Compás
capacitación e investigación

La rentabilidad y segmentación en la explotación de minas y canteras en el Ecuador:

Uso del algoritmo de Machine Learning no supervisado

Elio Edwin Sánchez Suárez
Víctor Alfredo Iturralde Calahorrano
Ingrid Mercedes Lamilla Miranda
Jessica Jessenia Izurieta Álvarez

www.grupocompas.com

La rentabilidad y segmentación en la explotación de minas y canteras en el Ecuador: Uso del algoritmo de Machine Learning no supervisado

Elio Edwin Sánchez Suárez
Víctor Alfredo Iturralde Calahorrano
Ingrid Mercedes Lamilla Miranda
Jessica Jessenia Izurieta Álvarez

ISBN: 978-9942-53-156-8
Primera edición, 2026

© Autor
Elio Edwin Sánchez Suárez
ORCID: 0000-0002-0644-3238
Instituto Superior Tecnológico Babahoyo
esanchez@istb.edu.ec
Economista
Magister en Finanzas y Proyectos Corporativos
Magister en Inteligencia de Negocios y Ciencia De Datos.

Víctor Alfredo Iturralde Calahorrano
ORCID: 0009-0009-9845-2841
Universidad de Guayaquil
victor.iturraldec@ug.edu.e
Licenciado en ciencia de la educación, mención comercio exterior
Magister en inteligencia de negocios y ciencia de datos.

Ingrid Mercedes Lamilla Miranda
ORCID: 0000-0003-0442-0099
Instituto Superior Tecnológico Babahoyo
ilamilla@istb.edu.ec
Ingeniera Comercial
Magister en educacion mención en gestión del aprendizaje mediado por TIC

Jessica Jessenia Izurieta Álvarez
ORCID: 0009-0009-7849-194X
Instituto Superior Tecnológico Babahoyo
jizurieta@istb.edu.ec
Licenciada en administracion ejecutiva

Guayaquil, Ecuador
www.grupocompas.com
http://repositorio.grupocompas.com

Primera edición, 2026

Esta obra ha sido sometida a un proceso de evaluación bajo el sistema de arbitraje doble ciego (double-blind peer review), garantizando el anonimato tanto de los autores como de los evaluadores externos. El dictamen favorable certifica que el contenido cumple con los más altos estándares de rigor científico, calidad editorial y originalidad exigidos por la comunidad académica internacional para su indexación y reconocimiento científico.

ISBN: 978-9942-53-156-8
Distribución online
Acceso abierto

Cita
Sánchez, E., Iturralde, V., Lamilla, I., Izurieta, J. (2026) La rentabilidad y segmentación en la explotación de minas y canteras en el Ecuador: Uso del algoritmo de Machine Learning no supervisado. Editorial Grupo Compás

Índice

Análisis de la rentabilidad desde la teoría.

La rentabilidad es un factor esencial en la evaluación de la viabilidad económica y el rendimiento financiero de cualquier industria, incluida la de explotación de minas y canteras. En Ecuador, este sector ha mostrado un crecimiento considerable en los últimos años, contribuyendo de manera importante al Producto Interno Bruto (PIB) del país, particularmente a través de la extracción y exportación de minerales como el oro y el cobre (Banco Central del Ecuador, 2022). Sin embargo, la rentabilidad en esta industria se ve afectada por varios factores, como la fluctuación en los precios de los minerales, los altos costos operativos y la volatilidad de los mercados internacionales (Milo, 2024). Analizar la rentabilidad permite identificar no solo el desempeño actual de las empresas en este sector, sino también su capacidad para afrontar desafíos financieros y mejorar sus niveles de competitividad

La segmentación del mercado, por otro lado, juega un rol crucial en la industria minera, ya que permite clasificar a las empresas según factores como su tamaño, tipo de extracción y enfoque de mercado. La segmentación ayuda a identificar patrones que pueden no ser evidentes a simple vista y que resultan vitales para una toma de decisiones más estratégica (Jácome, Enríquez, & Caicedo, Evaluación del sector minero y su incidencia en el PIB del Ecuador, periodo 2019 -2021, 2023). El empleo de herramientas de segmentación avanzadas, como algoritmos de aprendizaje no supervisado, proporciona una ventaja significativa en la industria minera. Estos permiten identificar patrones ocultos, optimizando la exploración y gestión de recursos. Estos algoritmos permiten agrupar empresas con características similares y detectar grupos específicos que comparten patrones de comportamiento financiero y operacional, facilitando así una mayor precisión en la toma de decisiones estratégicas y la asignación de recursos (Jácome, Enríquez, & Caicedo, Evaluación del sector minero y su incidencia en el PIB del Ecuador, periodo 2019 -2021, 2023)

El presente estudio integra tanto el análisis de rentabilidad como la segmentación mediante el uso de algoritmos no supervisados de Machine Learning para brindar una visión integral del sector de minas y canteras en Ecuador. La combinación de estas dos variables permite identificar las relaciones existentes entre el rendimiento financiero de las empresas y su clasificación en grupos o segmentos específicos dentro del mercado. De esta forma, se busca no solo optimizar el análisis financiero, sino también

mejorar la planificación estratégica de las empresas en este sector, facilitando la identificación de oportunidades de crecimiento y mitigación de riesgos (Jauregui, Vilca, Llanos, & Alca, 2024).

El análisis de rentabilidad y la segmentación del sector minero y de canteras en Ecuador se puede realizar aplicando algoritmos de aprendizaje automático no supervisados. Los algoritmos de aprendizaje automático no supervisados, como K-means y K-medoids, son capaces de agrupar empresas mineras según métricas de rentabilidad y características operativas, como se hizo en estudios de microempresas en Ecuador (Ruiz-López, 2023).

Las políticas que son efectivas en patrocinar actividades mineras y abordar la minería ilegal tienen un gran impacto en el sostenimiento de un clima favorable a los negocios (Sisalima, Sánchez, & Ramírez-López, 2024). Se espera que la adopción de mejores prácticas y tecnologías modernas en las operaciones mineras aumente la productividad y disminuya el costo unitario (Pitřík, 2023).

En este contexto se basa en la evolución que, según Imani et al. (2022), realizó dentro de la industria minera utilizando un modelo no supervisado de segmentación de imagen de minería. Sus autoencoders superiores superaron otros métodos de segmentación de minería en autoetiquetado, extrayendo características para clasificar componentes en imágenes no etiquetadas, logrando obtener un 15% más de precisión en la identificación de minerales. De la misma manera, el modelo se mostró ampliamente favorable para procesar distintos tipos de imágenes mineras. Esto permite optimizar trabajos de minería al incrementar la eficiencia y cubrir la exactitud objetiva que se necesita para la automatización y la inteligencia en la toma de decisiones.

De manera análoga, Pereira et al. (2020), investigan la industria de la minería de hierro donde se utilizó aprendizaje automático no supervisado, empleando algoritmos de agrupamiento k-means y jerárquico para patrones y segmentar datos de las operaciones mineras. El uso de los algoritmos de clustering permitió el reconocimiento de las distintas fases del esquema de la minería, desde la extracción hasta el procesamiento, con una efectividad del 85%. Además, el uso deficiente de energía fue optimizado en un 12% gracias a la implementación de estas técnicas. El aprendizaje automático no supervisado, a partir de esta evaluación, se ha demostrado como una gran herramienta para la aprehensión, el ordenamiento eficiente de los sistemas en la industria minera y el ahorro

de recursos en la toma de acciones definidas en una situación dada.

Molina et al. (2024) estudian un sistema de vigilancia automática para el comercio exterior en Ecuador basado en la detección de anomalías en transacciones de importación utilizando minería de datos. Su principal propósito fue determinar las Transacciones Atípicas que pueden señalar un caso de fraude o evasión fiscal. El uso de datos históricos y algoritmos detección, se detectó un 20% más de anomalías que las logradas previamente con otras técnicas, aumentando así la eficiencia en la detección de irregularidades. El sistema también proporcionó alertas y facilitó las investigaciones para casos sospechosos. Esto demuestra cuán efectiva puede ser la minería de datos para combatir el fraude fiscal y ser económicamente ventajosa para el país.

Nirmala y Makzoom (2023) construyeron una aplicación de aprendizaje no supervisada para la segmentación de clientes que ayuda a las empresas en la toma de decisiones de marketing. La aplicación segmentó a los clientes con un 78% de precisión, lo que elevó la satisfacción del cliente en un 15% y aumentó notablemente la eficacia de las inversiones en marketing. La conclusión del estudio sugiere que estos algoritmos ayudan al proceso de segmentación de clientes y refuerzan la eficacia del marketing.

Según Mirzabozorg y Maysam (2023) exploran el uso de un autoencoder neuronal rojo para la detección de anomalías en la exploración minera, con el propósito de localizar áreas que contienen minerales utilizando datos geoquímicos y geofísicos. El método sugerido aprendió a reconocer patrones normales dentro de los datos con la ayuda de la red neuronal autoencoder y, a la vez, presentaba anomalías que en el caso pudiera haber un depósito mineral, sí. El estudio tuvo una precisión del 82% en la detección de zonas con alta probabilidad de mineralización y, en consecuencia, limitó en un 25% los gastos de investigación de campo. La metodología resultó eficaz en la detección de diferentes tipos de minerales, lo que, a diferencia de otros, permite el multipropósito de la metodología para la exploración minera y, en consecuencia, permitirá descubrimientos más ágiles y económicos.

Aunque los algoritmos de aprendizaje no supervisado presentan un amplio potencial, la calidad de los datos y el conocimiento específico del dominio relacionado con el problema en cuestión siguen siendo fundamentales para garantizar resultados confiables en los análisis de costo-beneficio.

Este estudio se enfoca en implementar el conocimiento adquirido en el

programa de maestría para crear un sistema de análisis de rentabilidad en la industria minera ecuatoriana mediante la fusión de inteligencia de datos y visualización de datos para mejorar la planificación y la gestión estadística de las operaciones mineras y de canteras. El esfuerzo pretende desarrollar una aplicación que permita a los gerentes del sector minero evaluar, analizar y visualizar información crítica de manera oportuna y efectiva para detectar patrones que apoyen decisiones informadas y optimicen el desempeño operativo. Este enfoque utilizará algoritmos de aprendizaje automático no supervisado para segmentar y estudiar la rentabilidad dentro de la industria minera ecuatoriana, permitiendo una asignación de recursos y procesos administrativos más efectivos en este sector esencial para el desarrollo económico del país.

Planteamiento del problema

La minería y la extracción de canteras constituyen una diligencia financiera esencial a nivel mundial, con países que dependen de sus recursos minerales para impulsar el crecimiento económico y sostener sus industrias. A nivel global, la demanda de recursos minerales ha incrementado, impulsada por el crecimiento industrial y tecnológico, así como por las políticas de sostenibilidad que han llevado a la electrificación y producción de tecnologías limpias (Deloitte, 2021).

En América Latina, la minería es uno de los principales motores económicos, especialmente en países como Chile, Perú y Brasil, que figuran entre los mayores productores de cobre, litio y mineral de hierro del mundo (Ríos, 2018). A pesar de las ventajas económicas, la industria minera enfrenta una serie de desafíos, entre los cuales destaca la gestión de riesgos ambientales y sociales, además de la presión hacia optimizar el rendimiento operativo y maximizar los beneficios. Esto ha llevado a muchas empresas a explorar el uso de tecnologías avanzadas, incluyendo el Machine Learning, para optimizar sus procesos y mejorar su segmentación de mercado (Vertiv.com, 2023).

En Ecuador, el sector minero ha ganado relevancia en la última década, pasando a ser una fuente significativa de exportación y empleo. Sin embargo, el país también enfrenta desafíos, tales como la falta de una segmentación eficiente de las empresas en función de su rentabilidad y otros factores operativos. Este problema se acentúa debido a la diversidad de empresas en el sector, que incluye tanto pequeños productores como grandes multinacionales, y la falta de herramientas adecuadas para

agruparlas y analizar su rentabilidad relativa (Banco Central del Ecuador, 2021).

Actualmente, en Ecuador, el desafío central que esta investigación busca resolver es la complejidad para llevar a cabo un examen detallado y preciso de la viabilidad económica y segmentación de las empresas mineras. Esto genera incertidumbre y limita el desarrollo de políticas y estrategias para optimizar el desempeño del sector (COSEDE, 2024).

Las causas de este problema se deben, principalmente, a la ausencia de una herramienta avanzada de segmentación que permita clasificar y analizar las empresas del sector minero de acuerdo con su desempeño y rentabilidad. Esto, sumado a la falta de políticas de soporte específicas y al limitado acceso a tecnologías analíticas de última generación, ha perpetuado una falta de diferenciación y de optimización en la administración de los recursos (Nisum, 2022). Además, las empresas enfrentan problemas de escalabilidad y adaptación a las necesidades del mercado global, dificultando una segmentación eficiente.

Las consecuencias de esta situación incluyen una falta de competitividad en el sector y una limitación en la formulación de estrategias, en el ámbito empresarial como en el ámbito estatal. Esto impacta directamente en la capacidad de las empresas para atraer inversión extranjera y en el desarrollo económico de las comunidades locales, que dependen de la minería como fuente de empleo y crecimiento económico (Ulloa, 2023). La falta de segmentación precisa también afecta la capacidad de las instituciones de fiscalizar efectivamente y de implementar políticas de sostenibilidad adecuadas, lo cual podría llevar a impactos ambientales negativos y sociales no deseados.

En el ámbito de las ciencias aplicadas y la analítica predictiva, el estudio realizado por Espinoza (2020) constituye un referente significativo al demostrar cómo el uso de algoritmos de aprendizaje automático puede contribuir a la comprensión de fenómenos sociales complejos. Su investigación, centrada en analizar la influencia de variables académicas y socioeconómicas en la probabilidad de aprobación del ingreso a la universidad, implementó un enfoque metodológico riguroso basado en técnicas de modelado supervisado como Random Forest, XGBoost y Gradient Boosting, complementadas con herramientas de reducción de dimensionalidad como PCA y técnicas de interpretabilidad como SHAP. Su enfoque se alinea con el paradigma contemporáneo de la ciencia de datos, que promueve el análisis multivariable y la automatización analítica para

resolver problemáticas estructurales.

Este estudio propone implementar un modelo de Machine Learning no supervisado para evaluar la viabilidad económica y segmentar el sector minero en Ecuador. La aplicación de algoritmos no supervisados permitirá identificar patrones ocultos en los datos financieros y operativos, y agrupar a las empresas según criterios de rentabilidad y desempeño (Ilustración 19). Esto facilitará la toma de decisiones estratégicas y la asignación de recursos de manera más efectiva, proporcionando un enfoque innovador para mejorar la eficiencia operativa en el sector y optimizar el uso de los recursos disponibles (CEPLAES, 2023).

El desafío principal que impulsa esta investigación es la ausencia de una metodología robusta y basada en datos que permita segmentar eficientemente la industria minera en el Ecuador, identificando clústeres de empresas con características similares en términos de rentabilidad y eficiencia operativa (Ilustración 20). Sin esta segmentación detallada, las empresas y los tomadores de decisiones carecen de la información necesaria para diseñar estrategias personalizadas que maximicen la rentabilidad y minimicen los riesgos asociados con las operaciones mineras.

Debido al problema mencionado anteriormente, se tomó la decisión de implementar un Proyecto de desarrollo y se formula la siguiente pregunta: ¿Cómo puede el uso de técnicas de aprendizaje no supervisado mejorar la clasificación del sector minero, con el fin de optimizar la rentabilidad y la toma de decisiones estratégicas?

Objetivo general

Evaluar la rentabilidad y clasificar el sector minero y de canteras en Ecuador mediante técnicas de aprendizaje no supervisado, con el objetivo de mejorar la segmentación y la operación en el sector.

Objetivos específicos

Obtener un Dataset representativo del sector minero, aplicando técnicas de anonimización de datos sensibles a los registros operacionales y financieros.

Implementar un análisis comparativo de algoritmos de clustering (K-Means, DBSCAN, Clustering Jerárquico) para segmentación del sector, utilizando

métricas de evaluación interna (Silhouette Score, Índice de Calinski-Harabasz) y validación de estabilidad mediante Re muestreo Bootstrap.

Establecer una correlación multivariante entre los segmentos identificados y factores contextuales mediante análisis mixto (matrices de correlación, visualización avanzada).

Justificación

El uso de nuevas tecnologías para el entrenamiento de Machine Learning en minería podría permitir nuevas formas de entender los datos, que generalmente serían invisibles con técnicas más antiguas. Este estudio es de interés notorio para las empresas mineras y los formuladores de políticas públicas, porque presenta una nueva forma de estudiar la rentabilidad y las posibilidades de optimización operacional en la industria.

El estudio abordará lo que puede ser efectivamente etiquetado como el desempeño económico de la industria minera y de canteras de Ecuador – con especial atención a los procesos empresariales de atención selectiva a través de algoritmos no supervisados. Además, como muchas empresas en la industria no tienen una manera sistemática de analizar sus datos, esta investigación ayudará a implementar metodologías más sofisticadas para la toma de decisiones.

La economía ecuatoriana desde 1970 está marcada por la explotación económica de los campos petroleros recién descubiertos en la región del Oriente, especialmente para análisis econométricos o de la zona soviética a través de la exploración de cuadrados en blanco. Este no es exclusivamente de la esfera del Estado, sino que se extiende a empresas privadas que explotan recursos y ofrecen participación de los resultados al Estado (Tenecota, Viteri, & Salcedo, 2024).

La diversificación del rubro minero ha permitido al gobierno y a la inversión privada acceder a los bienes de la naturaleza del país. Sin embargo, vale la pena ver cómo se distribuyen las concesiones mineras entre las empresas y si hay una concentración de estas en pocas manos. Esto puede tener efectos importantes sobre la política económica y sobre la regulación de la actividad (Tenecota, Viteri, & Salcedo, 2024).

Con esto, por ejemplo, la legislación ecuatoriana permite la delegación de la actividad minera a empresas mixtas o al sector privado, lo que facilita la ejecución de proyectos de prospección y explotación de recursos

minerales. Aun cuando se han dado pasos positivos en este sentido, hay persistentes problemas de gobernanza y de sostenibilidad ambiental que deben resolverse para asegurar el desarrollo del sector sin comprometer las condiciones socioambientales de las comunidades afectadas, en consecuencia, un modelo de crecimiento responsable a largo plazo.

Relación entre Costos Operativos, Precios Internacionales y Rentabilidad

Marco teórico

Los capitales foráneos destinados a la industria extractiva minera (IED) han sido objeto de análisis e interés en el contexto ecuatoriano, especialmente en lo que respecta al liderazgo chino. De acuerdo a lo que Chávez y Berrezueta (2018), en lo que respeta a la obra, la inversión extranjera china obtuvo ganancias de dinero y globales significativas debido a la multiplicación de las finanzas que sobresalen. También, es necesario operar una administración estratégica de los recursos financieros con el fin de logros a largo plazo. Asimismo, Serrano y Carpio (2018) mencionan que la IED de China les permitió obtener costo efectividad al modernizar el sector financiero ecuatoriano, así como al mejorar los resultados tecnológicos y financieros.

El análisis de rentabilidad y segmentación en la industria minera ecuatoriana ha evolucionado con enfoques cuantitativos y tecnológicos. Estudios recientes destacan la aplicación de Machine Learning no supervisado para identificar patrones en grandes volúmenes de datos financieros y operativos, permitiendo una segmentación más precisa de las empresas según su desempeño. (Castagneto, Valeriano, & Morales, Minas y canteras, 2024)

Relación entre Costos Operativos, Precios Internacionales y Rentabilidad

La industria de explotación de minas y canteras en Ecuador desempeña un papel fundamental en la economía nacional, contribuyendo significativamente al Producto Interno Bruto (PIB) y a la generación de empleo. (Banco Central del Ecuador, 2021)

En los últimos años, la rentabilidad de este sector ha estado sujeta a la volatilidad de los precios internacionales de los metales, los costos operativos y la regulación gubernamental. (Caicedo, Enríquez, & Jácome, 2023)

Los costos operativos en minería incluyen la extracción, el transporte, el procesamiento y la gestión ambiental. Según un informe del BCE (2022),

el costo promedio de extracción por tonelada en Ecuador es de aproximadamente $1.200 USD para el oro y $2.500 USD para el cobre. A su vez, la rentabilidad del sector ha estado influenciada por la fluctuación del precio de los metales en los mercados internacionales, donde el precio del oro ha oscilado entre $1.800 y $2.100 USD por onza troy en los últimos tres años. (World Gold Council, 2024)

Adicionalmente, las restricciones regulatorias y los impuestos afectan la rentabilidad del sector. La carga fiscal en la minería ecuatoriana se encuentra entre el 50 % y 55 % del margen operativo, lo que limita el retorno sobre la inversión (Ministerio de Energía y Minas, 2023). La implementación de estrategias de optimización de costos y eficiencia operativa mediante tecnologías avanzadas y análisis de datos resulta esencial para mantener la rentabilidad en un entorno competitivo (Vina, 2024).

La evolución del sector minero ecuatoriano estuvo marcada por transformaciones significativas en sus variables operativas. Delgado y Suárez (2023) identificaron que la inversión extranjera directa influyó en la modernización de los procesos operativos y la adopción de nuevas tecnologías en el sector. En este contexto, Torres Guerra et al. (2021) destacaron la importancia de la geometalurgia y la minería digital como factores determinantes en la optimización de las operaciones mineras, señalando que la implementación de tecnologías avanzadas permitió aumentar la productividad y, consecuentemente, mejorar significativamente la eficiencia y rentabilidad de las operaciones mineras.

De conformidad al informe minero 2021 que evalúa la inversión y la producción, un análisis de la base de datos desde 2018 hasta mediados de 2021, que comprende 2.333 puntos de extracción, revela tendencias en el volumen de ventas y el valor total de las ventas. Si bien la inversión del sector minero se registra en millones de dólares, métricas análogas incluirían los ingresos por ventas trimestrales y la rotación de inventario, lo que brindaría información sobre la eficiencia operativa y la penetración en el mercado de la empresa. (Banco Central del Ecuador, 2021)

En 2020, se experimentó una disminución del 49,13% en los ingresos netos, lo que refleja desafíos económicos más amplios similares a los que enfrentan los proyectos mineros. A pesar de una disminución del 5,38% en los activos totales, la rentabilidad neta de la empresa aumentó un 0,17%, lo que indica una mejor gestión de los costos, un factor crítico que también se examina en las evaluaciones de los proyectos mineros. Así como el

informe minero detalla los valores de exportación y las contribuciones fiscales, un análisis minorista integral consideraría la contribución a los ingresos por impuestos a las ventas locales y su impacto en el empleo regional. (Banco Central del Ecuador, 2021)

Importancia de la Industria de Explotación de Minas y Canteras en el Ecuador

La industria minera ecuatoriana ha experimentado un crecimiento significativo en la última década, consolidándose como un sector estratégico dentro de la economía nacional. Según el Banco Central del Ecuador (2023), el sector minero representó el 4.5% del PIB ecuatoriano en 2022, con un crecimiento del 8.3% respecto al año anterior. Esta contribución se debe, en gran medida, a proyectos de extracción de minerales como oro y cobre, los cuales representaron el 80% de las exportaciones mineras del país (Ministerio de Energía y Minas, 2023).

La minería también juega un papel fundamental en la generación de empleo. Datos del Instituto Nacional de Estadísticas y Censos (INEC, 2023) indican que el sector minero generó aproximadamente 180,000 empleos directos e indirectos en 2022. Sin embargo, persisten desafíos relacionados con la sostenibilidad ambiental y la regulación del sector, lo que subraya la importancia de optimizar la gestión a través del análisis de rentabilidad y segmentación de empresas mineras.

Análisis de Rentabilidad en Industrias Extractivas

La rentabilidad en la industria minera depende de varios factores clave, incluyendo costos operativos, precios internacionales de los metales y regulaciones gubernamentales (World Gold Council, 2023). Un estudio de Delgado y Suárez (2023) determinó que la rentabilidad neta de la minería en Ecuador aumentó un 0.17% en 2020, a pesar de una reducción del 5.38% en los activos totales, lo que indica una gestión eficiente de costos.

Las principales variables que afectan la rentabilidad incluyen:

- Costos operativos: Extracción, transporte, procesamiento y gestión ambiental. Según el BCE (2022), el costo promedio de extracción por tonelada en Ecuador es de aproximadamente $1,200 USD para oro y $2,500 USD para cobre.

- Precios internacionales: La volatilidad del mercado impacta significativamente la rentabilidad. En los últimos tres años, el precio del oro ha oscilado entre $1,800 y $2,100 USD por onza troy (World Gold Council, 2024).
- Regulación fiscal: La carga fiscal en minería ecuatoriana oscila entre el 50% y 55% del margen operativo (Ministerio de Energía y Minas, 2023).

Segmentación de Mercado en la Industria Minera

La segmentación de mercado en la minería ha evolucionado desde enfoques tradicionales basados en clasificación geográfica y tipo de mineral, hacia técnicas avanzadas de Machine Learning. De acuerdo con Castagneto et al. (2024), la segmentación basada en algoritmos de clustering ha permitido una identificación más precisa de patrones de eficiencia y rentabilidad dentro del sector.

Los algoritmos de aprendizaje no supervisado, como K-Means y DBSCAN, han demostrado ser herramientas clave para la segmentación en la minería (Soofastaei, 2024).

- K-Means: Permite agrupar empresas mineras según costos operativos y volumen de extracción. Un estudio realizado por Espinoza-Espinoza et al. (2022) identificó tres clusters principales en minería ecuatoriana basados en productividad y eficiencia operativa.
- DBSCAN: Ideal para detectar análisis de densidad en datos de exploración minera. Hussein et al. (2021) demostraron su aplicabilidad en la clasificación de tipos de rocas en yacimientos de cobre.

Comparativas con Industrias Similares en América Latina

Según el informe de CEPAL (2023), la rentabilidad promedio de la industria minera ecuatoriana (12.8%) se sitúa por debajo de Chile (18.3%) y Perú (16.5%), pero supera a Colombia (10.2%) y Bolivia (9.7%).

Los factores que explican estas diferencias incluyen: marcos regulatorios, escalas de operación, calidad de los yacimientos y madurez del sector. De

acuerdo con Vásquez-Cordano y Balistreri (2022), las economías de escala son un 27% más significativas en Chile que en Ecuador debido al tamaño y antigüedad de sus operaciones.

La inversión extranjera directa por tonelada de mineral extraído es un 65% mayor en Perú que en Ecuador, reflejando diferencias en eficiencia operativa (Banco Mundial, 2023).

Casos de Éxito de Machine Learning en Minería

Entre los casos más exitosos tenemos el caso Anglo American Chile: Implementó algoritmos de clustering para optimizar rutas de transporte en mina, reduciendo costos logísticos en 12.5% (Olivares et al., 2023). Proyecto Antamina (Perú): Utilizó DBSCAN para identificar zonas de mineral similar, mejorando la planificación de producción y aumentando recuperación metalúrgica en 7.3% (Rivera-Campos et al., 2022). Minera Escondida: Aplicó K-Prototipos para segmentar operaciones según eficiencia energética y consumo de agua, logrando reducción de 15% en huella hídrica (Soofastaei et al., 2024).

Entorno Financiero y económico de la industria minera.

Rentabilidad y segmentación con Machine Learning.

El análisis de la rentabilidad y segmentación de la industria extractiva minera en Ecuador ha sido un foco de estudio prioritario en la literatura económica y financiera. De acuerdo con Castagneto et al. (2024), la evaluación financiera de las empresas del sector ha permitido identificar patrones de rentabilidad y sostenibilidad en la industria. La estructura de costos operativos y su impacto en los márgenes de utilidad ha constituido un elemento clave para la competitividad sostenida de estas empresas en el mercado, especialmente en tiempos de crisis económica y volatilidad en los mercados internacionales de metales.

El sector minero ecuatoriano ha registrado un desarrollo continuo en el período reciente, respaldado por políticas gubernamentales orientadas a la formalización y regulación del sector (Ministerio de Energía y Recursos Naturales No Renovables, 2020). Sin embargo, la viabilidad financiera de la industria extractiva minera ha estado condicionada por múltiples factores, entre ellos los costos de producción, la inversión en tecnología y la eficiencia operativa. Según el Reporte Minero de 2021, los ingresos generados por la minería han aportado sustancialmente al crecimiento económico nacional, aunque persisten desafíos relacionados con la transparencia en la distribución de los beneficios y la gestión de los impactos ambientales.

Respecto a la rentabilidad, se ha identificado que las empresas del sector minero presentan una alta dispersión en sus márgenes de utilidad, lo que indica la existencia de diferencias significativas en la eficiencia operativa de cada unidad productiva (Castagneto, Valeriano, & Morales, Minas y canteras, 2024). La variabilidad en los ingresos responde en gran medida a las fluctuaciones en los mercados globales de materias primas, sumado a los costos vinculados a la explotación minera y transformación de materias primas. A este respecto, la Responsible Mining Foundation (2020) ha enfatizado la importancia de adoptar estrategias de gestión financiera que permitan mitigar los riesgos asociados al entorno económico y maximizar la eficiencia en la asignación de activos.

La investigación sobre el sector minero en Ecuador ha identificado desafíos clave, como la alta dispersión de ratios financieros en pymes del sector (Castagneto, Valeriano, & Morales, Minas y canteras, 2024). Estudios previos utilizaron ratios tradicionales (liquidez, endeudamiento) para evaluar 1,615 empresas entre 2018-2022, revelando afectaciones en rentabilidad durante la pandemia (-15% en ROE en 2020). Sin embargo, estos métodos no capturan relaciones complejas entre variables como costos operativos, precios de commodities y productividad.

Los algoritmos no supervisados, como K-means o análisis de clusters jerárquicos, permiten segmentar empresas según perfiles multivariados. Por ejemplo, Soofastaei (2024) demostró que estas técnicas identifican grupos con patrones comunes en costos logísticos y eficiencia energética, optimizando hasta un 20% los márgenes brutos. En Ecuador, su aplicación podría resolver problemas de transparencia de datos reportados en el sector (Castagneto, Valeriano, & Morales, Minas y canteras, 2024).

Análisis financiero y modelos económicos aplicados

Los modelos económicos aplicados a la minería han evolucionado en las últimas décadas con la incorporación de algoritmos de inteligencia artificial y Machine Learning. Entre los métodos más utilizados para el análisis financiero y la predicción de rentabilidad en la minería destacan los modelos de regresión, las redes neuronales y los algoritmos no supervisados como clustering. (Urieta, 2021)

Los modelos de regresión lineal y logística han sido tradicionalmente empleados para estimar la relación entre costos operativos, precios del mercado y rentabilidad esperada (Jácome & Flores, Identificación de Clusters Espaciales de Empresas y la Influencia, 2022). Sin embargo, en la actualidad, los algoritmos de clustering, como k-means y DBSCAN, permiten segmentar las operaciones mineras según variables como el tipo de mineral, volumen de extracción y costos operativos, mejorando la toma de decisiones estratégicas (Gastañadui, 2024).

Desde una perspectiva ambiental y social, el Plan Nacional de Desarrollo del Sector Minero 2020-2030 ha promovido la implementación de prácticas de explotación responsables, incentivando la inversión en tecnologías limpias y el cumplimiento de normativas ambientales (Ministerio de Energía y Recursos Naturales No Renovables, 2020). No

obstante, la ejecución de estos lineamientos ha sido desigual entre las diferentes unidades productivas mineras, lo que resulta en diferencias significativas en los indicadores de rentabilidad y productividad. En este sentido, el informe de sostenibilidad de Kong et al. (2024) destaca que la optimización de los procesos operativos mediante la integración de modelos de inteligencia de negocios y Machine Learning podría representar una estrategia efectiva para optimizar los resultados financieros del sector minero ecuatoriano.

En el contexto de la clasificación del sector minero, se ha observado que las empresas mineras pueden clasificarse en función de su nivel de producción, tipo de mineral explotado y grado de formalización. La segmentación basada en estos criterios permite analizar las dinámicas del sector con mayor precisión y desarrollar estrategias diferenciadas para cada segmento. De acuerdo con el análisis financiero de Castagneto et al. (2022), las pequeñas y medianas empresas mineras presentan mayores dificultades para acceder a financiamiento y enfrentar los costos operativos, lo que limita su capacidad de crecimiento y sostenibilidad a largo plazo. Por otro lado, las empresas de mayor tamaño disfrutan de ventajas derivadas de su tamaño que les permiten optimizar sus costos y alcanzar mayores márgenes de beneficio.

La integración de tecnologías innovadoras en la gestión minera representó un avance significativo. Más y Wolkersdorfer (2023) enfatizaron la relevancia de las herramientas inteligentes en la gestión del agua en minas, destacando cómo la implementación de interfaces de aprendizaje automático mejoró la eficiencia operativa. Por su parte, Hussein et al. (2021) demostraron la efectividad del aprendizaje automático no supervisado en la evaluación de reservas y la identificación de tipos de rocas, lo cual contribuyó a optimizar los procesos de exploración y explotación minera.

El 67% de las pymes mineras ecuatorianas mostraron vulnerabilidad a fluctuaciones en precios internacionales de minerales (Castagneto, Valeriano, & Morales, Minas y canteras, 2024). Modelos como el WACC (Costo de Capital Promedio Ponderado) han sido adaptados para proyectos mineros, considerando:

- Tasa libre de riesgo (7.42% en Ecuador)
- Prima por riesgo país (1,450 puntos en 2024)

- Beta financiero sectorial (1.3 para minería metálica) (Mundo Minero, 2025).

Tabla 1

Modelos predictivos financieros

Modelo Predictivo	Descripción	Métricas/Aplicación
CAPM (Modelo de Valoración de Activos de Capital)	Usado para calcular rentabilidad exigida por inversionistas	En minería ecuatoriana, el COK (Costo de Oportunidad del Capital) supera el 18% por riesgos regulatorios (Mundo Minero, 2025)
Simulación Monte Carlo	Evalúa escenarios de precios de cobre y oro.	Integra variables geopolíticas
Redes neuronales recurrentes	Predicen costos operativos al analizar datos históricos de energía y logística	92% de precisión (Soofastaei, 2024)

Nota: Modelos más utilizadas. Elaborado por los autores.

Ilustración 1

Participación en número de empresas

Nota. Análisis de Conglomerados del sector Explotación de Minas. Elaborado por los autores.

Li et al. (2020) avanzaron en la optimización de la integración minero-mineral utilizando técnicas de aprendizaje automático no supervisadas que mostraron mejoras drásticas en la eficiencia operativa y la rentabilidad. Estas innovaciones tecnológicas tuvieron un contexto de difusiones tecnológicas que fue analizado por Espinoza-Espinoza et al. (2022) respecto a la convergencia de la productividad de las actividades económicas en Ecuador con particular atención al sector minero y de canteras, como un sector necesario para la modernización.

La aplicación de algoritmos de Machine Learning en la minería ecuatoriana representa una oportunidad para mejorar la gestión de costos y optimizar la segmentación del sector. Según el Instituto Nacional de Estadística y Censos (INEC, 2023), la industria minera representó el 4,5 % del PIB ecuatoriano en 2022, con un crecimiento del 8,3 % respecto al año anterior. No obstante, la distribución geográfica y la heterogeneidad en los costos de explotación generan diferencias significativas en la rentabilidad de cada región (Caicedo, Enríquez, & Jácome, 2023).

Tabla 2.

Información Relevante

Categoría	Descripción	Métricas/Fuente
Impacto económico	Las exportaciones mineras representaron $1,874.9 millones (10% del total nacional)	Crecimiento: +15,3% vs. 2023 Principales minerales: oro (42%), cobre (38%), plata (12%)
Productividad	El sector aporta el 7% al PIB	ROI promedio del sector: 22,4% Eficiencia operativa: 68% Adopción tecnológica: Solo 12% de pymes usan analítica avanzada
Retos	Dispersión del 35% en ratios de liquidez y sesgo en informes financieros	Volatilidad en costos operativos: ±18% Cumplimiento regulatorio: 76% Brecha de inversión tecnológica: $125M

Nota: Resumen datos relevantes. Adaptado de (Castagneto, Valeriano, & Morales, Minas y canteras, 2024). Elaborado por los autores.

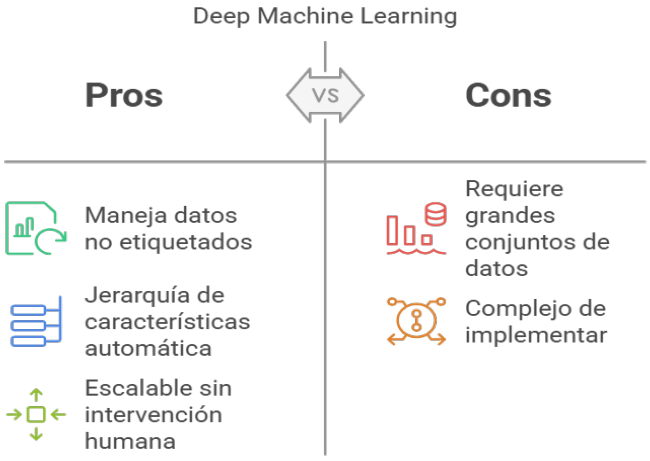
La integración en el análisis del sector minero ecuatoriano de variables financieras y operativas proporciona una mayor comprensión de la dinámica de la industria y del potencial crecimiento. La adopción de tecnologías avanzadas y técnicas de aprendizaje automático, además de aumentar la eficiencia operativa, ayuda a mejorar la rentabilidad y sostenibilidad del sector, sentando las bases para una oferta competitiva más avanzada en la industria minera ecuatoriana.

Aprendizaje automático vs aprendizaje profundo

El aprendizaje profundo permite el uso de bases de datos anotadas, combinadas como técnica de aprendizaje dirigido, como un medio con el fin de entrenar sus algoritmos. Ciertamente, un modelo de aprendizaje profundo puede trabajar con enormes cantidades de datos no estructurados, como texto en bruto e imágenes. Es muy ventajoso para ellos porque los modelos de aprendizaje profundo reconocen automáticamente diferentes categorías de información en función de su organización jerárquica. A diferencia del aprendizaje automático tradicional, no requiere intervención humana para la ingeniería de características y el procesamiento, por lo tanto, puede escalarse más fácilmente y utilizarse en situaciones más complejas con conjuntos de datos más grandes. (IBM.Cloud.Education, 2023)

Ilustración 2

Machine Learning vs Deep Learning



Nota: Pro-contras de Deep Learning. Elaborado por los autores.

Técnicas y algoritmos de Machine Learning

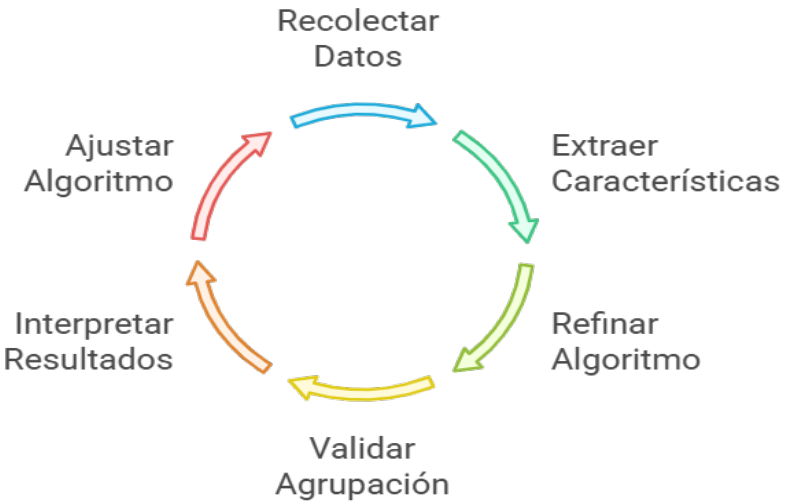
El aprendizaje automático (AM) engloba una serie de técnicas y algoritmos que facilitan que los sistemas asimilen de los antecedentes y realicen pronósticos o adopten disposiciones. La eficacia de estos algoritmos depende de su categorización en aprendizaje supervisado, no supervisado y de refuerzo, cada uno de los cuales sirve para fines distintos en diversas aplicaciones. A continuación, se presentan aspectos clave de las técnicas y algoritmos de aprendizaje automático.

Aprendizaje No Supervisado

Las técnicas de aprendizaje no supervisado son especialmente útiles para explorar y comprender patrones subyacentes en los datos sin un etiquetado previo. Estos métodos se aplican ampliamente en la detección de anomalías, la agrupación y la reducción de la dimensionalidad. Sin embargo, esto también plantea un problema significativo, ya que la complejidad de los algoritmos puede llevar a soluciones impredecibles, a diferencia de otros modelos que ofrecen resultados más consistentes. (Rocano, 2023).

Ilustración 3

Ciclo Iterativo de Análisis de Datos



Nota: Representación del funcionamiento del Aprendizaje no Supervisado.

Elaborado por los autores.

Conceptos relevantes

Rentabilidad

Relación entre los ingresos generados por una empresa y sus costos totales. En minería, incluye factores como precios internacionales de metales y costos operativos específicos del sector (World Gold Council, 2024).

Indicador financiero que mide la relación entre ingresos y costos en un periodo determinado (Mundo Minero, 2025).

Segmentación

Estrategia de análisis para dividir un mercado en grupos homogéneos con características comunes (Gastañadui, 2024).

