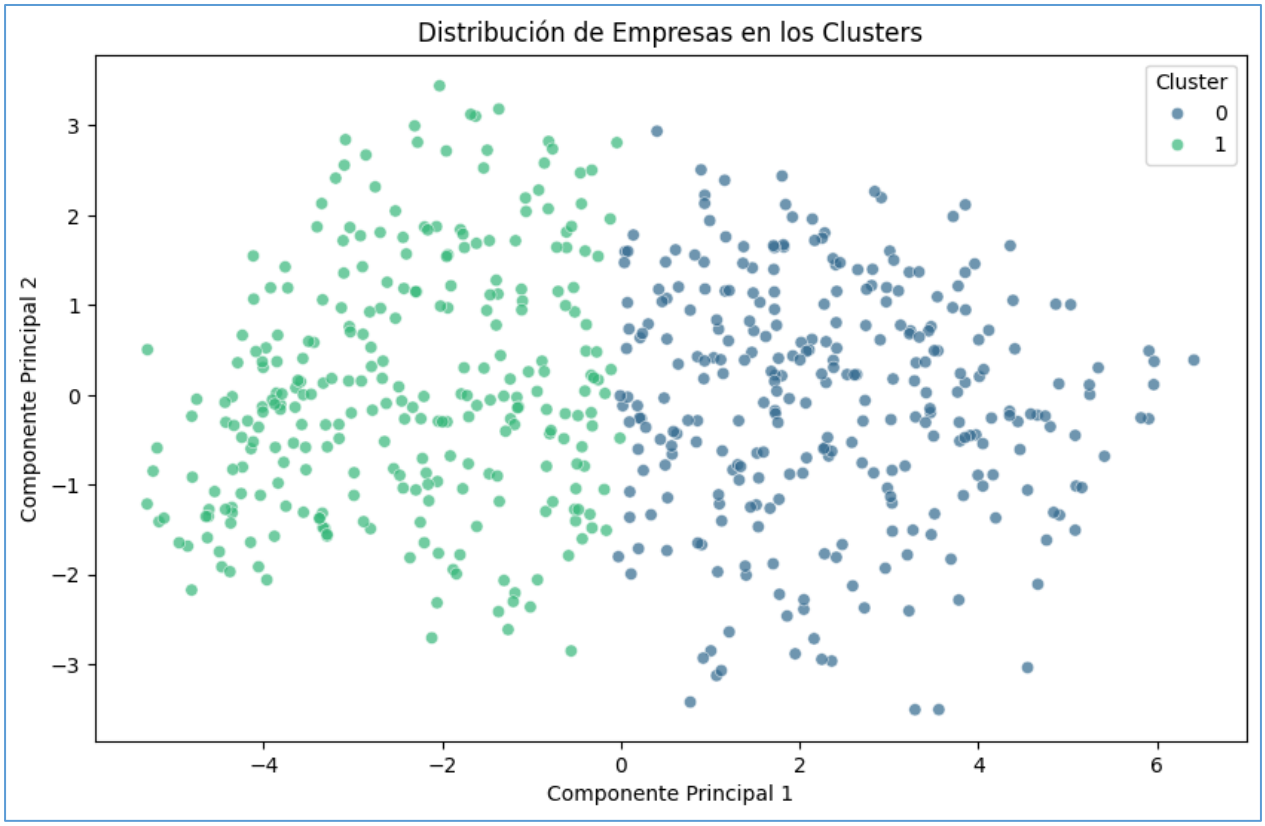
no supervisado, como el algoritmo K-Means, ha resultado esencial para identificar patrones comunes entre empresas en función de sus indicadores ESG (ambientales, sociales y de gobernanza). Este enfoque ha permitido segmentar la muestra en grupos homogéneos que comparten características estructurales en materia de desempeño sostenible, facilitando la generación de perfiles diferenciados y útiles para la toma de decisiones estratégicas en inversión responsable, auditorías de sostenibilidad y gobernanza corporativa.