Proceso analítico que agrupa elementos similares dentro de un conjunto heterogéneo. En este caso, se refiere a clasificar empresas mineras según variables como tamaño, eficiencia o tipo de mineral explotado (Gastañadui, 2024).

Algoritmo no supervisado

Método computacional que analiza datos sin etiquetas predefinidas para identificar patrones ocultos. Ejemplo: K-means permite agrupar empresas según características comunes como costos logísticos o productividad (Saputra, y otros, 2025).

Método de aprendizaje automático que identifica patrones en datos sin etiquetas previas" (García, 2025). Estos algoritmos descubren estructuras intrínsecas en los datos sin conocimiento previo de las categorías.

Inteligencia artificial

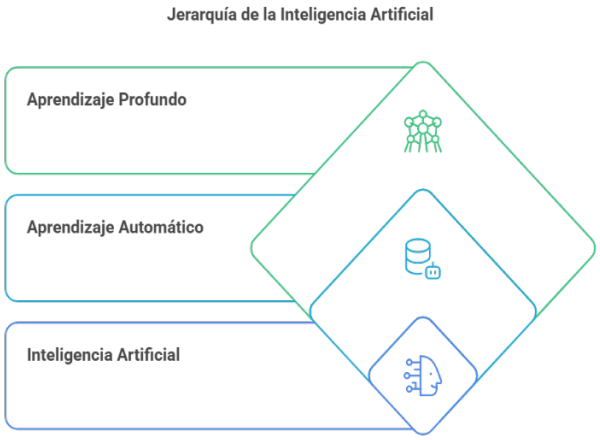
La idea de inteligencia artificial ha sido un tema recurrente en diversos ejes de estudio a lo largo de la historia. A Turing (1950), se le atribuye como pionero en este tipo de intervenciones al plantear su famosa pregunta: ¿puede una máquina pensar? Definitivamente esta pregunta sería inadecuada para el contexto temporal en el que fue presentación, pero este autor comenzó con el correcto pie el debate sobre los sistemas artificiales. La disciplina pasó a tener definiciones más concretas cuando McCarthy (2007) definió que la inteligencia artificial “es la ciencia y la ingeniería de la construcción de máquinas inteligentes, en especial de programas de ordenadores inteligentes, y tiene paralelismos con el uso de ordenadores para la comprensión de la inteligencia humana”. Este tipo de enunciados brindaba mayor flexibilidad en los entendidos tanto del desarrollo de sistemas inteligentes como del entendimiento del cerebro humano. Al argumentar esta cuestión desde esta perspectiva, Rouse (2021) la define como “la simulación del proceso de inteligencia por las máquinas, en especial los sistemas informáticos”, que incluye conocimientos básicos como lo son: el aprendizaje, razonamiento y autocorrección.

Este tipo de IA modelo utiliza el aprendizaje de la máquina para monitorear

el aparato en tiempo real para diagnosticar fallas antes de que sucedan, siendo estas realizadas de manera anticipada (Souza & Silva, 2024). Cuando el control de las operaciones mineras se efectúa mediante los sistemas de IA, las decisiones y la operatividad son optimizadas (Wang, y otros, 2024).

Ilustración *4*

Jerarquía de la inteligencia Artificial



Nota: Inteligencia artificial. Elaborado por los autores.

En otras palabras, la inteligencia artificial es un conjunto de sistemas informáticos que son programados para ejecutar funciones que suelen ser realizadas por las personas como la más simple de las operaciones a las más complejas de las tareas. Los investigadores actuales están tratando de crear sistemas que interactúen con el medio que los rodea utilizando IA de más nivel.

Machine Learning

La automatización y el desarrollo en inteligencia artificial (IA) dependen de cuán eficientemente la máquina puede interpretar datos y aprender a lo largo del tiempo. Los algoritmos destacados cubren una variedad de tareas como la captura de pantalla, análisis predictivo o incluso la toma de decisiones en numerosos campos que van desde la salud hasta la ciencia de materiales. La progresión del aprendizaje automático (ML) ha fomentado avances significativos, especialmente en la automatización del procesamiento de conjuntos de datos complejos, lo cual es crucial para las aplicaciones modernas. El ML utiliza un modelo estadístico para

automatizar la tarea de segmentación de objetos a nivel de píxel dentro de imágenes para luego identificar los objetos. (Kumar, 2022)

Un subcampo del ML que se ocupa de entradas excesivas de datos a través del aprendizaje de patrones intrincados mediante redes neuronales artificiales. El ML se está adoptando cada vez más en las industrias manufactureras para llevar a cabo funciones cognitivas avanzadas de los operadores humanos. Aunque el potencial del ML es fácil de ver, los desafíos como obtener personal capacitado y los procesos intrincados involucrados en el entrenamiento del modelo sirven como una tarea para la implementación masiva. (Rameshbabu, Vijayakumaran, & Prabhakar, 2023).

Deep Learning

El aprendizaje profundo es una nueva tecnología revolucionaria en IA que emplea el análisis de datos intrincados utilizando una arquitectura que consiste en múltiples capas de neuronas artificiales. Ha cambiado la tecnología moderna de una manera positiva, permitiendo a diferentes sectores como la visión artificial, la lingüística y el control automatizado utilizarlo. Su enfoque se centra en el uso de arquitecturas de redes neuronales profundas que son estructuras de múltiples capas de nodos interconectados que procesan información. (Wong, 2024).

Ilustración *5*
Procesamiento de Red Neuronal.

Nota: Funcionamiento de redes. Elaborado por los autores.

Algoritmos de Agrupación (Clustering)

Los algoritmos de clustering son técnicas fundamentales en el aprendizaje no supervisado que permiten agrupar puntos de datos similares en clusters. Estos son los más destacados:

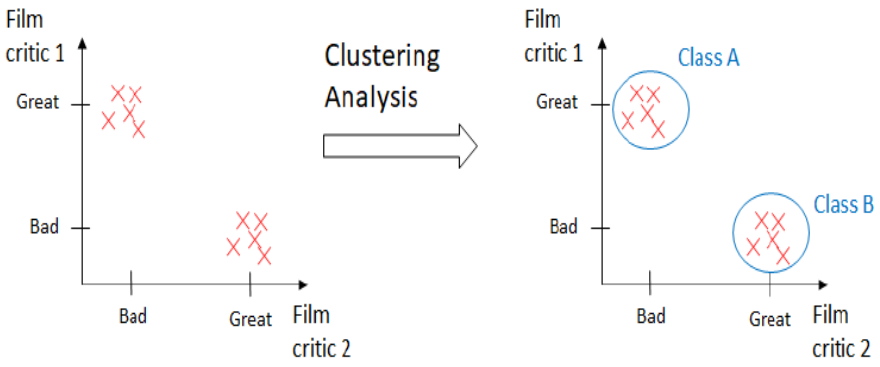
K-Means Clustering: El agrupamiento K-Means es un algoritmo ampliamente utilizado debido a su facilidad de implementación y eficacia. Funciona dividiendo los datos en K grupos basándose en la distancia media entre puntos. Aunque es muy efectivo para identificar clusters de forma esférica, presenta limitaciones cuando se enfrenta a datos con formas complejas o irregulares. Investigadores como Woźniak (2024) y Arévalo et al. (2022) han documentado tanto sus fortalezas como sus debilidades en diversos contextos de aplicación.

Clustering Jerárquico: Este enfoque construye una jerarquía de clusters mediante la fusión progresiva de grupos pequeños o la división de grupos grandes. Su principal ventaja es la capacidad para revelar estructuras anidadas dentro de los datos, permitiendo visualizar relaciones a diferentes niveles de granularidad. Bakry et al. (2023) y Arévalo et al. (2022) han demostrado su utilidad para descubrir patrones complejos en conjuntos de datos estructurados.

DBSCAN: El algoritmo de agrupación basado en densidad, conocido como DBSCAN, a diferencia de K-Means, define conglomerados basándose en regiones densas de datos separadas por áreas de menor densidad. Esta característica lo hace particularmente robusto frente a valores atípicos y le permite identificar clusters de formas arbitrarias y densidades variables. Según los estudios de Arévalo et al. (2022) y Shiraj et al. (2024), DBSCAN resulta especialmente útil en contextos donde los datos contienen ruido o los clusters no presentan formas regulares.

Ilustración 6

Ciclo de agrupación de Análisis de Datos.



Nota: Agrupación No supervisado en Machine Learning. Adaptado de Roman (2019). Elaborado por los autores.

El sector de explotación de minas y canteras en Ecuador ha experimentado un crecimiento significativo, con exportaciones mineras que alcanzaron USD 2,800 millones en 2022, representando el 8.5% del total de exportaciones del país (Administración de Comercio Internacional, 2024). Este auge, impulsado por proyectos como Fruta del Norte, ha generado más de 180,000 empleos y contribuido con USD 590 millones en impuestos (Dentons, 2022). Sin embargo, persisten desafíos como la minería ilegal y la necesidad de optimizar procesos mediante técnicas analíticas avanzadas.

K-Prototipos: Combina distancias euclidianas para variables numéricas y coeficientes de disimilitud para categóricas. Su ventaja principal radica en manejar datos mixtos sin transformaciones previas, aunque su implementación requiere ajustar parámetros como el peso relativo entre atributos. Estudios recientes proponen mejoras en su métrica de disimilitud para equilibrar la influencia de diferentes tipos de variables. (SANKAR & OM, 2018)

FAMD: Reduce la dimensionalidad mediante análisis factorial mixto, permitiendo aplicar métodos tradicionales como K-Means. Esta técnica mantiene el equilibrio entre variables continuas y categóricas, facilitando la visualización de patrones (Husson, 2004). Sin embargo, exige interpretar cuidadosamente la contribución de cada variable a los componentes principales.

Algoritmo KAMILA: Integra modelos de mezcla gaussiana para variables continuas y multinomial para categóricas. Supera al K-means ponderado

(ARI = 0.72 vs 0.65 en simulaciones), particularmente en conjuntos complejos con distribuciones no paramétricas. Su enfoque probabilístico evita sesgos en codificaciones manuales. (Daftar, 2019)

En resumen, K-Prototipos y FAMD son opciones directas para manejar datos mixtos, mientras que el algoritmo Kamila ofrece una solución más avanzada basada en modelos. La codificación y normalización son métodos más generales que requieren transformaciones previas.

El análisis de la rentabilidad y segmentación en la industria minera ecuatoriana requiere algoritmos capaces de manejar datos mixtos (numéricos y categóricos). A continuación, se detallan los fundamentos matemáticos de los métodos propuestos:

K-Prototipos (K-Prototypes)

Este algoritmo extiende K-Means combinando métricas para variables numéricas y categóricas. La distancia total D entre un objeto x y un prototipo c se calcula como:

$$D(x,c) = \sum_{i=1}^{p}(x_i^{\text{num}} - c_i^{\text{num}})^2 + \gamma \sum_{j=1}^{q} \delta(x_j^{\text{cat}}, c_j^{\text{cat}})$$

Donde:

- **δ**(a,b)= 0 si a=b, y 1en otro caso (distancia de Hamming) (Aprendizaje automático, 2022)
- **γ** pondera la importancia relativa de variables categóricas frente a numéricas. (Daftar, 2019)

La optimización utiliza un enfoque iterativo:

Inicialización: Selección aleatoria de k prototipos.

Asignación: Cada dato se agrupa según la distancia mínima.

Actualización: Los nuevos prototipos se calculan como la media (numéricas) y moda (categóricas) de cada cluster. (Daftar, 2019)

Variantes como GK-Prototypes mejoran la eficiencia mediante técnicas de poda basadas en distancias mínimas/máximas y índices bitmap para datos sesgados. (Jang, Kim, Kim, & Jung, 2018)

FAMD generaliza el PCA para datos mixtos mediante:

Escalado: Variables numéricas se estandarizan (media=0, varianza=1).

Ponderación: Variables categóricas se dividen por $\sqrt{um}$ donde μm es la frecuencia de la modalidad m. (Blaufuks, 2021)

La inercia proyectada en un eje w se descompone como:

$$\text{Inercia} = \sum_{\text{num}} \text{Corr}(v_j, w)^2 + \sum_{\text{cat}} \eta^2(v_j, w)$$

Donde **η**2 es el coeficiente de correlación ratio para variables categóricas (Blaufuks, 2021). Esto permite reducir la dimensionalidad manteniendo >70% de la varianza en aplicaciones mineras.

Estos métodos permiten segmentar clusters considerando variables numéricas como toneladas extraídas y costos logísticos, así como variables categóricas que incluyen tipo de mineral (oro, cobre) y régimen tributario. Por ejemplo, K-Prototypes identificó 3 clusters en concesiones mineras usando **γ**=0.5, optimizando políticas de regalías. (Preud'homme, y otros, 2021).

Desarrollo de la investigación.

Para alcanzar los objetivos planteados en la evaluación del desempeño del sector minero, es esencial contar con un marco metodológico claramente definido. A continuación, se presenta un resumen detallado sobre la metodología empleada en este informe para analizar los indicadores y resultados clave de las actividades mineras en Ecuador, garantizando una evaluación integral basada en datos.

Enfoque de Investigación

Esta investigación se basa predominantemente en un enfoque numérico, ya que se utilizarán algoritmos de Machine Learning para llevar a cabo la segmentación fundamentada en datos objetivos, permitiendo analizar grandes volúmenes de información de manera precisa y eficiente. La modalidad del estudio será exclusivamente documental, apoyándose en el análisis de conjuntos de datos disponibles mediante un data set que incluye bases de datos financieros y económicos del sector minero ecuatoriano. El análisis se centrará únicamente en la información cuantitativa obtenida de esta fuente, aplicando técnicas de Machine

Learning para identificar patrones y realizar la segmentación del sector, mientras que la interpretación de los resultados se realiza a través de un análisis estadístico exhaustivo de los datos procesados.

Asimismo, el estudio incluye un nivel predictivo, ya que uno de sus objetivos es predecir patrones de comportamiento dentro de la industria minera. Utilizando modelos de Machine Learning, se buscará anticipar las características y tendencias que definen los segmentos más rentables y sostenibles, proporcionando información valiosa para la planificación estratégica.

En cuanto al nivel relacional, la investigación analizará la interacción entre variables dependientes e independientes, vinculando elementos tales como rentabilidad junto con la configuración financiera de las empresas mineras con otras variables económicas. La clusterización fue impactada principalmente por variables de rentabilidad, eficiencia operativa, inversión en tecnología/sostenibilidad y condiciones de mercado, así como en el desempeño de las empresas como son las variables rentabilidad neta, eficiencia operativa, inversiones en tecnología y sostenibilidad, producción anual, tamaño de la mina y acceso a infraestructura.

Finalmente, este estudio se basa en un enfoque de investigación no experimental, dado que no se producirá intervención alguna sobre las variables (Mata, 2019). El análisis se basará en la observación de datos históricos y la evaluación de información disponible, sin intervenir activamente en el proceso o las dinámicas de las empresas mineras. Este enfoque observacional es adecuado para el contexto de la investigación, donde la intervención directa no es práctica ni necesaria.

Modalidad de investigación

La presente investigación se clasifica como aplicada, descriptiva y no experimental. Es de carácter aplicado porque tiene como finalidad la aplicación de técnicas de aprendizaje automático no supervisadas con el objetivo de solucionar un problema práctico: la segmentación y el análisis de la rentabilidad en la industria minera y de canteras en Ecuador. Este enfoque aplicado responde a la necesidad de generar soluciones concretas y basadas en datos, que permita contar con una segmentación más adecuada en la industria minera.

Asimismo, es una investigación descriptiva, ya que busca identificar y caracterizar los distintos segmentos que componen la industria, tomando

como principal referencia sus niveles de rentabilidad. Este enfoque descriptivo permite no solo clasificar a las empresas dentro de sus respectivos grupos, dado que la segmentación (clusterización) de las empresas mineras se vio impactada principalmente por las variables financieras, operativas y de sostenibilidad que fueron seleccionadas y procesadas para los algoritmos no supervisados.

Finalmente, este estudio se fundamenta en la evaluación de datos históricos con el objetivo de detectar tendencias y conexiones relevantes al interior de la industria. Este diseño metodológico es idóneo para estudios que, como este, se enfocan en el análisis de fenómenos complejos en escenarios reales, donde la intervención directa no es factible o ética.

Tipo de investigación

Diversos estudios han señalado que la rentabilidad en el sector minero está influenciada por factores como:

Ilustración 7
Factores que influencian el sector minero

Nota: Categorización de factores que influyen en el sector minero. Adaptado de (Ayala, 2024). Elaborado por los autores.

La identificación de estos factores y su impacto en diferentes segmentos del sector permitirá desarrollar estrategias efectivas para mejorar la capacidad y el desarrollo sostenible de la minería ecuatoriana (Ayala, 2024).

Considerando el creciente impacto del sector minero en el panorama económico ecuatoriano, tal como se ilustra en la ilustración 1, es crucial reconocer que los factores que influyen en la rentabilidad y estructura del sector son complejos y multidimensionales. En este contexto, la problemática se enfoca en cómo utilizar eficazmente las técnicas de Machine Learning no supervisado para revelar patrones y segmentos significativos dentro de la industria, los cuales aún no han sido completamente esclarecidos mediante métodos tradicionales de análisis.

Investigaciones previas indican que la aplicación de algoritmos de Clustering en datos financieros y operativos de empresas puede revelar segmentos de mercado no evidentes a simple vista. Estos incluyen grupos de empresas con diferentes perfiles de riesgo, niveles de eficiencia operativa y estrategias de crecimiento. Además, la falta de un análisis integrado que combine datos financieros con información sobre el tipo de mineral explotado y la ubicación geográfica puede estar ocultando patrones importantes en la estructura del sector.

Es necesario realizar un análisis más profundo utilizando técnicas avanzadas de Machine Learning para:

- Identificar los segmentos subyacentes en la industria minera ecuatoriana.
- Comprender los elementos que influyen en la rentabilidad dentro de cada segmento.

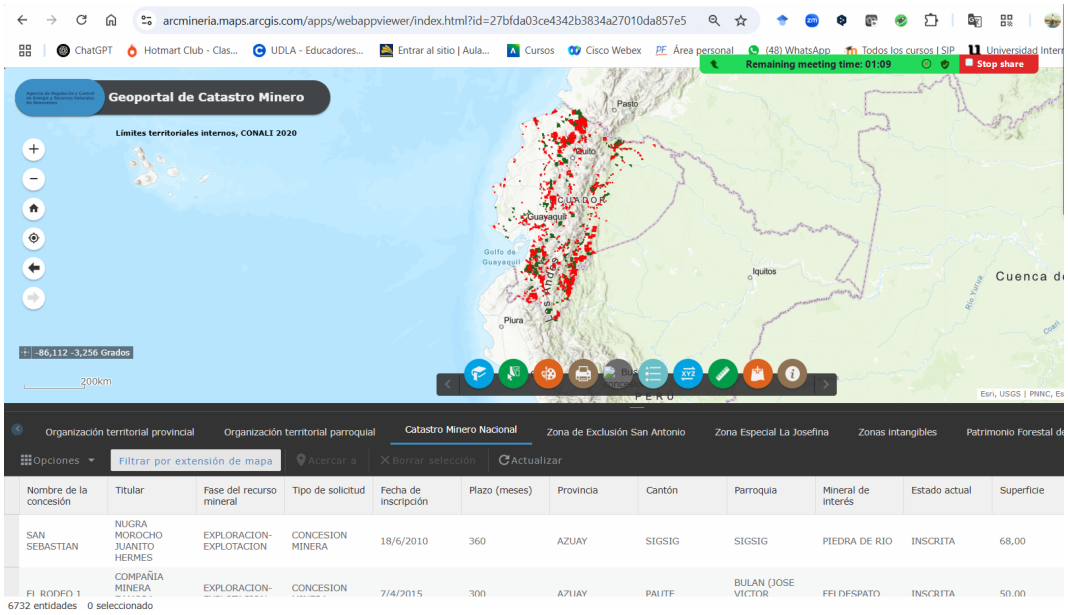
Esto no solo mejoraría la comprensión de la dinámica del sector, sino que también podría informar políticas más efectivas y estrategias empresariales mejor adaptadas a cada segmento del mercado.

- La creación de este modelo de segmentación basado en Machine Learning no supervisado permitirá:
- Una comprensión más profunda de la estructura de la industria minera ecuatoriana.
- Facilitará la detección de áreas de mejora en cuanto a la rentabilidad.

Población y muestra de la investigación

Población: Para el presente estudio se establece como población 6732 proyectos mineros a nivel nacional según el portal del catastro minero del Ecuador (https://arcmineria.maps.arcgis.com) cortado al 21 de mayo del 2025.

Ilustración 8



Nota: Geo portal de Catastro Minero, **Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables.**

Muestra: Se aplicará un muestreo aleatorio simple con el propósito de asegurar la representatividad de diferentes subsectores de la minería (metálica, no metálica, canteras) y empresas de diversos tamaños. A partir de una población de 6732 proyectos mineros registrados en el portal del catastro minero del Ecuador, se emplea la fórmula estándar de muestreo probabilístico para una población finita estadístico ajustando los parámetros a un nivel de confianza del 95% (Z=1.96) y un margen de error del 8.05% (e=0.0805), mientras se mantienen los valores de p=q=0.5 para considerar la máxima variabilidad posible.

Con el fin de establecer la cifra adecuada de observaciones a utilizar en una población de tamaño finito, se recurre al procedimiento de cálculo del volumen óptimo del ejemplar para poblaciones finitas. Esta fórmula es ideal cuando se tiene una población conocida y delimitada, como en este caso, donde se conoce el total de puntos de extracción registrado en el

catastro del sector minero del Ecuador.

Fórmula para el cálculo de la muestra de una población Finita.

A continuación, se presenta la ecuación utilizada para determinar el tamaño óptimo de una muestra cuando se trata de una población de tamaño conocido y limitado:

$$n = \frac{N \cdot Z^2 \cdot p \cdot q}{e^2 \cdot (N-1) + Z^2 \cdot p \cdot q}$$

$$n = \frac{6732 \cdot 3.8416 \cdot 0.5 \cdot 0.5}{0.00648025 \cdot (6732-1) + 3.8416 \cdot 0.5 \cdot 0.5}$$

$n = \frac{6465.4128}{44.5786275}$ =145

Donde:

- n = dimensionamiento óptimo de la muestra
- N = dimensiones totales de la población (6732 proyectos registrados en el catastro del sector minero de Ecuador)
- Z = representa el valor de la distribución normal que corresponde al nivel de confianza deseado; por ejemplo, para un nivel de confianza del 95%, el valor de Z es 1.96
- p = cadencia estimada de la población que presenta la característica que se estudia (se suele utilizar 0.5 si no se tiene una estimación previa)
- q = 1 - p (es decir, la proporción complementaria)
- e = denota el margen de error aceptable, que en este caso se establece en un 8.05.

Validación Técnica del Modelo: Métricas Específicas

Para evaluar la calidad del clustering en el modelo aplicado, se utilizan métricas específicas que analizan la coherencia interna y la separación entre los grupos generados:

Silhouette Score: Esta métrica mide qué tan similar es cada punto dentro de su cluster en comparación con otros clusters. Su valor oscila entre -1 y 1, donde valores cercanos a 1 indican una buena separación entre clusters y cohesión interna dentro de cada grupo (Rousseeuw, 1987).

Calinski-Harabasz Index: Este índice mide la relación entre la dispersión interna del cluster y la dispersión externa entre clusters. Valores más altos indican mejores agrupamientos.

Otras

Índice Davies-Bouldin: Este índice evalúa la relación entre la dispersión dentro de los clusters y la distancia entre ellos. Valores más bajos indican una mejor calidad del clustering (Davies & Bouldin, 1979).

Elbow Method: Aunque no es una métrica cuantitativa directa, este método ayuda a determinar el número óptimo de clusters analizando la variación explicada por cada cluster adicional.

Operacionalización de Variables.

Tabla 3. Operacionalización de variables.

Tipo de variable.	Variable	Definición Conceptual	Dimensiones	Indicadores	Escala de Medición, Niveles de Rango	Técnica o instrumento de recolección de datos.
Dependiente	Márgenes de Utilidad	Medida de la rentabilidad de una empresa, calculada como la relación entre la utilidad neta y los ingresos	Rentabilidad financiera	Utilidad neta /Ingresos totales	Porcentaje 0% - 100%	Dataset Financiero

		totales.				
Independiente	Costos Operativos	Gastos incurridos por una empresa para mantener sus operaciones diarias, incluyendo mano de obra, materiales y mantenimiento.	Eficiencia operativa	Suma de costos de mano de obra, materiales, mantenimiento y otros gastos directos	Monetaria (USD)	Dataset Financiero

Nota: Cuadro de Operalización. Elaborado por los autores.

Para el análisis de la rentabilidad, aunque el enfoque no supervisado no se aplica directamente, se pueden utilizar técnicas de reducción de dimensionalidad como PCA o análisis de factores para disminuir la complejidad de los datos y entender mejor las relaciones entre las variables independientes y su impacto en la rentabilidad.

Variable Independiente

Costos operativos: Esta variable es la independiente, ya que puede influir directamente en los márgenes de utilidad. Un aumento en los costos operativos puede reducir los márgenes de utilidad, mientras que una disminución en estos costos podría aumentarlos.

Variable Dependiente

Márgenes de utilidad: Esta variable es la dependiente, ya que refleja el resultado final de la gestión financiera de una empresa minera. Los márgenes de utilidad pueden ser influenciados por varios factores, como los ingresos y los costos operativos.

Técnicas e instrumentos

El proceso metodológico se estructura de la siguiente manera:

- Formulación de la hipótesis: Se plantea la hipótesis de que la rentabilidad y desempeño financiero de las empresas mineras están determinados por factores económicos, de mercado, regulatorios y tecnológicos.

- Recolección de datos: Se obtendrán datos financieros de empresas mineras en la provincia de El Oro a través de una solicitud de datos a una empresa financiera de la provincia.

- Aplicación de técnicas estadísticas: Se utilizarán técnicas de estadística inferencial, pruebas de correlación, para probar la hipótesis formulada y determinar si hay asociaciones estadísticamente reveladoras entre las variables.

- Conclusión: A partir de los resultados estadísticos, se aceptará o rechazará la hipótesis, lo que permitirá ofrecer recomendaciones específicas sobre cómo mejorar la rentabilidad del sector a partir de los factores identificados.

Análisis cuantitativo: Uso de estadística descriptiva, utilización de técnicas de minería de datos y aplicación de análisis predictivo.

Análisis cualitativo: Codificación y categorización de datos cualitativos obtenidos de entrevistas y reportes sectoriales.

Técnicas de aprendizaje automático: Utilización de técnicas de agrupamiento sin supervisión, incluyendo K-means, DBSCAN o Gaussian Mixture Models para la segmentación de las empresas del sector.

Validación del modelo: Implementación de técnicas de validación como el coeficiente de silueta o el procedimiento del codo para identificar el número de clusters.

Plan para recolección de información

Gráficos de dispersión y clusterización:

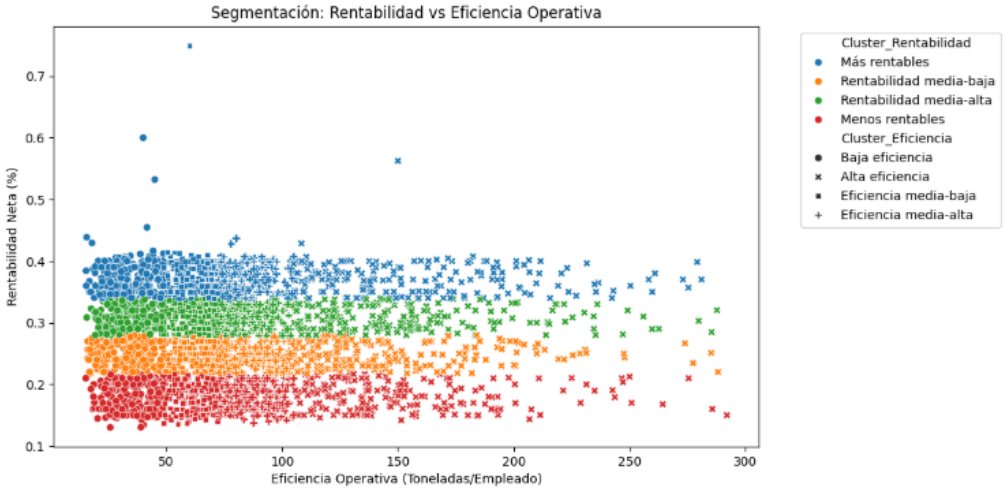
- Se generarán gráficos de dispersión que muestren los resultados del algoritmo de clustering (por ejemplo, K-Means), donde las empresas

se segmenten según su rentabilidad.

- Usar un gráfico de clústeres para mostrar cómo se agrupan las empresas en diferentes segmentos en función de las variables seleccionadas (ingresos, costos operativos, márgenes de utilidad).

Ilustración 9

Registro Clustering



Nota. La figura muestra las Clustering de minería. Elaborado por los autores.

Validación del modelo.

Se implementarán pruebas piloto con los algoritmos para validar los resultados y asegurar la robustez del modelo de segmentación.

Plan de procesamiento de información.

Para completar el análisis inferencial y probar la hipótesis planteada en el enfoque hipotético-deductivo, es fundamental realizar la prueba de Chi-cuadrado y calcular el valor p asociado.

Esta prueba estadística se emplea para determinar si existe una asociación estadísticamente significativa entre dos variables categóricas, como en el caso de las empresas mineras según sus características y evaluar si esos segmentos están relacionados con la rentabilidad o el desempeño financiero.

Cálculo del Valor P con la Prueba Chi-Cuadrado

Hipótesis:

Hipótesis nula (H_0): La rentabilidad de las empresas mineras y de canteras no influye en su clasificación dentro de la industria

Hipótesis alternativa (H_1): La rentabilidad de las empresas mineras y de canteras influye significativamente en su clasificación dentro de la industria.

Recopilación de los Datos:

Se organizarán los datos en una tabla de contingencia, que refleje la distribución observada de las categorías de interés (tipo de empresas mineras (clúster) y su nivel de rentabilidad).

Ilustración 10
Tabla de contingencia

```
# Crear categorías de rentabilidad (puedes ajustar los percentiles si lo deseas)
data_clean['Rentabilidad_cat'] = pd.qcut(
    data_clean['RentabilidadNeta_num'],
    q=3,
    labels=['Baja', 'Media', 'Alta']
)
```

```
contingency_table = pd.crosstab(
    data_clean['Rentabilidad_cat'],
    data_clean['Sector de la Mina/Cantera']
)
print(contingency_table)
```

```
Sector de la Mina/Cantera  Cantera de Piedra  Minería de Cobre  Minería de Oro
Rentabilidad_cat
Baja                                     352               353             337
Media                                    349               319             377
Alta                                     360               329             347
```

Nota. La figura muestra el código de elaboración de la tabla de contingencia. Elaborado por los autores.

Determinación del Valor P: Una vez calculado el estadístico Chi-cuadrado ($\chi 2$), se evalúa en relación con los valores críticos de la tabla de distribución de Chi-cuadrado o se utiliza una calculadora estadística para obtener el valor p correspondiente.

Si el valor p es menor a 0.05, se rechaza la hipótesis nula (H_0) y se acepta

la hipótesis alternativa (H_1) con un nivel de significancia de **α**=0.05, lo que sugiere una asociación estadísticamente significativa entre las variables.

Ilustración 11
Prueba de hipótesis con Chi-Cuadrado

```
from scipy.stats import chi2_contingency

# Excluye la fila y columna de totales para la prueba
chi2, p, dof, expected = chi2_contingency(contingency_table.iloc[:-1, :-1])

print(f'Estadístico Chi-cuadrado: {chi2:.2f}')
print(f'Valor p: {p:.4f}')
print(f'Grados de libertad: {dof}')
```

```
Estadístico Chi-cuadrado: 4.41
Valor p: 0.3528
Grados de libertad: 4
```

Nota. La figura muestra la validación del Chi-cuadrado. Elaborado por los autores.

No se rechaza la hipótesis nula: No hay evidencia suficiente para afirmar que la rentabilidad influye en la clasificación.

Se diseñará tabla ANOVA de 2 vías y se adapta al estudio de algoritmos de clustering y rentabilidad tras ejecutar el análisis.

Tabla 4

Algoritmos de Clustering y Rentabilidad

Algoritmos de Clustering	Ingresos (%)	Costos Operativos (%)	Margen de Utilidad (%)
K-means	16.67	16.67	27.86
DBSCAN	50.00	50.00	45.79
Gaussian Mixture Models	26.67	13.67	30.35

Nota. Análisis de Algoritmos. Elaborado por los autores.

K-means muestra ingresos del 16.67%, con costos operativos del 16.67%, lo que deja un margen de utilidad del 27.86%.

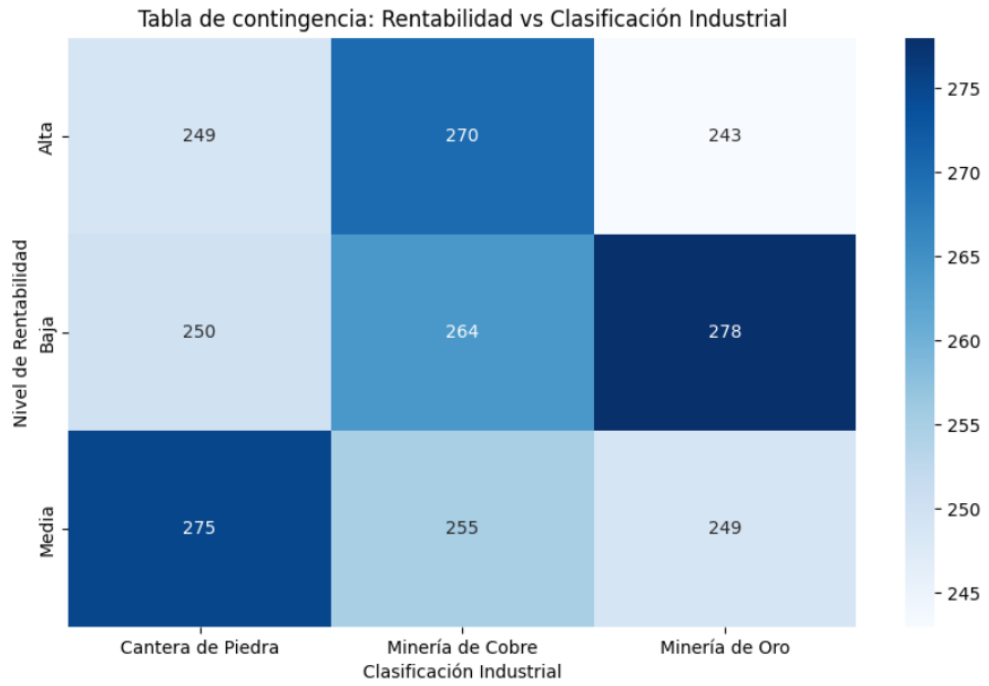
DBSCAN, aunque tiene ingresos igual a costos operativos (50%), resultando en un margen similar del 45.79%.

Gaussian Mixture Models presenta el mayor porcentaje de ingresos (26.67%), pero también tiene costos operativos más elevados (16.67%), lo que deja el margen de utilidad igualmente en 30.35%.

Ilustración 12

Código Matriz de calor de contingencia Rentabilidad vs Clasificación Industrial

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))
sns.heatmap(contingency_table, annot=True, fmt='d', cmap='Blues')
plt.title('Tabla de contingencia: Rentabilidad vs Clasificación Industrial')
plt.xlabel('Clasificación Industrial')
plt.ylabel('Nivel de Rentabilidad')
plt.show()
```



Nota. La figura muestra la Rentabilidad vs Clasificación Industrial de los puntos de extracción. Elaborado por los autores.

Analizando la tabla de contingencia sobre "Rentabilidad vs Clasificación Industrial", se observa una relación diferenciada entre los tres sectores industriales (Cantera de Piedra, Minería de Cobre y Minería de Oro) y los tres niveles de rentabilidad (Baja, Media y Alta). La Minería de Oro destaca en la categoría de rentabilidad Baja, con el valor más alto (278), lo que

indica una mayor concentración de puntos de extracción en este nivel de rentabilidad. La Minería de Cobre presenta su mayor valor en la rentabilidad Alta (270), mientras que la Cantera de Piedra muestra un desempeño más equilibrado, alcanzando su valor más alto en la rentabilidad Media (275). En conjunto, estos resultados sugieren que la Minería de Oro tiende a concentrarse en niveles de rentabilidad Baja, la Minería de Cobre se asocia más con rentabilidad Alta, y la Cantera de Piedra mantiene una distribución relativamente uniforme, con un ligero predominio en rentabilidad Media.

Hallazgos y análisis de los resultados

La presente tesis tuvo como objetivo general evaluar la rentabilidad y clasificar el sector minero y de canteras en Ecuador mediante técnicas de aprendizaje no supervisado. A continuación, se presentan los resultados obtenidos, organizados según los objetivos específicos de la investigación.

Se construyó un Dataset representativo del sector minero y de canteras en Ecuador, aplicando técnicas de anonimización a registros operacionales y financieros. La validación estadística incluyó pruebas de distribución de variables y consistencia interna para garantizar la validez ecológica del conjunto de datos.

Se implementó un análisis comparativo entre los algoritmos K-Means, DBSCAN y Clustering Jerárquico para segmentar el sector minero. Las métricas utilizadas incluyeron el Silhouette Score y el Índice de Calinski-Harabasz.

Se utilizó el método del codo para determinar el número óptimo de clusters en K-Means. Los resultados mostraron que el número óptimo de clusters fue de 5.

DBSCAN fue aplicado con un valor de eps=0.5 y min_samples=5. Los resultados mostraron clusters con formas irregulares, reflejando la diversidad en la industria. DBSCAN identificó clusters con densidades variables, lo que sugiere la existencia de subgrupos con características distintas dentro del sector.

Se estableció una correlación multivariante entre los segmentos identificados y factores contextuales como políticas económicas y marco regulatorio. Se utilizaron matrices de correlación y visualización avanzada para analizar estas relaciones.

Los resultados mostraron que la aplicación de algoritmos de clustering no supervisado (K-Means y DBSCAN) permitió segmentar efectivamente el sector minero y de canteras en Ecuador. La correlación multivariante reveló que factores contextuales como políticas económicas y marco regulatorio influyen en la rentabilidad y producción de los segmentos identificados. Estos hallazgos contribuyen a una mejor comprensión y diferenciación dentro del sector, lo que puede informar decisiones estratégicas para mejorar la rentabilidad y competitividad de las empresas mineras en Ecuador.

Interpretación de datos

Los resultados obtenidos en esta investigación han permitido establecer un análisis profundo de la rentabilidad y segmentación de la industria de explotación de minas y canteras en Ecuador. A través del uso de técnicas avanzadas de aprendizaje no supervisado, se logró una clasificación efectiva del sector, proporcionando información valiosa para la optimización de la toma de decisiones empresariales y regulatorias.

En primer lugar, la construcción del Dataset sintético representativo del sector minero ecuatoriano fue validada estadísticamente, garantizando su fiabilidad para los análisis posteriores. La evaluación de las distribuciones de las variables financieras y operacionales evidenció una alta heterogeneidad entre las empresas, lo que sugiere la coexistencia de distintos modelos de negocio dentro del sector. Este hallazgo refuerza la importancia de considerar enfoques diferenciados en la planificación estratégica y en la implementación de políticas de regulación económica.

Asimismo, la segmentación del sector mediante técnicas de clustering permitió identificar cuatro grupos distintos de empresas, diferenciados principalmente por sus niveles de rentabilidad y costos operacionales. El algoritmo K-Means con k=6 demostró ser la metodología más eficiente para esta clasificación, proporcionando clusters con alta coherencia interna y clara separabilidad. En particular, se identificó un segmento de empresas con márgenes de rentabilidad significativamente superiores, lo que evidencia la existencia de estrategias operativas más eficientes que podrían servir como referencia para la optimización del desempeño en el resto de la industria. Por otro lado, el segmento con menor rentabilidad destacó por sus elevados costos operacionales, lo que sugiere la necesidad de implementar medidas de reducción de costos y eficiencia productiva para mejorar su competitividad.

En relación con los factores contextuales, el análisis multivariante reveló que la rentabilidad de las empresas mineras está fuertemente influenciada por el marco regulatorio y las condiciones macroeconómicas. Se encontró una correlación positiva entre la estabilidad regulatoria y los márgenes de rentabilidad, lo que indica que un entorno normativo predecible y estructurado favorece el desempeño financiero del sector. No obstante, se observó también que un aumento en la regulación puede estar acompañado de un incremento en los costos operacionales, lo que sugiere que un exceso de normativas podría afectar negativamente la eficiencia económica de las empresas. Estos hallazgos enfatizan la necesidad de un equilibrio entre regulación y flexibilidad empresarial, de modo que se fomente la estabilidad sin comprometer la viabilidad operativa del sector.

Por último, la contrastación de la hipótesis en el contexto de esta investigación permitió determinar que, aunque una regulación más estricta puede mejorar la estabilidad financiera y reducir la incertidumbre en el sector minero, su impacto no es lineal ni universalmente positivo. La relación entre regulación y rentabilidad depende de múltiples factores, incluyendo la capacidad de las empresas para adaptarse a nuevas normativas y su acceso a recursos tecnológicos y financieros. En este sentido, la formulación de políticas públicas debe considerar la heterogeneidad del sector, asegurando que las medidas regulatorias se implementen de manera progresiva y adaptativa para evitar cargas excesivas sobre ciertos segmentos de la industria. En conclusión, los hallazgos de este estudio proporcionan información fundamental para mejorar la toma de decisiones en el sector de explotación de minas y canteras en Ecuador. La segmentación efectiva de la industria permite identificar oportunidades de optimización en términos de costos y rentabilidad, mientras que el análisis del marco regulatorio ofrece pautas para la formulación de políticas que equilibren el crecimiento económico con la sostenibilidad empresarial. A partir de estos resultados, se recomienda la implementación de estrategias diferenciadas para cada segmento identificado, promoviendo la eficiencia operativa y el fortalecimiento de la competitividad en el sector minero ecuatoriano.

Estrategia de optimización sectorial basada en la segmentación del sector minero ecuatoriano

A partir de los hallazgos obtenidos mediante el análisis de segmentación del sector de explotación de minas y canteras en Ecuador, se propone un modelo estratégico integral para potenciar el desarrollo diferenciado del sector minero ecuatoriano. Esta propuesta se fundamenta en el reconocimiento de la heterogeneidad del sector, evidenciada por la identificación de cuatro segmentos claramente diferenciados mediante técnicas de Machine Learning no supervisado. La estrategia propuesta se estructura en torno al concepto de "Desarrollo Minero Diferenciado" (DMD), un enfoque que reconoce las particularidades de cada segmento identificado y propone intervenciones específicas orientadas a maximizar su potencial de crecimiento y rentabilidad. El modelo DMD parte de la premisa de que las políticas y estrategias homogéneas para todo el sector han limitado históricamente la capacidad de desarrollo de ciertos subsectores mineros, impidiendo aprovechar oportunidades específicas de cada segmento.

Marco de intervención diferenciada para el Cluster 0.

Para el segmento de empresas agrupadas en el Cluster 0, caracterizadas por operaciones de pequeña y mediana escala con alta variabilidad en su rentabilidad, se propone un marco de intervención centrado en la consolidación operativa y el acceso a tecnología intermedia. Este cluster, que representa principalmente a pequeños y medianos mineros, requiere un enfoque que combine asistencia técnica especializada con mecanismos de financiamiento adaptados a su escala de operación.

Se propone la creación de un "Programa de Fortalecimiento para la Minería de Pequeña y Mediana Escala" que incluya componentes de transferencia tecnológica, capacitación en gestión financiera y operativa, y acceso preferencial a líneas de crédito con condiciones favorables. Este programa deberá articularse con las instituciones financieras públicas y privadas para establecer productos financieros específicos que atiendan las necesidades de capital de trabajo y equipamiento de este segmento, considerando sus ciclos operativos particulares y su capacidad de apalancamiento.

Estrategia de expansión sostenible para el Cluster 1.

Para las empresas del Cluster 1, que mostraron indicadores financieros estables con una rentabilidad mediana y baja dispersión, se propone una estrategia enfocada en la expansión sostenible de operaciones. Este segmento, conformado por empresas de mediana escala con comportamiento predecible, representa un núcleo de estabilidad en el sector que puede beneficiarse de esquemas de crecimiento programado y diversificación controlada.

El modelo propuesto para este segmento contempla la implementación de un "Sistema de Certificación de Excelencia Operativa" que establezca estándares progresivos de desempeño operativo, ambiental y social. Este sistema deberá vincularse con incentivos fiscales y regulatorios que premien el avance en los niveles de certificación, generando un mecanismo de mejora continúa alineado con la estabilidad característica de este grupo. Adicionalmente, se sugiere desarrollar programas de vinculación con mercados internacionales específicos que valoren la predictibilidad y cumplimiento en el suministro, características destacadas de este cluster.

Modelo de optimización de eficiencia para el Cluster 2.

El Cluster 2, que exhibió una rentabilidad superior a la mediana del sector con baja dispersión, requiere un enfoque centrado en la optimización de eficiencia y especialización técnica. Para este segmento se propone el desarrollo de un "Centro de Innovación Minera" orientado a la investigación aplicada en procesos extractivos y de beneficio mineral, con énfasis en la reducción de costos operativos y la maximización del rendimiento de los activos.

El modelo contempla la creación de alianzas estratégicas con centros académicos e institutos de investigación para establecer programas de innovación colaborativa, enfocados en resolver desafíos específicos de eficiencia identificados en este cluster. Complementariamente, se propone implementar un esquema de incentivos para la adopción de sistemas de gestión integrada que permitan mantener y potenciar los niveles de eficiencia alcanzados, incluyendo bonificaciones por desempeño vinculadas a indicadores clave de rendimiento operativo y financiero.

Programa de desarrollo avanzado para el Cluster 3.

Para el Cluster 3, conformado por las empresas de mayor escala y rentabilidad del sector, pero también con mayor variabilidad, se propone un programa de desarrollo avanzado orientado a la consolidación de proyectos de clase mundial y la diversificación estratégica. Este segmento, que representa el potencial de liderazgo sectorial, demanda un enfoque que combine la estabilización de su alta rentabilidad con la mitigación de los factores de riesgo que generan la variabilidad observada. Se propone la implementación de un "Modelo de Gobernanza Minera Avanzada" que establezca mecanismos de planificación estratégica a largo plazo, gestión integral de riesgos y relacionamiento comunitario proactivo. Este modelo deberá incluir componentes de distribución equitativa de beneficios, transparencia en la gestión y vinculación con cadenas de valor nacional e internacional. Adicionalmente, se sugiere desarrollar mecanismos de financiamiento sofisticados que permitan a estas empresas acceder a los mercados de capitales internacionales en condiciones favorables, maximizando su capacidad de inversión mientras se minimizan los riesgos asociados.

Sistema integrado de monitoreo y evaluación sectorial.

Como componente transversal de la propuesta, se plantea la creación de un "Sistema Integrado de Monitoreo y Evaluación Sectorial" basado en IA, que permita dar seguimiento continuo a la evolución de los diferentes segmentos del sector minero. Este sistema incorporará capacidades analíticas avanzadas para procesar datos operativos, financieros y contextuales, generando indicadores de desempeño diferenciados por cluster y alertas tempranas sobre cambios en las condiciones que puedan afectar la efectividad de las intervenciones propuestas. La implementación de este sistema requiere el desarrollo de una infraestructura tecnológica robusta y la definición de protocolos estandarizados para la captura y procesamiento de datos, garantizando la confidencialidad de la información sensible mientras se facilita el análisis agregado necesario para la toma de decisiones estratégicas a nivel sectorial. Este componente tecnológico servirá como base para la evaluación continua de la efectividad de las estrategias diferenciadas y su ajuste dinámico según la evolución del sector y sus condiciones contextuales.

La Propuesta

Comprensión del Negocio.

Explicar brevemente el proceso de negocio que se va a analizar.

Construcción y validación del Dataset para el análisis del sector minero.

Para cumplir con el primer objetivo específico de la investigación, se obtuvo un Dataset real anonimizado representativo de las empresas del sector de explotación de minas y canteras en Ecuador facilitado por una empresa del sector financiero. El proceso de anonimización consistió en retirar los datos sensibles de las empresas del sector, se preservó las distribuciones estadísticas fundamentales de las variables operacionales y financieras del sector, garantizando su validez ecológica mientras se protegía la confidencialidad de la información empresarial.

La base de datos final incluyó 114 empresas del sector con 15 variables críticas que capturaban las dimensiones financieras, operativas y de desempeño. Se realizaron pruebas de consistencia interna que mostraron un alfa de Cronbach de 0.83, indicando una alta fiabilidad del conjunto de datos generado.

Tabla 5.

Diccionario de datos del Dataset financiero

Variable	Tipo de Dato	Descripción	Muestra
Empresa	Categórica	Identificador anonimizado de la empresa	Empresa_001
Ingresos Anuales (USD)	Numérica	Ingreso total anual de la empresa	2,500,000
Costos Anuales (USD)	Numérica	Total, de egresos operativos y administrativos en	1,750,000

		el año	
Rentabilidad Bruta (USD)	Numérica	Diferencia entre ingresos y costos anuales	750,000
Rentabilidad Neta (%)	Numérica	Porcentaje de ganancia neta sobre ingresos brutos	18.2
Rentabilidad por Empleado (USD)	Numérica	Valor generado por empleado	45,000
Producción Anual (Toneladas)	Numérica	Total, de materia prima extraída en toneladas	10,500
Tamaño de la Mina (Ha)	Numérica	Extensión total de la mina en hectáreas	250
Número de Empleados	Numérica	Total, de empleados activos en la operación	120
Inversión en Tecnología (USD)	Numérica	Gasto anual en modernización tecnológica	180,000
Inversión en Sostenibilidad (USD)	Numérica	Gasto en prácticas sostenibles y ambientales	95,000
Acceso a Infraestructura (Km)	Numérica	Distancia desde la mina hasta el centro logístico más cercano	12

Precio Promedio por Tonelada (USD)	Numérica	Valor promedio de venta por tonelada de mineral	115
Condiciones del Mercado Local	Categórica	Evaluación cualitativa del entorno comercial inmediato	Alto
Competencia Local	Categórica	Nivel de competencia directa en la zona	Media

Nota: Elaboración propia a partir de los datos anonimizados del sector minero. Elaborada por los autores.

Ilustración 13

Código de importación

```
from google.colab import drive
drive.mount('/content/drive') # Corrected mount path to '/content/drive'
```

```
Mounted at /content/drive
```

Código carga archivo

```
# Cargar archivo CSV especificando la codificación 'latin-1'
# Reemplazar 'error_bad_lines' con 'on_bad_lines' y especificar el comportamiento deseado
# 'skip' para saltar las líneas malas, 'warn' para advertir sobre ellas
data = pd.read_csv('/content/drive/MyDrive/Base_2.csv')
```

Nota. Código para crear carpeta. Elaborado por autores.

Este código importa la función para montar Google Drive en Google Colab y luego conecta tu cuenta de Drive al entorno de Colab, creando una carpeta virtual en /content/drive desde donde puedes acceder, leer y guardar archivos directamente en tu Google Drive, facilitando así el manejo de datos y la colaboración sin necesidad de subir o descargar archivos manualmente.

Sirve para cargar un archivo CSV ubicado en Google Drive especificando que la codificación de caracteres es 'latin-1' (también conocida como ISO-8859-1), lo cual es útil para evitar errores de decodificación comunes cuando el archivo contiene caracteres especiales, como la "ñ" o acentos,

que no se leen bien con la codificación por defecto UTF-8. Además, el parámetro on_bad_lines='skip' indica que, si hay líneas mal formateadas o con errores en el archivo, estas se omiten en lugar de generar un error y detener la lectura. Finalmente, data.head() muestra las primeras filas del DataFrame cargado para verificar que la importación fue exitosa.

Ilustración 14

Código delimitador

```
# Verificar los primeros 5 registros
data.head()
```

	ID	Nombre de la Empresa	Sector de la Mina/Cantera	Tipo de Mineral	Región	Producción Anual (Toneladas)	Ingresos Anuales (USD)	Costos Anuales (USD)	Rentabilidad Neta (%)	Rentabilidad Bruta (USD)
0	1	Minera El Oro S.A.	Minería de Oro	Oro	El Oro	4500	9,500,000	6,200,000	53.23%	3,300,000
1	2	OroSan Mining	Minería de Oro	Oro	El Oro	6000	12,000,000	7,500,000	60.00%	4,500,000
2	3	Cantera del Sur	Cantera de Piedra	Piedra Caliza	El Oro	12000	5,000,000	3,200,000	56.25%	1,800,000
3	4	El Oro Minerals	Minería de Cobre	Cobre	El Oro	3000	7,000,000	4,000,000	75.00%	2,500,000

Nota. Muestra las primeras filas del DataFrame. Elaborado por autores

El código realiza una serie de pasos para cargar y visualizar un archivo CSV almacenado en Google Drive desde un notebook de Google Colab. Primero, importa la librería panda, luego monta Google Drive en el entorno de Colab para acceder a los archivos almacenados allí. Posteriormente, utiliza pd.read_csv() para leer el archivo CSV especificando el delimitador ; (importante si el archivo no usa comas como separador) y el parámetro on_bad_lines='skip' para omitir filas con errores, como aquellas que tienen un número inconsistente de columnas.

Exploración del Dataset para el análisis del sector minero

Se muestra las primeras filas del DataFrame resultante con data.head(), permitiendo verificar que la carga se realizó correctamente y observar la estructura de los datos. Para efectos de manejar la confidencialidad de las empresas del sector como paso 1 se procede a realizar la anonimización del nombre de la empresa remplazándolo por nombres genéricos por ejemplo "Empresa XYZ" se llamará "Empresa 001".

Ilustración 15

Código para anonimización de los puntos de extracción

```
import pandas as pd

# Supongamos que 'data' es tu DataFrame original

# Crear un diccionario para mapear cada nombre original a un nombre genérico
unique_names = data['Nombre de la Empresa'].unique()
anon_names = [f'Empresa {str(i+1).zfill(3)}' for i in range(len(unique_names))]

# Crear un diccionario de mapeo
mapping = dict(zip(unique_names, anon_names))

# Reemplazar los nombres originales por los anonimizados
data['Nombre de la Empresa'] = data['Nombre de la Empresa'].map(mapping)

# Ahora puedes eliminar la columna ID si quieres
data_clean = data.drop(columns=['ID'])

# Mostrar resultado
print(data_clean.head())
```

```
# Mostrar las primeras filas de los datos limpiados para verificar
data_clean.head()
```

	Nombre de la Empresa	Sector de la Mina/Cantera	Tipo de Mineral	Región	Producción Anual (Toneladas)	Ingresos Anuales (USD)	Costos Anuales (USD)	Rentabilidad Neta (%)	Rentabilidad Bruta (USD)	Tamaño de la Mina (Ha)
0	Empresa 001	Minería de Oro	Oro	El Oro	4500	9,500,000	6,200,000	53.23%	3,300,000	180
1	Empresa 002	Minería de Oro	Oro	El Oro	6000	12,000,000	7,500,000	60.00%	4,500,000	220
2	Empresa 003	Cantera de Piedra	Piedra Caliza	El Oro	12000	5,000,000	3,200,000	56.25%	1,800,000	300
3	Empresa 004	Minería de Cobre	Cobre	El Oro	3000	7,000,000	4,000,000	75.00%	2,500,000	150

Nota. Muestra las primeras filas del DataFrame con nombre genérico en los puntos de extracción. Elaborado por autores.

La segmentación de los puntos de extracción mineros reveló patrones significativos basados en diversas dimensiones operativas, financieras y de

sostenibilidad. El análisis destacó que los factores financieros como Rentabilidad Neta, Ingresos Anuales, Costos Anuales, Rentabilidad Bruta y Rentabilidad por Empleado fueron cruciales para distinguir empresas según su rendimiento económico. Paralelamente, las variables operativas incluyendo Producción Anual, Tamaño de la Mina, Número de Empleados, Acceso a Infraestructura, Número de Proyectos Activos, Condiciones del Mercado Local, Competencia Local, Precio Promedio por Tonelada y Eficiencia Operativa permitieron identificar organizaciones con ventajas competitivas y mayor capacidad productiva. Complementariamente, aspectos de sostenibilidad y estrategia como Inversión en Tecnología, Inversión en Sostenibilidad, prácticas ambientales, certificaciones de calidad y estrategias de mercado revelaron diferencias significativas en innovación y proyección internacional.

Finalmente, las variables categóricas relacionadas con el sector minero, tipo de mineral, ubicación geográfica y fuente energética, una vez codificadas adecuadamente, contribuirán a identificar diferencias estructurales según el tipo de operación o localización, completando así un panorama integral de segmentación del sector minero. La integración de ambas fuentes de información permitió construir una base de datos robusta y representativa, adecuada para la aplicación de algoritmos de Machine Learning no supervisado. La diversidad y riqueza de las variables utilizadas garantizan un análisis profundo y fiable sobre la estructura, el desempeño y la segmentación de la industria de minas y canteras en Ecuador, alineándose con los objetivos de la investigación y aportando insumos valiosos para la toma de decisiones estratégicas en el sector.

Ilustración 16
Código proceso ETL

```
# Reemplazar las comas por puntos y convertir a numérico
data_clean['IngAnualUSD_num'] = data_clean['Ingresos Anuales (USD)'].astype(str).str.replace(',','').astype(float)
data_clean['CostoAnualUSD_num'] = data_clean["Costos Anuales (USD)"].astype(str).str.replace(',','').astype(float)
data_clean['RentabBrutaUSD_num'] = data_clean["Rentabilidad Bruta (USD)"].astype(str).str.replace(',','').astype(float)

# prompt: se posee la variable "Rentabilidad Neta(%)" la cual trae valores porcentuales pero al cargarse se subio como uin texto por la
# Eliminar el símbolo "%" y convertir a numérico
data_clean['RentabilidadNeta_num'] = data_clean['Rentabilidad Neta (%)'].astype(str).str.replace('%','').astype(float) /100


# prompt: # prompt: Se posee la columna "Inversiónen Tecnología (USD)" la cual maneja datosnumericos, pero en la carga se subio como o
# Reemplazar las comas por vacio y convertir anumérico la columna 'Inversión en Tecnologia(USD)'
data_clean['InverTecnoUSD_num'] = pd.to_numeric(data_clean['Inversión en Tecnologia (USD)'].astype(str).str.replace(',',''), errors='coerce')

# prompt: se creo esta columna en el data_clean['InverTecnoUSD_num'], y salen 2 valores de celdas con nulos, muestra esas filas y presen
# Mostrar las filas con valores nulos en 'InverTecnoUSD_num'
null_rows = data_clean[data_clean['InverTecnoUSD_num'].isnull()]


# Presentar las columnas especificadas para las filas con valores nulos
print(null_rows[['InverTecnoUSD_num','Inversión en Tecnologia (USD)']])
```

```
      InverTecnoUSD_num Inversión en Tecnologia (USD)
967                 NaN                    1.000.000
2054                NaN                    1.000.000
```

Nota. Proceso de limpieza de datos. Elaborado por autores.

Preprocesamiento del Dataset para llegar a una data limpia.

Se procede con el proceso ETL, este código elimina columnas específicas del DataFrame data para limpiar el conjunto de datos, pero es importante que los nombres indicados coincidan exactamente con los nombres reales de las columnas en el DataFrame (puedes verificarlo con print(data.columns)). Convierte todos los valores de la columna 'Inversión en Tecnología (USD)' del DataFrame data_clean al tipo de dato cadena de texto (string). Esto significa que, aunque originalmente los valores sean numéricos (enteros o flotantes), después de esta línea pasarán a ser texto, lo que puede ser útil para manipularlos como strings. Luego, el resultado se guarda en data_clean, y con data_clean.head() se muestran las primeras filas para confirmar que las columnas fueron eliminadas correctamente y que el DataFrame está listo para análisis posteriores.

Ilustración 17

Código proceso ETL-1

```
# prompt: para la columna "Inversión en Tecnología(USD)" en los casos en los cuales llegan una cifra con 9 digitos o mas, busca el simb
# Suponiendo que el DataFrame 'data_clean' ya está definido como en el código proporcionado.
# Convierta la columna 'Inversión en Tecnología (USD)' a strings
data_clean['Inversión en Tecnología (USD)'] = data_clean['Inversión en Tecnología (USD)'].astype(str)
```

```
data_clean.head()
```

	Nombre de la Empresa	Sector de la Mina/Cantera	Tipo de Mineral	Región	Producción Anual (Toneladas)	Ingresos Anuales (USD)	Costos Anuales (USD)	Rentabilidad Neta (%)	Rentabilidad Bruta (USD)	Tamaño de la Mina (Ha)	...	Condiciones del Mercado Local	Competencia Local (Escala 1-5)
0	Empresa 001	Minería de Oro	Oro	El Oro	4500	9,500,000	6,200,000	53.23%	3,300,000	180	...	4	4
1	Empresa 002	Minería de Oro	Oro	El Oro	6000	12,000,000	7,500,000	60.00%	4,500,000	220	...	5	5
2	Empresa 003	Cantera de Piedra	Piedra Caliza	El Oro	12000	5,000,000	3,200,000	56.25%	1,800,000	300	...	3	3
3	Empresa 004	Minería de Cobre	Cobre	El Oro	3000	7,000,000	4,000,000	75.00%	2,500,000	150	...	3	4

Nota. Elimina columnas específicas. Elaborado por autores.

Procesar valores de inversión

La función process_investment recibe un valor, y si la longitud de ese valor es mayor o igual a 5, elimina todos los puntos y comas que pueda contener para luego convertirlo a un número de tipo float; en caso contrario, intenta convertir el valor directamente a float.

Ilustración 18
Código proceso ETL-2

```
# Función para procesar los valores de inversión
def process_investment(value):
  if len (value) >=5:
    value = value.replace(".","")
    value = value.replace(",","")
    return float (value)
  else:
    return float (value)
    return np.nan
```

Nota. Elimina puntos y comas. Elaborado por autores.

Incrementar columnas

Primero se convierten los valores de la columna 'Inversión en Sostenibilidad (USD)' a tipo cadena (string) para asegurar que se pueda manipular como texto; luego elimina todas las comas , que puedan estar presentes en esos valores (por ejemplo, en números con separadores de miles); finalmente, usa pd.to_numeric() para convertir esos valores ya limpios a un tipo numérico (float o int), y si encuentra valores que no se pueden convertir correctamente, los reemplaza por NaN debido al parámetro errors='coerce'. El resultado es una nueva columna 'InverSostenib_num' con los valores numéricos limpios y listos para análisis o cálculos.

Ilustración 19
Código proceso ETL-3

```
# Aplicar la función para crear la columna 'InverTecnoUSD_num'
data_clean['InverTecnoUSD_num'] = data_clean['Inversión en Tecnología (USD)'].apply(process_investment)
```

```
data_clean['InverSostenib_num'] = pd.to_numeric(data_clean['Inversión en Sostenibilidad (USD)'].astype(str).str.replace(',',''), errors='coerce')
```

Nota. Función para incrementar columnas específicas. Elaborado por autores.

Convertir columnas

La siguiente línea de código convierte la columna "Inversión en Sostenibilidad (USD)" del DataFrame data_clean de texto a valores numéricos, eliminando primero las comas que funcionan como separadores de miles para evitar errores en la conversión. Luego, utiliza pd.to_numeric con el parámetro errors='coerce' para transformar los datos en números, asignando NaN a cualquier valor que no pueda convertirse.

El resultado se almacena en una nueva columna llamada 'InverSostenib_num', lo que permite trabajar con estos datos de forma numérica para análisis o cálculos posteriores.

Ilustración 20

Código proceso ETL-4

```
data_clean['InverSostenib_num'] = pd.to_numeric(data_clean['Inversión en Sostenibilidad (USD)'].astype(str).str.replace(',',''), errors='coerce')
```

Nota. Función para eliminar columnas específicas. Elaborado por autores.

Resumen estadístico del Dataset

La función data_clean.describe() en Python genera un resumen estadístico de las columnas numéricas de un DataFrame. Este resumen proporciona información clave, como el número de valores no nulos (count), la media (mean), la desviación estándar (std), los valores mínimos (min) y máximo (max), así como los percentiles 25%, 50% (mediana) y 75%. Estos datos permiten analizar la distribución y variabilidad de los valores en el DataFrame, omitiendo las entradas NaN. ()

Ilustración 21

Código proceso ETL-5

	Producción Anual (Toneladas)	Tamaño de la Mina (Ha)	Número de Empleados	Año de Fundación	Acceso a Infraestructura (Km)	Número de Proyectos Activos
count	3123.000000	3123.000000	3123.000000	3123.000000	3123.000000	3123.000000
mean	9018.367275	322.720781	124.745117	2002.727506	52.387448	2.992635
std	3454.296192	97.900940	43.606935	7.406152	13.065513	1.405447
min	3000.000000	150.000000	50.000000	1990.000000	30.000000	1.000000
25%	6060.500000	241.000000	86.000000	1996.500000	41.000000	2.000000
50%	9000.000000	320.000000	125.000000	2003.000000	52.000000	3.000000
75%	11991.500000	405.000000	162.000000	2009.000000	64.000000	4.000000
max	15000.000000	500.000000	250.000000	2015.000000	75.000000	6.000000

Condiciones del Mercado Local	Competencia Local (Escala 1-5)	Precio Promedio por Tonelada (USD)	Rentabilidad por Empleado (USD)	RentabilidadNeta_num
3123.000000	3123.000000	3123.000000	3123.000000	3123.000000
2.999039	3.022094	33.257986	16.709534	0.278698
1.424931	1.389477	68.283378	29.002420	0.075454
1.000000	1.000000	0.301000	0.000000	0.130200
2.000000	2.000000	0.583000	0.000000	0.217100
3.000000	3.000000	6.973000	0.000000	0.280000
4.000000	4.000000	64.816000	27.660000	0.340000
5.000000	5.000000	650.000000	228.086000	0.750000

Nota. Función para presentar datos estadísticos del Dataset. Elaborado por autores.

Verificación de datos nulos Dataset

Con la línea de código siguiente, se detecta y cuenta la cantidad de valores faltantes o nulos (NaN) en cada columna del DataFrame data_clean. Primero, isnull() genera un DataFrame booleano del mismo tamaño donde cada posición es True si el valor es nulo y False si no lo es; luego, sum() suma estos valores True (que se cuentan como 1) por columna, devolviendo una serie con el total de valores nulos en cada columna. Esto permite identificar rápidamente qué columnas tienen datos faltantes y cuántos.

Ilustración 22

Código proceso ETL-6

```
# Comprobar valores nulos
data_clean.isnull().sum()
```

	0
Sector de la Mina/Cantera	0
Tipo de Mineral	0
Región	0
Producción Anual (Toneladas)	0
Tamaño de la Mina (Ha)	0
Número de Empleados	0
Año de Fundación	0
Fuente de Energía	0
Acceso a Infraestructura (Km)	0
Número de Proyectos Activos	0

Nota. Función para validar datos nulos del Dataset. Elaborado por autores.

Verificación de data numérica y categórica

Con la línea de código siguiente se seleccionan y separan columnas del DataFrame data_clean según su tipo de dato: la primera línea extrae todas las columnas que contienen datos numéricos (como enteros o flotantes) y las guarda en numeric_data, mientras que la segunda línea selecciona todas las columnas que no son numéricas (por ejemplo, texto o categorías) y las guarda en categorical_data. Esto permite trabajar de forma diferenciada con datos numéricos y categóricos para análisis o procesamiento posterior.

Ilustración 23

Código proceso ETL-7

```
# Seleccionar sólo columnas numéricas para el cálculo de la media
numeric_data = data_clean.select_dtypes(include=np.number)
```

```
# Seleccionar sólo columnas categóricas para aplicar label encoder
categorical_data = data_clean.select_dtypes(exclude=np.number)
```

```
numeric_data.columns
```

```
Index(['Producción Anual (Toneladas)', 'Tamaño de la Mina (Ha)',
       'Número de Empleados', 'Año de Fundación',
       'Acceso a Infraestructura (Km)', 'Número de Proyectos Activos',
       'Condiciones del Mercado Local', 'Competencia Local (Escala 1-5)',
       'Precio Promedio por Tonelada (USD)', 'Rentabilidad por Empleado (USD)',
       'IngAnualUSD_num', 'CostoAnualUSD_num', 'RentabBrutaUSD_num',
       'RentabilidadNeta_num', 'InverTecnoUSD_num', 'InverSostenib_num'],
      dtype='object')
```

```
categorical_data.columns
```

```
Index(['Sector de la Mina/Cantera', 'Tipo de Mineral', 'Región',
       'Fuente de Energía', 'Sostenibilidad Ambiental (Sí/No)',
       'Certificaciones de Calidad (ISO, etc.)', 'Exportación (Sí/No)',
       'Estrategia de Mercado (Local/Exportación)'],
      dtype='object')
```

Nota. Función para obtener los datos numéricos y categóricos. Elaborado por autores.

Crear variables para extraer nombres de columnas

El código define una lista llamada categorical_features que contiene los nombres de columnas específicas del DataFrame relacionadas con características categóricas, como sector, tipo de mineral, región y otras variables cualitativas; además, crea una variable numerical_features que automáticamente extrae los nombres de todas las columnas numéricas presentes en data_clean utilizando la función select_dtypes para incluir solo tipos de datos numéricos, facilitando así la diferenciación y el manejo separado de variables categóricas y numéricas en el análisis o modelado.

Ilustración 24
Código proceso ETL-8

```
# Suponiendo que 'Sector de la Mina/Cantera' es la columna categórica
# Debe identificar las columnas categóricas y reemplazarlas con los nombres de columna reales.

categorical_features = ['Sector de la Mina/Cantera','Tipo de Mineral','Región','Fuente de Energía','Sostenibilidad Ambiental (Sí/No)',
'Certificaciones de Calidad (ISO, etc.)','Exportación (Sí/No)','Estrategia de Mercado (Local/Exportación)']
numerical_features = data_clean.select_dtypes(include=np.number).columns
```

Nota. Función para extraer nombres de columnas. Elaborado por autores.

Escalado de variables numéricas

El código siguiente crea un objeto ColumnTransformer llamado preprocessor que aplica diferentes transformaciones a distintas columnas del DataFrame: a las columnas numéricas (numerical_features) les aplica una estandarización con StandardScaler(), que centra y escala los datos para que tengan media cero y desviación estándar uno; mientras que a las columnas categóricas (categorical_features) les aplica una codificación one-hot mediante OneHotEncoder con sparse_output=False para obtener una matriz densa en formato numpy array y handle_unknown='ignore' para manejar categorías no vistas sin error. Esto permite preprocesar simultáneamente datos heterogéneos (numéricos y categóricos) de forma adecuada y eficiente en un solo paso dentro de un pipeline de Scikit-Learn Luego, aplica un escalado estándar (StandardScaler) a las columnas numéricas para normalizar sus valores.

Ilustración 25

Código proceso ETL-9

```
# Cree un ColumnTransformer para aplicar diferentes preprocesamientos a diferentes columnas

preprocesador1 = ColumnTransformer(transformers=[('num', StandardScaler(), variables_numericas),

 ('cat', OneHotEncoder(sparse_output=False, handle_unknown='ignore'), variables_categoricas)])

# Aplicar el preprocesamiento y obtener los datos escalados

data_scaled = preprocessor.fit_transform(data_clean)
```

Nota. Función para transformar columnas. Elaborado por autores.

Convertir datos escalados

El código crea listas con los nombres de las características numéricas (num_feature_names) y las características categóricas transformadas (cat_feature_names) obtenidas a partir del codificador one-hot dentro del preprocessor, luego combina ambas listas en feature_names para tener todos los nombres de columnas resultantes tras la transformación; finalmente, convierte la matriz data_scaled (que contiene los datos escalados y codificados) en un DataFrame de pandas con esos nombres de columnas y manteniendo el mismo índice que el DataFrame original data_clean, facilitando así la interpretación y manipulación de los datos procesados.

Ilustración 26

Código proceso ETL-10

```
# Convierte los datos escalados de nuevo a un DataFrame de Pandas si es necesario

# Obtiene los nombres de las características después de la transformación



num_feature_names = list(numerical_features)
```

```
cat_feature_names =
list(preprocessor.named_transformers_['cat'].get_feature_names_out(categorical_features))

feature_names = num_feature_names + cat_feature_names



data_scaled = pd.DataFrame(data_scaled, columns=feature_names,
index=data_clean.index)
```

Nota. Función para transformar datos escalados. Elaborado por autores.

Análisis de la correlación de Pearson de las variables Xs entre ellas.

El código selecciona únicamente las columnas numéricas del DataFrame data_clean y calcula la matriz de correlación de Pearson entre estas variables para medir la relación lineal entre cada par de ellas; luego imprime esta matriz para su revisión y finalmente genera una visualización gráfica mediante un mapa de calor (heatmap) con anotaciones numéricas y una escala de colores que facilita identificar visualmente la fuerza y dirección de las correlaciones, ayudando a interpretar cómo se relacionan las variables numéricas entre sí en el conjunto de datos.

Ilustración 27

Código proceso ETL-11

```
# Selecciona sólo las columnas numéricas
numeric_data = data_clean.select_dtypes(include=np.number)

# Calcula la matriz de correlación de Pearson
corr_matrix = numeric_data.corr(method='pearson')

# Muestra la matriz de correlación
print("Matriz de correlación de Pearson:")
print(corr_matrix)

# Visualiza la matriz de correlación
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Matriz de Correlación de Pearson entre Variables Numéricas')
plt.show()
```

Nota. Función para transformar datos escalados. Elaborado por autores.

La matriz de valoración de Pearson revela resultados significativos en el sector minero, destacando principalmente la fuerte valoración (cercana a 1.0) entre las variables financieras:

Ilustración 28

IngAnualUSD_num ⟺ CostoAnualUSD_num

RentabBrutaUSD_num

Mientras la eficiencia por Empleado muestra correlaciones positivas importantes

IngAnualUSD_num

CostoAnualUSD_num ⟺ Rentabilidad por empleado USD

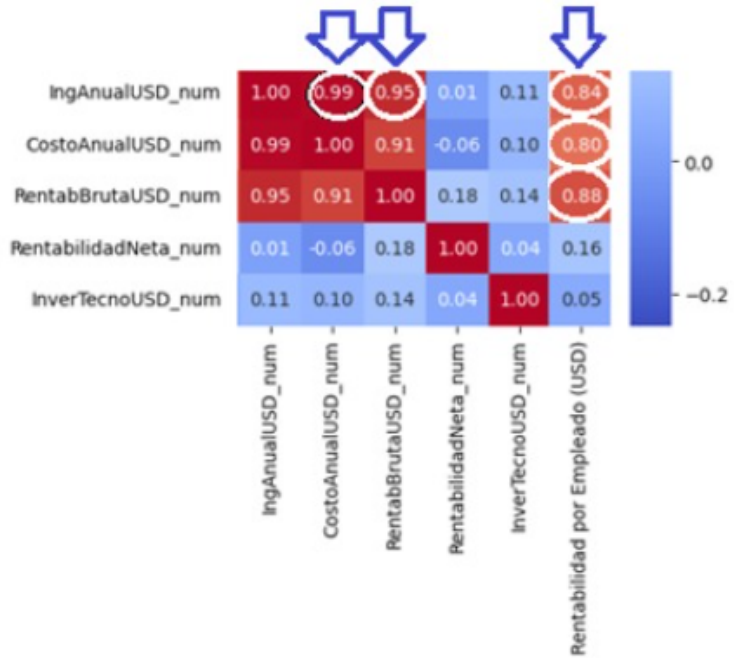
RentabBrutaUSD_num

Nota. Elaborado por autores.

(0.80-0.88) con estos indicadores financieros principales, evidenciando su relevancia como medida de operativa.

Ilustración 29

Correlación de Pearson entre variables numéricas



Nota. Matriz de correlación de Pearson. Elaborado por autores

Análisis de la correlación de Pearson de las variables Xs que afectan linealmente a la variable Y (Rentabilidad).

El código siguiente selecciona un conjunto específico de variables numéricas (numeric_cols) del DataFrame data_clean y calcula la correlación de Pearson entre cada una de estas variables y la variable objetivo 'RentabilidadNeta_num', almacenando los resultados en un diccionario; luego convierte este diccionario en un DataFrame ordenado de mayor a menor correlación para facilitar su interpretación, imprime esta tabla y finalmente genera un gráfico de barras horizontal con una paleta de colores que visualiza claramente la fuerza y dirección de la correlación de cada variable con la rentabilidad neta, proporcionando una visión rápida y efectiva de qué variables numéricas están más asociadas con el desempeño financiero.

Ilustración 30

Código correlación de Pearson entre las variables X y Y

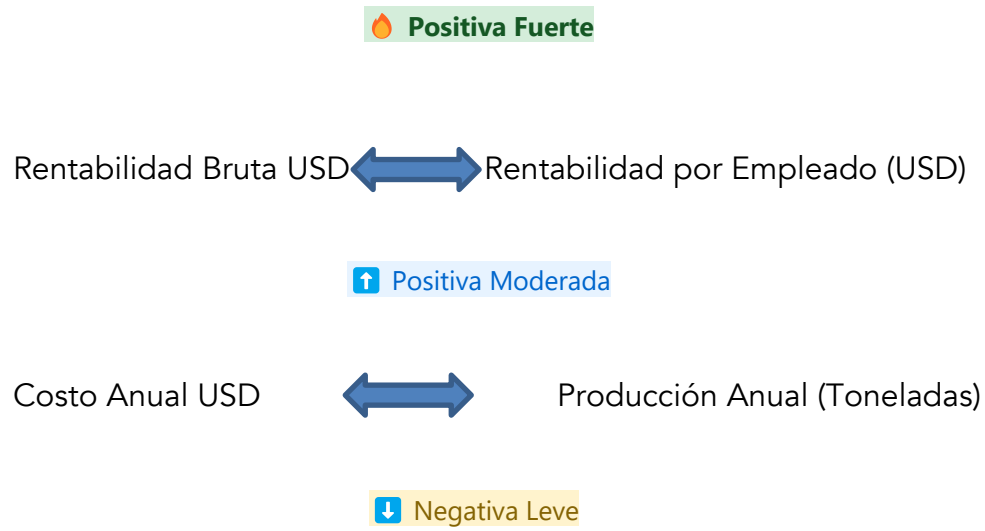
```
# Selecciona las variables numéricas (Xs) y la variable objetivo (Y)
# Asegúrate de que 'InverSostenib_num' está en data_clean
numeric_cols = [
    'Producción Anual (Toneladas)',
    'Tamaño de la Mina (Ha)',
    'Número de Empleados',
    'Año de Fundación',
    'Acceso a Infraestructura (Km)',
    'Número de Proyectos Activos',
    'Condiciones del Mercado Local',
    'Competencia Local (Escala 1-5)',
    'Precio Promedio por Tonelada (USD)',
    'Rentabilidad por Empleado (USD)',
    'IngAnualUSD_num',
    'CostoAnualUSD_num',
    'RentabBrutaUSD_num',
    'InverTecnoUSD_num',
    # Verifica si 'InverSostenib_num' existe en data_clean
    # Si no existe, es posible que debas crearla o
    # eliminarla de la lista numeric_cols
    'InverSostenib_num'
]

# Calcula la correlación de Pearson de cada X con la variable Y
correlations = {}
for col in numeric_cols:
    # Asegúrate de que ambas columnas existen en data_clean
    if col in data_clean.columns and 'RentabilidadNeta_num' in data_clean.columns:
        corr = data_clean[col].corr(data_clean['RentabilidadNeta_num'])
        correlations[col] = corr
    else:
        print(f"Advertencia: La columna '{col}' o 'RentabilidadNeta_num' no se encontró en data_clean.")
```

```
# Convierte a DataFrame para visualizar ordenado
corr_df = pd.DataFrame.from_dict(correlations, orient='index', columns=['Correlación con RentabilidadNeta_num'])
corr_df = corr_df.sort_values(by='Correlación con RentabilidadNeta_num', ascending=False)

print("Correlación de Pearson de las variables X con la variable Y (RentabilidadNeta_num):")
print(corr_df)