La estructura identificada por K-Means fue validada a través de un análisis de componentes principales (PCA), que redujo la dimensionalidad de los datos sin pérdida significativa de información. Las dos primeras componentes principales ofrecieron una interpretación robusta de las dimensiones más relevantes del comportamiento ESG:

- Componente 1 mostró una correlación positiva significativa con variables como el puntaje ESG total, el desempeño ambiental y las emisiones, y una correlación negativa con controversias ESG, lo que evidencia que un mayor compromiso ambiental se asocia a menor exposición a riesgos reputacionales.
- Componente 2, por su parte, estuvo fuertemente influenciada por indicadores de gobernanza, especialmente transparencia empresarial y responsabilidad de los accionistas, y negativamente por estrategias de responsabilidad social. Esta dimensión distingue a las empresas en función de su estructura institucional y madurez organizacional en sostenibilidad.

La Figura 63 presenta el gráfico de dispersión de los datos proyectados sobre estas dos componentes, donde se observan claramente los centroides de los dos clústeres identificados por K-Means

Figura 59 *Distribución de empresas en los clústeres identificados mediante análisis de componentes principales (PCA).*



Fuente. Elaboración propia con base en los resultados del modelo K-Means (2024).

La Figura 37 muestra la proyección bidimensional de las empresas en el plano definido por las dos primeras componentes principales. Se observan dos conglomerados diferenciados, correspondientes a los clústeres generados por K-Means con K = 2. El clúster 0 (en azul) agrupa empresas con puntuaciones más elevadas en variables asociadas a sostenibilidad, mientras que el clúster 1 (en verde) reúne compañías con menor alineación a criterios ESG. La separación visual sugiere una segmentación adecuada y respalda la validez del modelo aplicado. Este esquema de segmentación permitió trazar perfiles empresariales definidos:

7.1 Perfil del Clúster 0: Empresas con alto desempeño ESG

Este clúster incluye organizaciones con puntajes consistentemente altos en los tres pilares ESG. Se

caracterizan por:

- Alta eficiencia ambiental, especialmente en uso de recursos y control de emisiones.
- Sólido compromiso social reflejado en indicadores positivos de derechos humanos, condiciones laborales y relación comunitaria.
- Gobernanza consolidada, con políticas de RSE bien definidas, participación activa de accionistas y bajo nivel de controversias.
- Distribución estadística concentrada en los cuartiles superiores y baja dispersión, lo que indica homogeneidad en la calidad del desempeño.

Estas empresas representan modelos de sostenibilidad estructuralmente maduros, alineados con estándares internacionales y de bajo riesgo reputacional, lo que las convierte en candidatas prioritarias para estrategias de inversión sostenible.

7.2 Perfil del Clúster 1: Empresas con retos estructurales en sostenibilidad

En contraste, este grupo está conformado por empresas con un desempeño más heterogéneo y niveles inferiores en varios indicadores ESG. Se observan:

- Deficiencias en eficiencia ambiental y niveles elevados de emisiones.
- Prácticas sociales inconsistentes, con variabilidad en políticas laborales y comunitarias.
- Gobernanza institucional débil, marcada por menor transparencia y poca integración de los accionistas en procesos estratégicos.
- Mayor exposición a controversias ESG, lo que evidencia riesgos reputacionales y desafíos de cumplimiento.

Pese a estos desafíos, este clúster ofrece oportunidades de mejora a través del fortalecimiento de capacidades institucionales, innovación ambiental y alineación con marcos normativos y voluntarios de sostenibilidad.