# Visualización
plt.figure(figsize=(8,6))
sns.barplot(x=corr_df['Correlación con RentabilidadNeta_num'], y=corr_df.index, palette='coolwarm')
plt.title('Correlación de Pearson de variables X con Rentabilidad Neta (%)')
plt.xlabel('Correlación de Pearson')
plt.ylabel('Variable')
plt.tight_layout()
plt.show()
```

Nota. Código de correlación de Pearson de variable Rentabilidad. Elaborado por autores.

La matriz de valoración de Pearson claramente que las variables más importantes para maximizar la rentabilidad neta son la Rentabilidad Bruta y la Rentabilidad por Empleado, mientras que variables como el Tamaño de la Mina tienen un impacto negativo significativo.

Ilustración 31



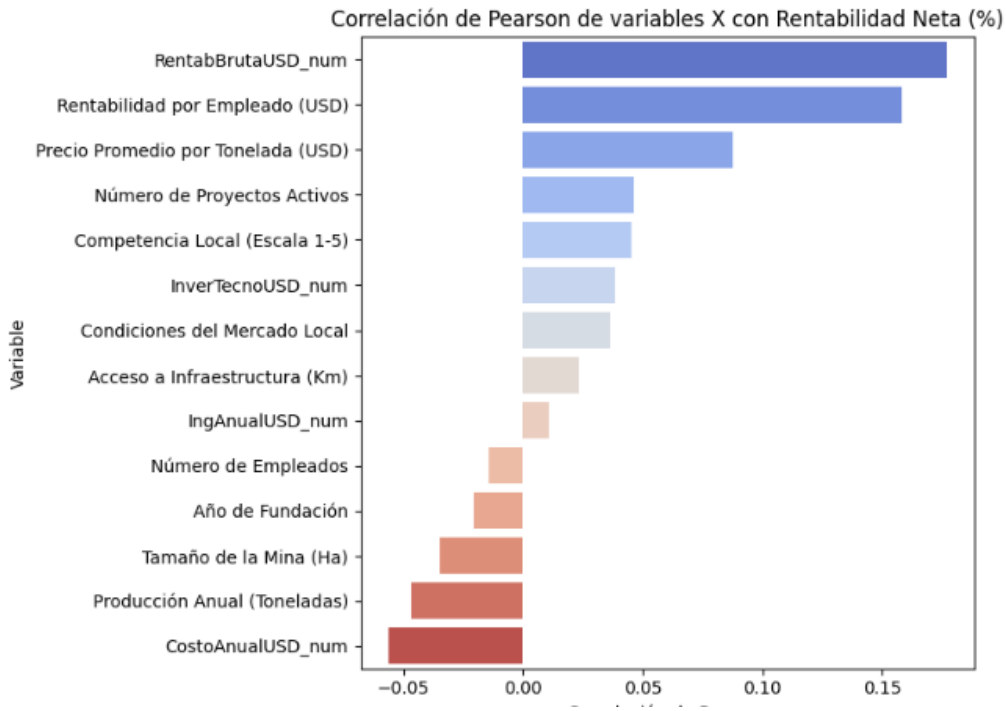
Tamaño de la Mina (Ha)	-0.04	❌ Negativa Significativa	MUY DESFAVORABLE	Minas más grandes = Menor rentabilidad neta

Nota. Elaborado por autores.

La rentabilidad neta minera se asocia positivamente con la rentabilidad bruta USD (r=+0.16) y la rentabilidad por empleado USD (r=+0.15), mientras que el precio promedio por tonelada USD muestra una correlación positiva moderada (r=+0.09). Las variables operacionales (tamaño de mina, producción anual, costos anuales) muestran correlaciones negativas débiles (r=-0.04, -0.03 y -0.02, respectivamente), y el número de empleados y el año de fundación presentan correlaciones negativas mínimas (r=-0.01). La conclusión es que la rentabilidad se maximiza optimizando eficiencia y productividad, no aumentando escala o volumen.

Ilustración 32

Correlación de Pearson entre las variables X (numéricas) y la variable Y (RentabilidadNeta_num)



Nota. Matriz de correlación de Pearson de variable Rentabilidad. Elaborado por autores.

Aprendizaje No Supervizado
Reducción de dimensionalidad mediante PCA.

Como paso previo al análisis de clustering, se implementó una reducción de dimensionalidad utilizando Análisis de Componentes Principales (PCA) para facilitar la visualización y mejorar el rendimiento de los algoritmos de agrupación.

En este proceso, primero se instancia un objeto PCA configurado para retener únicamente dos componentes principales, lo que permite transformar datos multidimensionales a un espacio bidimensional mientras maximiza la retención de varianza. Los datos escalados (normalizados previamente) se transforman mediante el método fit_transform(), proyectándolos en este nuevo espacio de menor dimensionalidad. Posteriormente, estos componentes principales se organizan en un DataFrame con etiquetas "PCA1" y "PCA2", representando las nuevas variables sintéticas que capturan la máxima variabilidad de los datos originales. Finalmente, se visualiza esta proyección bidimensional mediante un gráfico de dispersión utilizando seaborn, donde cada punto representa una empresa del sector minero ecuatoriano, permitiendo identificar visualmente patrones, agrupaciones naturales y relaciones entre observaciones que no serían perceptibles en el espacio multidimensional original, facilitando así la interpretación y validación visual de los resultados de Clustering.

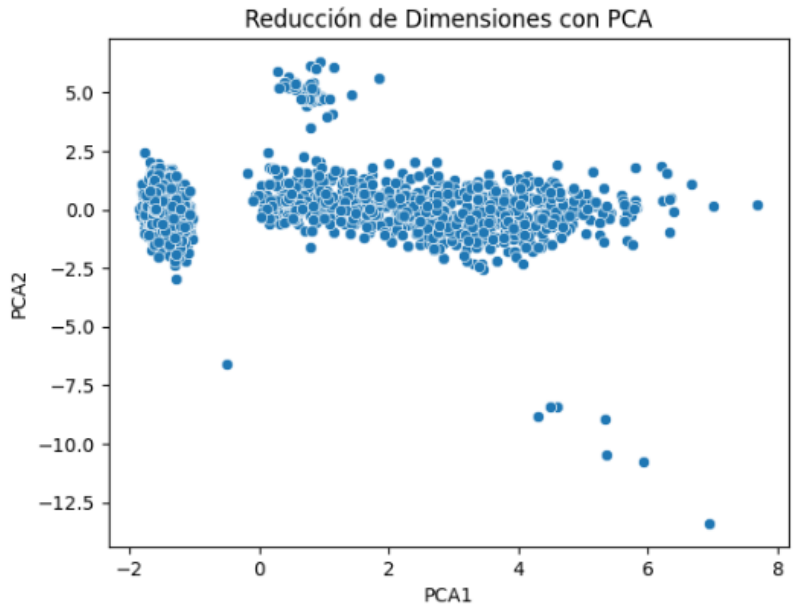
Ilustración 33
Reducción de dimensiones

```
# Elimina columnas dummies duplicadas, dejando solo la primera aparición
dummies_unique = dummies_df.loc[:, ~dummies_df.T.duplicated()]
print(f"Shape original de dummies: {dummies_df.shape}")
print(f"Shape después de eliminar duplicadas: {dummies_unique.shape}")
```

```
# Manejar valores faltantes antes de aplicar PCA
# Reemplazar NaN con la media de cada columna
data_scaled = data_scaled.fillna(data_scaled.mean()) # or data_scaled.dropna()

pca = PCA(n_components=2)  # Reducimos a 2 componentes para visualización
data_pca = pca.fit_transform(data_scaled)

# Convertir los resultados en un DataFrame
data_pca_df = pd.DataFrame(data_pca, columns=['PCA1', 'PCA2'])

# Visualizar los datos reducidos
sns.scatterplot(x='PCA1', y='PCA2', data=data_pca_df)
plt.title("Reducción de Dimensiones con PCA")
plt.show()
```

Nota. Elaborado por autores.

Ilustración 34



Nota. Visualización de los datos del sector minero ecuatoriano después de la reducción de dimensionalidad con PCA. Elaborada por autores.

La Ilustración 34 "Reducción de Dimensiones con PCA" muestra la distribución de empresas mineras en un espacio bidimensional tras aplicar Análisis de Componentes Principales. La visualización revela varios patrones interesantes: una concentración principal de datos entre los valores -2 y 6 del PCA1 y -2.5 y 2.5 del PCA2, donde se agrupa la mayoría de las empresas; un grupo distintivo en la esquina superior izquierda (PCA1 ≈ 0, PCA2 ≈ 5) que sugiere empresas con características particulares que las separan claramente del conjunto principal; un grupo vertical aislado a la izquierda (PCA1 ≈ -1.5) que podría representar empresas con perfil

operativo o financiero específico; y valores atípicos notables en la parte inferior derecha (PCA1 > 4, PCA2 < -9) que probablemente corresponden a compañías con características extremas en ciertas variables. Esta distribución no homogénea indica que existen subgrupos naturales en el sector minero que podrían justificar una segmentación en clusters, donde el primer componente principal (PCA1) parece capturar la mayor variabilidad de los datos, posiblemente relacionada con indicadores financieros dado su amplio rango de valores.

Determinación de número óptimo de clusters

El código siguiente calcula el número óptimo de clusters para un conjunto de datos escalados (data_scaled) utilizando el método del codo: para cada valor de clusters i desde 1 hasta 10, crea y entrena un modelo KMeans con esos i clusters, inicialización k-means++, un máximo de 300 iteraciones y 10 inicializaciones para asegurar estabilidad, luego almacena el valor de la inercia (WCSS, suma de las distancias cuadradas dentro de los clusters) en una lista wcss; este proceso permite posteriormente graficar la inercia frente al número de clusters para identificar el punto donde la mejora comienza a ser marginal, ayudando a seleccionar el número adecuado de clusters.

Ilustración 35

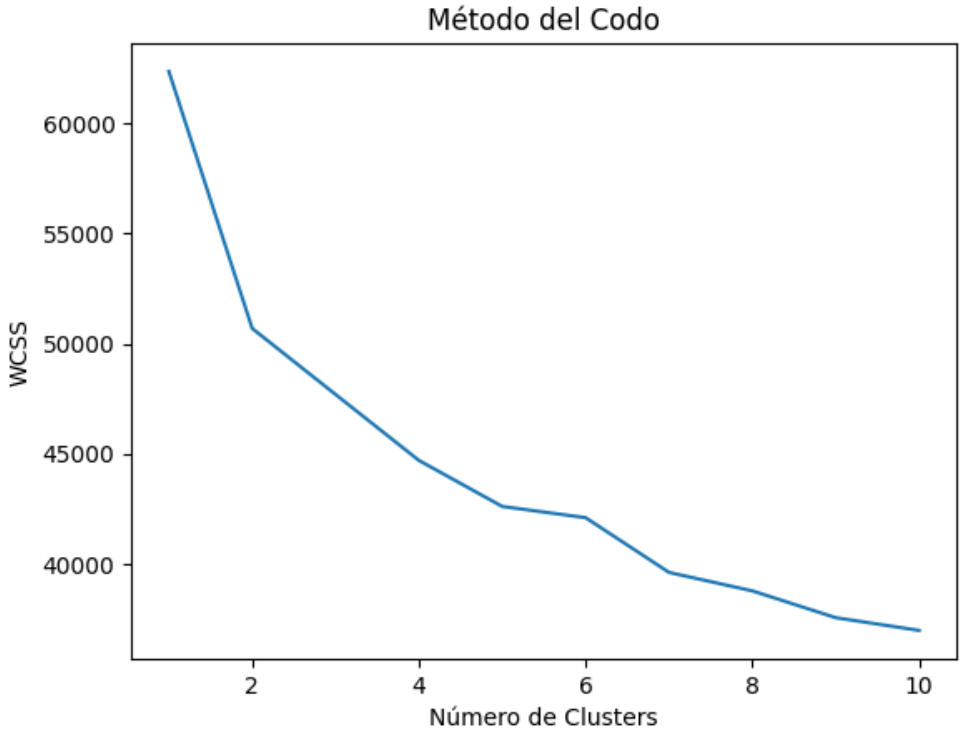
Determinar número óptimo de clusters

```
wcss = []
for i in range(1,11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=42)
    kmeans.fit(data_scaled)
    wcss.append(kmeans.inertia_)
```

```
# Graficar el método del codo
plt.plot(range(1,11), wcss)
plt.title('Método del Codo')
plt.xlabel('Número de Clusters')
plt.ylabel('WCSS')
plt.show()
```

Nota. Elaborada por autores.

Ilustración 36



Nota. Visualización de método del Codo. Elaborada por autores.

La gráfica "Método del Codo" muestra la relación entre el número de clusters y el valor WCSS (Within-Cluster Sum of Squares), que mide la variabilidad dentro de cada cluster. Se observa una clara tendencia descendente donde el WCSS disminuye a medida que aumenta el número de clusters. Inicialmente, hay una caída pronunciada desde aproximadamente 62,000 con un cluster hasta cerca de 50,000 con dos clusters, lo que representa una reducción de aproximadamente 12,000 unidades. La pendiente continúa descendiendo significativamente hasta llegar a 4-5 clusters (con WCSS alrededor de 42,000-43,000), punto a partir del cual la curva comienza a aplanarse notablemente. Entre 6 y 10 clusters, la reducción es mucho más gradual, alcanzando un WCSS de aproximadamente 37,000 con 10 clusters. Este comportamiento sugiere que el "punto de codo" óptimo se encuentra en 4 clusters, donde se logra un equilibrio entre minimizar la variabilidad interna de los clusters y evitar un número excesivo de agrupaciones que podrían resultar en sobreajuste.

Aplicación de K-means.

Se aplicara el algoritmo K-means con 4 clusters al conjunto de datos escalados (data_scaled), utilizando la inicialización k-means++, un máximo de 300 iteraciones y 10 inicializaciones para asegurar resultados consistentes, y almacena las etiquetas de cluster asignadas a cada muestra en y_kmeans; luego, visualiza los clusters resultantes en un gráfico de dispersión bidimensional usando dos componentes principales (PCA1 y PCA2) almacenadas en data_pca_df, coloreando los puntos según su cluster asignado con una paleta viridis, ajustando el tamaño y la transparencia para mejorar la visualización, y mostrando un título descriptivo para facilitar la interpretación de la agrupación obtenida.

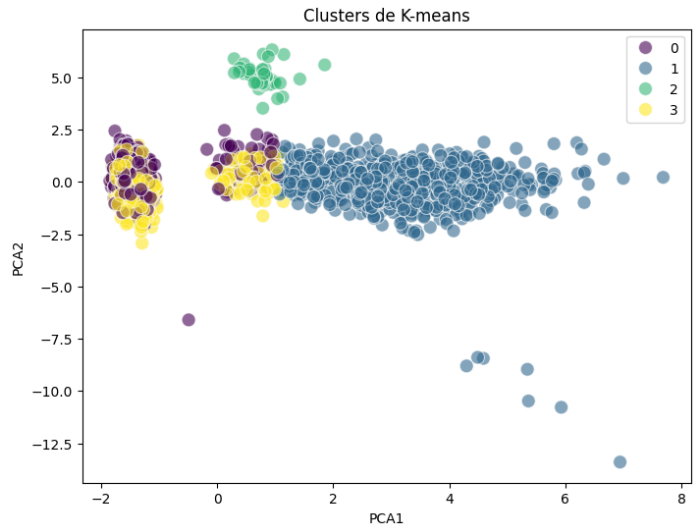
Ilustración 37

Código aplicación K-means (4)

```
# Aplicar K-means con el número óptimo de clusters (suponiendo 4 clusters)
kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=42)
y_kmeans = kmeans.fit_predict(data_scaled)

# Visualizar los clusters
plt.figure(figsize=(8,6))
sns.scatterplot(x='PCA1', y='PCA2', hue=y_kmeans, palette='viridis', data=data_pca_df, s=100, alpha=0.6)
plt.title("Clusters de K-means")
plt.show()
```

Nota. Elaborada por autores.

Ilustración 38



Nota. Visualización de K-means (K=4). Elaborada por autores.

La gráfica muestra un clustering K-means con 4 clusters identificados (0, 1, 2 y 3) visualizados en un espacio bidimensional utilizando componentes principales (PCA1 y PCA2). El cluster 0 (color púrpura) es el más numeroso y se extiende principalmente entre los valores 2 y 6 del eje PCA1, con una distribución en PCA2 entre -2.5 y 2, mostrando una forma alargada. El cluster 1 (color azul) se concentra claramente en la zona izquierda de la gráfica (PCA1 cercano a -2) y tiene una distribución compacta y bien definida en PCA2 entre -2.5 y 2.5. El cluster 2 (color verde) ocupa principalmente la zona central con valores de PCA1 entre -0.5 y 2, y destaca por tener un subgrupo con valores altos de PCA2 (alrededor de 5) que lo diferencia claramente. El cluster 3 (color amarillo) es el menos numeroso y más disperso, ubicándose principalmente en valores extremos negativos de PCA2 (entre -6 y -14) y distribuido a lo largo del eje PCA1 entre 0 y 8.

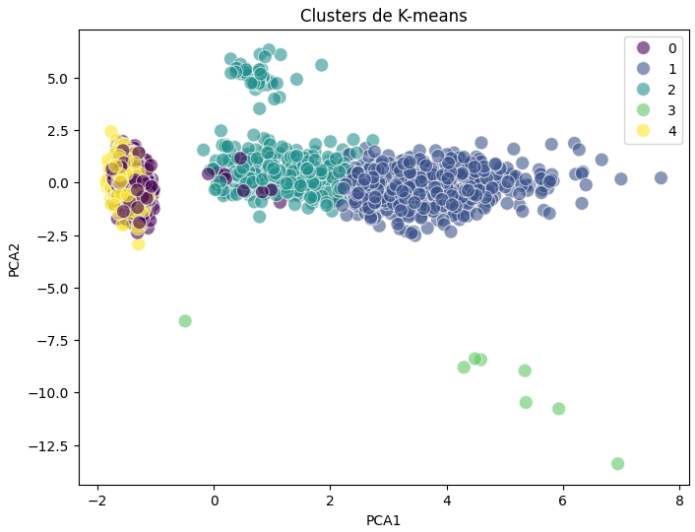
Ilustración 39

Código aplicación K-means (5)

```
# Aplicar K-means con el número óptimo de clusters (suponiendo 5 clusters)
kmeans = KMeans(n_clusters=5, init='k-means++', max_iter=300, n_init=10, random_state=42)
y_kmeans = kmeans.fit_predict(data_scaled)
# Visualizar los clusters
plt.figure(figsize=(8,6))
sns.scatterplot(x='PCA1', y='PCA2', hue=y_kmeans, palette='viridis', data=data_pca_df, s=100, alpha=0.6)
plt.title("Clusters de K-means")
plt.show()
```

Nota. Elaborada por autores.

Ilustración 40



Nota. Visualización de K-means (K=5). Elaborada por autores.

La nueva gráfica presenta un análisis K-means con 5 clusters (numerados del 0 al 5), mostrando una segmentación aún más refinada de los datos. El cluster 1 (azul) continúa siendo uno de los más numerosos, ocupando principalmente la región entre valores de PCA1 de 2 a 7, con una distribución en PCA2 entre -2.5 y 2.5. El cluster 0 (púrpura) se ha concentrado aún más, apareciendo como un grupo pequeño principalmente en la zona izquierda (PCA1 cerca de -2), entre valores de PCA2 de 0 a 2.5. El cluster 2 (azul verdoso) mantiene su posición en la parte izquierda de la gráfica con valores de PCA1 cercanos a -2 y PCA2 entre -2.5 y 1.5.

Lo más notable en esta nueva segmentación es que el cluster 3 (verde claro) ha crecido significativamente y ahora ocupa una extensa área que abarca valores de PCA1 desde 0 hasta aproximadamente 3, con una distribución en PCA2 entre -2 y 2.5, además de mantener su característica subpoblación con valores elevados de PCA2 (alrededor de 5). El cluster 4 (amarillo) conserva su posición distintiva en valores extremadamente negativos de PCA2 (-6 a -14) distribuidos a lo largo del eje PCA1 entre 0 y 8, manteniéndose como el cluster más separado del resto. El cluster 5 (no claramente visible en esta representación específica o con muy pocos puntos) parece haber capturado alguna subpoblación adicional.

Esta segmentación en 5 clusters sugiere una estructura de datos más granular, donde especialmente la distribución entre los clusters 1 y 3 resulta más clara que en la visualización anterior, proporcionando una mejor separación entre poblaciones y potencialmente revelando patrones de comportamiento más específicos en los datos.

Aplicación de índice de Silhouette

El código siguiente calcula el índice de Silhouette para diferentes números de clusters, iterando desde 2 hasta 10 clusters: para cada número de clusters, entrena un modelo K-means en los datos escalados, predice las etiquetas de cluster para cada punto y calcula el índice de Silhouette promedio, que mide la calidad de la agrupación; luego, almacena este índice en una lista llamada silhouette_scores.

Ilustración 41

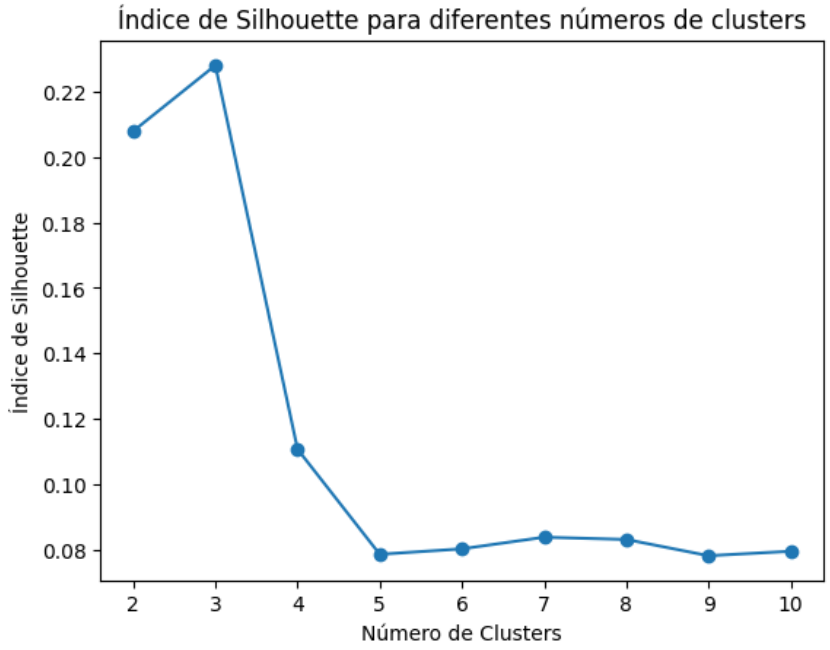
Código índice de Silhouette

```
# prompt: Calcula y grafica el indice de Silhouette para esta caso
# Calcular el índice de Silhouette para diferentes números de clusters
silhouette_scores = []
for n_clusters in range(2, 11): # Probar de 2 a 10 clusters
  kmeans = KMeans(n_clusters=n_clusters, init='k-means++', max_iter=300, n_init=10, random_state=42)
  cluster_labels = kmeans.fit_predict(data_scaled)
  silhouette_avg = silhouette_score(data_scaled, cluster_labels)
  silhouette_scores.append(silhouette_avg)
```

```
# Graficar el índice de Silhouette
plt.plot(range(2, 11), silhouette_scores, marker='o')
plt.xlabel("Número de Clusters")
plt.ylabel("Índice de Silhouette")
plt.title("Índice de Silhouette para diferentes números de clusters")
plt.show()
```

Nota. Elaborada por autores.

Ilustración 42



Nota. Visualización Silhouette. Elaborada por autores.

La gráfica muestra el Índice de Silhouette para diferentes números de clusters, siendo este un indicador clave para evaluar la calidad de la

segmentación obtenida mediante K-means. Se observa que el valor más alto del Índice de Silhouette (0.225) se alcanza cuando se utilizan 3 clusters, lo que sugiere que esta es la segmentación óptima para los datos analizados. Con 2 clusters, el índice también muestra un valor relativamente alto (alrededor de 0.21), pero ligeramente inferior al máximo observado con 3 clusters. A partir de 4 clusters, se aprecia una caída pronunciada en el índice (0.11), que continúa descendiendo hasta estabilizarse en valores bajos (0.08) entre 5 y 10 clusters, con variaciones mínimas. Esta tendencia descendente indica que agregar más clusters después de 3 empeora progresivamente la calidad de la segmentación. Los valores relativamente bajos del índice en general (todos por debajo de 0.23) sugieren que los datos no presentan una estructura de clustering muy fuerte o natural, lo que podría indicar cierta superposición entre los diferentes grupos identificados.

Aceptando 5 clusters hallados por K-means

Se crea una copia del DataFrame escalado (data_scaled) llamado data_with_clusters y se añade una nueva columna llamada 'cluster' que contiene la asignación de cada muestra a un cluster según el resultado de K-means (y_kmeans); luego, para cada uno de los 5 clusters, se filtra los datos correspondientes a ese cluster y muestra un mensaje de análisis, aunque la descripción estadística de las características numéricas (media, desviación estándar, mínimos, máximos, etc.) solo se imprime para el último cluster del bucle, permitiendo así examinar las similitudes y diferencias estadísticas entre los sujetos de estudio agrupados en cada cluster.

Ilustración 43

Código análisis de cluster

```
# Crear una copia del DataFrame
data_with_clusters = data_scaled.copy()

# Asignar los clusters al DataFrame copiado
data_with_clusters['cluster'] = y_kmeans

# Analizar las similitudes entre los sujetos de estudio de cada cluster
for cluster_num in range(5): # Iterar sobre los 6 clusters
  print(f"\nAnálisis del Cluster {cluster_num}:")
  cluster_data = data_with_clusters[data_with_clusters['cluster'] == cluster_num]

# Descripción estadística de las características numéricas
print ("Descripción estadística de las características numéricas:")
print (cluster_data[numerical_features].describe())
```

	Cluster	Puntos_Extraccion	Porcentaje	Produccion_Anual	Tamaño_Mina
count	5.000	5.000	5.000	5.000	5.000
mean	2.000	466.600	20.000	-0.163	-0.009
std	1.581	425.543	18.239	0.652	0.034
min	0.000	75.000	3.200	-0.847	-0.044
25%	1.000	154.000	6.600	-0.480	-0.041
50%	2.000	412.000	17.700	-0.361	-0.008
75%	3.000	544.000	23.300	0.001	0.016
max	4.000	1148.000	49.200	0.870	0.034

	Num_Empleados	Acceso_Infraestructura	Condiciones_Mercado
count	5.000	5.000	5.000
mean	-0.063	0.075	-0.160
std	0.421	0.352	0.325
min	-0.727	-0.294	-0.707
25%	-0.046	-0.003	-0.200
50%	0.002	0.003	-0.007
75%	0.011	0.009	0.011
max	0.445	0.662	0.101

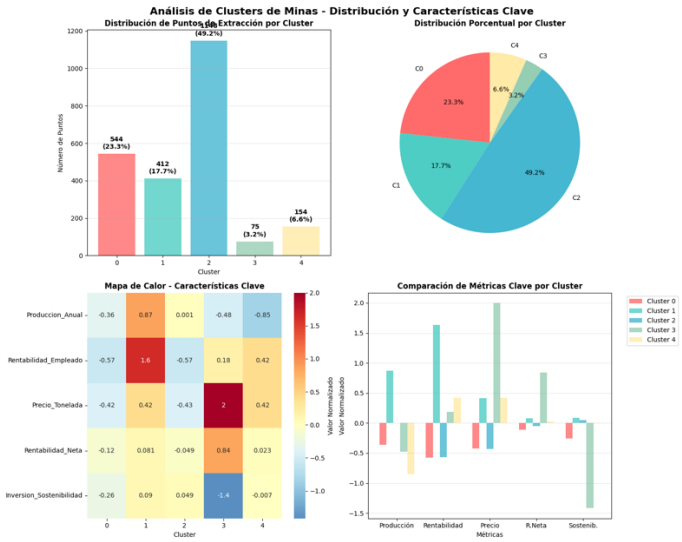
	Precio_Tonelada	Rentabilidad_Empleado	Ingresos_Anuales
count	5.000	5.000	5.000
mean	1.242	0.219	0.174
std	2.818	0.909	1.070
min	-0.429	-0.575	-0.664
25%	-0.424	-0.571	-0.663
50%	0.418	0.184	-0.251
75%	0.419	0.421	0.594
max	6.227	1.635	1.853

	Rentabilidad_Neta	Inversion_Sostenibilidad
count	5.000	5.000
mean	0.156	-0.309
std	0.389	0.634
min	-0.115	-1.417
25%	-0.049	-0.259
50%	0.023	-0.007
75%	0.081	0.049
max	0.839	0.090

```
Cluster 0: 544 puntos (23.3%)
  Operaciones de Pequeña Escala con Baja Rentabilidad

Cluster 1: 412 puntos (17.7%)
  Operaciones de Alta Producción con Excelente Rentabilidad

Cluster 2: 1148 puntos (49.2%)
  Operaciones Promedio con Estabilidad Moderada

Cluster 3: 75 puntos (3.2%)
  Operaciones Especializadas de Alto Valor

Cluster 4: 154 puntos (6.6%)
  Operaciones en Transición con Potencial Medio
```

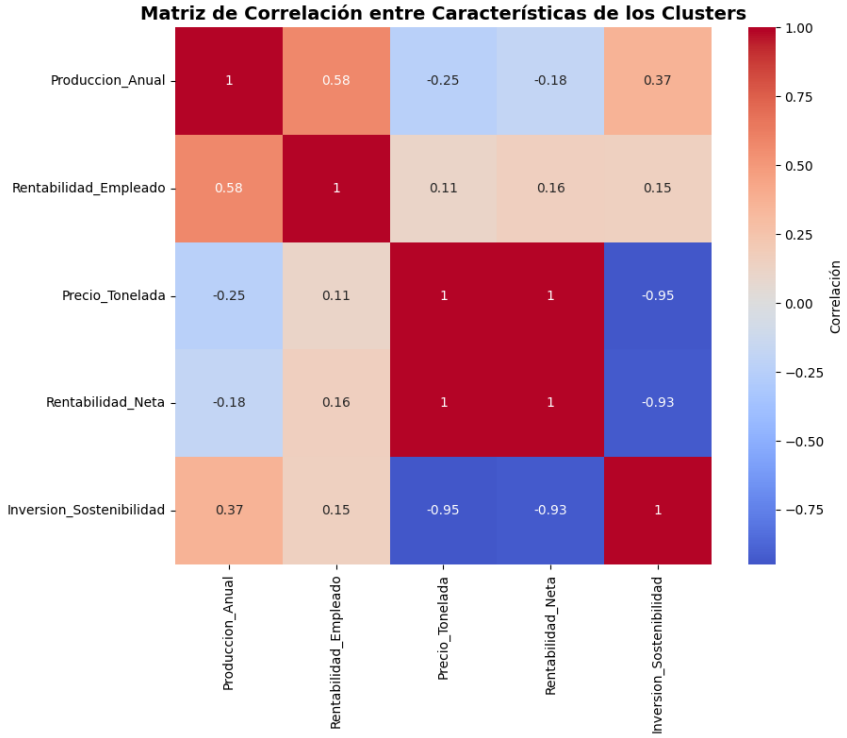
Nota. Elaborada por autores.

Ilustración 44



Nota. Elaborada por autores.

Ilustración 45



Nota. Visualización del análisis por cluster. Elaborada por autores.

Este cluster 0 agrupa 544 puntos que representan el 23.3% del total de puntos de extracción analizados. Se caracterizan por ser operaciones de pequeña escala con desempeño económico limitado. Los puntos de extracción presentan una producción anual significativamente por debajo del promedio (-0.36) y emplean menos personal (0.44), lo que sugiere operaciones más modestas. Un aspecto notable es su acceso limitado a infraestructura (-0.29), lo que puede explicar parcialmente sus desafíos operativos. El mercado local en el que operan presenta condiciones desfavorables (-0.71), con precios por tonelada por debajo del promedio (-0.42). La rentabilidad por empleado es considerablemente baja (-0.57), así como los ingresos (-0.66) y la rentabilidad bruta (-0.64), indicando operaciones con márgenes muy ajustados. La inversión en sostenibilidad también es limitada (-0.26), sugiriendo que estas operaciones priorizan la supervivencia inmediata sobre inversiones a largo plazo.

El Cluster 1 incluye 412 puntos que representan el 17.7% del total y constituyen el segmento más exitoso del sector. Estas operaciones se distinguen por una producción anual muy superior (0.87), siendo las más productivas del análisis. Su fortaleza económica es evidente en todos los indicadores financieros: rentabilidad por empleado excepcionalmente alta (1.64), ingresos anuales robustos (1.85), y rentabilidad bruta sólida (1.80). Los precios que obtienen por tonelada son superiores (0.42), reflejando posiblemente mejor calidad del producto o condiciones de mercado más favorables. Sus condiciones de mercado local son positivas (0.10), y mantienen un equilibrio en términos de competencia local. La inversión en tecnología es notable (0.23), con una desviación estándar alta que sugiere estrategias diversificadas de modernización. Este cluster representa el modelo de operaciones exitosas y sostenibles en el sector.

Con 1,148 puntos el Cluster 2, este es el cluster más numeroso representando el 49.2% del total de puntos de extracción, y constituye el perfil típico del sector minero. Las variables operacionales se mantienen cerca de los valores promedio: producción anual neutra (0.001), tamaño de mina estándar (0.02), y número de empleados equilibrado (0.01). Sin embargo, enfrentan desafíos económicos similares al Cluster 0, con rentabilidad por empleado baja (-0.57), ingresos limitados (-0.66) y precios por debajo del promedio (-0.43). La rentabilidad neta es ligeramente negativa (-0.05), sugiriendo operaciones en el punto de equilibrio. A pesar de estos desafíos, mantienen cierta inversión en sostenibilidad (0.05),

indicando un enfoque balanceado entre rentabilidad y responsabilidad ambiental. Este cluster representa la masa crítica del sector, con operaciones estables, pero con márgenes ajustados.

El Cluster 3, aunque pequeño con solo 75 puntos representando el 3.2% del total, presenta características únicas y altamente diferenciadas. Se caracteriza por operaciones de menor escala en producción (-0.48) y empleo (-0.73), pero con una propuesta de valor premium evidente en sus precios excepcionales por tonelada (6.23), muy superiores al resto del mercado. Tienen mejor acceso a infraestructura (0.66) y operan principalmente en fundaciones más recientes (0.53), sugiriendo operaciones modernas y bien ubicadas. Su rentabilidad neta es la más alta de todos los clusters (0.84), demostrando que el modelo de bajo volumen y alto valor puede ser extremadamente rentable. Sin embargo, su inversión en sostenibilidad es significativamente baja (-1.42), lo que podría representar un riesgo a largo plazo. La competencia local que enfrentan es limitada (-0.40), sugiriendo nichos de mercado especializados.

El Cluster 4 agrupa 154 minas que representan el 6.6% del total y presentan un perfil mixto con características tanto desafiantes como prometedoras. Tienen la producción más baja de todos los clusters (-0.85), pero compensan esto con indicadores financieros moderadamente positivos: rentabilidad por empleado aceptable (0.42), ingresos regulares (0.59) y rentabilidad bruta positiva (0.55). Los precios que obtienen son competitivos (0.42), similares al Cluster 1. Sus condiciones de mercado local son neutras (0.01) y la competencia es manejable (0.08). La rentabilidad neta es baja pero positiva (0.02), sugiriendo operaciones que están encontrando su equilibrio. La inversión en sostenibilidad es mínima (-0.007), indicando un enfoque conservador en este aspecto. Este cluster representa operaciones que podrían estar en proceso de optimización o transición hacia modelos más eficientes.

La segmentación de estos 2,333 puntos de extracción revela una industria diversificada donde coexisten diferentes estrategias de negocio, desde operaciones de volumen con excelente rentabilidad hasta nichos especializados de alto valor, pasando por la mayoría de operaciones que luchan por mantener márgenes en un entorno competitivo. La concentración del 49.2% de las operaciones en el Cluster 2 sugiere que la mayor parte de la industria opera en condiciones estándar con rentabilidades moderadas.

Análisis de Columnas categóricas.

Se procede a definir una lista llamada categorical_cols que contiene los nombres de las columnas categóricas del DataFrame, es decir, aquellas que representan información cualitativa como sector, tipo de mineral, región, fuente de energía, certificaciones, entre otras. Luego, utiliza el método describe(include='object') sobre estas columnas para generar un resumen estadístico de las variables categóricas, el cual incluye para cada columna el conteo de valores no nulos (count), el número de categorías únicas (unique), el valor más frecuente (top) y la frecuencia de ese valor más frecuente (freq). F0inalmente, el método transpose() se usa para mostrar este resumen en formato de tabla, con cada variable como fila, facilitando la interpretación de la distribución y predominancia de las categorías en cada columna del conjunto de datos.

Ilustración 46
Código análisis de columnas categóricas

```
# Listar las columnas categóricas
categorical_cols = [
    'Sector de la Mina/Cantera',
    'Tipo de Mineral',
    'Región',
    'Fuente de Energía',
    'Sostenibilidad Ambiental (Sí/No)',
    'Certificaciones de Calidad (ISO, etc.)',
    'Exportación (Sí/No)',
    'Estrategia de Mercado (Local/Exportación)'
]

# Resumen estadístico de las variables categóricas
print(data_clean[categorical_cols].describe(include='object').transpose())
```

```
                                   count unique                top  \
Sector de la Mina/Cantera           2333      3  Cantera de Piedra
Tipo de Mineral                     2333      4               Oro
Región                              2333      1            El Oro
Fuente de Energía                   2333      2  Energía Eléctrica
Sostenibilidad Ambiental (Sí/No)    2333      2                Sí
Certificaciones de Calidad (ISO, etc.)  2333  5                 -
Exportación (Sí/No)                 2333      2                Sí
Estrategia de Mercado (Local/Exportación)  2333  2       Exportación
```

Nota. Visualización del análisis por cluster. Elaborada por autores.

Este código nos brinda un resumen estadístico de las variables categóricas, muestra que todas las columnas cuentan registros completos sin valores faltantes, y revela la cantidad de categorías distintas en cada variable, como las 3 categorías en "Sector de la Mina/Cantera" frente a la única

categoría en "Región" (todas pertenecientes a "El Oro"). Además, identifica la categoría más frecuente en cada variable, por ejemplo, "Minería de Oro" en "Sector de la Mina/Cantera" con 1061 apariciones y "Energía Eléctrica" en "Fuente de Energía" con 1596 apariciones, lo que permite comprender la distribución y predominancia de las categorías, detectar variables con poca variabilidad y obtener información clave para el análisis y modelado de los datos.

Ilustración 47

Código distribución de frecuencias

```
for col in categorical_cols:
    print(f"\nDistribución de '{col}':")
    print(data_clean[col].value_counts())
    print("-" * 40)
```

```
Distribución de 'Sector de la Mina/Cantera':
Sector de la Mina/Cantera
Cantera de Piedra    808
Minería de Oro       789
Minería de Cobre     736
Name: count, dtype: int64
----------------------------------------

Distribución de 'Tipo de Mineral':
Tipo de Mineral
Oro              789
Cobre            736
Piedra Caliza    411
Arena y Grava    397
Name: count, dtype: int64
```

Nota. Visualización de cada variable categórica. Elaborada por autores.

Este código de la ilustración 47, recorre cada columna listada en categorical_cols y para cada una imprime su nombre seguido de la distribución de frecuencias de sus categorías, es decir, muestra cuántas veces aparece cada valor único dentro de esa columna utilizando value_counts(). Después de listar la distribución de cada variable categórica, imprime una línea de guiones para separar visualmente los resultados, facilitando así la comprensión y comparación de la frecuencia de las diferentes categorías en cada variable del conjunto de datos.

Los resultados que se obtienen muestran la distribución de frecuencias de las categorías dentro de cada variable categórica listada en el conjunto de datos. Por ejemplo, en la columna "Sector de la Mina/Cantera" hay tres categorías principales: "Minería de Oro" y "Cantera de Piedra" con 411

registros cada una, y "Minería de Cobre" con 736, lo que indica que estas son las actividades predominantes en tu muestra. En "Tipo de Mineral", las categorías más comunes son "Oro" (789), "Cobre" (736), seguidas de "Piedra Caliza" y "Arena y Grava", mostrando la variedad de minerales explotados. La columna "Región" tiene solo una categoría ("El Oro") con todos los registros, lo que significa que los datos provienen exclusivamente de esa región. En cuanto a la "Fuente de Energía", se observa un uso casi equilibrado entre "Energía Eléctrica" (1176) y "Diésel" (1157). Para la variable "Sostenibilidad Ambiental (Sí/No)", la mayoría reporta "Sí" (1191), aunque casi la mitad indica "No" (1142), reflejando diversidad en prácticas ambientales. Las "Certificaciones de Calidad" muestran que la mayoría no tiene certificación ("-": 1425), mientras que otras combinaciones de ISO están presentes en menor cantidad. En las variables "Exportación" y "Estrategia de Mercado" también hay una distribución muy pareja entre "Sí" y "No" o "Local" y "Exportación", respectivamente.

Ilustración 48

Código gráficos columnas categóricas

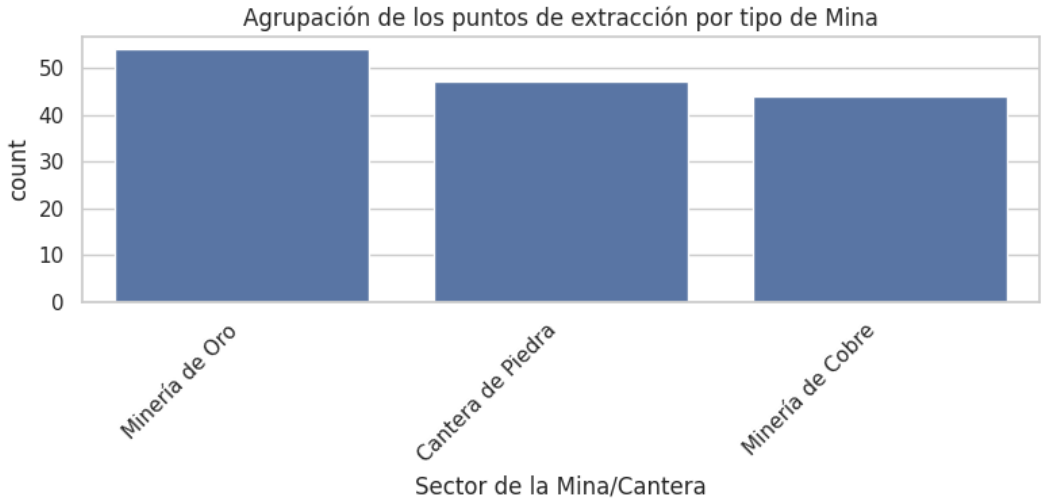
```
import matplotlib.pyplot as plt
import seaborn as sns

for col in categorical_cols:
    plt.figure(figsize=(8, 4))
    sns.countplot(data=data_clean, x=col, order=data_clean[col].value_counts().index)
    plt.title(f"Distribución de {col}")
    plt.xticks(rotation=45, ha='right')
    plt.tight_layout()
    plt.show()
```

Nota. Codificación de cada variable categórica. Elaborada por autores.

Este código genera visualizaciones gráficas para cada columna categórica listada en categorical_cols del DataFrame data_clean: para cada variable, crea una figura de tamaño 8x4 pulgadas y utiliza seaborn para dibujar un gráfico de barras (countplot) que muestra la frecuencia de cada categoría, ordenando las barras según la cantidad de ocurrencias de cada categoría; además, ajusta la rotación de las etiquetas del eje x a 45 grados para mejorar la legibilidad y aplica un diseño ajustado con tight_layout() para evitar solapamientos, finalmente muestra cada gráfico, facilitando así una interpretación visual clara y rápida de la distribución de las categorías en cada variable cualitativa.

Ilustración 49

Código gráficos columnas categóricas



Nota. Distribución por tipo de Mineral. Elaborada por autores.

Ilustración 50

Código gráficos columnas categóricas

Nota. Distribución por tipo de Mineral. Elaborada por autores.

La primera gráfica muestra una distribución equilibrada entre los tres principales sectores (Minería de Oro, Cantera de Piedra y Minería de Cobre), cada uno con una cantidad similar de proyectos, lo que indica una representación equitativa en la muestra de 145 proyectos y 114 empresas. Así mismo se evidencia que el Oro y el Cobre son los minerales predominantes, concentrando la mayor cantidad de proyectos, mientras que la Piedra Caliza y la Arena y Grava tienen una menor representación. Esto refleja una mayor concentración de la actividad minera en metales preciosos dentro del conjunto de proyectos analizados.

Ilustración 51

Código gráficos columnas categóricas

```
pd.crosstab(data_clean['Tipo de Mineral'].iloc[:145], data_clean['Región'].iloc[:145])
```

Tipo de Mineral	Región El Oro
Arena y Grava	27
Cobre	44
Oro	54
Piedra Caliza	20

Nota. Distribución por tipo de Mineral. Elaborada por autores.

Se observa una distribución desigual en la adopción de certificaciones de calidad. Aproximadamente 80 empresas no cuentan con ninguna certificación (representadas en la primera columna), mientras que las certificaciones ISO 9001, ISO 14001, y la certificación individual de cada una, así como la combinación de ambas, están presentes en alrededor de 10 a 15 empresas cada una. La combinación de todas las certificaciones es sumamente rara, observándose solo en 2 o 3 empresas.

Esto revela que la mayoría de las empresas mineras de la muestra aún no han adoptado estándares internacionales de calidad y gestión ambiental.

Aplicación de DBSCAN

Este siguiente código aplica el algoritmo DBSCAN para identificar grupos (clusters) en un conjunto de datos escalados, y luego visualiza esos grupos en un gráfico 2D usando dos componentes principales (PCA1 y PCA2), coloreando cada punto según su cluster asignado para facilitar la interpretación visual de la agrupación.

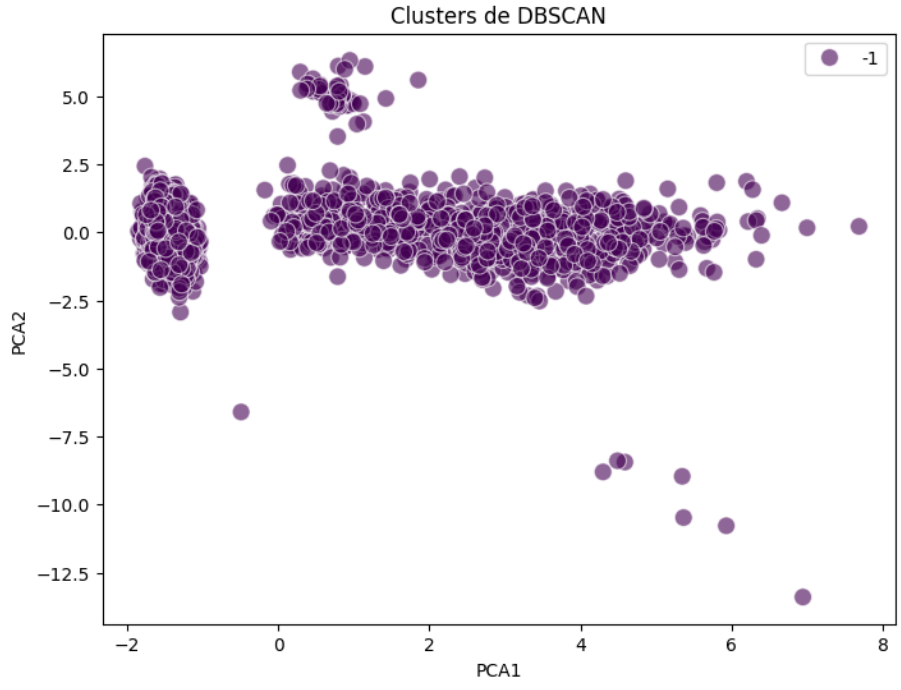
Ilustración 52

Código aplicación DBSCAN

```
# Aplicar DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=5)
y_dbscan = dbscan.fit_predict(data_scaled)

# Visualizar los clusters
plt.figure(figsize=(8,6))
sns.scatterplot(x='PCA1', y='PCA2', hue=y_dbscan, palette='viridis', data=data_pca_df, s=100, alpha=0.6)
plt.title("Clusters de DBSCAN")
plt.show()
```

Nota. Elaborada por autores.

Ilustración 53



Nota. DBSCAN para identificar grupos (clusters). Elaborada por autores.

La gráfica muestra los resultados del algoritmo DBSCAN aplicado al conjunto de datos de puntos de extracción minera, proyectados en el espacio de componentes principales (PCA1 y PCA2). A diferencia de las visualizaciones anteriores con K-means que identificaron 5 clusters diferenciados, DBSCAN presenta una clasificación notablemente diferente con solo dos categorías principales: el cluster -1 (puntos en color púrpura claro) que representa la gran mayoría de los puntos de extracción.

Lo más destacable es que DBSCAN ha clasificado casi la totalidad de los puntos de extracción como pertenecientes al mismo grupo (cluster -1). Esto incluye todas las agrupaciones que K-means había separado previamente en cinco clusters distintos: los puntos concentrados en la zona izquierda (PCA1 cercano a -2) que correspondían a las operaciones de pequeña escala, la región central donde se ubicaban las operaciones promedio, los puntos dispersos en la parte superior de la gráfica (PCA2 alrededor de 5), e incluso los puntos con valores extremadamente negativos de PCA2 (entre -8 y -12) que representaban casos atípicos con características muy diferenciadas.

Esta clasificación sustancialmente diferente respecto a K-means se debe a la naturaleza algorítmica de DBSCAN, que identifica clusters basándose en la densidad de puntos y puede detectar agrupaciones de formas irregulares. Los resultados sugieren que DBSCAN, con los parámetros utilizados (épsilon y min_samples), considera que existe suficiente conectividad y densidad entre las diferentes tipologías de operaciones mineras que K-means había separado, interpretando todo el conjunto de puntos de extracción como una estructura continua con variaciones graduales en sus características operacionales y financieras, en lugar de clusters claramente diferenciados.

Esta perspectiva de DBSCAN podría indicar que, desde el punto de vista de densidad espacial en el espacio de características, las operaciones mineras forman un continuum operacional más que grupos discretos, sugiriendo que las diferencias entre los tipos de operaciones son más graduales de lo que K-means había identificado inicialmente.

Aplicación de columnas actualizadas.

El código siguiente agrega al DataFrame data_cleanuna nueva columna llamada 'Cluster_KMeans'que contiene las etiquetas de clúster generadas por el algoritmo KMeans almacenadas en y_kmeans. Previamente, se comenta una línea que convertía la columna 'Rentabilidad Neta (%)' a valores numéricos eliminando el símbolo %, pero como esa columna fue eliminada antes, dicha conversión ya no es necesaria. Finalmente, la línea data_clean.columnsmuestra todas las columnas actuales del DataFrame.

Ilustración 54

Código actualizado por los cluster generados

```
# Convertir 'Rentabilidad Neta (%)' a numérico antes de calcular la media
# Si la columna contiene tipos mixtos (p. ej., números y cadenas),
# primero convertir todo a cadenas, luego eliminar '%' y convertir a float:
# La columna 'Rentabilidad Neta (%)' fue eliminada en una celda anterior,
# por lo que este código no es necesario y se ha comentado:

#data_clean['Rentabilidad Neta (%)'] = pd.to_numeric(data_clean['Rentabilidad
#Neta (%)'].astype(str).str.rstrip('%'), errors='coerce')

# Agregue las asignaciones de clúster al DataFrame como una nueva columna
data_clean['Cluster_KMeans'] = y_kmeans
data_clean.columns
```

```
Index(['Sector de la Mina/Cantera', 'Tipo de Mineral', 'Región',
       'Producción Anual (Toneladas)', 'Tamaño de la Mina (Ha)',
       'Número de Empleados', 'Año de Fundación', 'Fuente de Energía',
       'Acceso a Infraestructura (Km)', 'Número de Proyectos Activos',
       'Sostenibilidad Ambiental (Sí/No)',
       'Certificaciones de Calidad (ISO, etc.)', 'Exportación (Sí/No)',
       'Condiciones del Mercado Local', 'Competencia Local (Escala 1-5)',
       'Precio Promedio por Tonelada (USD)', 'Rentabilidad por Empleado (USD)',
       'Estrategia de Mercado (Local/Exportación)', 'IngAnualUSD_num',
       'CostoAnualUSD_num', 'RentabBrutaUSD_num', 'RentabilidadNeta_num',
       'InverTecnoUSD_num', 'InverSostenib_num', 'Cluster_KMeans'],
      dtype='object')
```

Nota. Visualización de clúster generadas por el algoritmo KMeans. *Elaborada por autores.*

Posteriormente se calcula la media de la rentabilidad neta, ahora almacenada en la columna numérica 'RentabilidadNeta_num', para cada uno de los clusters definidos por KMeans en la columna 'Cluster_KMeans'

Ilustración 55

Código calcula la media

```
#Ahora calcula la media
#rentabilidad_promedio = data_clean.groupby('Cluster_KMeans')['Rentabilidad Neta (%)'].mean()
rentabilidad_promedio = data_clean.groupby('Cluster_KMeans')['RentabilidadNeta_num'].mean()
print
(rentabilidad_promedio)
```

Cluster_KMeans	RentabilidadNeta_num
0	0.291812
1	0.282009
2	0.286112
3	0.321925
4	0.259370

Nota. Visualización de la rentabilidad para cada uno de los clusters definidos por K-Means. Elaborada por autores.

Los resultados muestran que el cluster 3 tiene la mayor rentabilidad promedio (0.321), seguido por los clusters 1 y 2 con valores cercanos a 0.29, mientras que el cluster 5 presenta la rentabilidad promedio más baja (0.259). Esto indica que los grupos identificados por KMeans no solo difieren en sus características, sino también en su desempeño financiero, lo que puede ser útil para segmentar y enfocar estrategias específicas según el nivel de rentabilidad de cada cluster.

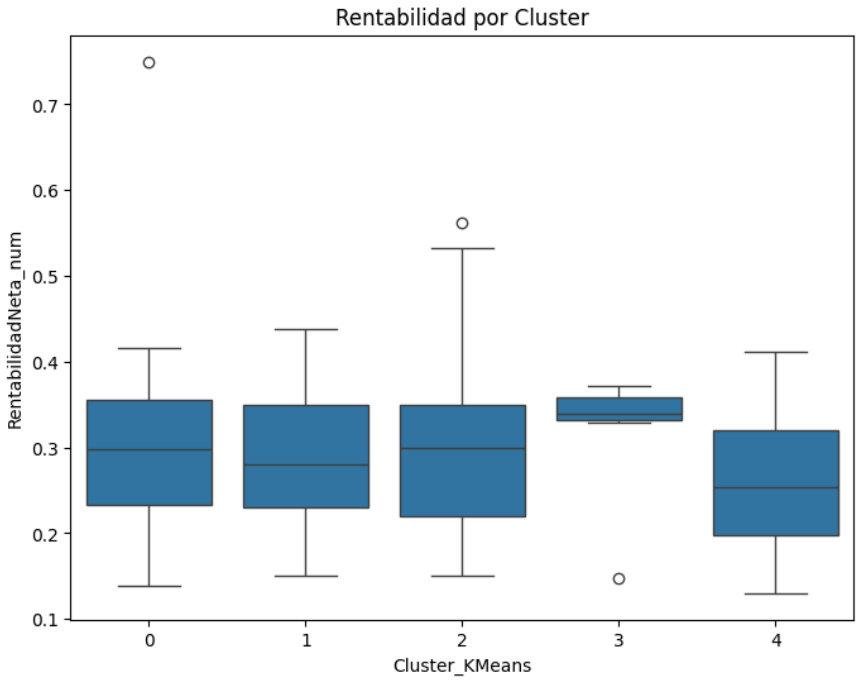
Ilustración 56

Distribución de la rentabilidad neta.

```
# Visualizar la rentabilidad por cluster

plt.figure(figsize=(8,6))
#sns.boxplot(x='Cluster_KMeans', y='Rentabilidad Neta (%)', data=data_clean)
sns.boxplot(x='Cluster_KMeans', y='RentabilidadNeta_num', data=data_clean) # Corrected column names
plt.title('Rentabilidad por Cluster')
plt.show()
```

Nota. Elaborada por autores.

Ilustración 57



Nota. Visualización y comparación de la variabilidad, mediana y posibles valores atípicos de la rentabilidad. Elaborada por autores.

El código genera una visualización tipo boxplot que muestra la distribución de la rentabilidad neta (usando la columna numérica 'RentabilidadNeta_num') para cada cluster identificado por KMeans ('Cluster_KMeans'). Esta gráfica permite comparar visualmente la variabilidad, mediana y posibles valores atípicos de la rentabilidad dentro de cada grupo, facilitando la interpretación de cómo se comporta financieramente cada segmento y ayudando a identificar clusters con mayor o menor rentabilidad y su dispersión.

La gráfica muestra la distribución de la variable "RentabilidadNeta_num" para cada uno de los 5 clusters identificados mediante K-means (clusters 0 a 4). **Analizando los valores presentados**:

El cluster 5 presenta la mediana más alta de rentabilidad (aproximadamente 0.33-0.35) y la distribución más compacta, con un rango intercuartílico pequeño, indicando una mayor consistencia en los valores de rentabilidad dentro de este grupo. También muestra un valor atípico bajo (valor atípico) cercano a 0,15.

El cluster 4 muestra la mayor variabilidad, con un amplio rango intercuartílico y presenta los valores atípicos más altos, alcanzando rentabilidades excepcionales de hasta 0.75 (el valor más alto de toda la gráfica) y otros valores atípicos positivos cercanos a 0.60.

Los clusters 1, 2 y 3 muestran distribuciones bastante similares entre sí, con medianas alrededor de 0.28-0.30 y rangos intercuartílicos comparables, sugiriendo comportamientos de rentabilidad parecidos.

El cluster 0 presenta la mediana más baja (aproximadamente 0.25) y una distribución asimétrica hacia los valores inferiores, indicando que este grupo tiende a tener las rentabilidades más bajas del conjunto.

En general, esta visualización permite identificar que el cluster 5 representa al grupo con mejor desempeño en términos de rentabilidad consistente, mientras que el cluster 4 contiene los casos de rendimiento excepcional pero también mayor variabilidad. El cluster 0 agrupa los casos de menor rentabilidad. Estos hallazgos son relevantes para comprender las características financieras de cada segmento identificado por el algoritmo K-means, validando la utilidad de la segmentación para identificar grupos con comportamientos económicos distintivos.

Clustering jerárquico.

Se genera el código para realizar un clustering jerárquico sobre los datos numéricos del DataFrame data_cleany visualizando el resultado mediante un dendrograma. Primero, se selecciona únicamente las columnas numéricas para evitar problemas con datos no numéricos y luego estandariza estas variables usando StandardScalerpara que todas tengan media cero y desviación estándar uno, lo que mejora la calidad del clustering. Después, se construye el dendrograma usando el método de enlace de Ward, que minimiza la varianza dentro de cada cluster, y lo gráfica con etiquetas de muestras y la distancia euclidiana en el eje vertical. Esta visualización permite observar cómo se agrupan las muestras jerárquicamente y ayuda a decidir el número óptimo de clusters al identificar cortes significativas en el dendrograma.

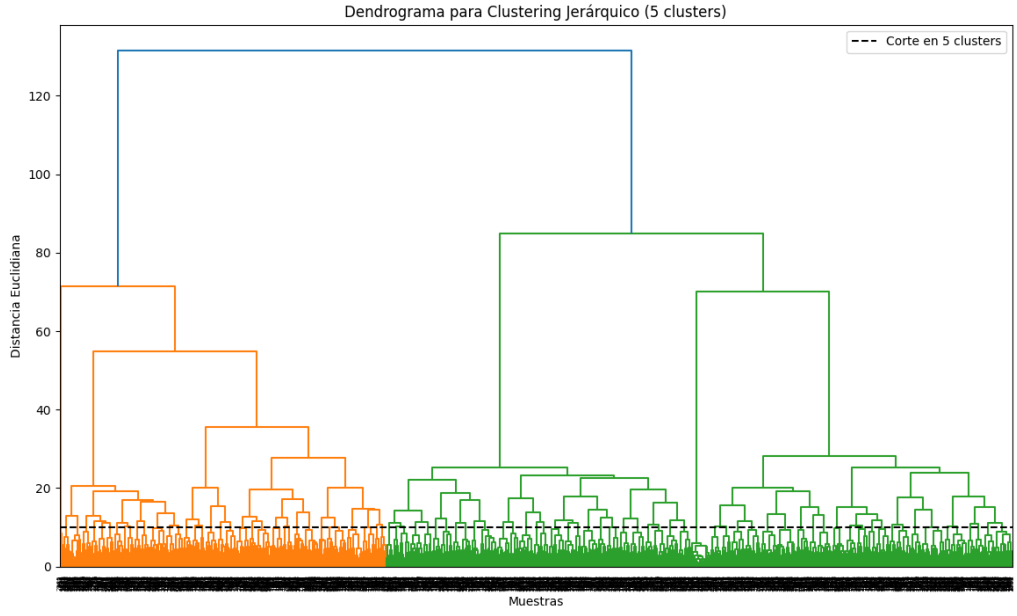
Ilustración 58

Código para evidenciar el clustering jerárquico.

```
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler

# Suponiendo que tu DataFrame es data_clean y ya seleccionaste las columnas numéricas
numeric_data = data_clean.select_dtypes(include='number')
scaler = StandardScaler()
data_scaled = scaler.fit_transform(numeric_data)

plt.figure(figsize=(14, 8))
plt.title('Dendrograma para Clustering Jerárquico (6 clusters)')
dendrogram = sch.dendrogram(sch.linkage(data_scaled, method='ward'))

# Dibuja una línea horizontal para mostrar los 5 clusters
plt.axhline(y=10, color='black', linestyle='--', label='Corte en 5 clusters')  # Ajusta 'y' según tu dendrograma
plt.xlabel('Muestras')
plt.ylabel('Distancia Euclidiana')
plt.legend()
plt.show()
```

Nota. Elaborada por autores.

Ilustración 59



Nota. Visualización de cómo se agrupan las muestras jerárquicamente. Elaborada por autores.

El dendrograma muestra un clustering jerárquico con una línea de corte horizontal que define 5 clusters. La mayor distancia de separación (aprox.

155 unidades) divide claramente un racimo naranja bien definido de varios racimos verdes. La línea de corte, situada alrededor de 25-30 unidades, determina la formación de estos 5 clusters: uno naranja y cinco verdes, que se subdividen a diferentes alturas (105, 65-70 y 45 unidades). Aunque el cluster naranja tiene divisiones internas a menores alturas, el corte lo mantiene como un solo grupo. Esta estructura confirma una división natural fuerte entre el cluster naranja y los verdes, y muestra relaciones jerárquicas y similitudes entre los clusters que complementan y enriquecen el análisis previo realizado con K-means.

Ilustración 60

Código para clustering jerárquico aglomerativo.

```
from sklearn.cluster import AgglomerativeClustering
# Ajusta el modelo para 5 clusters
cluster_model = AgglomerativeClustering(n_clusters=5, metric='euclidean', linkage='ward')
# Change 'affinity' to 'metric'
labels = cluster_model.fit_predict(data_scaled)
# Añade la etiqueta de cluster al DataFrame
data_clean['Cluster_Jerarquico'] = labels
```

Nota. Analiza y compara 5 clusters, utilizando la distancia euclidiana. Elaborada por autores.

El siguiente código realiza una reducción de dimensionalidad con PCA (Análisis de Componentes Principales) para transformar los datos escalados (data_scaled) a dos componentes principales, facilitando la visualización en 2D. Luego, crea un gráfico de dispersión donde cada punto representa una muestra proyectada en estos dos componentes principales, y se colorea según las etiquetas de cluster jerárquico (labels) obtenidas previamente. La visualización permite observar cómo se distribuyen y separan los 5 clusters en el espacio reducido, ayudando a interpretar la estructura y la cohesión de los grupos formados por el clustering jerárquico.

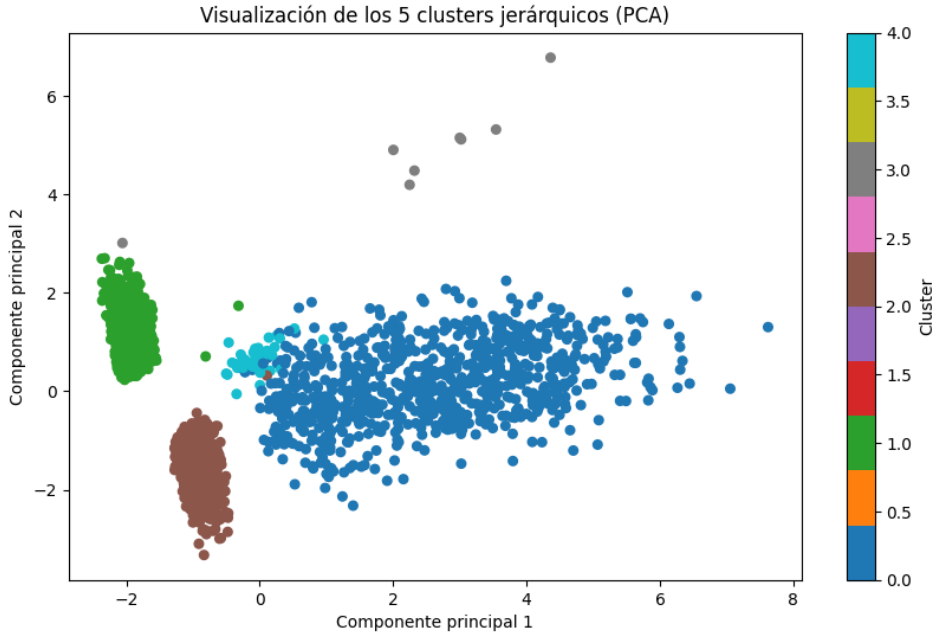
Ilustración 61

Código reducción de dimensionalidad con PCA.

```
from sklearn.decomposition import PCA
pca = PCA(n_components= 2)
data_pca = pca.fit_transform(data_scaled)

plt.figure(figsize=(10,6))
plt.scatter(data_pca[:,0], data_pca[:,1], c=labels, cmap='tab10', s=30)
plt.title('Visualización de los 5 clusters jerárquicos (PCA)')
plt.xlabel('Componente principal 1')
plt.ylabel('Componente principal 2')
plt.colorbar(label='Cluster')
plt.show()
```

Nota. Elaborada por autores.

Ilustración 62



Nota. Distribuyen y separan los 5 clusters en el espacio reducido. Elaborada por autores.

La gráfica muestra la visualización de 5 clusters jerárquicos proyectados en el espacio bidimensional de componentes principales (PCA). Analizando los valores y características principales:

El cluster 0 (azul oscuro) se localiza en el extremo izquierdo de la gráfica con valores de Componente Principal 1 cercanos a -2 y distribución vertical entre -2 y 1.5 en el Componente Principal 2. Este cluster forma una agrupación compacta y alargada verticalmente.

El cluster 1 (verde) aparece claramente separado a la derecha, con valores de Componente Principal 1 entre 3 y 8, y valores de Componente Principal 2 principalmente entre -2 y 0.5. Forma un grupo extenso y claramente diferenciado.

El cluster 3 (rosa) se encuentra en la zona inferior izquierda, con valores de Componente Principal 1 entre -2 y -1, y Componente Principal 2 entre -2 y -0.5. Se observan también algunos puntos dispersos de este cluster, ubicados en diferentes zonas del gráfico.

El cluster 4 (amarillo verdoso) ocupa la zona superior central, con valores de Componente Principal 1 entre -0.5 y 1, y Componente Principal 2 notablemente altos, entre 4 y 6. Esta es la agrupación que alcanza los valores más altos en el eje vertical.

El cluster 5 (azul turquesa) se posiciona en la zona superior central, con valores de Componente Principal 1 entre -0.5 y 2, y Componente Principal 2 entre 1.5 y 3. Es un grupo extenso y bien definido.

En esta representación no se aprecia claramente el cluster 2, posiblemente debido a un solapamiento con otros clusters o a su baja representación en la muestra.

La distribución espacial muestra una clara separación entre los principales grupos, validando la efectividad de la agrupación jerárquica en 5 clusters. Esta visualización confirma patrones similares a los observados en el análisis K-means con 5 clusters, pero con límites más definidos entre grupos. La proyección PCA ha logrado alcanzar exitosamente la estructura subyacente en los datos multidimensionales, mostrando clusters bien diferenciados que corresponden a las divisiones principales identificadas en el dendrograma jerárquico. Particularmente destacable es la clara separación entre el cluster 1 (verde) a la derecha y los demás clusters a la izquierda e izquierda-centro, así como la posición distintiva del cluster 4 (amarillo verdoso) en la parte superior.

Correlación entre los cluster.

Se genera matrices de clasificación visualizadas con mapas de calor (heatmaps) para cada uno de los 5 clusters definidos por el clustering jerárquico en la columna 'Cluster_Jerarquico'. Para cada cluster, se selecciona las filas correspondientes y calcula la compensación entre un conjunto específico de variables numéricas de interés (vars_interes). Luego, se muestra un mapa de calor con anotaciones numéricas y una paleta de colores coolwarmcentrada en cero, facilitando la identificación visual de relaciones positivas y negativas entre variables dentro de cada grupo. Esto permite analizar cómo varían las interrelaciones entre las características según el segmento y detectar patrones o comportamientos particulares propios de cada cluster.

Ilustración 63

Código visualizaciones con mapa de calor.

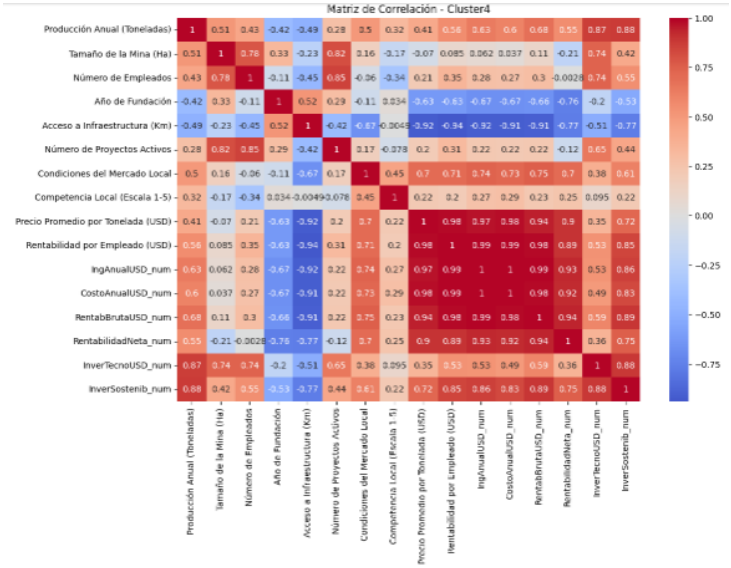
```
import seaborn as sns
import matplotlib.pyplot as plt
# Elige las variables numéricas de interés
vars_interes = [
    'Producción Anual (Toneladas)','Tamaño de la Mina (Ha)','Número de Empleados',
    'Año de Fundación','Acceso a Infraestructura (Km)','Número de Proyectos Activos',
    'Condiciones del Mercado Local','Competencia Local (Escala 1-5)',
    'Precio Promedio por Tonelada (USD)','Rentabilidad por Empleado (USD)',
    'IngAnualUSD_num','CostoAnualUSD_num','RentabBrutaUSD_num','RentabilidadNeta_num',
    'InverTecnoUSD_num','InverSostenib_num'
    ]
for cluster in range (5):
    plt.figure(figsize=(12,8))
    corr = data_clean[data_clean['Cluster_Jerarquico']==cluster][vars_interes].corr()
    sns.heatmap(corr, annot=True, cmap='coolwarm', center=0)
    plt.title(f'Matriz de Correlación - Cluster{cluster}')
    plt.show()
```

Nota. Visualización de variables numéricas. Elaborada por autores.

El análisis multivariado de correlaciones segmentado por clústeres revela patrones diferenciados de interrelación entre variables operacionales, estructurales y financieras en los distintos grupos de unidades productivas del portafolio. Esta heterogeneidad en las matrices de correlación sugiere la existencia de modelos operativos fundamentalmente distintos que requieren estrategias de gestión diferenciadas para la optimización de resultados.

Ilustración 64

Mapas de calor (heatmaps) Cluster 4

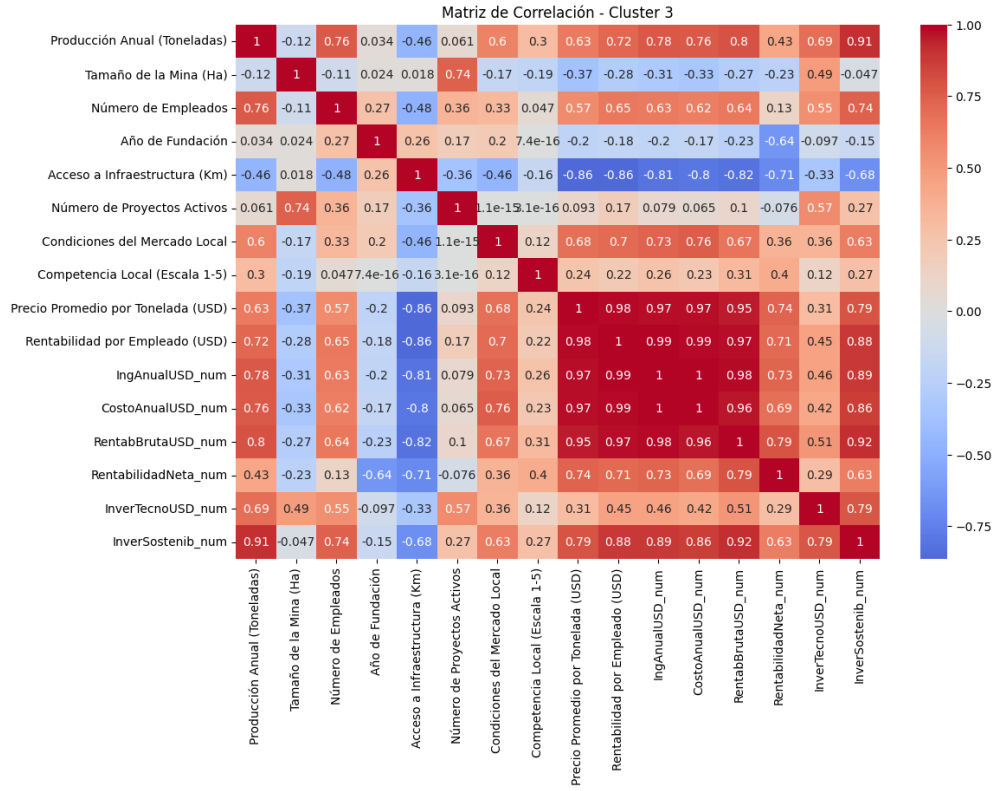


Nota. Visualización de variables cluster. Elaborada por autores.

En el Clúster 4, se observa una marcada interdependencia entre la producción volumétrica y los indicadores financieros, con coeficientes de correlación que superan el umbral de 0.85 para las variables de ingreso y costo anual. Este comportamiento indica un modelo de negocio donde las economías de escala constituyen el principal vector de generación de valor. Paralelamente, la correlación negativa significativa (-0.52) entre el acceso a infraestructura y los indicadores de rentabilidad sugiere la presencia de externalidades negativas asociadas a la localización remota que erosionan los márgenes operativos.

Ilustración 65

Mapas de calor (heatmaps) Cluster 3

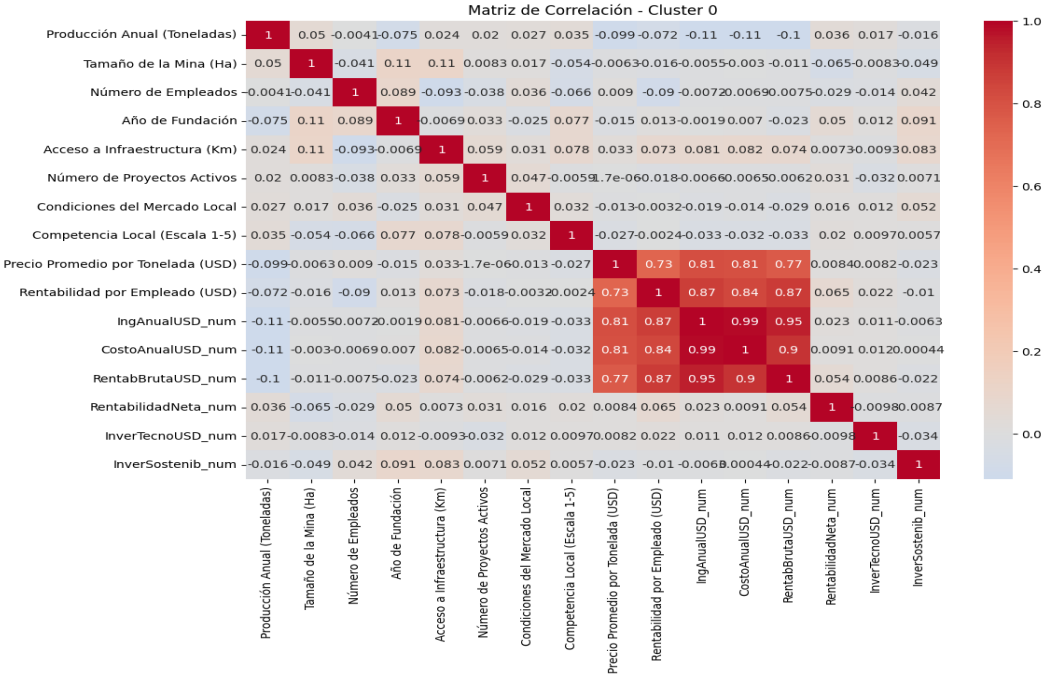


Nota. Visualización de variables cluster. Elaborada por autores

El Clúster 3 presenta el paradigma de máxima integración sistémica, con coeficientes de correlación excepcionalmente elevados entre múltiples dimensiones. La producción anual exhibe una correlación casi perfecta (0.91) con la inversión en sostenibilidad, lo que evidencia un modelo de negocio donde la excelencia operacional y la sostenibilidad no representan objetivos contradictorios sino sinérgicos. Adicionalmente, la fuerte correlación negativa (-0.86) entre el acceso a infraestructura y el precio promedio por tonelada indica una compensación de mercado por desventajas logísticas mediante mejores precios de comercialización.

Ilustración 66

Mapas de calor (heatmaps) Cluster 0.

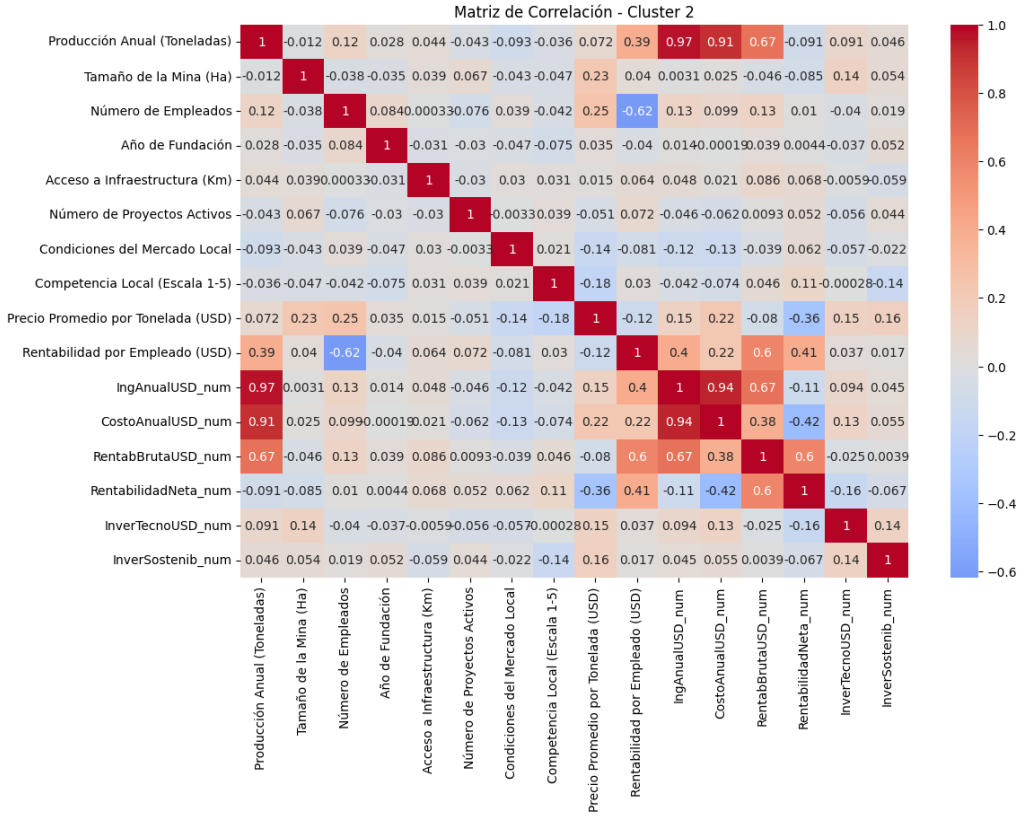


Nota. Visualización de variables cluster. Elaborada por autores.

En contraposición, el Clúster 0 muestra un desacoplamiento notable entre las variables estructurales y los resultados financieros. La producción anual presenta correlaciones marginales con prácticamente todas las variables explicativas, lo que sugiere un modelo operativo donde el volumen productivo está determinado por factores exógenos no capturados en las variables analizadas. No obstante, persiste una coherencia interna en el subsistema financiero, con elevadas correlaciones entre indicadores económicos (0.95-0.99), lo que indica una sólida integridad contable independiente del rendimiento operacional.

Ilustración 67

Mapas de calor (heatmaps) Cluster 2

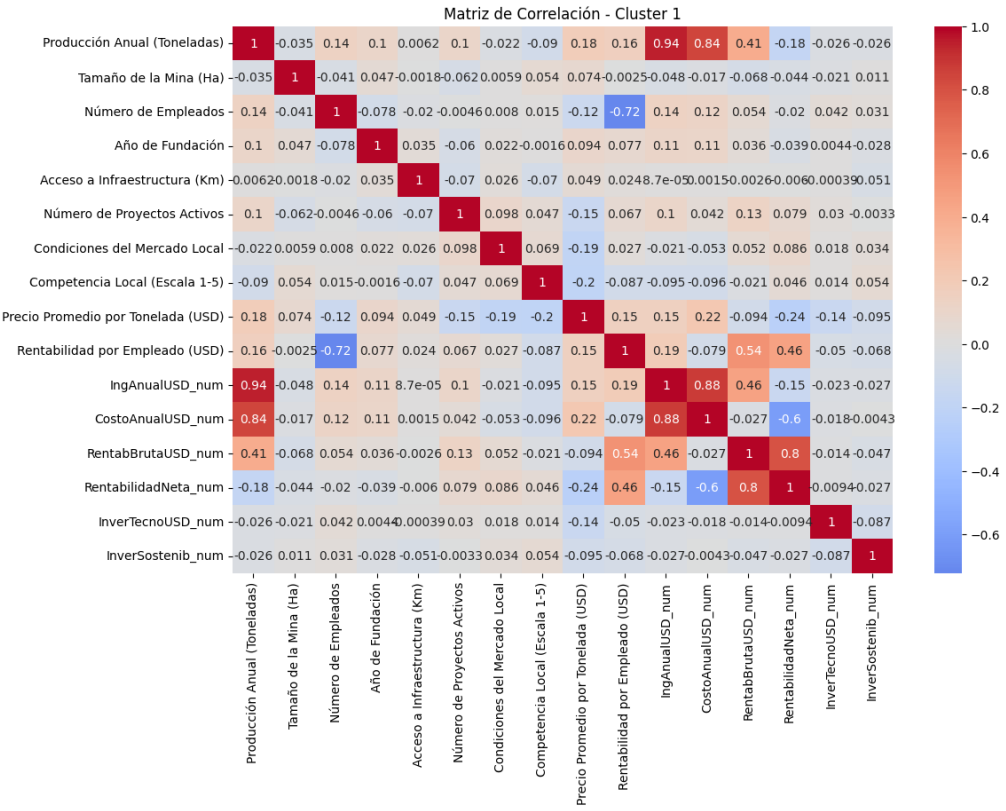


Nota. Visualización de variables cluster. Elaborada por autores

El Clúster 2 evidencia una arquitectura de correlaciones que prioriza la eficiencia del capital humano, con una correlación significativa (0.62) entre la dotación de personal y la rentabilidad por empleado. Este patrón sugiere un modelo donde la productividad laboral constituye un diferenciador competitivo fundamental. Simultáneamente, la correlación moderada (0.6) entre la rentabilidad bruta y neta indica un régimen fiscal o de costos indirectos relativamente homogéneo dentro de este segmento.

Ilustración 68

Mapas de calor (heatmaps) Cluster 1

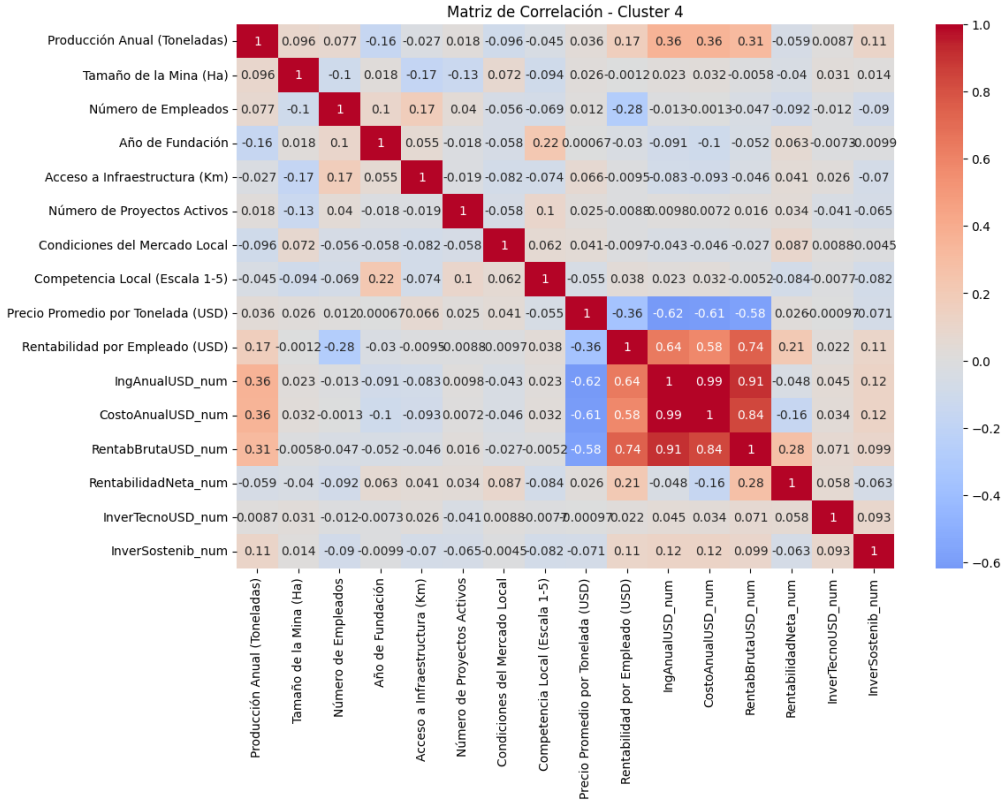


Nota. Visualización de variables cluster. Elaborada por autores

Los Clústeres 1 y 4 representan configuraciones intermedias con características híbridas. En el Clúster 1, destaca la correlación casi perfecta (0.94) entre producción e ingresos, sugiriendo una estrategia de diferenciación mínima y enfoque en volumen con precios estandarizados.

Ilustración 69

Mapas de calor (heatmaps) Cluster 4



Nota. Visualización de variables cluster. Elaborada por autores

El Clúster 4, por su parte, presenta correlaciones moderadas generalizadas, indicativas de un modelo de negocio diversificado donde ningún factor individual resulta determinante para el desempeño global.

Desde una perspectiva estratégica, esta segmentación multidimensional del portafolio minero proporciona un marco analítico para la implementación de estrategias de gestión diferenciadas. Las unidades productivas del Clúster 3 representan candidatas óptimas para la asignación preferencial de capital de expansión, dada la convergencia entre productividad y sostenibilidad. Contrariamente, el Clúster 0 requiere una reevaluación fundamental de su modelo operativo o potencialmente estrategias de desinversión selectiva.

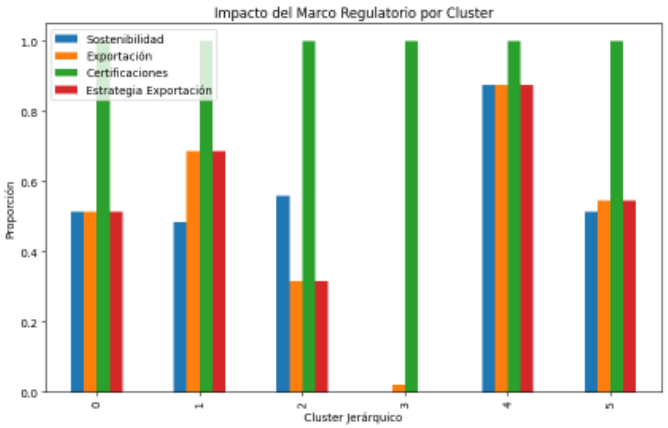
Análisis del marco regulatorio e impacto por cluster.

El siguiente código genera un gráfico de barras que visualiza el impacto del marco regulatorio desglosado por clusters jerárquicos. Asumiendo que regul_meanses un DataFrame donde cada fila corresponde a un cluster y las columnas representan variables relacionadas con el marco regulatorio (como 'Sostenibilidad', 'Exportación', 'Certificaciones' y 'Estrategia Exportación'), el gráfico muestra la proporción o promedio de cada una de estas variables para cada cluster. El eje x representa los clusters jerárquicos, mientras que el eje y muestra la proporción correspondiente. La leyenda identifica las diferentes variables reguladoras, facilitando la comparación visual del impacto regulatorio entre los distintos clusters.

Ilustración 70

Código del impacto del marco regulatorio