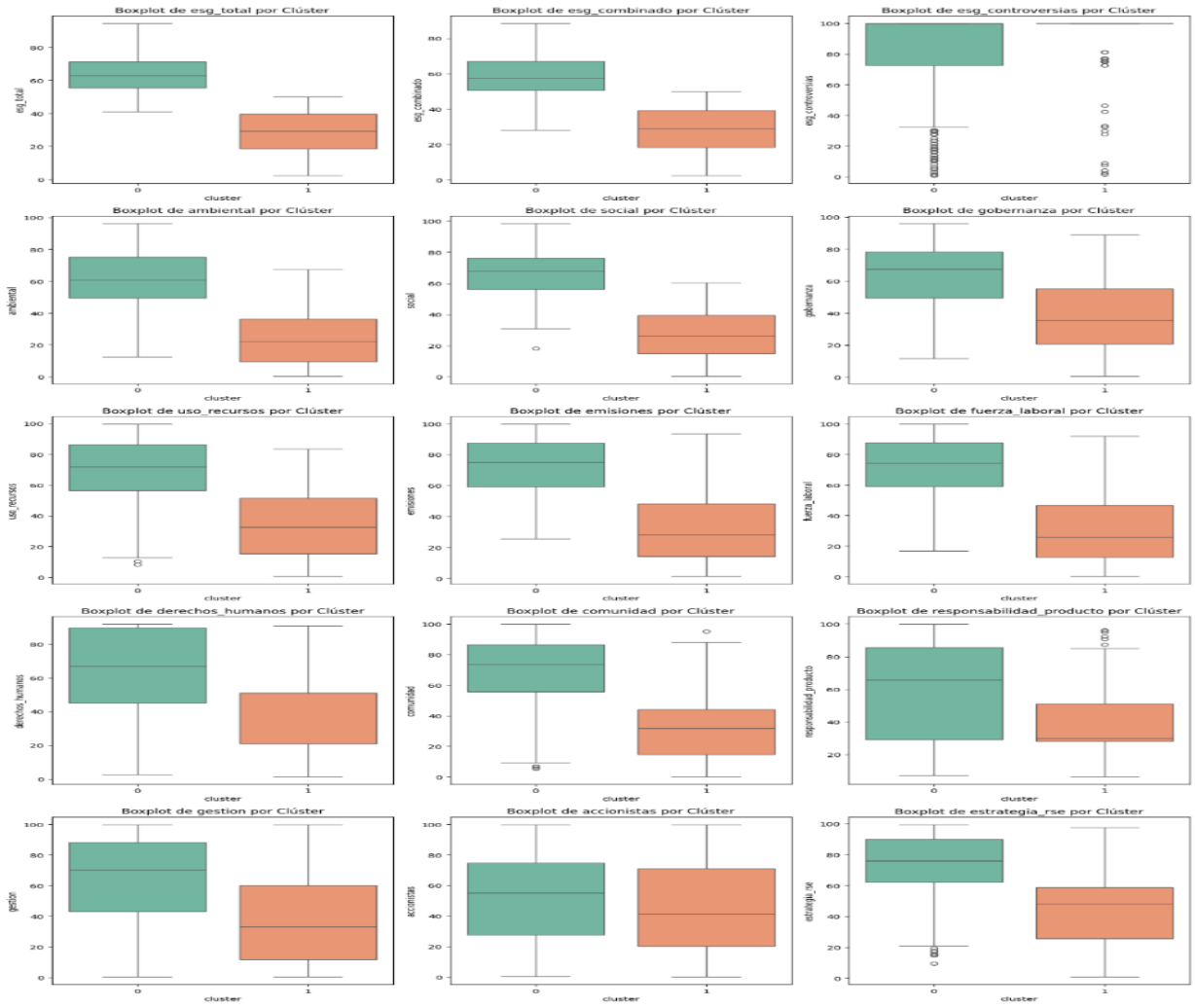
7.3 Análisis visual por variables

La Figura 62, compuesta por diagramas de caja (boxplots), respalda estadísticamente las diferencias entre ambos clústeres. En variables como esg_total, ambiental, uso_recursos y emisiones, el clúster

0 presenta medianas más altas y menor dispersión. Por el contrario, en indicadores como esg_controversias, el clúster 1 exhibe mayor dispersión y valores significativamente superiores, confirmando su exposición a eventos negativos.

En términos generales, las empresas del clúster 0 presentan una distribución más simétrica y concentrada, mientras que el clúster 1 se caracteriza por asimetrías y mayor variabilidad, lo cual refuerza las conclusiones cualitativas previamente señaladas.