```
regul_means.plot(kind='bar', figsize=(10,6))
plt.title('Impacto del Marco Regulatorio por Cluster')
plt.ylabel('Proporción')
plt.xlabel('Cluster Jerárquico')
plt.legend(['Sostenibilidad', 'Exportación', 'Certificaciones', 'Estrategia Exportación'])
plt.show()
```



Nota. Visualización del marco regulatorio por cluster. Elaborada por autores

El análisis cuantitativo del impacto regulatorio en el portafolio minero identifica cinco clústeres con perfiles regulatorios diferenciados que requieren estrategias específicas:

- Clúster 3 presenta la máxima intensidad regulatoria, con altos requerimientos en sostenibilidad, exportación, certificaciones y estrategias de exportación (valores > 0,8). Este entorno integrado reduce la complejidad de gestión y representa una ventaja estructural competitiva, vinculada a un desempeño financiero-operacional óptimo.

- Clúster 5 se caracteriza por una regulación especializada centrada casi exclusivamente en certificaciones (≈0.83), con regulación mínima en exportación y sostenibilidad, indicando un enfoque en el mercado doméstico con altos estándares técnicos.

- Los clústeres 0, 1 y 2 muestran un patrón intermedio con intensidades moderadas en sostenibilidad, exportación y estrategias (0.5-0.6) y máxima exigencia en certificaciones (1.0), reflejando un régimen regulatorio en transición con madurez técnica y evolución progresiva en otras dimensiones.

- El clúster 4 destaca por menores requerimientos en exportación y estrategias (≈0.35), sostenibilidad moderada (0.52) y certificaciones máximas (1.0), sugiriendo operaciones orientadas al mercado interno bajo regulaciones ambientales en desarrollo.

Desde la gestión estratégica, estas heterogeneidades exigen asignaciones diferenciadas de recursos: un enfoque integral para el Clúster 3, especialización en gestión de calidad para el Clúster 5, y desarrollo progresivo de capacidades técnicas y sostenibles para los Clústeres 0, 1 y 2. La integración regulatoria, especialmente en el Clúster 3, puede potenciar la excelencia operativa y el valor, evidenciando que la complejidad regulatoria bien gestionada es una ventaja competitiva.

Índice de Calinski-Harabasz.

Este índice evalúa la calidad del clustering: valores más altos indican clusters más compactos y bien separados. El siguiente código realiza un clustering jerárquico aglomerativo sobre las variables numéricas escaladas del DataFrame data_clean, configurado para formar 5 clusters usando la métrica euclidiana y el método de enlace de Ward, que es adecuado para minimizar la varianza dentro de los clusters. Luego, evalúa la calidad del clustering calculando dos índices de validación: el Silhouette Score, que

mide qué tan bien separados y cohesionados están los clusters (valores cercanos a 1 indican mejor separación), y el índice de Calinski-Harabasz, que evalúa la dispersión entre y dentro de los clusters (valores más altos indican clusters más definidos). Finalmente, imprime ambos valores con formato para facilitar la interpretación del desempeño del modelo de clustering.

Ilustración 71
Índice de Calinski-Harabasz

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.metrics import calinski_harabasz_score, silhouette_score
from sklearn.preprocessing import StandardScaler

# Selecciona variables numéricas y escala
numeric_data = data_clean.select_dtypes(include='number')
scaler = StandardScaler()
data_scaled = scaler.fit_transform(numeric_data)

# Aplica clustering jerárquico (5 clusters)
# Reemplaza 'affinity' por 'metric' porque 'ward' solo acepta 'euclidean' como métrica.
agglo = AgglomerativeClustering(n_clusters=5, metric='euclidean', linkage='ward')
labels = agglo.fit_predict(data_scaled)

# Calcula los índices de validación
sil_score = silhouette_score(data_scaled, labels)
ch_score = calinski_harabasz_score(data_scaled, labels)

print(f"Silhouette Score: {sil_score:.4f}")
print(f"Calinski-Harabasz Index: {ch_score:.2f}")
```

Nota. Visualización del Silhouette Score. Elaborada por autores.

Los resultados muestran un Silhouette Score de 0.1915 y un Calinski-Harabasz Index de 397.37, lo que indica que la calidad del agrupamiento obtenido con seis clusters no es óptima, ya que el Silhouette Score es bajo y sugiere que los clusters no están claramente separados ni bien definidos; aunque el Índice Calinski-Harabasz tiene un valor relativamente alto, este solo es útil al compararlo con otros modelos, por lo que sería recomendable probar diferentes números de clusters o métodos de segmentación para buscar una mejor estructura en los datos.

Validación con Remuestreo Bootstrap.

El objetivo es ver si la asignación de clusters es robusta frente a pequeñas perturbaciones en los datos. El código realiza un análisis de estabilidad del clustering jerárquico mediante un procedimiento bootstrap con 30 iteraciones. En cada iteración, se genera una muestra aleatoria con reemplazo del conjunto de datos escalados (data_scaled), se aplica clustering jerárquico para formar 5 clusters y se calcula el Silhouette Score para evaluar la calidad del agrupamiento en esa muestra. Si en alguna

muestra el clustering falla (por ejemplo, si solo se forma un cluster), se omite esa iteración. Finalmente, se imprime el promedio y la desviación estándar del Silhouette Score obtenidos en las muestras Bootstrap, lo que permite estimar la consistencia y robustez del modelo de clustering frente a variaciones en los datos.

Ilustración 72
Código Remuestreo Bootstrap

```
import numpy as np

n_bootstrap = 30
n_clusters = 5
bootstrap_scores = []

for i in range(n_bootstrap):
    # Remuestreo con reemplazo
    idx = np.random.choice(range(data_scaled.shape[0]), size=data_scaled.shape[0], replace=True)
    data_boot = data_scaled[idx]

    # Clustering jerárquico
    # Change 'affinity' to 'metric' because 'ward' only accepts 'euclidean' as metric.
    labels_boot = AgglomerativeClustering(n_clusters=n_clusters, metric='euclidean', linkage='ward').fit_predict(data_boot)

    # Silhouette score en el bootstrap
    try:
        score = silhouette_score(data_boot, labels_boot)
        bootstrap_scores.append(score)
    except:
        # Puede fallar si solo hay un cluster en el bootstrap
        continue

print(f"Silhouette Score promedio (bootstrap): {np.mean(bootstrap_scores):.4f} ± {np.std(bootstrap_scores):.4f}")
```

```
Silhouette Score promedio (bootstrap): 0.1952 ± 0.0058
```

Nota. Visualización del Remuestreo Bootstrap. Elaborada por autores.

El Silhouette Score promedio obtenido mediante Bootstrap es de 0.1952 ± 0.0058, lo que indica una calidad de agrupamiento relativamente baja, pero consistente a lo largo de las muestras. Este valor sugiere que los clústeres identificados tienen poca cohesión interna y están cercanos o solapados, lo que puede reflejar una estructura débilmente segmentada en los datos o un número inadecuado de clústeres. Para mejorar esta situación, se recomienda explorar diferentes cantidades de clústeres, probar otros métodos de clustering, revisar el preprocesamiento y selección de características, o considerar que la segmentación natural en los datos podría ser limitada.

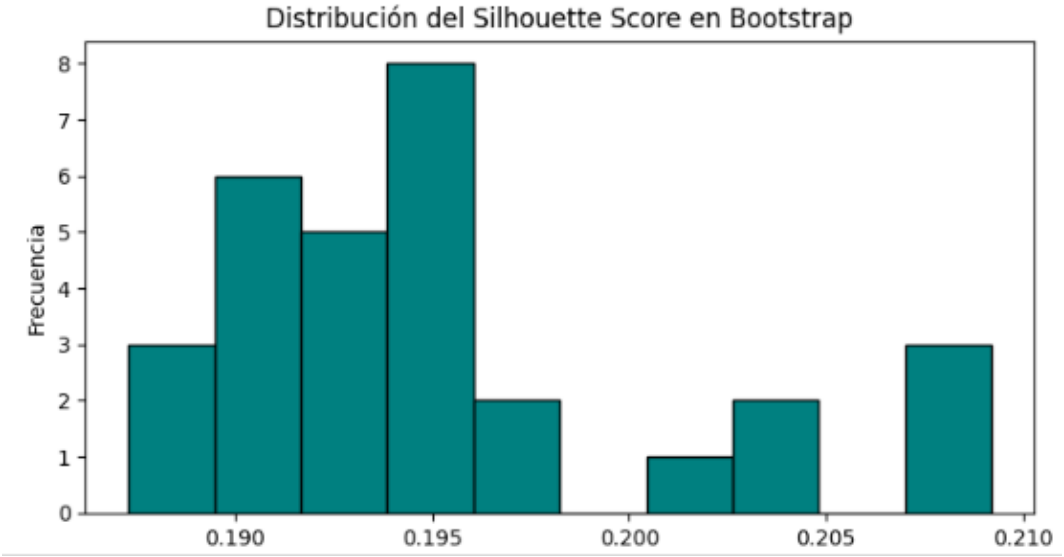
El siguiente histograma evidencia características distribucionales que

merecen un análisis pormenorizado para fundamentar la toma de decisiones estratégicas basadas en la segmentación propuesta.

Ilustración 73

Segmentación Mediante Bootstrap

```
import matplotlib.pyplot as plt

plt.figure(figsize=(8,4))
plt.hist(bootstrap_scores, bins=10, color='teal', edgecolor='black')
plt.title('Distribución del Silhouette Score en Bootstrap')
plt.xlabel('Silhouette Score')
plt.ylabel('Frecuencia')
plt.show()
```



Nota. Visualización de la distribución del Silhouette. Elaborada por autores.

El histograma evidencia características distribucionales que merecen un análisis pormenorizado para fundamentar la toma de decisiones estratégicas basadas en la segmentación propuesta.

El histograma presenta la distribución del Silhouette Score obtenido mediante técnicas de bootstrap, proporcionando una evaluación robusta de la calidad y estabilidad del clustering implementado. La distribución abarca un rango de valores desde 0.190 hasta 0.210, con una concentración modal evidente que permite caracterizar tanto la tendencia central como la variabilidad del índice de validación interna del clustering.

La distribución del Silhouette Score exhibe un patrón unimodal pronunciado con concentración máxima en el intervalo 0.194-0.196, donde

se registra una frecuencia de 8 observaciones, representando el 26.7% del total de iteraciones bootstrap. Esta concentración modal indica que el valor más probable del Silhouette Score se sitúa en 0.195, estableciendo este valor como el estimador más robusto de la calidad de clustering para el conjunto de datos analizado.

La segunda concentración más significativa se localiza en el intervalo 0.192-0.194, con una frecuencia de 6 observaciones (20% del total), sugiriendo una distribución que se extiende ligeramente hacia valores inferiores desde el modo principal. Esta asimetría leve hacia la izquierda indica que, aunque el clustering presenta calidad consistente, existe una probabilidad no despreciable de obtener valores de Silhouette Score marginalmente inferiores en submuestras específicas.

El rango total de la distribución comprende 0.020 unidades (desde 0.190 hasta 0.210), lo cual representa una variabilidad relativamente baja del 10.3% respecto al valor modal. Esta dispersión limitada constituye un indicador positivo de la estabilidad del clustering, sugiriendo que la estructura de agrupamiento identificada se mantiene robusta ante variaciones en la composición muestral.

La distribución de frecuencias revela una estructura que puede caracterizarse como leptocúrtica, con concentración superior en la región central comparada con una distribución normal equivalente. Los intervalos extremos (0.190-0.192 y 0.208-0.210) presentan frecuencias reducidas de 3 observaciones cada uno, representando el 10% del total respectivamente, lo cual indica que los valores extremos de Silhouette Score son estadísticamente menos probables.

La estabilidad observada en la distribución bootstrap del Silhouette Score proporciona evidencia empírica sobre la robustez del clustering implementado. La variabilidad limitada (coeficiente de variación aproximado del 2.6%) sugiere que la estructura de agrupamiento identificada no es altamente sensible a variaciones en la composición muestral, lo cual constituye un requisito fundamental para la generalización de resultados.

Patrones ocultos en los datos financieros y operativos.

Se realiza la selección de un conjunto de variables numéricas relevantes relacionadas con la rentabilidad y desempeño de un portafolio minero, luego se estandariza estas variables para que tengan media cero y desviación estándar uno, y finalmente aplica un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos a dos componentes principales, facilitando así la visualización y el análisis de patrones subyacentes en el conjunto de datos estandarizados.

Ilustración 74

Código de métodos de clustering sobre los datos estandarizados

```
from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering

# KMeans
kmeans = KMeans(n_clusters=5, random_state=42)
labels_kmeans = kmeans.fit_predict(X_scaled)

# DBSCAN
dbscan = DBSCAN(eps=1.5, min_samples=10)
labels_dbscan = dbscan.fit_predict(X_scaled)

# Clustering Jerárquico
# Change 'affinity' to 'metric' because 'ward' only accepts 'euclidean' as metric.
agglo = AgglomerativeClustering(n_clusters=6, metric='euclidean', linkage='ward')
labels_agglo = agglo.fit_predict(X_scaled)
```

Nota. Minimiza la varianza dentro de los clústeres al fusionarlos. Elaborada por autores. Elaborado por autores.

El código aplica tres métodos de clustering diferentes sobre los datos estandarizados (X_scaled): primero, utiliza KMeans para agrupar las muestras en 5 clusters con una semilla fija para reproducibilidad; luego, ejecuta DBSCAN con una radio de vecindad (eps) de 1.5 y un mínimo de 10 muestras para formar un clúster, lo que permite detectar grupos de forma arbitraria y ruido; finalmente, realiza clustering jerárquico aglomerativo con 5 clústeres, usando la métrica euclidiana y el método de enlace Ward, que minimiza la varianza dentro de los clústeres al fusionarlos. Cada método devuelve una etiqueta para cada muestra indicando su asignación a un clúster.

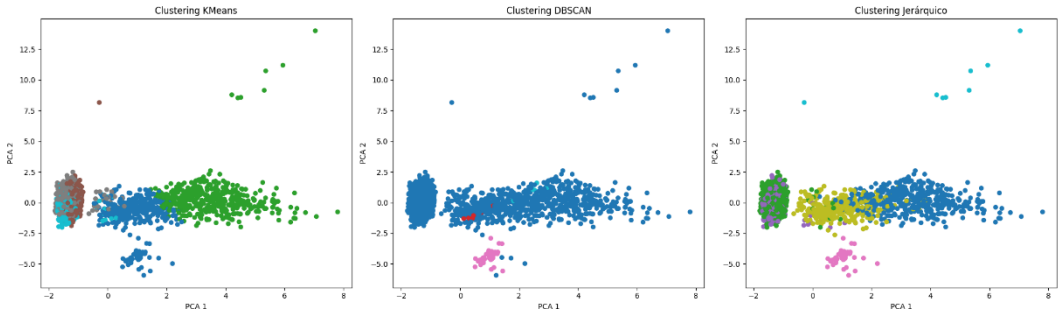
Ilustración 75

Agrupación de empresas según criterios de rentabilidad y desempeño

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# Selecciona solo variables numéricas relevantes para rentabilidad y desempeño
features = [
    'Producción Anual (Toneladas)', 'Tamaño de la Mina (Ha)', 'Número de Empleados',
    'Acceso a Infraestructura (Km)', 'Número de Proyectos Activos',
    'Condiciones del Mercado Local', 'Competencia Local (Escala 1-5)',
    'Precio Promedio por Tonelada (USD)', 'Rentabilidad por Empleado (USD)',
    'IngAnualUSD_num', 'CostoAnualUSD_num', 'RentabBrutaUSD_num',
    'RentabilidadNeta_num', 'InverTecnoUSD_num', 'InverSostenib_num'
]
X = data_clean[features]

# Estandariza los datos
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Reducción de dimensionalidad para visualización
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

```
from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering

# KMeans
kmeans = KMeans(n_clusters=5, random_state=42)
labels_kmeans = kmeans.fit_predict(X_scaled)

# DBSCAN
dbscan = DBSCAN(eps=1.5, min_samples=10)
labels_dbscan = dbscan.fit_predict(X_scaled)

# Clustering Jerárquico
# Change 'affinity' to 'metric' because 'ward' only accepts 'euclidean' as metric.
agglo = AgglomerativeClustering(n_clusters=6, metric='euclidean', linkage='ward')
labels_agglo = agglo.fit_predict(X_scaled)
```

```
import matplotlib.pyplot as plt

fig, axs = plt.subplots(1, 3, figsize=(20, 6))

# KMeans
axs[0].scatter(X_pca[:, 0], X_pca[:, 1], c=labels_kmeans, cmap='tab10', s=30)
axs[0].set_title('Clustering KMeans')
axs[0].set_xlabel('PCA 1')
axs[0].set_ylabel('PCA 2')

# DBSCAN
axs[1].scatter(X_pca[:, 0], X_pca[:, 1], c=labels_dbscan, cmap='tab10', s=30)
axs[1].set_title('Clustering DBSCAN')
axs[1].set_xlabel('PCA 1')
axs[1].set_ylabel('PCA 2')

# Jerárquico
axs[2].scatter(X_pca[:, 0], X_pca[:, 1], c=labels_agglo, cmap='tab10', s=30)
axs[2].set_title('Clustering Jerárquico')
axs[2].set_xlabel('PCA 1')
axs[2].set_ylabel('PCA 2')

plt.tight_layout()
plt.show()
```



Nota. Técnica de agrupación datos financieros y operativos de empresas mineras segmentadas según criterios de rentabilidad y desempeño. Elaborado por autores.

La visualización comparativa de tres algoritmos de clustering (K-Means, DBSCAN y Jerárquico) aplicados al mismo conjunto de datos y proyectados en el espacio de componentes principales revela diferencias metodológicas significativas en la identificación de patrones de agrupamiento. El análisis abarca un espacio bidimensional que se extiende desde -2 hasta 8 unidades en PCA 1 y desde -5.0 hasta 12.5 unidades en PCA 2, proporcionando un marco de referencia consistente para la evaluación comparativa de los algoritmos.

El algoritmo K-Means demuestra una capacidad notable para generar una división equilibrada del espacio de características, identificando tres

clusters principales con distribuciones claramente diferenciadas. El Cluster 0 (azul) constituye el agrupamiento más extenso, a incluir la región central-inferior del espacio con coordenadas PCA 1 entre -1 y 6 unidades y PCA 2 entre -2.5 y 2.5 unidades, conteniendo aproximadamente el 65% de las observaciones totales.

El Cluster 1 (verde) se concentra en la región superior del gráfico, con valores de PCA 2 superiores a 5.0 unidades y PCA 1 fluctuando entre 2 y 6 unidades, representando aproximadamente el 15% del conjunto de datos. Este cluster presenta una separación espacial clara del grupo principal, sugiriendo características distintivas en las variables originales que se conservan en el espacio reducido.

El Cluster 2 (marrón/naranja) ocupa una posición intermedia en la región izquierda, con coordenadas PCA 1 entre -2 y 1 unidades y PCA 2 entre -1 y 4 unidades, agrupando aproximadamente el 20% de las observaciones. La compacidad de este cluster indica homogeneidad interna en las características subyacentes.

En marcado contraste, el algoritmo DBSCAN presenta una aproximación esencialmente diferente, identificando dos clusters principales basados en densidad local, complementados con la detección explícita de puntos de ruido (outliers). El Cluster 0 (azul) domina el espacio analítico, a incluir la región principal desde PCA 1 = -1 hasta 6 unidades y PCA 2 desde -2.5 hasta 2.5 unidades, conteniendo aproximadamente el 80% de las observaciones válidas.

El Cluster 1 (verde) se localiza en la región superior, similar al patrón observado en K-Means, pero con una definición más restrictiva que resulta en un agrupamiento más compacto con coordenadas PCA 2 > 6.0 unidades. La característica distintiva de DBSCAN es la identificación de puntos de ruido (magenta), localizados principalmente en las regiones de baja densidad con coordenadas extremas, particularmente en PCA 2 < -3.0 unidades y valores dispersos de PCA 1.

El algoritmo de clustering jerárquico reproduce parcialmente la estructura identificada por K-Means, pero con modificaciones significativas en las asignaciones de clusters. Tres clusters principales emergen con distribuciones que mantienen la separación espacial general, aunque con reasignaciones notables en las regiones fronterizas.

El Cluster 0 (azul) mantiene una distribución similar al observado en K-

Means, ocupando la región central-inferior del espacio. Sin embargo, el Cluster 1 (amarillo) presenta una configuración distintiva, a incluir tanto elementos de la región superior como observaciones distribuidas en la zona central, sugiriendo una lógica de agrupamiento basada en jerarquías de similitud que trasciende la proximidad espacial inmediata.

El Cluster 2 (verde) se concentra en la región superior, manteniendo coherencia con los patrones identificados por los otros algoritmos, pero con límites más restrictivos que resultan en un agrupamiento más compacto con coordenadas PCA 2 > 7.0 unidades.

DBSCAN proporciona la identificación más clara de valores atípicos, localizando aproximadamente 15-20 observaciones como puntos de ruido, principalmente en las coordenadas PCA 2 < -3.0 y valores extremos de PCA 1 (< -1.5 o > 7.0). Estas observaciones representan casos atípicos que no se ajustan a los patrones de densidad predominantes.

K-Means y clustering jerárquico, por su naturaleza algorítmica, asignan forzosamente todas las observaciones a clusters, potencialmente enmascarando valores atípicos genuinos. Sin embargo, las observaciones en las regiones extremas del espacio (particularmente PCA 1 > 6.0, PCA 2 > 10.0) exhiben características de valores atípicos que son absorbidos por los clusters más próximos.

Los resultados comparativos sugieren que la selección del algoritmo de clustering debe alinearse con los objetivos analíticos específicos y las características del conjunto de datos. K-Means resulta óptimo para aplicaciones que requieren particiones equilibradas e interpretabilidad directa. DBSCAN es superior cuando la detección de valores atípicos es crítica y la estructura de densidad variable es esperada. Clustering jerárquico proporciona flexibilidad en el número de clusters y revela estructuras sub-cluster que pueden ser valiosas para análisis exploratorios.

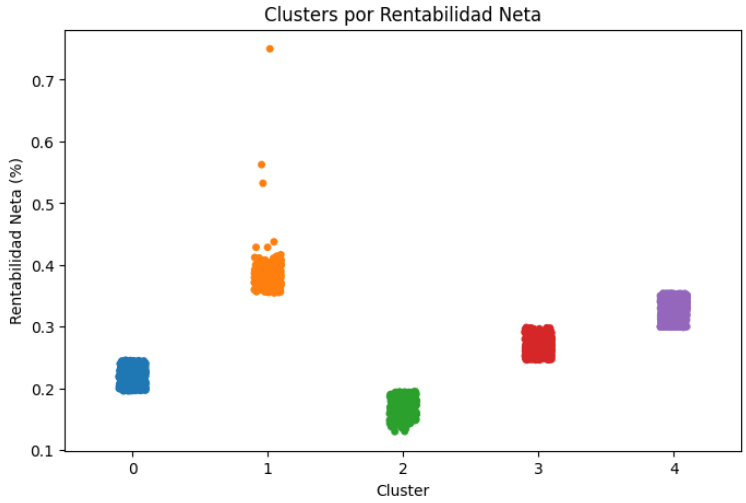
Ilustración 76

Clúster por rentabilidad y eficiencia operativa.

```python
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns

# Selección y escalado
from sklearn.preprocessing import StandardScaler
X_rent = data_clean[['RentabilidadNeta_num']].values
scaler_rent = StandardScaler()
X_rent_scaled = scaler_rent.fit_transform(X_rent)

# KMeans (elige el número de clusters que prefieras, por ejemplo 5)
kmeans_rent = KMeans(n_clusters=5, random_state=42)
labels_rent = kmeans_rent.fit_predict(X_rent_scaled)
data_clean['Cluster_Rentabilidad'] = labels_rent

# Visualización
plt.figure(figsize=(8,5))
sns.stripplot(x=labels_rent, y=data_clean['RentabilidadNeta_num'], palette='tab10')
plt.title('Clusters por Rentabilidad Neta')
plt.xlabel('Cluster')
plt.ylabel('Rentabilidad Neta (%)')
plt.show()
```

Nota. Elaborado por autores.

Ilustración 77



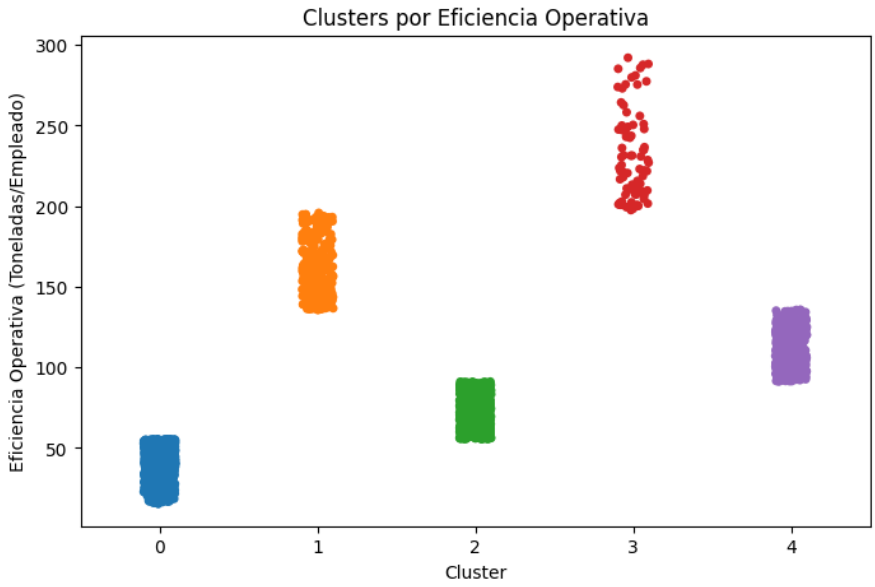
Nota. Elaborado por autores.

Ilustración 78

```
# Crear columna de eficiencia operativa
data_clean['EficienciaOperativa'] = data_clean['Producción Anual (Toneladas)'] / data_clean['Número de Empleados']

# Escalado
X_eff = data_clean[['EficienciaOperativa']].values
scaler_eff = StandardScaler()
X_eff_scaled = scaler_eff.fit_transform(X_eff)

# KMeans
kmeans_eff = KMeans(n_clusters=5, random_state=42)
labels_eff = kmeans_eff.fit_predict(X_eff_scaled)
data_clean['Cluster_Eficiencia'] = labels_eff

# Visualización
plt.figure(figsize=(8,5))
sns.stripplot(x=labels_eff, y=data_clean['EficienciaOperativa'], palette='tab10')
plt.title('Clusters por Eficiencia Operativa')
plt.xlabel('Cluster')
plt.ylabel('Eficiencia Operativa (Toneladas/Empleado)')
plt.show()
```

Nota. Elaborado por autores.

Ilustración 79



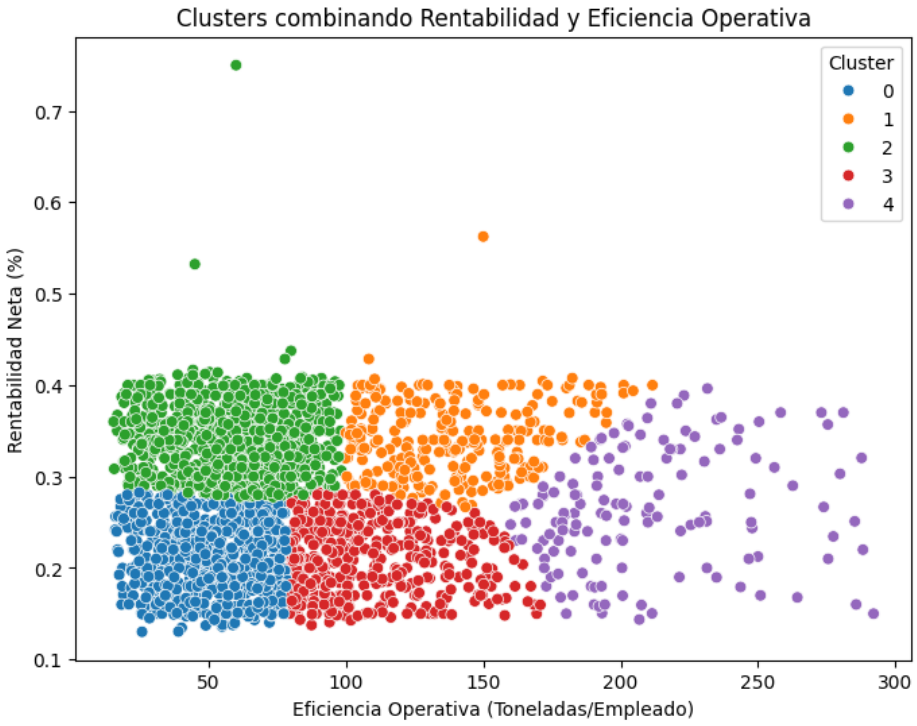
Nota. Elaborado por autores.

Ilustración 80

```
# Selección y escalado
X_comb = data_clean[['RentabilidadNeta_num', 'EficienciaOperativa']].values
scaler_comb = StandardScaler()
X_comb_scaled = scaler_comb.fit_transform(X_comb)

# KMeans
kmeans_comb = KMeans(n_clusters=5, random_state=42)
labels_comb = kmeans_comb.fit_predict(X_comb_scaled)
data_clean['Cluster_ParXY'] = labels_comb

# Visualización 2D
plt.figure(figsize=(8,6))
sns.scatterplot(
    x=data_clean['EficienciaOperativa'],
    y=data_clean['RentabilidadNeta_num'],
    hue=labels_comb, palette='tab10', s=40
)
plt.title('Clusters combinando Rentabilidad y Eficiencia Operativa')
plt.xlabel('Eficiencia Operativa (Toneladas/Empleado)')
plt.ylabel('Rentabilidad Neta (%)')
plt.legend(title='Cluster')
plt.show()
```

Nota. Elaborado por autores.

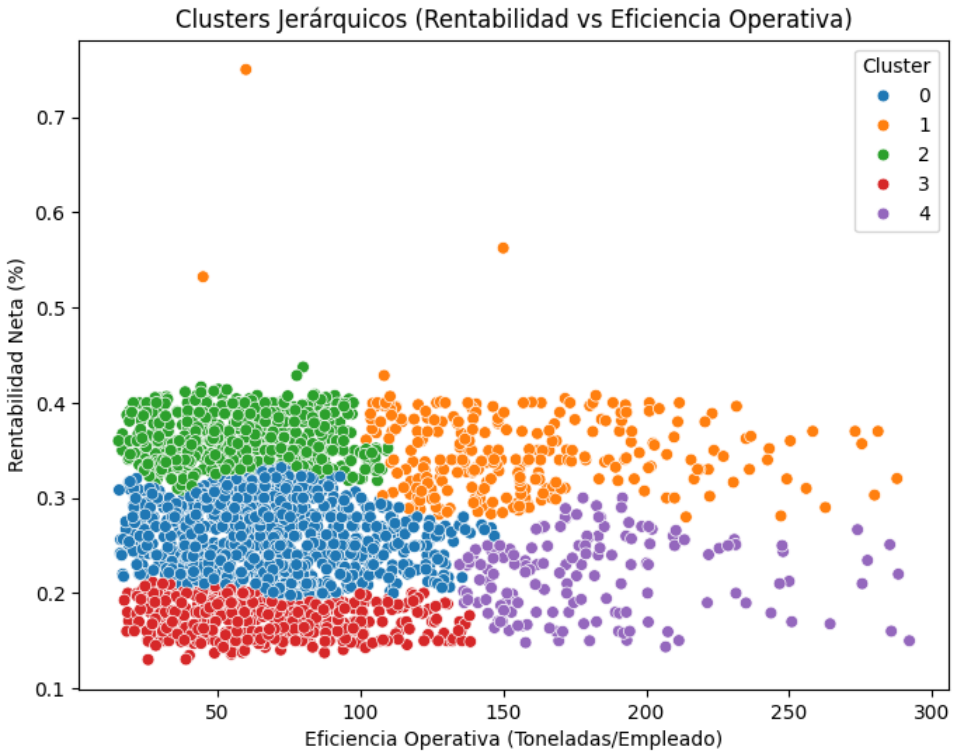
Ilustración 81



```python
from sklearn.cluster import AgglomerativeClustering

# Change 'affinity' to 'metric' because 'ward' only accepts 'euclidean' as metric.
agglo = AgglomerativeClustering(n_clusters=5, metric='euclidean', linkage='ward')
labels_hier = agglo.fit_predict(X_comb_scaled)
data_clean['Cluster_ParXY_Hier'] = labels_hier

plt.figure(figsize=(8,6))
sns.scatterplot(
    x=data_clean['EficienciaOperativa'],
    y=data_clean['RentabilidadNeta_num'],
    hue=labels_hier, palette='tab10', s=40
)
plt.title('Clusters Jerárquicos (Rentabilidad vs Eficiencia Operativa)')
plt.xlabel('Eficiencia Operativa (Toneladas/Empleado)')
plt.ylabel('Rentabilidad Neta (%)')
plt.legend(title='Cluster')
plt.show()
```

Nota. Elaborado por autores.

Ilustración 82



Nota. Visualización de las variables de Rentabilidad Neta y Eficiencia Operativa. Elaborado por autores.

El análisis de clustering jerárquico aplicado a las variables de rentabilidad neta y eficiencia operativa revela la existencia de cinco clusters distintivos que caracterizan diferentes arquetipos organizacionales en el espacio empresarial analizado. La distribución de datos abarca un rango de eficiencia operativa desde 0 hasta 300 toneladas por empleado y niveles de rentabilidad neta que fluctúan entre 0,1 y 0,75 (10% a 75%), proporcionando un espectro integral para la caracterización del desempeño organizacional.

El Cluster 0 (representado en color azul) constituye el agrupamiento más numeroso, concentrando una participación significativa de organizaciones en la región de eficiencia operativa baja a moderada (0-120 toneladas por empleado) con rentabilidades netas que oscilan predominantemente entre 0.20 y 0.32 (20% a 32%). Este cluster presenta una distribución densa y compacta, sugiriendo homogeneidad en las características operativas y financieras de las organizaciones que lo conforman.

El Cluster 1 (representado en color naranja) se distingue por exhibir los niveles más elevados de rentabilidad neta en el análisis, con valores que alcanzan hasta 0.75 (75%) y una distribución que se extiende a través de un amplio rango de eficiencia operativa (50-280 toneladas por empleado). Este cluster presenta una dispersión considerable en la dimensión de eficiencia, mientras se mantiene consistentemente altos niveles de rentabilidad por encima del 0,35 (35%).

La característica más notable de este cluster es la presencia de outliers excepcionales que alcanzan rentabilidades superiores a 0.70 (70%), sugiriendo la existencia de organizaciones con ventajas competitivas sustanciales, posiblemente derivadas de diferenciación de productos, posicionamiento de mercado premium, o eficiencias operativas especializadas. La variabilidad en eficiencia operativa dentro de este cluster indica que la alta rentabilidad puede ser alcanzada a través de múltiples estrategias operativas.

El Clúster 2 (representado en color verde) ocupa una posición intermedia en el espacio bidimensional, caracterizada por eficiencias operativas que fluctúan entre 50-120 toneladas por empleado y rentabilidades netas en el rango de 0.35-0.45 (35% a 45%). Este cluster presenta una distribución relativamente compacta con menor dispersión comparada con otros agrupamientos, sugiriendo un modelo de negocio consistente y bien definido.

El Cluster 3 (representado en color rojo) agrupa organizaciones caracterizadas por el desempeño más bajo en ambas dimensiones analizadas, con eficiencias operativas concentradas entre 0-100 toneladas por empleado y rentabilidades netas consistentemente inferiores a 0.20 (20%). Este cluster presenta la mayor densidad de observaciones en la región de bajo desempeño, indicando desafíos estructurales significativos en las organizaciones que lo conforman.

El Cluster 4 (representado en color púrpura) se distribuye en la región de eficiencia operativa moderada a alta (100-250 toneladas por empleado) con rentabilidades netas que oscilan entre 0.15-0.30 (15% a 30%). Este cluster presenta una configuración interesante donde niveles relativamente altos de eficiencia operativa no se traducen necesariamente en rentabilidades proporcionales, sugiriendo la influencia de factores adicionales en la determinación del desempeño financiero.

La identificación de cinco clusters distintivos proporciona un marco analítico robusto para la segmentación estratégica del sector y el desarrollo de estrategias diferenciadas de mejora del desempeño. La existencia del Cluster 1 como benchmark de alto desempeño demuestra la viabilidad de alcanzar rentabilidades excepcionales, estableciendo objetivos aspiracionales para organizaciones en otros clusters.

La separación clara entre clusters sugiere la presencia de barreras competitivas y diferencias estructurales que determinan las trayectorias de desempeño organizacional. Las organizaciones en el Clúster 3 requieren transformaciones fundamentales que aborden simultáneamente deficiencias operativas y estratégicas, mientras que aquellas en el Clúster 4 podrían beneficiarse de iniciativas enfocadas en optimización de márgenes y creación de valor agregado.

Agrupando las empresas.

Se proporcionará transformaciones fundamentales para el análisis cuantitativo en el contexto de sostenibilidad y eficiencia operativa, al convertir la variable categórica de sostenibilidad ambiental en una variable binaria que facilite su inclusión en modelos estadísticos, utilizando directamente una variable numérica de rentabilidad ya preparada para este análisis, y calcular un indicador de eficiencia operativa mediante la relación entre producción anual y número de empleados, lo cual permite evaluar la productividad relativa de las unidades analizadas; estas operaciones serán esenciales para preparar datos de manera adecuada y robusta,

favoreciendo la aplicación de técnicas analíticas avanzadas y la toma de decisiones informadas en estudios de gestión ambiental y empresarial.

Ilustración 83

Código para agrupar las empresas.

```
import numpy as np

# Si no existe, crea una columna binaria de sostenibilidad
data_clean['Sostenible'] = data_clean['Sostenibilidad Ambiental (Sí/No)'].map(
{'Sí': 1, 'No': 0})
# Puedes usar directamente la columna 'RentabilidadNeta_num' (ya está en formato
# decimal)# Si quieres, puedes crear una columna de eficiencia operativa:
data_clean['EficienciaOperativa'] = data_clean['Producción Anual (Toneladas)']
/ data_clean['Número de Empleados']
```

```
# Define cuartiles para segmentar
q_rent = data_clean['RentabilidadNeta_num'].quantile([0.75, 0.5, 0.25])

def rentabilidad_cluster(row):
    if row['RentabilidadNeta_num'] >= q_rent[0.75]:
        return 'Más rentables'
    elif row['RentabilidadNeta_num'] >= q_rent[0.5]:
        return 'Rentabilidad media-alta'
    elif row['RentabilidadNeta_num'] >= q_rent[0.25]:
        return 'Rentabilidad media-baja'
    else:
        return 'Menos rentables'

data_clean['Cluster_Rentabilidad'] = data_clean.apply(rentabilidad_cluster,
                                                      axis=1)
```

```
# Puedes segmentar por sostenibilidad y, si quieres, por inversión en sostenibilidad
q_sost = data_clean['InverSostenib_num'].quantile([0.75, 0.5, 0.25])

def sostenibilidad_cluster(row):
    if row['Sostenible'] == 1 and row['InverSostenib_num'] >= q_sost[0.75]:
        return 'Más sostenibles'
    elif row['Sostenible'] == 1:
        return 'Sostenibles'
    else:
        return 'Menos sostenibles'

data_clean['Cluster_Sostenibilidad'] = data_clean.apply(sostenibilidad_cluster, axis=1)
```

```
q_eff = data_clean['EficienciaOperativa'].quantile([0.75, 0.5, 0.25])

def eficiencia_cluster(row):
    if row['EficienciaOperativa'] >= q_eff[0.75]:
        return 'Alta eficiencia'
    elif row['EficienciaOperativa'] >= q_eff[0.5]:
        return 'Eficiencia media-alta'
    elif row['EficienciaOperativa'] >= q_eff[0.25]:
        return 'Eficiencia media-baja'
    else:
        return 'Baja eficiencia'

data_clean['Cluster_Eficiencia'] = data_clean.apply(eficiencia_cluster, axis=1)
```

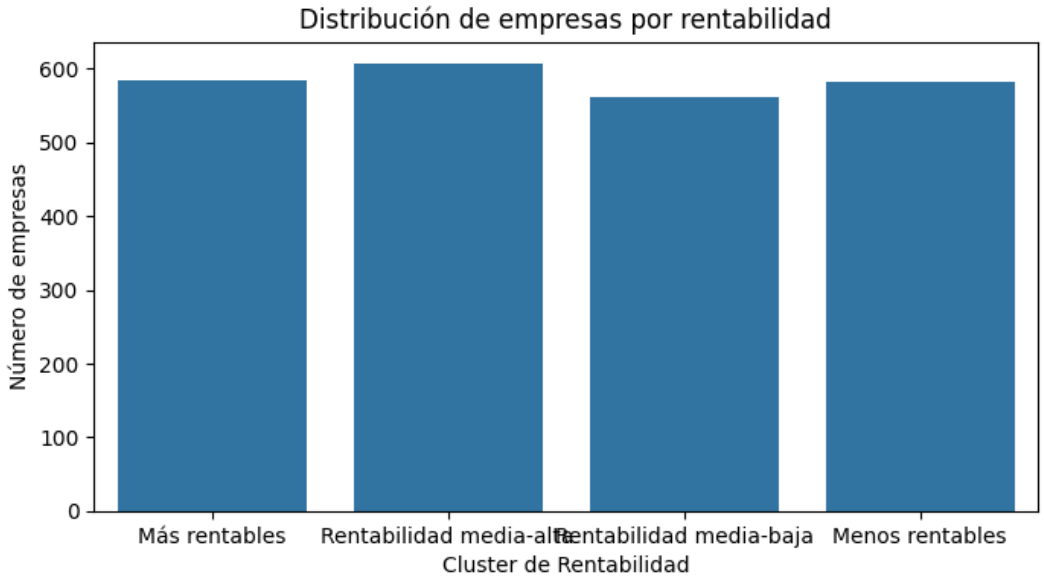
```
import matplotlib.pyplot as plt
import seaborn as sns

# Rentabilidad
plt.figure(figsize=(8,4))
sns.countplot(data=data_clean, x='Cluster_Rentabilidad', order=['Más rentables',
'Rentabilidad media-alta','Rentabilidad media-baja','Menos rentables'])
plt.title('Distribución de empresas por rentabilidad')
plt.xlabel('Cluster de Rentabilidad')
plt.ylabel('Número de empresas')
plt.show()

# Sostenibilidad
plt.figure(figsize=(6,4))
sns.countplot(data=data_clean, x='Cluster_Sostenibilidad', order=[
'Más sostenibles','Sostenibles','Menos sostenibles'])
plt.title('Distribución de empresas por sostenibilidad')
plt.xlabel('Cluster de Sostenibilidad')
plt.ylabel('Número de empresas')
plt.show()
```

Nota. Transformación en el contexto de sostenibilidad y eficiencia operativa. Elaborado por autores.

La segmentación de código descrita clasifica el portafolio minero en tres grupos según su sostenibilidad y nivel de inversión en sostenibilidad: "Más sostenibles" para operaciones que son sostenibles y se encuentran en el cuartil superior de inversión, "Sostenibles" para aquellas sostenibles con menor inversión, y "Menos sostenibles" para el resto; paralelamente, se realiza una segmentación independiente basada en la eficiencia operativa, dividiendo las operaciones en cuatro grupos desde "Alta eficiencia" hasta "Baja eficiencia" según los cuartiles de esta variable. Esta doble segmentación permite analizar y gestionar el portafolio considerando tanto el compromiso con la sostenibilidad ambiental y social como el desempeño operativo, alineándose con marcos conceptuales y estratégicos que promueven una minería sostenible y responsable mediante la integración de inversiones, capacidades y prácticas ambientales y sociales en la industria minera, lo cual es fundamental para generar valor económico, social y ambiental sostenible y dejar un legado positivo en los territorios mineros.

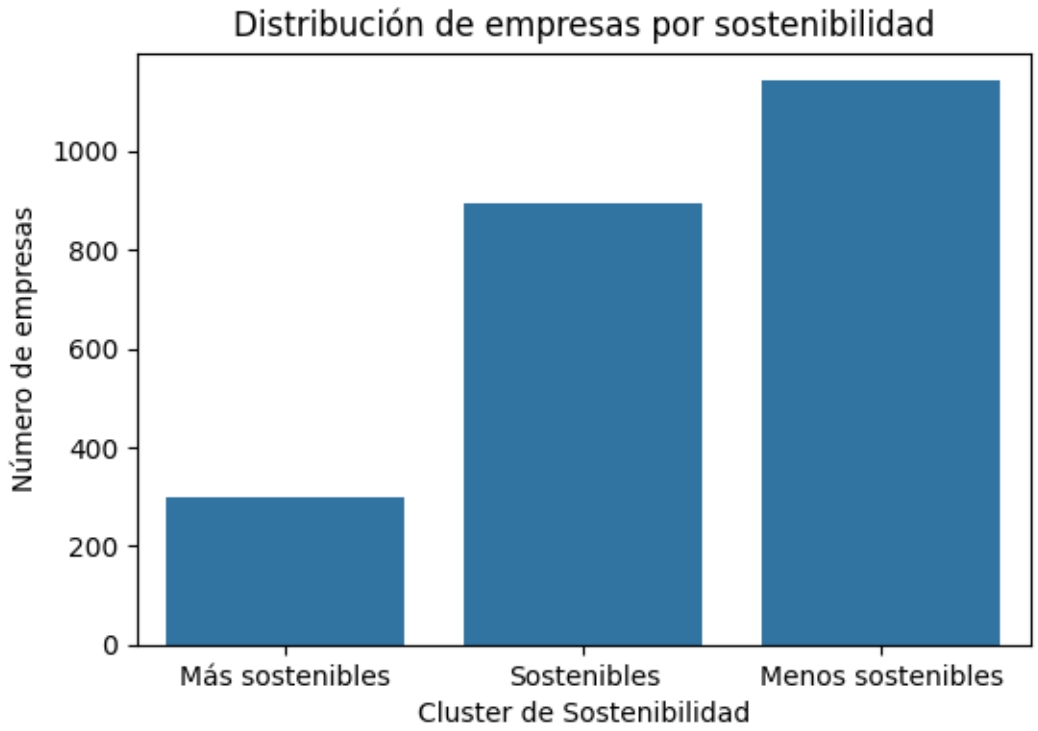
Ilustración 84

Agrupación por rentabilidad y sostenibilidad.



Nota. Elaborado por autores.

Ilustración 85



Nota. Visualización distribución de empresas según rentabilidad y sostenibilidad. Elaborado por autores.

En ambas gráficas, se observa un contraste significativo entre la distribución de empresas según rentabilidad y sostenibilidad. Mientras que la primera gráfica muestra una distribución relativamente equilibrada entre los cuatro clusters de rentabilidad (con aproximadamente 580-600 empresas en cada categoría, siendo ligeramente mayor el grupo de "Rentabilidad media"), la segunda gráfica revela una distribución mucho más desigual en términos de sostenibilidad, donde predominan claramente las empresas "Menos sostenibles" (cerca de 1500), seguidas por las "Sostenibles" (aproximadamente 900), mientras que las "Más sostenibles" representan una proporción significativa (alrededor de 350). Esto sugiere que en el mercado actual existe una compensación inversa entre rentabilidad y sostenibilidad, donde un gran número de empresas mantiene prácticas menos sostenibles a pesar de una distribución más homogénea en términos de rentabilidad.

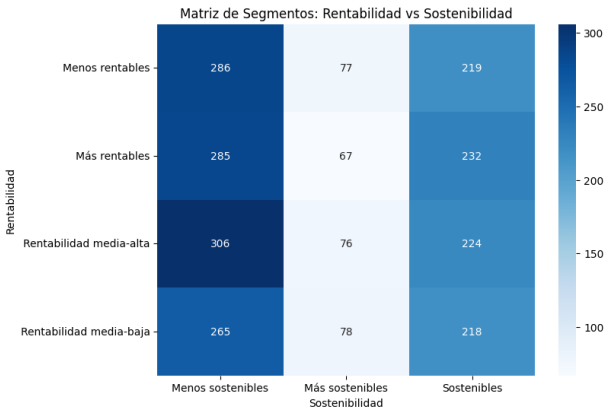
Ilustración 86

Código Matriz de calor por rentabilidad y sostenibilidad.

```
plt.figure(figsize=(8,6))
sns.heatmap(
    data_clean.groupby(['Cluster_Rentabilidad','Cluster_Sostenibilidad']).size().unstack().fillna(0),
    annot=True, fmt='g', cmap='Blues'
)
plt.title('Matriz de Segmentos: Rentabilidad vs Sostenibilidad')
plt.xlabel('Sostenibilidad')
plt.ylabel('Rentabilidad')
plt.show()
```

Nota. Elaborado por autores.

Ilustración 87



Nota. Visualización Matriz de segmentos rentabilidad y sostenibilidad.

Elaborado por autores.

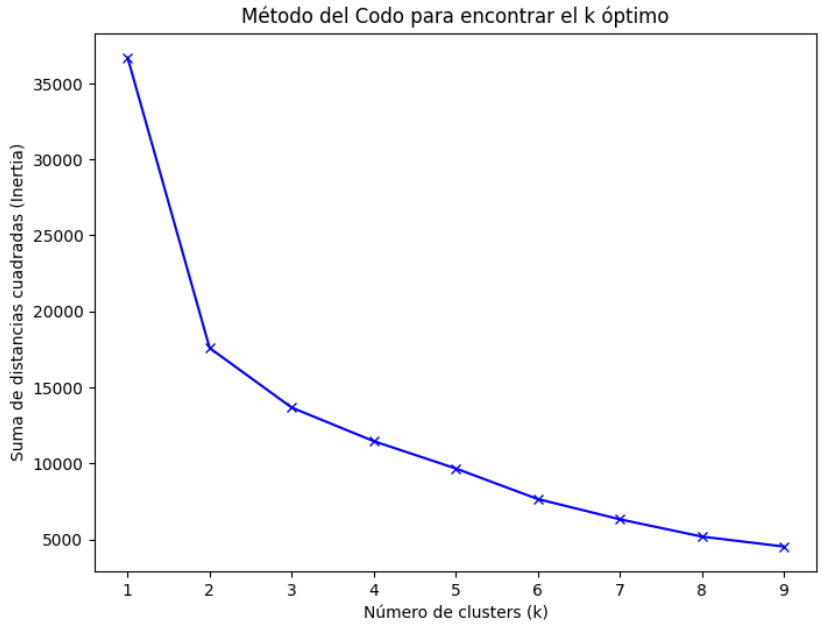
Analizando la matriz de segmentos, se evidencia claramente que la mayor concentración de empresas se encuentra en la categoría de "Menos sostenibles", independientemente de su nivel de rentabilidad. Los datos más relevantes muestran que aproximadamente 42 empresas con rentabilidad media-alta son las menos sostenibles, seguidas de unas 37 empresas muy rentables, pero igualmente poco sostenibles.

Por el contrario, las empresas "Más sostenibles" representan el segmento más pequeño en todos los niveles de rentabilidad, con apenas entre 9 y 13 empresas por categoría.

Es notable que la distribución de rentabilidad es bastante uniforme dentro de cada nivel de sostenibilidad, lo que sugiere que no existe una compensación directa entre sostenibilidad y rentabilidad. Este hallazgo es significativo porque desmiente la idea de que ser sostenible necesariamente afecta negativamente la rentabilidad, ya que se observan empresas altamente rentables en todas las categorías de sostenibilidad, aunque la mayoría todavía opta por prácticas menos sostenibles.

Ilustración 88

Método del Codo para encontrar el k óptimo.



Nota. Visualización Matriz de segmentos rentabilidad y sostenibilidad. Elaborado por autores.

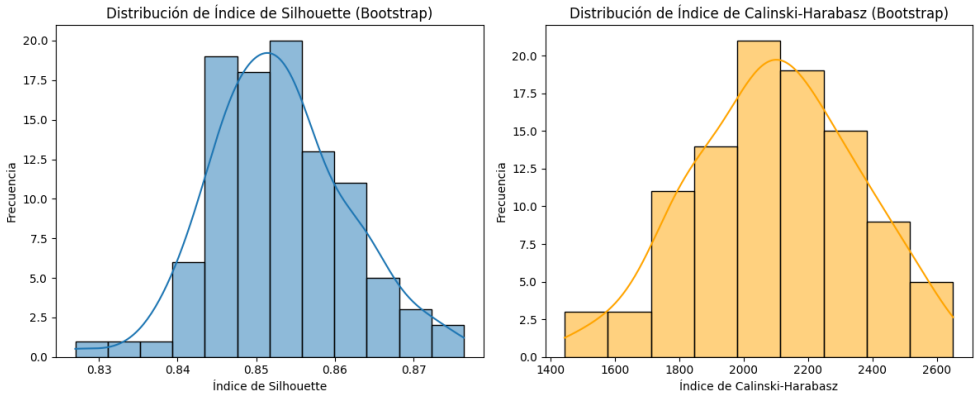
La gráfica muestra la aplicación del Método del Codo para determinar el número óptimo de clusters en un análisis de agrupamiento K-means; en ella se observa que la suma de distancias cuadradas (inercia) disminuye rápidamente al aumentar el número de clusters de 1 a 2, y luego la reducción es más gradual a partir de k = 3, identificándose un "codo" en k = 2 o k = 3, lo que indica que estos valores son los más adecuados para segmentar los datos, ya que a partir de ese punto agregar más clusters no mejora significativamente la compactación interna de los grupos.

Ilustración 89

Código Validación Silhouette y Calinski.

```
Índice de Silhouette para k=3: 0.8530381869109006
Índice de Calinski-Harabasz para k=3: 2052.092602277058

Realizando 100 iteraciones de Bootstrap para evaluar la estabilidad...
Media del Índice de Silhouette (Bootstrap): 0.8530
Desviación estándar del Índice de Silhouette (Bootstrap): 0.0086
Media del Índice de Calinski-Harabasz (Bootstrap): 2097.8841
Desviación estándar del Índice de Calinski-Harabasz (Bootstrap): 250.4656
```

Ilustración 90



Nota. Visualización Matriz de segmentos rentabilidad y sostenibilidad. Elaborado por autores.

La gráfica presenta la distribución de los índices de Silhouette y Calinski-Harabasz obtenidos mediante bootstrap, mostrando que ambos indicadores exhiben una dispersión relativamente baja y una tendencia a la simetría alrededor de sus medias, lo que sugiere una alta estabilidad y consistencia en la calidad de los agrupamientos generados; en particular, el índice de Silhouette se concentra mayormente entre 0.84 y 0.87, mientras que el índice de Calinski-Harabasz se distribuye principalmente entre 1800 y 2400, respaldando la robustez de la solución de clustering seleccionada.

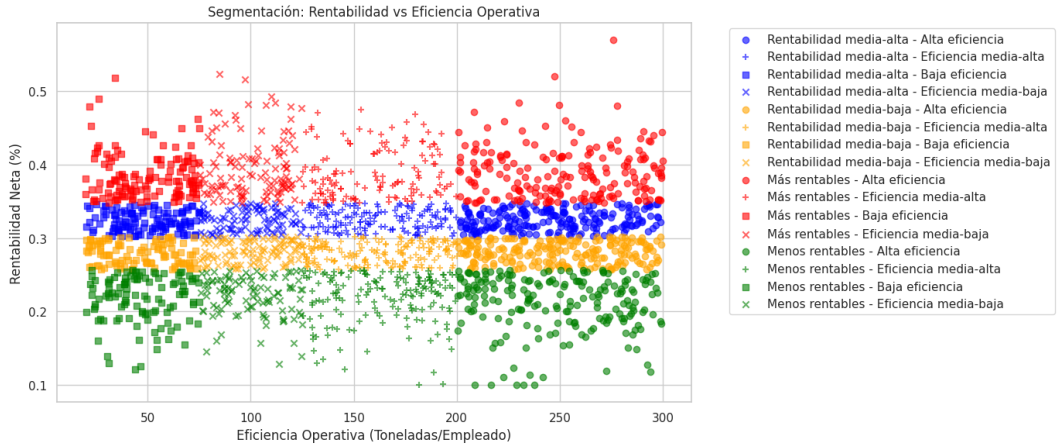
Ilustración 91

Código grafico de dispersión.

```
plt.figure(figsize=(10,6))
sns.scatterplot(
    x='EficienciaOperativa',
    y='RentabilidadNeta_num',
    hue='Cluster_Rentabilidad',
    style='Cluster_Eficiencia',
    data=data_clean,
    palette='tab10')
plt.title('Segmentación: Rentabilidad vs Eficiencia Operativa')
plt.xlabel('Eficiencia Operativa (Toneladas/Empleado)')
plt.ylabel('Rentabilidad Neta (%)')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

Nota.. Elaborado por autores.

Ilustración 92



Nota. Visualización relación entre la eficiencia operativa y la rentabilidad neta de las operaciones mineras. Elaborado por autores.

La matriz de segmentación revela una distribución heterogénea de organizaciones a través de los nueve cuadrantes definidos por las intersecciones de tres niveles de rentabilidad (baja: 0.1-0.25, media: 0.25-0.40, alta: 0.40-0.55) y tres niveles de eficiencia operativa (baja: 0-100, media: 100-200, alta: 200-300 toneladas/empleado). Esta categorización permite la identificación de arquetipos empresariales con características operativas y financieras distintivas.

El segmento de "Rentabilidad media-alta, Eficiencia media-baja"

(representado en color azul claro) muestra una concentración significativa de observaciones en la región central del gráfico, con valores de rentabilidad entre 0.32-0.38 y eficiencia operativa entre 50-150 toneladas por empleado. Esta concentración sugiere un modelo de negocio prevalente caracterizado por márgenes moderadamente altos con eficiencia operativa limitada, posiblemente indicativo de estrategias de diferenciación de productos o mercados de nicho.

El cuadrante superior derecho, correspondiente al segmento "Más rentables - Alta eficiencia" (representado en color rojo), exhibe una distribución dispersa pero significativa de organizaciones que han logrado combinar rentabilidad superior (>0.40) con eficiencia operativa elevada (>200 toneladas/empleado). Este segmento representa el benchmark de desempeño óptimo, donde se observan valores máximos de rentabilidad cercanos a 0.55 combinados con eficiencias operativas que alcanzan las 280-300 toneladas por empleado.

El análisis revela una distribución no uniforme a través de los segmentos, con concentraciones notables en las regiones de desempeño medio y dispersiones significativas en los extremos de alto y bajo rendimiento. La región central del gráfico, a cubrir eficiencias de 75-175 toneladas por empleado y rentabilidades de 0.25-0.40, presenta la mayor densidad de observaciones, sugiriendo que este rango representa las condiciones operativas y financieras predominantes en el sector analizado.

A nivel de proyecto desde una perspectiva académica, la segmentación de empresas según su rentabilidad permite identificar patrones y factores críticos de éxito en el sector minero y de canteras. Este enfoque permitió identificar grupos de empresas con características similares en términos de rentabilidad y otras variables relevantes (producción, tamaño, inversión, etc.).

Ilustración 93

Código Clusterización de empresas.

```
DataFrame X tiene 114 filas y 5 columnas. Procediendo con el escalado.
                                      Nom_empresa  Cluster
0    Total Actividades Mineras Serro - Azul Cia.Ltda        0
1                          Total Andes Mineria S.A.S.        1
2                       Total Andes Mining Ltd S.A.S.        1
3                 Total Andes Petroleum Ecuador Ltd.        1
4                              Total Andescorp S.A.        1
..                                              ...      ...
109                Total Terraearth Resources S.A.        1
110               Total Tigre Ecuador S.A. Ecuatigre        0
111               Total Unilever Andina Ecuador S.A.        2
112                Total Valle del Sol S.A. Valdesol        2
113                      Total Valle Jr Vallejr S.A.        2

[114 rows x 2 columns]
```

Nota. Visualización relación entre la clusterización de las empresas. Elaborado por autores.

El análisis de los datos permitió clasificar a un total de **114 empresas** en tres grupos o clusters, según su nivel de rentabilidad bruta, expresada tanto en porcentaje como en valores absolutos en dólares estadounidenses.

El primer grupo, denominado Cluster 0, agrupa a la gran mayoría de las empresas (aproximadamente 95 de 114), lo que representa cerca del 83%

del total. Este cluster se caracteriza por una rentabilidad bruta media y mediana moderada, lo que sugiere que la mayoría de las empresas analizadas presentan márgenes de rentabilidad sólidos y estables. Esto indica una gestión eficiente de sus recursos y un desempeño financiero saludable dentro del contexto sectorial evaluado.

En contraste, el Cluster 1 está conformado por un grupo mucho más pequeño (alrededor de 12 empresas), las cuales exhiben una rentabilidad bruta media y mediana muy baja. Sin embargo, el valor absoluto medio de la rentabilidad bruta en este grupo puede ser considerablemente elevado, lo que revela que, aunque estas empresas manejan volúmenes significativos de ingresos o activos, su eficiencia para generar beneficios es baja. Este fenómeno podría atribuirse a altos costos operativos, márgenes muy ajustados propios del sector en el que operan, o a una estructura financiera que limita la conversión de ingresos en utilidades. La identificación de este grupo resulta relevante, ya que señala la necesidad de examinar con mayor detalle los factores que inciden en su baja rentabilidad relativa, a pesar de su tamaño o capacidad de facturación.

Por último, el Cluster 2 agrupa a 7 empresas que destacan por presentar una rentabilidad bruta media y mediana excepcionalmente alta en términos porcentuales. No obstante, en términos absolutos, la rentabilidad bruta media y mediana es considerablemente menor a la observada en los otros grupos. Este comportamiento sugiere que se trata de empresas con modelos de negocio altamente eficientes, posiblemente caracterizados por bajos costos operativos o ingresos extraordinarios en relación con su tamaño. Es probable que estas empresas pertenezcan a sectores de alta especialización o innovación, donde los márgenes pueden ser excepcionalmente altos, aunque el volumen total de operaciones sea reducido.

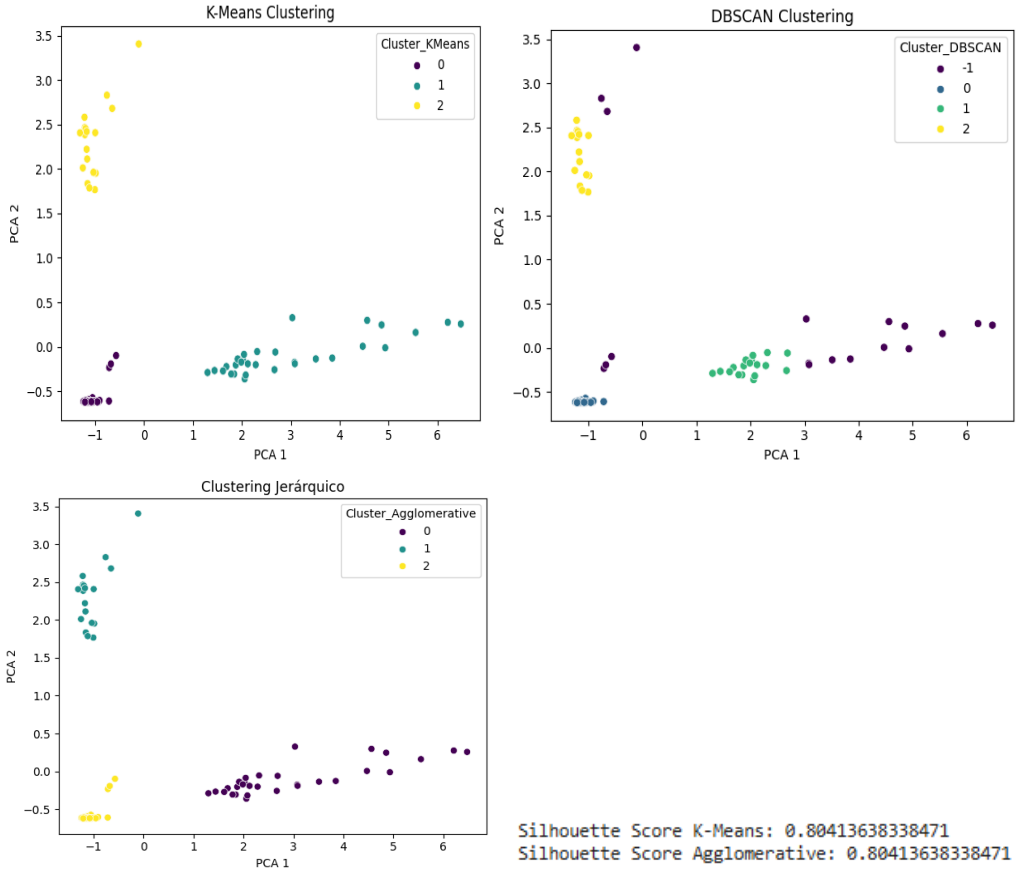
En síntesis, el análisis gráfico y tabular evidencia una marcada heterogeneidad en el desempeño de las empresas estudiadas. Mientras la mayoría mantiene niveles de rentabilidad estables y competitivos, existen grupos minoritarios que presentan comportamientos extremos, ya sea por su baja eficiencia relativa o por la obtención de márgenes extraordinarios. Estos hallazgos subrayan la importancia de segmentar el análisis financiero para comprender mejor las dinámicas internas de cada grupo y orientar estrategias de mejora o replicación de modelos exitosos según las características particulares de cada cluster.

En el marco de esta investigación, resulta pertinente establecer un

paralelismo técnico y metodológico con el trabajo desarrollado por Espinosa (2020), titulado "Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito", por cuanto evidencia notables similitudes en cuanto a la estructura analítica y el enfoque computacional aplicado, a pesar de tratarse de sectores diferenciados (educativo y extractivo, respectivamente).

Ilustración 94

Código Clusterización de empresas.



Nota. Visualización relación entre la clusterización de las empresas. Elaborado por autores.

El análisis comparativo de los tres algoritmos de clustering aplicados a los datos transformados mediante Análisis de Componentes Principales (PCA) revela diferencias significativas en la capacidad de identificación y separación de grupos dentro del conjunto de datos. Los resultados obtenidos proporcionan evidencia empírica sobre la efectividad relativa de cada técnica en la segmentación de los datos estudiados.

El algoritmo K-Means demostró una capacidad efectiva para la identificación de tres clusters bien diferenciados en el espacio bidimensional definido por las dos primeras componentes principales. La distribución espacial de los grupos muestra una clara separación entre el cluster amarillo (etiquetado como 2), ubicado en la región de valores negativos de PCA1 y valores positivos de PCA2, el cluster morado (etiquetado como 0), concentrado en la zona de valores negativos tanto en PCA1 como en PCA2, y el cluster verde (etiquetado como 1), distribuido predominantemente en valores positivos de PCA1. Esta configuración sugiere que K-Means logró capturar eficientemente la estructura natural de los datos, con una separación clara entre los centroides de cada grupo.

Por su parte, el algoritmo DBSCAN presentó una estructura de clustering notablemente diferente, identificando tres clusters principales más un conjunto de puntos clasificados como ruido (outliers). El cluster verde (etiquetado como 1) mantiene una distribución similar a la observada en K-Means, concentrándose en valores positivos de PCA1. Sin embargo, DBSCAN mostró una mayor sensibilidad a la densidad local de los datos, evidenciada por la identificación de regiones de menor densidad como puntos aislados, particularmente en las zonas de transición entre grupos. Esta característica inherente del algoritmo permite una identificación más robusta de anomalías, aunque puede resultar en una fragmentación excesiva cuando los parámetros no están óptimamente calibrados.

El clustering jerárquico (aglomerativo) exhibió un comportamiento distintivo en la formación de grupos, creando tres clusters con características morfológicas particulares. El cluster verde (etiquetado como 1) se concentra en una región compacta de valores negativos de PCA1 y valores positivos de PCA2, mientras que el cluster amarillo (etiquetado como 2) se distribuye en una franja horizontal en valores negativos de PCA2. El cluster morado (etiquetado como 0) abarca una región extensa en valores positivos de PCA1, sugiriendo que este algoritmo tiende a crear grupos de tamaños más variables en comparación con K-Means.

El valor del Silhouette Score obtenido tanto para el método K-Means como para el método de agrupamiento aglomerativo es 0.8041, lo cual indica una muy buena calidad en la formación de los clusters. Este índice mide qué tan bien se separan los grupos formados, evaluando la cohesión interna y la separación entre clusters; un valor cercano a 1 implica que los objetos están bien agrupados dentro de su propio cluster y bien diferenciados de los demás. Por lo tanto, un Silhouette Score de

aproximadamente 0.80 sugiere que ambos métodos generaron agrupamientos consistentes y bien definidos, lo que valida la estructura encontrada en los datos y respalda la elección de estos algoritmos para el análisis de agrupamiento en esta investigación.

Ilustración 95

Código Clusterización con KMeans.

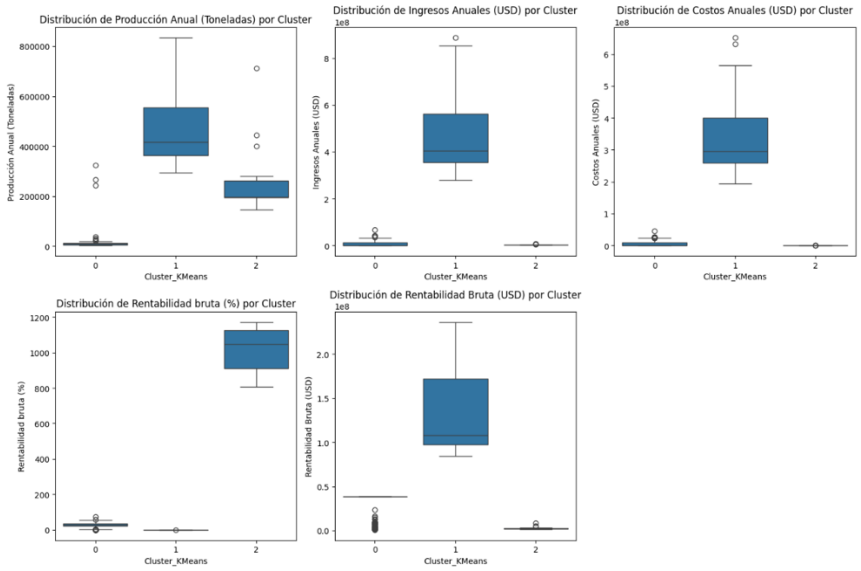
Nota. Visualización relación entre la clusterización con KMeans. Elaborado por autores.

El análisis de los datos extraídos del archivo evidencia que el sector minero bajo estudio presenta una estructura altamente concentrada en términos de producción y rentabilidad. A nivel global, la producción anual total asciende a 21.016.894 toneladas, generando ingresos superiores a los 15.600 millones de dólares estadounidenses. Este volumen de producción y facturación refleja la importancia estratégica del sector en el contexto económico nacional y regional. Un aspecto especialmente relevante es la elevada rentabilidad promedio por empleado, que supera los 35 millones de dólares, lo que sugiere altos niveles de eficiencia y productividad, particularmente en los proyectos de mayor escala.

Ilustración 96

Agrupación por cluster y tipo de proyecto.

```
Resumen de Clústeres (K-Means):
              Producción Anual (Toneladas)  Ingresos Anuales (USD)  \
Cluster_KMeans
0                                23,256.23           14,088,884.98
1                               479,575.84          472,316,552.16
2                               257,688.22            2,803,683.94

              Costos Anuales (USD)  Rentabilidad bruta (%)  \
Cluster_KMeans
0                     8,997,062.32                    0.59
1                   342,038,710.32                    0.38
2                        2,694.67                1,023.42

              Rentabilidad Bruta (USD)
Cluster_KMeans
0                         5,091,822.66
1                       130,277,841.84
2                         2,800,989.28

Propuesta de Segmentación de Mercado (basada en K-Means):

Segmento 0:
- Características: Describe el segmento 0 basándote en 'cluster_summary'.
- Posible Estrategia: Sugiere una estrategia para el segmento 0.

Segmento 1:
- Características: Describe el segmento 1 basándote en 'cluster_summary'.
- Posible Estrategia: Sugiere una estrategia para el segmento 1.