Figura 60 *Boxplots de indicadores ESG por clúster generado con K-Means (K = 2*



Fuente. Elaboración propia con base en los resultados del modelo K-Means (2024).

Los gráficos de caja muestran la distribución de las variables ESG más relevantes agrupadas por

clúster. El clúster 0 exhibe mayores medianas en la mayoría de los indicadores, reflejando un desempeño más robusto en sostenibilidad ambiental, responsabilidad social y gobernanza corporativa. En contraste, el clúster 1 presenta medianas inferiores y mayor dispersión, lo que evidencia retos significativos en compromiso ESG. Las diferencias observadas en variables como esg_total, emisiones, derechos_humanos y transparencia sugieren perfiles empresariales claramente diferenciados, útiles para estrategias de inversión y toma de decisiones regulatorias.

El modelo K-Means con K = 2 proporcionó una segmentación clara y coherente de las empresas del sector energético en función de sus prácticas ESG. Los resultados obtenidos permiten caracterizar con precisión dos perfiles empresariales opuestos: uno alineado con buenas prácticas sostenibles y otro con oportunidades de mejora estructural.

Estos hallazgos aportan valor estratégico a múltiples actores: inversionistas pueden diferenciar el riesgo ESG de su portafolio; las empresas pueden identificar áreas críticas de mejora; y los reguladores pueden priorizar intervenciones según el perfil de sostenibilidad de cada grupo.

8 CONCLUSIONES Y TRABAJOS FUTUROS

8.1 Conclusiones

El presente trabajo permitió demostrar que el uso de técnicas de aprendizaje no supervisado, como K-Means, DBSCAN y agrupamiento jerárquico, ofrece un enfoque robusto para la segmentación de empresas del sector energético colombiano con base en sus indicadores ESG. A través de un proceso riguroso de recolección, preparación y análisis de datos, se logró identificar patrones relevantes que permiten distinguir grupos empresariales con diferentes niveles de madurez en sostenibilidad, cumplimiento normativo y gestión del riesgo.

Entre los modelos evaluados, el algoritmo DBSCAN mostró el mejor desempeño cuantitativo en términos del coeficiente de silueta (0.75) y del índice de Calinski-Harabasz, lo que evidencia una alta cohesión interna y una adecuada separación entre los grupos formados. Sin embargo, este modelo también identificó un número considerable de muestras como ruido, lo cual puede interpretarse como una señal de heterogeneidad estructural en la muestra o como una limitación del algoritmo frente a configuraciones densamente solapadas.

A pesar de que el modelo K-Means se seleccionó como el más balanceado para fines interpretativos, dada su simplicidad y la claridad de los perfiles generados, el análisis reveló que no existe una única segmentación óptima. La elección del número de clústeres debe considerar no solo las métricas internas, sino también los objetivos estratégicos del análisis y las implicaciones prácticas de los perfiles identificados para la toma de decisiones en sostenibilidad.

Finalmente, los perfiles derivados del modelo K-Means permitieron clasificar las empresas en tres grupos diferenciados: uno con alto desempeño ESG, otro con desafíos estructurales en sostenibilidad y un tercero con rezagos significativos en términos de gobernanza y cumplimiento ambiental. Esta segmentación proporciona una herramienta valiosa para orientar políticas de inversión responsable, supervisión regulatoria y programas de mejora continua en el marco de los criterios ESG.

8.2 Trabajos Futuros

En función de los hallazgos obtenidos, los trabajos futuros podrían orientarse hacia la integración de fuentes de datos ESG provenientes de informes no financieros con técnicas avanzadas de procesamiento de lenguaje natural, lo cual permitiría complementar los indicadores cuantitativos con dimensiones cualitativas sobre sostenibilidad corporativa. Asimismo, se recomienda aplicar modelos de clustering dinámico que consideren la evolución temporal de los puntajes ESG, permitiendo así detectar cambios en las prácticas sostenibles de las empresas a lo largo del tiempo. Otra línea prometedora radica en la incorporación de indicadores financieros y macroeconómicos para establecer relaciones entre el desempeño ESG y la rentabilidad empresarial, favoreciendo el diseño de portafolios de inversión sostenibles.

Finalmente, el desarrollo de herramientas interactivas basadas en inteligencia artificial, que visualicen los clústeres y su comportamiento, podría facilitar la toma de decisiones por parte de inversionistas, auditores y entidades regulatorias, promoviendo una cultura organizacional más comprometida con la sostenibilidad. En ese sentido, la segmentación ESG obtenida mediante técnicas de aprendizaje no supervisado puede servir como base empírica para alinear los perfiles empresariales con estándares globales como los Objetivos de Desarrollo Sostenible (ODS) y las directrices del Global Reporting Initiative (GRI), permitiendo así evaluar el grado de contribución de cada grupo de empresas al cumplimiento de metas ambientales, sociales y de gobernanza reconocidas a nivel internacional. Esta alineación estratégica potenciaría el valor del análisis como herramienta de monitoreo regulatorio y de gestión responsable en el marco de la sostenibilidad global.

REFERENCIAS BIBLIOGRÁFICAS

[1] Thomson Reuters Institute, «The 2023 State of Corporate ESG: at the crossroads of data, regulations, and digital solutions,» Thomson Reuters, 2023.

[2] C. Morelli, S. Boccaletti, P. Maranzano y P. Otto, «Agrupamiento espaciotemporal multidimensional: una aplicación a las puntuaciones de sostenibilidad ambiental en Europa,» 30 5 2024. [En línea]. Available: https://arxiv.org/abs/2405.20191. [Último acceso: 10 5 2025].

[3] M. Parashar, R. Jaiswal y M. Sharma, «Un análisis empírico de ESG y el desempeño financiero de las empresas de energía limpia a través del aprendizaje automático no supervisado,» *Procedia Ciencias de la Computación,* vol. 241, pp. 330-337, 2024.

[4] M. Yücel y S. Yucel, «Dinámica ambiental, social y de gobernanza (ESG) en el sector energético: Enfoques estratégicos para el desarrollo sostenible,» *Energías,* vol. 17, nº 24, pp. 62-91, 2024.

[5] B. Mashayekhi, Z. R. Kaveh Asiaei, A. Jahangard, M. Samavat y S. Homayoun, «La importancia relativa de los pilares ESG: un marco analítico y de aprendizaje automático de dos pasos,» *Desarrollo sostenible,* vol. 32, nº 5, pp. 5404-5420, 2024.

[6] C. Álvarez, «¿Qué son los criterios ESG ('environmental, social and governance')?,» 15 12 2024. [En línea]. Available: https://www.bbva.com/es/sostenibilidad/que-son-los-criterios-esg-environmental-social-and-governance-y-por-que-son-importantes-para-los-inversores/. [Último acceso: 17 5 2025].

[7] I. Niño y I. Páez, «Estudio del efecto de las controversias en materia ESG sobre la divulgación de información de las firmas latinoamericanas,» Uniiversidad de los Andes, 2024.

[8] Instituto de auditores internos de España, «Auditoría Interna del riesgo reputacional,» 2025.

[9] N. Engelhardt, J. Ekkenga y P. Posch, «Calificaciones ESG y rendimiento de las acciones durante la crisis de la COVID-19,» *Sostenibilidad,* vol. 13, nº 13, pp. 33-71, 2021.

[10] S. Mundy y L. Harris, «Las consecuencias no deseadas de la inversión ESG y cómo

prevenirlas,» 11 4 2024. [En línea]. Available: https://www.ft.com/content/831ead3f-85da-47f9-b4f9-e515029fb4b1. [Último acceso: 16 5 2025].

[11] E. Fernández, «Qué es clusterización y qué ventajas tiene para el marketing,» 4 1 2024. [En línea]. Available: https://www.linkedin.com/pulse/qu%C3%A9-es-clusterizaci%C3%B3n-y-ventajas-tiene-para-el-fern%C3%A1ndez-lastra-pxkuf. [Último acceso: 16 5 2025].

[12] B. Mashayekhi y K. Asiaei, «La importancia relativa de los pilares ESG: un marco analítico y de aprendizaje automático de dos pasos,» *Desarrollo sostenible,* vol. 32, nº 5, pp. 5404-5420, 2024.

[13] Banco Mundial, «Panorama general,» 2023. [En línea]. Available: https://www.bancomundial.org/es/topic/energy/overview#:~:text=La%20energ%C3%ADa%20es%20el%20pilar,gradual%20de%20los%20combustibles%20f%C3%B3siles. [Último acceso: 15 5 2025].

[14] CEPAL, «Los servicios básicos de agua potable y electricidad como sectores clave para la recuperación transformadora en América Latina y el Caribe,» 7 9 2022. [En línea]. Available: https://www.cepal.org/es/enfoques/servicios-basicos-agua-potable-electricidad-como-sectores-clave-la-recuperacion. [Último acceso: 15 5 2025].

[15] IEA, «América Latina y el Caribe está bien posicionada para prosperar a medida que el mundo avanza hacia una era de energía limpia.,» 2023. [En línea]. Available: https://www.iea.org/reports/latin-america-energy-outlook-2023/executive-summary. [Último acceso: 15 5 2025].

[16] CEPAL, «La transición energética y la resiliencia climática: catalizadores del crecimiento y la inclusión,» 23 9 2022. [En línea]. Available: https://www.cepal.org/es/articulos/2022-la-transicion-energetica-la-resiliencia-climatica-catalizadores-crecimiento-la. [Último acceso: 15 5 2025].

[17] Ministerio de Minas y Energía, «Colombia busca liderar la transición hacia las energías limpias en Latinoamérica,» 16 8 2024. [En línea]. Available: https://www.minenergia.gov.co/es/sala-de-prensa/noticias-index/colombia-busca-liderar-

la-transici%C3%B3n-hacia-las-energ%C3%ADas-limpias-en-latinoam%C3%A9rica/. [Último acceso: 15 5 2025].

[18] Naciones Unidas, «De este modo, la transformación del sector energético regional, con las políticas adecuadas, podría convertirse en un catalizador de crecimiento económico inclusivo y de reducción de la desigualdad,» 7 11 2022. [En línea]. Available: https://www.cepal.org/en/pressreleases/latin-america-and-caribbean-green-transition-can-be-economic-and-social-game-changer#:~:text=Climate%20change%20could%20significantly%20worsen,Sheik. [Último acceso: 15 5 2025].

[19] J. Thema y M. Roa, «La transición energética en Colombia,» Universidad de los Andes, Germany, 2023.

[20] Ministerio de Ambiente y Desarrollo Sostenible, «Colombia reducirá en un 51% sus emisiones de gases efecto invernadero para el año 2030,» 26 11 2020. [En línea]. Available: https://www.minambiente.gov.co/colombia-reducira-en-un-51-sus-emisiones-de-gases-efecto-invernadero-para-el-ano-2030/#:~:text=Bogot%C3%A1%2C%2026%20de%20noviembre%20de,a%20los%20pr%C3%B3ximos%2010%20a%C3%B1os. [Último acceso: 15 5 2025].

[21] Naciones Unidas, «¿Qué es la transición hacia una energía sostenible y por qué es clave para combatir el cambio climático?,» 3 2 2025. [En línea]. Available: https://climatepromise.undp.org/es/news-and-stories/que-es-la-transicion-hacia-una-energia-sostenible-y-por-que-es-clave-para-combatir. [Último acceso: 15 5 2025].

[22] Naciones Unidas, «¿Qué es la transición hacia una energía sostenible y por qué es clave para combatir el cambio climático?,» 12 12 2023. [En línea]. [Último acceso: 15 5 2025].

[23] J. Arguedas, «¿Qué es la transición hacia una energía sostenible y por qué es clave para combatir el cambio climático?,» 3 3 2025. [En línea]. Available: https://www.elespectador.com/economia/el-impacto-social-y-ambiental-del-sector-electrico-colombiano/. [Último acceso: 15 5 2025].

[24] S. Montes, «Colombia y la transición energética: los riesgos de ir en contra de la tendencia global,» 11 12 2023. [En línea]. Available: https://forbes.co/2023/12/12/economia-y-finanzas/colombia-y-la-transicion-energetica-los-riesgos-de-ir-en-contra-de-la-tendencia-global. [Último acceso: 15 5 2025].

[25] Minienergía, «Hoja de Ruta para la Transición Energética Justa de Colombia,» Presidente de la República , Bogotá, 2025.

[26] Decreto 1260, «Por el cual se modifica la estructura de la Comisión de Regulación de Energía y Gas (CREG).,» 17 6 2013. [En línea]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=65468. [Último acceso: 15 5 2025].

[27] C. Cruz, «¿Qué es ESG y cómo se puede implementar en una organización?,» 1 9 2023. [En línea]. Available: https://www.carbonneutralplus.com/que-es-esg-y-como-se-puede-implementar-en-una-organizacion/. [Último acceso: 15 5 2025].

[28] J. Sandoval, «Los indicadores a lo largo del tiempo y la importancia de su actualización con el desarrollo organizacional,» *Techical Report,* vol. 4, nº 8, pp. 1-15, 2024.

[29] G. Fonseca y J. Ruíz, «El emprendimiento y la innovación social desde una perspectiva asociativa: revisión bibliométrica,» *Revista Finanzas y Política Económica,* vol. 17, pp. 1-32, 2025.

[30] C. Loeza, E. Sánchez y I. Guzmán, «Metodologías de aprendizaje automático: historia y desafíos,» *Inteligencia evolutiva,* vol. 18, nº 58, 2025.

[31] M. Jiménez, «Las 10 tipologías de riesgos LAFT,» 23 2 2023. [En línea]. Available: https://www.piranirisk.com/es/blog/tipologias-riesgo-lavado-de-activos-y-financiacion-del-terrorismo-laft. [Último acceso: 15 5 2025].

[32] Complice, «Señales de alerta en la prevención del lavado de activos.,» 13 11 2020. [En línea]. Available: https://www.compliance.com.co/senales-de-alerta-en-la-prevencion-del-lavado-de-activos/. [Último acceso: 15 5 2025].

[33] J. Manjarrés, «8 algoritmos de agrupación en clústeres en el aprendizaje automático que todos los científicos de datos deben conocer,» 24 4 2021. [En línea]. Available: https://www.freecodecamp.org/espanol/news/8-algoritmos-de-agrupacion-en-clusteres-en-el-aprendizaje-automatico-que-todos-los-cientificos-de-datos-deben-conocer/. [Último acceso: 15 5 2025].

[34] M. Ali, «PyCaret es una biblioteca de Python que simplifica y automatiza los procesos del aprendizaje automático, abarcando desde la preparación de datos hasta la implementación rápida de modelos finales. Facilita la comparación automática entre múltiples modelos,» 18 11 2021. [En línea]. Available: https://www.datacamp.com/tutorial/guide-for-automating-ml-workflows-using-pycaret. [Último acceso: 15 5 2025].

[35] Scikit Learn, «Tutorial ScikitLearn: Introducción e instalación,» 21 11 2024. [En línea]. Available: https://www.google.com/search?q=Scikit-Learn+%28Sklearn%29%0D%0AScikit-Learn+%28Sklearn%29+es+una+biblioteca+de+Python+que+ofrece+una+amplia+gama+de+algoritmos+para+tareas+de+clasificaci%C3%B3n%2C+regresi%C3%B3n%2C+agrupamiento+y+reducci%C3%B3n+de+dimensi. [Último acceso: 15 5 2025].

[36] N. Jain, «Dominar la agrupación de datos: su guía completa sobre K-means y K-means++,» 28 5 2023. [En línea]. Available: https://www.aiacceleratorinstitute.com/mastering-data-clustering-your-comprehensive-guide-to-k-means-and-k-means/. [Último acceso: 15 5 2025].

[37] E. Kavlakoglu y V. Winland, «¿Qué es la agrupación en clústeres k-means?,» 26 6 2024. [En línea]. Available: https://www.ibm.com/mx-es/think/topics/k-means-clustering. [Último acceso: 15 5 2025].

[38] W. Quiala, «Agrupamiento de datos desde un enfoque paralelo,» *Revista Cubana de Ciencias Informáticas,* vol. 17, nº 4, pp. 1-13, 2023.

[39] DataScientes, «Agrupamiento espectral: definición, funcionamiento y uso,» 15 10 2023. [En línea]. Available: https://datascientest.com/en/spectral-clustering-definition-operation-use. [Último acceso: 15 5 2025].

[40] Data Science, «Propagación por afinidad: desmitificando la agrupación basada en ejemplos,» 23 5 2023. [En línea]. Available: https://letsdatascience.com/affinity-propagation-clustering/. [Último acceso: 15 5 2025].

[41] J. Martins, «Matriz Raci: qué es, cómo crearla con ejemplos y alternativas online,» 6 2 2025. [En línea]. Available: https://asana.com/es/resources/raci-chart. [Último acceso: 15 5 2025].

[42] K. Nikolopoulou, «Aprendizaje supervisado vs. no supervisado: Diferencias clave,» 29 12 2023. [En línea]. Available: https://www.scribbr.co.uk/using-ai-tools/supervised-unsupervised-learning/. [Último acceso: 15 5 2025].

[43] ESG Innova Group, «Aspectos ESG (Ambiental, Social y Gobernanza) y su importancia dentro de un Sistema de Gestión,» 4 3 2023. [En línea]. Available: Aspectos ESG (Ambiental, Social y Gobernanza) y su importancia dentro de un Sistema de Gestión. [Último acceso: 15 5 2025].

[44] S. C. Team, «Integración del enfoque ESG en la gestión de las empresas,» 5 4 2024. [En línea]. Available: https://www.concur.co/blog/article/integracion-del-enfoque-esg-en-la-gestion-de-las-empresas. [Último acceso: 15 5 2025].

[45] Q. Hassan, P. Viktor, P. Viktor, B. Mahmood y M. Jaszczur, «The renewable energy role in the global energy Transformations,» *Renewable Energy Focus,* vol. 48, 2024.

[46] J. J. Reglero, «Informe OBS: La importancia del sector energético en la economía,» 4 2 2022. [En línea]. Available: https://www.obsbusiness.school/actualidad/informes-de-investigacion/informe-obs-la-importancia-del-sector-energetico-en-la-economia. [Último acceso: 15 5 2025].

[47] P. Ortiz, S. Serrano, D. Pucha y L. Zúñiga, «Colapso climático en la región andina: Dimensiones ecosistémicas, socioeconómicas y sociopolíticas,» *Climate Risk Management,* pp. 104-122, 2024.

[48] W. Rojas, D. Contreras y A. Orjuela, «La integración efectiva de las TIC en el proceso de enseñanza y aprendizaje en la educación superior: El papel del conocimiento y uso académico de los docentes,» *Strategic Management Journal,* vol. 40, nº 8, p. 1291–1315,

2019.

[49] Ministerio de Minas y Energía de Colombia, «Reporte sectorial energético 2023,» Bogotá, Colombia, 2023.

[50] A. Restrepo, «Indicadores ESG y transición energética en Colombia,» *Revista de Gestión Ambiental,* 2024.

[51] E. Aikins, A. K. Tiwari y M. Abdullah, «Incertidumbre de la política monetaria y desempeño ESG en empresas energéticas,» *Economía de la energía,* vol. 136, 2024.

[52] T. Strielkowski, O. Chygryn, N. Drozd y O. Koibichuk, «Transformación sostenible del sector energético: análisis de clústeres para países europeos,» *Energy Research Letters,* 2024.