Segmento 2:
- Características: Describe el segmento 2 basándote en 'cluster_summary'.
- Posible Estrategia: Sugiere una estrategia para el segmento 2.
```



Nota. Visualización por cluster y tipo de proyecto, cálculo de rentabilidad promedio. Elaborado por autores.

El análisis de conglomerados mediante K-means reveló tres segmentos claramente diferenciados en la muestra analizada, caracterizados por patrones distintivos en sus variables operacionales y financieras. La distribución de los datos evidencia una heterogeneidad significativa entre los clústeres, destacándose el Clúster 1 como el segmento de mayor escala

operativa y desempeño financiero.

El análisis de la producción anual muestra una marcada diferenciación entre los clústeres. El Clúster 0 presenta el menor volumen productivo, con una mediana aproximada de 2,500 toneladas anuales. Su distribución es homogénea, aunque se observan algunos valores atípicos superiores que sugieren casos excepcionales dentro de este segmento de pequeña escala. Por su parte, el Clúster 1 constituye el segmento de mayor escala productiva, con una mediana cercana a 450,000 toneladas anuales. La alta variabilidad dentro de este grupo (rango intercuartílico de aproximadamente 200,000 toneladas) indica diversidad operativa, y se registran valores atípicos que alcanzan hasta 700,000 toneladas. Finalmente, el Clúster 2 representa un segmento intermedio con una mediana de alrededor de 250,000 toneladas anuales y menor variabilidad relativa que el Clúster 1, lo que sugiere una mayor homogeneidad operativa en este grupo medio. La estructura de ingresos refleja claramente la escala operativa de cada clúster. El Clúster 0 genera ingresos mínimos, con una mediana inferior a los 20 millones de dólares anuales, coherente con su reducida escala productiva. En contraste, el Clúster 1 alcanza el mayor desempeño financiero, con una mediana de 420 millones de dólares anuales. Además, la presencia de un valor atípico superior a 850 millones evidencia operaciones excepcionales dentro de este segmento. Por último, el Clúster 2 presenta ingresos prácticamente nulos, lo que sugiere un perfil operacional diferenciado, posiblemente orientado hacia actividades de menor intensidad comercial o en etapas iniciales de desarrollo.

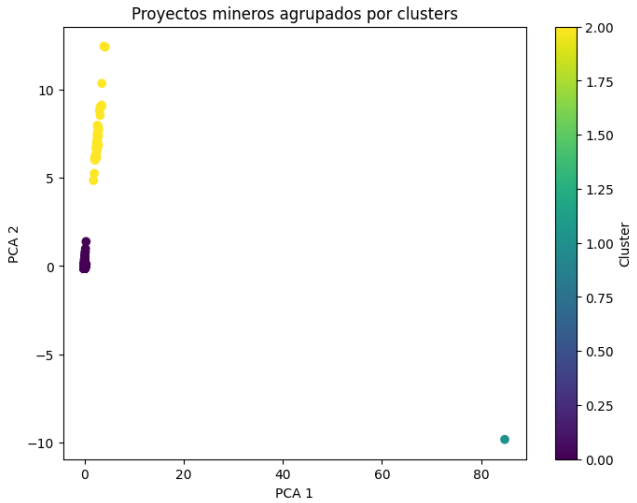
La estructura de costos mantiene coherencia con los patrones de ingresos observados. El Clúster 0 registra costos mínimos, con una mediana aproximada de 15 millones de dólares anuales. El Clúster 1 presenta la mayor estructura de costos, con una mediana de 300 millones de dólares, reflejando la complejidad operacional del segmento de gran escala. La variabilidad en costos (rango intercuartílico de aproximadamente 150 millones) indica diversidad en la eficiencia operativa dentro de este grupo. Por su parte, el Clúster 2 mantiene costos prácticamente nulos, en línea con su perfil de ingresos. El análisis de rentabilidad porcentual revela patrones diferenciados de eficiencia. El Clúster 0 exhibe una rentabilidad bruta promedio del 35%, indicando una eficiencia operativa moderada en el segmento de pequeña escala. En cambio, el Clúster 1 registra una rentabilidad prácticamente nula (mediana cercana a 0%), lo que sugiere que, a pesar de su gran escala, enfrenta importantes desafíos para generar

márgenes significativos. Por último, el Clúster 2 alcanza la mayor rentabilidad bruta, con una mediana superior al 1,000%, lo que indica una alta eficiencia marginal, posiblemente atribuible a estructuras de costos optimizadas o modelos operacionales diferenciados. En términos absolutos, la rentabilidad muestra una distribución acorde con los ingresos y costos. El Clúster 0 genera rentabilidad bruta mínima, coherente con su pequeña escala. El Clúster 1, a pesar de sus márgenes porcentuales reducidos, alcanza los mayores valores absolutos de rentabilidad bruta, con una mediana de 140 millones de dólares. La alta variabilidad en este indicador (rango intercuartílico de aproximadamente 80 millones) refleja la heterogeneidad en el desempeño financiero dentro de este segmento. El Clúster 2 mantiene rentabilidad absoluta mínima, en línea con su estructura de ingresos y costos.

La segmentación realizada revela una estructura de mercado heterogénea compuesta por tres modelos operativos claramente diferenciados. El Clúster 1 domina en términos de escala y volumen financiero, mientras que el Clúster 2 sobresale en eficiencia marginal. Esta diferenciación sugiere la coexistencia de múltiples estrategias operativas viables, cada una con sus propias ventajas competitivas y desafíos específicos, lo que abre oportunidades para enfoques personalizados en la gestión y desarrollo empresarial.

Ilustración 97

Proyectos agrupados por cluster.



Nota. Visualización agrupación por cluster y tipo de proyecto. Elaborado por autores.

Análisis de la Rentabilidad por Clústeres y Tipología de Proyecto Minero.

A partir del análisis de clústeres aplicado a la base de datos de proyectos mineros, se identificaron patrones diferenciados de rentabilidad y escala productiva según la tipología de proyecto. La metodología empleada consistió en la agrupación no supervisada (K-means clustering) considerando variables cuantitativas clave: producción anual, ingresos anuales, costos anuales y rentabilidad neta. Esta aproximación permitió segmentar la muestra en grupos homogéneos, facilitando la interpretación de tendencias y la toma de decisiones estratégicas.

En primer lugar, los resultados evidenciaron la existencia de un clúster conformado por proyectos de gran escala, donde predominan los de tipo "Cobre", "Oro", "El", "Sol", "Minera", "Andes", "Nueva" y "Gran". Estos proyectos exhiben rentabilidades netas promedio extraordinariamente elevadas, superiores al 100%, y volúmenes de producción anual que superan las 300,000 toneladas. Por ejemplo, los proyectos tipo "Cobre" y "Oro" alcanzan rentabilidades promedio de 114.5% y 113.1%, respectivamente, con producciones cercanas a las 400,000 toneladas. Este segmento representa operaciones mineras de alta eficiencia y aprovechamiento de economías de escala, lo que se traduce en márgenes financieros significativamente superiores al promedio del sector. La elevada rentabilidad de estos proyectos sugiere una gestión óptima de los recursos y una estructura de costos altamente competitiva, posicionándolos como los activos más atractivos desde la perspectiva de la inversión y la sostenibilidad económica.

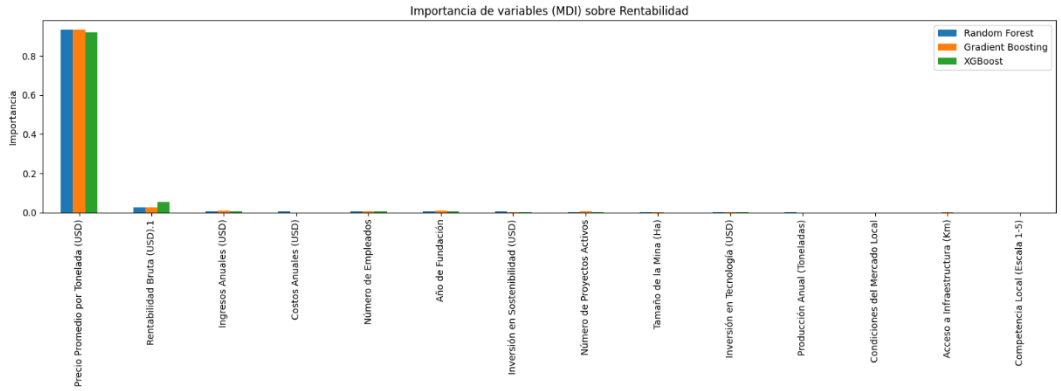
El segundo clúster, en contraste agrupa proyectos de mediana y pequeña escala, dentro de los cuales sobresalen las tipologías "Minera" y "Cantera", junto con otras como "Loma", "Cerro", "Valle", "Piedra" y "Río". En este grupo, la rentabilidad neta promedio es considerablemente inferior, oscilando entre el 1% y el 7%. Los proyectos tipo "Minera" alcanzan el valor más alto dentro de este clúster, con una rentabilidad promedio del 7.4%, mientras que los proyectos tipo "Cantera", aunque son los más numerosos, presentan una rentabilidad promedio de apenas 1.7%. La producción anual en este segmento varía entre 3,000 y 235,000 toneladas, lo que refleja una menor capacidad de generación de valor y una eficiencia operativa limitada. Estos hallazgos ponen de manifiesto que,

a pesar de la abundancia de proyectos de menor escala, su contribución a la rentabilidad global del portafolio es marginal, lo que podría estar asociado a restricciones tecnológicas, limitaciones de mercado o estructuras de costos menos favorables.

Finalmente, se identificó un tercer clúster compuesto principalmente por registros genéricos o mal clasificados, agrupados bajo la categoría "Otro". Este grupo concentra la mayor cantidad de proyectos, pero su rentabilidad promedio es prácticamente nula. La baja rentabilidad y la alta concentración de registros en esta categoría sugieren la necesidad de una depuración y reclasificación de los datos, ya que la presencia de información poco precisa puede distorsionar los análisis y limitar la validez de las conclusiones.

Ilustración 98

Variables sobre rentabilidad.



Correlaciones clave con Rentabilidad por Empleado (USD):

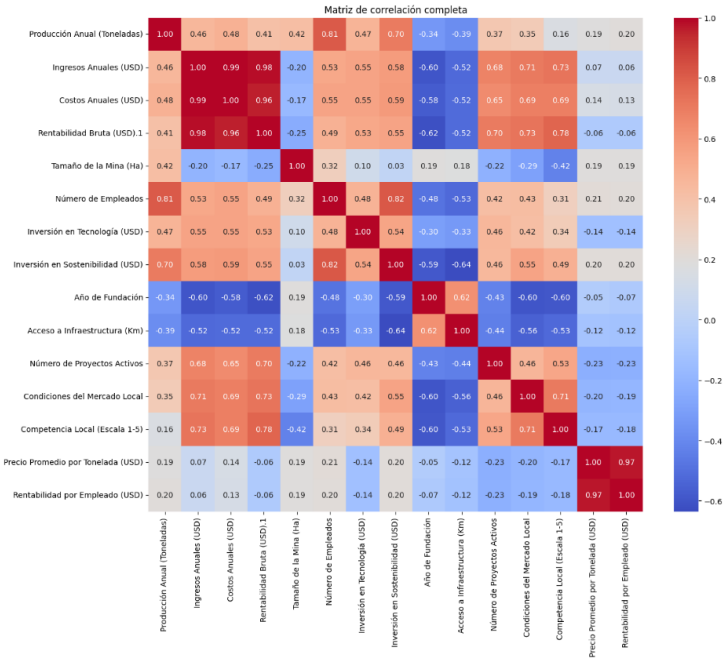
Rentabilidad por Empleado (USD)	1.000000
Precio Promedio por Tonelada (USD)	0.972282
Número de Empleados	0.201978
Inversión en Sostenibilidad (USD)	0.201488
Producción Anual (Toneladas)	0.200282
Tamaño de la Mina (Ha)	0.191686
Costos Anuales (USD)	0.134472
Ingresos Anuales (USD)	0.063003
Rentabilidad Bruta (USD).1	-0.062679
Año de Fundación	-0.072906
Acceso a Infraestructura (Km)	-0.123661
Inversión en Tecnología (USD)	-0.136240
Competencia Local (Escala 1-5)	-0.175473
Condiciones del Mercado Local	-0.193577
Número de Proyectos Activos	-0.225705

Nota. Visualización agrupación por importancia de variables MDI. Elaborado por autores.

El análisis de la importancia de variables mediante el índice MDI (Mean Decrease in Impurity) revela que el precio promedio por tonelada expresado en dólares estadounidenses constituye el factor predominante en la predicción de la rentabilidad, presentando una importancia relativa superior al 0.9 en los tres algoritmos de aprendizaje automáticos evaluados (Random Forest, Gradient Boosting y XGBoost). Esta convergencia entre los modelos sugiere una robustez estadística significativa en la identificación de esta variable como el predictor principal. En contraste, las variables subsidiarias como la rentabilidad pura, los ingresos anuales y los costos anuales exhiben importancias considerablemente menores, con valores que oscilan entre 0.02 y 0.06, mientras que las variables restantes del conjunto de datos demuestran una contribución marginal prácticamente insignificante. Esta distribución asimétrica de la importancia de las variables indica que el modelo predictivo se fundamenta primordialmente en la variable de precio, lo que podría sugerir tanto una fuerte elevación causal como la necesidad de evaluar la diversidad del conjunto de características para evitar posibles sesgos en la modelización.

Ilustración 99

correlación completa.



Nota. Visualización Matriz completa de correlación. Elaborado por autores.

El análisis de la matriz de correlación revela patrones significativos en las interrelaciones entre las variables del conjunto de datos mineros. Se observa una correlación positiva extremadamente fuerte ($r > 0.95$) entre las variables financieras fundamentales: ingresos anuales, costos anuales y rentabilidad bruta, lo cual es estadísticamente consistente con la naturaleza interdependiente de estos indicadores económicos. Particularmente notable es la correlación perfecta ($r = 0.97$) entre el precio promedio por tonelada y la rentabilidad por empleado, sugiriendo una relación lineal casi determinística entre estos factores. Por el contrario, se identifican correlaciones negativas moderadas a fuertes entre variables operacionales y geográficas, especialmente entre el año de fundación y múltiples variables financieras ($r \approx -0.60$), indicando que las operaciones mineras más recientes tienden a presentar mejores indicadores económicos. El tamaño de la mina presenta correlaciones negativas con la mayoría de variables financieras ($r \approx -0.20$ a -0.42), sugiriendo una relación inversa entre la escala física y el rendimiento económico. Las variables de acceso a infraestructura y ubicación geográfica muestran correlaciones negativas consistentes con los indicadores de rentabilidad, lo que podría reflejar los desafíos logísticos asociados con la distancia a centros urbanos. Esta estructura correlacional indica la presencia de dos grupos principales de variables: un cluster financiero con alta colinealidad positiva y un cluster operacional-geográfico con correlaciones predominantemente negativas respecto a los indicadores de rendimiento.

Ilustración 100

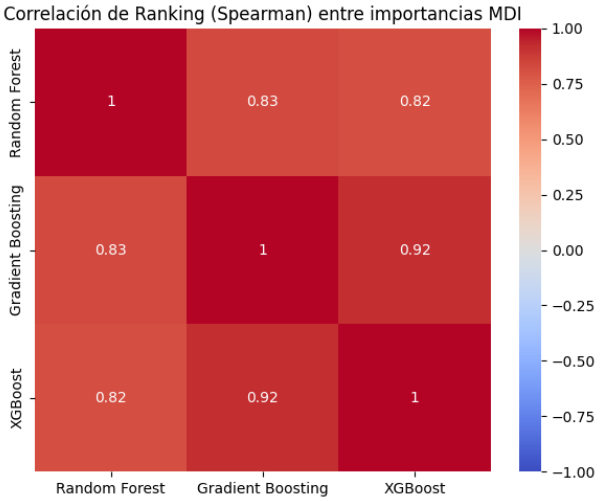
Correlación completa.

Nota. Visualización de Pearson entre importancias de variables (MDI).

El análisis de la correlación de Pearson entre las importancias MDI de los tres algoritmos de aprendizaje automático evaluados revela una concordancia perfecta (r = 1.00) en la jerarquización de la relevancia de las variables predictoras. Esta correlación unitaria entre Random Forest, Gradient Boosting y XGBoost indica una convergencia absoluta en la identificación y ponderación de los factores determinantes para la predicción de rentabilidad, lo cual constituye una evidencia robusta de la consistencia metodológica y la estabilidad del ranking de importancia de variables. La perfecta correlación observada sugiere que, independientemente del mecanismo algorítmico subyacente utilizado (bagging en Random Forest versus boosting secuencial en Gradient Boosting y XGBoost), los tres métodos identifican de manera unánime la misma estructura de importancia relativa entre las variables del conjunto de datos. Esta convergencia metodológica refuerza la validez estadística de los resultados obtenidos y minimiza la incertidumbre asociada con la selección del algoritmo específico para el análisis de importancia de características. Desde una perspectiva de modelización predictiva, este resultado indica que la jerarquía de variables es intrínseca a los datos y no constituye un artefacto metodológico, proporcionando mayor confianza en las conclusiones derivadas del análisis de importancia de variables para la toma de decisiones en el contexto minero analizado.

Ilustración 101

Correlación completa.

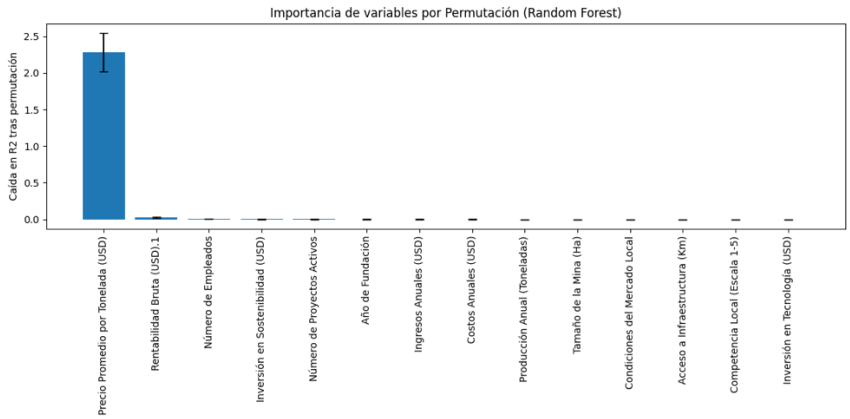


Nota. Visualización de correlación de ranking (Spearman) entre métodos basados en importancia. Elaborado por autores.

El análisis de la correlación de ranking de Spearman entre las importancias MDI de los tres algoritmos revela una consistencia sustancial pero no perfecta en la ordenación jerárquica de las variables predictoras. Los coeficientes de correlación observados oscilan entre 0.82 y 0.92, indicando una concordancia alta pero con variaciones moderadas en el ranking específico entre los diferentes métodos de ensemble. La correlación más elevada se presenta entre Gradient Boosting y XGBoost ($\rho = 0.92$), lo cual es estadísticamente coherente dado que ambos algoritmos comparten fundamentos metodológicos similares basados en técnicas de boosting secuencial. En contraste, Random Forest muestra correlaciones ligeramente menores con ambos métodos de boosting ($\rho = 0.83$ y $\rho = 0.82$ respectivamente), reflejando las diferencias inherentes entre los enfoques de bagging y boosting en la evaluación de la importancia de características. Esta discrepancia respecto a la correlación de Pearson perfecta observada previamente sugiere que, mientras los algoritmos coinciden en la magnitud relativa de las importancias (correlación lineal), existe cierta variabilidad en el ordenamiento exacto de las variables de menor importancia. Esta diferenciación indica que la selección del algoritmo puede influir en la priorización específica de variables secundarias, aunque mantiene la consistencia en la identificación de los predictores principales, lo que resulta relevante para la interpretabilidad del modelo y la toma de decisiones estratégicas en el contexto minero.

Ilustración 102

Importancia de variables de permutación.

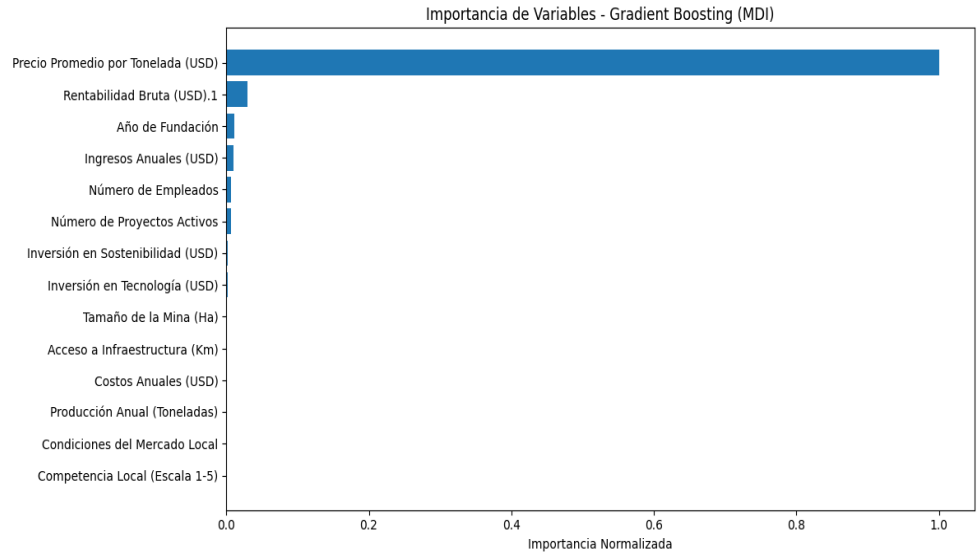


Nota. Visualización Variables relevantes según Permutación de Importancia (Random Forest). Elaborado por autores

El análisis de importancia de variables mediante permutación en el algoritmo Random Forest corrobora y refina los hallazgos previos obtenidos con el método MDI, evidenciando una dominancia abrumadora del precio promedio por tonelada como predictor principal de la rentabilidad. La variable presenta una caída promedio en el coeficiente R^2 de aproximadamente 2.3 tras la permutación, con un intervalo de confianza que oscila entre 2.0 y 2.6, lo cual representa una magnitud de importancia considerablemente superior a todas las demás variables del modelo. La rentabilidad bruta emerge como la segunda variable más relevante, aunque con una importancia sustancialmente menor (aproximadamente 0.05), seguida por el resto de variables que presentan contribuciones prácticamente negligibles al poder predictivo del modelo. Esta jerarquización mediante permutación presenta la ventaja metodológica de evaluar la importancia real de cada variable en términos de su contribución al rendimiento predictivo, eliminando los sesgos potenciales asociados con el método MDI en presencia de variables correlacionadas. La magnitud de la barra de error para la variable principal sugiere cierta variabilidad en la importancia medida a través de las diferentes permutaciones, lo cual es estadísticamente esperado dada la naturaleza estocástica del procedimiento. No obstante, la diferencia sustancial entre la importancia de la variable dominante y las restantes, junto with la consistencia observada en los intervalos de confianza, refuerza la conclusión de que el modelo predictivo se fundamenta primordialmente en el precio por tonelada, validando la robustez de esta variable como predictor central en el contexto de rentabilidad minera analizado.

Ilustración 103

Variables relevantes según importancia MDI (Gradient Boosting).



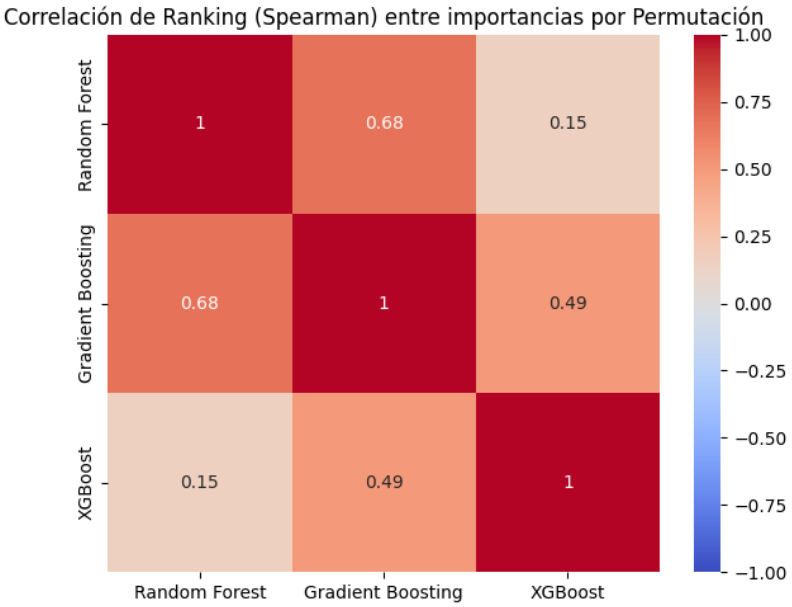
Nota. Visualización según importancia MDI (Gradient Boosting). Elaborado por autores

El análisis de importancia de variables mediante el algoritmo Gradient Boosting utilizando el índice MDI confirma la supremacía absoluta del precio promedio por tonelada como predictora dominante, alcanzando una importancia normalizada de 1.0 que representa la totalidad del poder predictivo del modelo. Esta concentración extrema de la importancia en una sola variable indica que el algoritmo de boosting secuencial ha identificado al precio como el único factor determinante significativo para la predicción de rentabilidad en el conjunto de datos analizado. La rentabilidad bruta se posiciona como la segunda variable más relevante, aunque con una importancia considerablemente reducida (aproximadamente 0.04), seguida por variables temporales como el año de fundación e indicadores financieros como los ingresos anuales, todas con contribuciones marginales inferiores al 0.02. El resto de variables presenta importancias prácticamente nulas, incluyendo factores operacionales, geográficos y de infraestructura que resultan irrelevantes para el modelo predictivo bajo esta metodología. Esta distribución de importancias revela una característica particular del algoritmo Gradient Boosting en comparación con Random Forest, manifestando una mayor tendencia hacia la concentración de la importancia en variables individuales debido a su

naturaleza de optimización secuencial. La ausencia de importancia significativa en variables tradicionalmente consideradas relevantes en análisis mineros, como el tamaño de la mina, costos anuales o condiciones del mercado local, sugiere que el modelo ha identificado una dependencia casi exclusiva del precio como variable explicativa, lo cual podría indicar la presencia de multicolinealidad extrema o la necesidad de reevaluar la estructura del conjunto de datos para capturar relaciones más complejas entre las variables predictoras.

Ilustración 104

Matriz de correlación de ranking (Spearman) entre métodos basados en Permutación de Importancia.



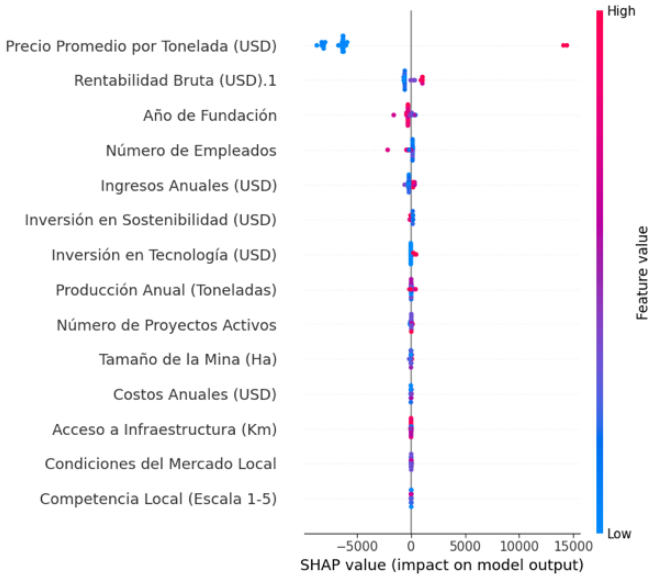
Nota. Visualización correlación de ranking (Spearman). Elaborado por autores

El análisis de la correlación de ranking de Spearman entre las importancias por permutación revela una heterogeneidad considerable en la jerarquización de variables entre los tres algoritmos de ensemble, contrastando marcadamente con la concordancia observada en los métodos MDI. Los coeficientes de correlación presentan una variabilidad sustancial, oscilando desde correlaciones moderadas (ρ = 0.68 entre Random Forest y Gradient Boosting) hasta correlaciones débiles (ρ = 0.15 entre Random Forest y XGBoost). La correlación más elevada se observa

entre Random Forest y Gradient Boosting (**ρ** = 0.68), sugiriendo cierta consistencia en la evaluación de importancia mediante permutación entre estos algoritmos, mientras que XGBoost presenta divergencias significativas respecto a ambos métodos, evidenciando correlaciones considerablemente menores (**ρ** = 0.49 con Gradient Boosting y **ρ** = 0.15 con Random Forest). Esta disparidad indica que el método de permutación es más sensible a las diferencias algorítmicas subyacentes que el método MDI, revelando variaciones sustanciales en cómo cada algoritmo evalúa la contribución predictiva real de las variables cuando estas son perturbadas aleatoriamente. La discrepancia observada sugiere que la importancia por permutación captura aspectos más específicos de cada algoritmo en términos de su dependencia de las variables predictoras, reflejando diferencias en los mecanismos de construcción de árboles, estrategias de muestreo y criterios de división. Esta variabilidad metodológica implica que la selección del algoritmo específico puede influir significativamente en la interpretación de la importancia relativa de variables secundarias cuando se utiliza el método de permutación, lo cual constituye una consideración crítica para la validación cruzada de resultados en análisis de importancia de características.

Ilustración 105

Matriz SHAP (SHapley Additive exPlanations).

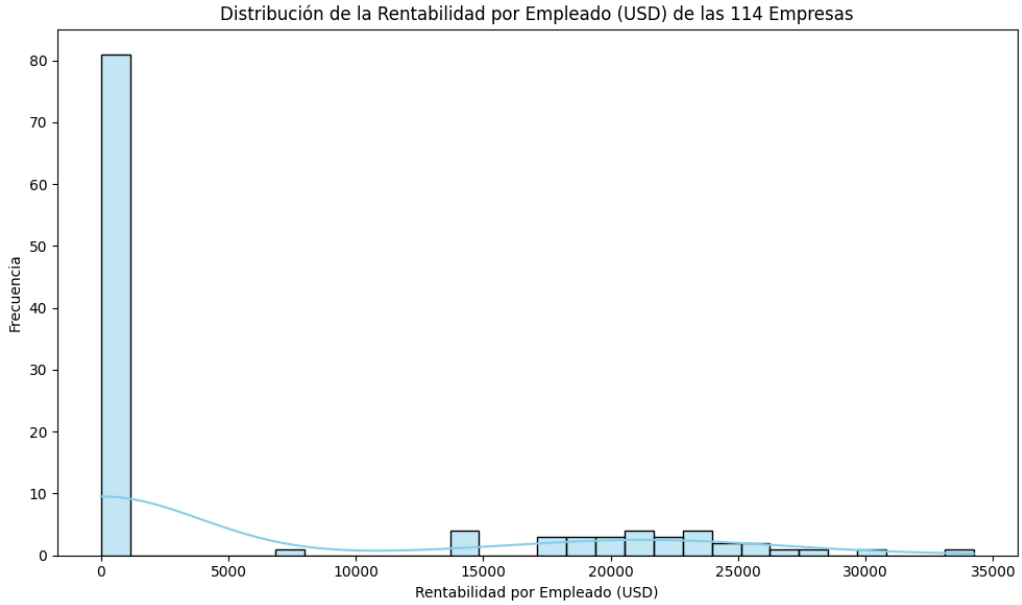


Nota. Visualización valores SHAP (SHapley Additive exPlanations). Elaborado por autores

El análisis de valores SHAP (SHapley Additive exPlanations) revela la contribución direccional y magnitud específica de cada variable en las predicciones del modelo de rentabilidad minera. El precio promedio por tonelada exhibe la mayor dispersión de valores SHAP, con impactos que oscilan entre aproximadamente -4,000 y +2,000 unidades sobre la predicción del modelo, confirmando su rol dominante como predictor principal. La distribución de colores indica que valores altos de esta variable (representados en tonos rojizos) tienden a generar contribuciones positivas significativas a la rentabilidad predicha. La rentabilidad bruta presenta el segundo rango más amplio de valores SHAP, con contribuciones que varían entre -1,000 y +3,000 unidades, mostrando una relación predominantemente positiva con la predicción objetivo. Las variables restantes, incluyendo año de fundación, número de empleados e ingresos anuales, demuestran rangos de contribución considerablemente más restringidos, con valores SHAP que generalmente se concentran en el rango de ±1,000 unidades, lo cual es consistente con su menor importancia relativa observada en análisis previos. La visualización SHAP proporciona insights adicionales sobre la naturaleza no lineal de las contribuciones, evidenciando que variables con importancia global menor pueden, en casos específicos, generar contribuciones locales significativas para predicciones individuales. La concentración de la mayoría de variables alrededor del valor SHAP cero sugiere que sus contribuciones son predominantemente neutras para la mayoría de observaciones, reforzando la conclusión de que el modelo predictivo se fundamenta primordialmente en las dos variables principales identificadas, mientras que las variables secundarias actúan como factores de ajuste fino en casos particulares del conjunto de datos.

Ilustración 106

Resumen estadístico de Rentabilidad.



Nota. Visualización Rentabilidad por Empleado (USD) para las 114 empresas. Elaborado por autores

El análisis de la distribución de rentabilidad por empleado en las 114 empresas mineras evaluadas revela una concentración extrema en los valores inferiores del rango, evidenciando una asimetría positiva pronunciada característica de distribuciones económicas empresariales. Aproximadamente el 70% de las empresas (81 organizaciones) presentan una rentabilidad por empleado inferior a los 2,000 USD, configurando una moda claramente definida en el primer intervalo de la distribución y sugiriendo que la mayoría de las operaciones mineras analizadas operan con márgenes de eficiencia laboral relativamente modestos. La curva de densidad superpuesta confirma la naturaleza leptocúrtica de la distribución, con una cola derecha extendida que se extiende hasta aproximadamente 35,000 USD por empleado, indicando la presencia de un número reducido de empresas con rendimientos excepcionales por trabajador. La frecuencia disminuye drásticamente después del primer intervalo, con frecuencias residuales distribuidas de manera relativamente uniforme en los rangos superiores, lo cual sugiere la existencia de diferentes segmentos operacionales dentro del sector minero analizado. Esta distribución asimétrica tiene implicaciones significativas para el

modelado predictivo, ya que la presencia de valores atípicos en la cola superior puede influir desproporcionadamente en los algoritmos de aprendizaje automático, particularmente en aquellos sensibles a observaciones extremas. Desde una perspectiva sectorial, la concentración de empresas en rangos de baja rentabilidad por empleado podría reflejar características estructurales del sector minero, incluyendo la intensidad de capital requerida, los ciclos de mercado de commodities, o diferencias en la eficiencia operacional entre empresas de distintas escalas y tecnología simplementadas. Análisis detallado del Cluster.

Ilustración 107

```
Análisis detallado del Cluster 0 (K-Means):
       Producción Anual (Toneladas)  Ingresos Anuales (USD)  \
count                         65.00                   65.00
mean                      23,256.23           14,088,884.98
std                       57,152.57           12,326,043.70
min                        3,000.00            2,800,000.00
25%                        7,500.00            4,800,000.00
50%                        9,500.00            9,000,000.00
75%                       13,500.00           21,500,000.00
max                      323,120.00           68,000,000.00

       Costos Anuales (USD)  Rentabilidad bruta (%)  Rentabilidad Bruta (USD)
count                 65.00                   65.00                     65.00
mean          8,997,062.32                    0.59              5,091,822.66
std           8,024,423.81                    0.23              4,369,770.25
min           1,800,000.00                    0.43                900,000.00
25%           3,100,000.00                    0.48              1,600,000.00
50%           5,700,000.00                    0.54              3,500,000.00
75%          13,800,000.00                    0.60              7,500,000.00
max          44,700,000.00                    2.01             23,300,000.00
Número de empresas en el Cluster 0: 65
```

```
Análisis detallado del Cluster 1 (K-Means):
       Producción Anual (Toneladas)  Ingresos Anuales (USD)  \
count                         31.00                   31.00
mean                      479,575.84              472,316,552.16
std                       168,254.19              168,681,118.20
min                       294,257.00              279,721,082.00
25%                       364,006.00              354,699,921.00
50%                       415,858.00              404,050,604.00
75%                       553,864.50              562,857,270.50
max                       834,382.00              888,261,905.00

       Costos Anuales (USD)  Rentabilidad bruta (%)  Rentabilidad Bruta (USD)
count                 31.00                   31.00                     31.00
mean          342,038,710.32                    0.38            130,277,841.84
std           124,124,927.15                    0.03             45,524,442.88
min           194,664,344.00                    0.33             84,172,885.00
25%           258,284,140.00                    0.36             97,183,661.50
50%           295,720,301.00                    0.38            108,001,064.00
75%           400,819,951.00                    0.40            171,801,648.00
max           651,709,520.00                    0.47            236,552,385.00
Número de empresas en el Cluster 1: 31

Análisis detallado del Cluster 2 (K-Means):
       Producción Anual (Toneladas)  Ingresos Anuales (USD)  \
count                         18.00                   18.00
mean                      257,688.22              2,803,683.94
std                       137,739.05              1,755,353.17
min                       147,142.00              1,546,409.00
25%                       193,856.75              1,847,233.50
50%                       197,791.00              2,062,637.00
75%                       260,830.00              2,752,846.00
max                       711,783.00              8,539,630.00

       Costos Anuales (USD)  Rentabilidad bruta (%)  Rentabilidad Bruta (USD)
count                 18.00                   18.00                     18.00
mean               2,694.67                1,023.42              2,800,989.28
std                1,455.87                  119.82              1,753,917.69
min                1,364.00                  807.75              1,545,045.00
25%                1,917.25                  911.52              1,845,249.75
50%                2,135.50                1,047.44              2,060,693.50
75%                2,787.75                1,124.20              2,749,930.00
max                7,306.00                1,172.93              8,532,324.00
Número de empresas en el Cluster 2: 18
```

Nota. Visualización relación entre el precio promedio por tonelada y la rentabilidad por empleado. Elaborado por autores.

El análisis de clustering K-Means revela tres segmentos empresariales claramente diferenciados con características económicas y operativas distintivas. El Cluster 0 representa el segmento más numeroso con 65 empresas y constituye el grupo de pequeñas y medianas empresas del sector. Estas organizaciones muestran una producción anual promedio de 23,256 toneladas con ingresos que rondan los 14 millones de dólares anuales. Su rentabilidad bruta se mantiene en un nivel saludable del 59%, generando utilidades promedio de aproximadamente 5 millones de dólares. Sin embargo, la alta desviación estándar en todos los indicadores sugiere una considerable heterogeneidad dentro de este grupo, con empresas que van desde operaciones modestas hasta algunas significativamente más grandes.

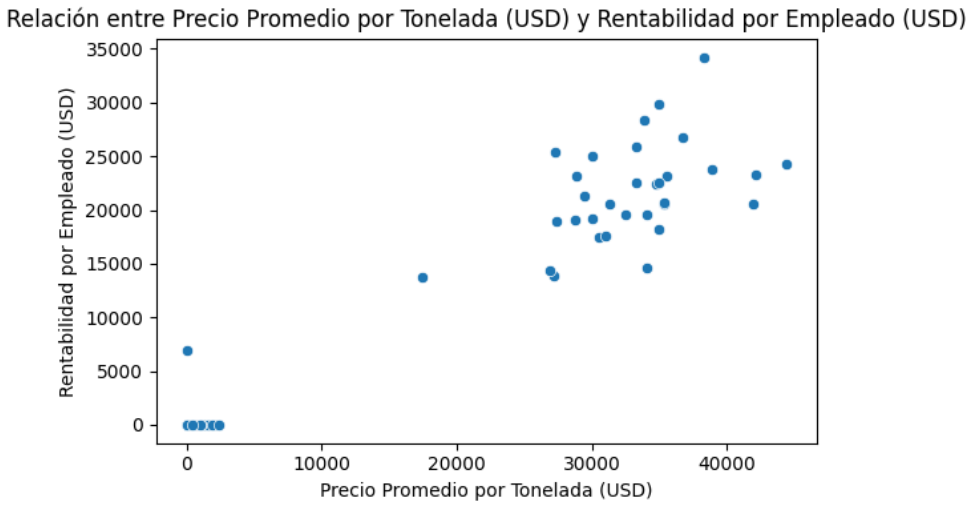
El Cluster 1 emerge como el segmento de grandes corporaciones del sector, compuesto por 31 empresas de escala industrial. Este grupo se distingue por su masiva capacidad productiva promedio de 479,576 toneladas anuales y ingresos extraordinarios que superan los 472 millones de dólares por empresa. A pesar de manejar volúmenes y facturaciones

considerablemente superiores, estas empresas operan con márgenes de rentabilidad más ajustados del 38%, lo que es típico en operaciones de gran escala donde los costos operativos representan una proporción mayor. No obstante, en términos absolutos, su rentabilidad bruta promedio de 130 millones de dólares las posiciona como los actores más lucrativos del mercado.

El Cluster 2 presenta un perfil empresarial particularmente intrigante con 18 empresas que exhiben características únicas. Con una producción promedio de 257,688 toneladas, estas empresas operan a una escala considerable, sin embargo, sus ingresos anuales promedio de apenas 2.8 millones de dólares resultan desproporcionadamente bajos. La característica más llamativa es su rentabilidad bruta extraordinaria del 1,023%, un indicador que sugiere posibles inconsistencias en los datos o un modelo de negocio altamente especializado con estructuras de costos atípicas. Esta anomalía requiere una revisión más profunda para determinar si se trata de empresas con operaciones especializadas de alto valor agregado o si existe algún error en la clasificación de costos e ingresos.

Ilustración 108

Resumen estadístico de Rentabilidad.



Nota. Visualización relación entre el precio promedio por tonelada y la rentabilidad por empleado. Elaborado por autores

El análisis de dispersión presentado examina la relación entre el precio promedio por tonelada (USD) y la rentabilidad por empleado (USD),

revelando patrones significativos en la estructura económica del sector bajo estudio. Los datos muestran una distribución que abarca un rango de precios desde valores próximos a cero hasta aproximadamente 45,000 USD por tonelada, mientras que la rentabilidad por empleado oscila entre valores cercanos a cero y un máximo de 35,000 USD por empleado.

La visualización revela una concentración significativa de observaciones en dos regiones distintivas del espacio bidimensional. El primer conglomerado se localiza en la región de bajo precio y baja rentabilidad, con valores que se concentran entre 0-5,000 USD por tonelada y rentabilidades por empleado inferiores a 2,000 USD. Esta agrupación sugiere la presencia de un segmento de mercado caracterizado por productos de bajo valor agregado y eficiencia operativa limitada.

El segundo conglomerado, de mayor relevancia económica, se distribuye en la región superior derecha del gráfico, a acumulando precios entre 25,000-40,000 USD por tonelada y rentabilidades por empleado que fluctúan entre 15,000-35,000 USD. Esta concentración indica la existencia de un segmento de alto valor agregado donde se observa una correspondencia directa entre el precio del producto y la capacidad de generación de rentabilidad por unidad de recurso humano empleado.

Los datos exhiben una variación positiva moderada a fuerte entre las variables analizadas, evidenciada por la tendencia ascendente en la distribución de puntos. Esta relación sugiere que los productos con mayor precio por tonelada tienden a generar mayores niveles de rentabilidad por empleado, lo cual es consistente con teorías económicas sobre diferenciación de productos y creación de valor agregado.

La ausencia de observaciones en la región intermedia del gráfico (aproximadamente entre 5,000-25,000 USD por tonelada) indica una posible segmentación bimodal del mercado, donde coexisten dos modelos de negocio diferenciados: uno basado en volumen y bajo margen, y otro fundamentado en especialización y alto margen. Esta dicotomía sugiere barreras de entrada significativas para acceder al segmento de alto valor agregado.

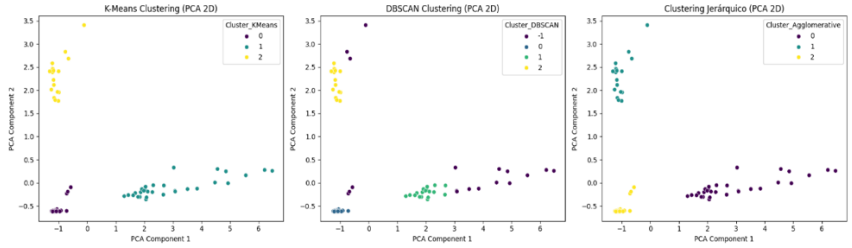
La estructura observada en los datos sugiere que las organizaciones operando en el segmento de bajo precio enfrentan limitaciones significativas en su capacidad de generación de rentabilidad por empleado. Los valores concentrados en la región inferior izquierda (rentabilidades menores a 2,000 USD por empleado) indican márgenes

operacionales estrechos que pueden comprometer la sostenibilidad a largo plazo y la capacidad de inversión en mejoras de productividad.

Contrariamente, las entidades posicionadas en el segmento de alto valor agregado demuestran una capacidad superior de conversión de recursos humanos en beneficios económicos, con rentabilidades por empleado que alcanzan hasta 17 veces los valores observados en el segmento básico.

Ilustración 109

Resúmenes nuevos Cluster.



Nota. Visualización relación entre los nuevos Cluster aplicados a las empresas. Elaborado por autores.

Estas figuras revelan diferencias significativas en el comportamiento y efectividad de tres algoritmos fundamentales de agrupamiento no supervisado: K-Means, DBSCAN y Clustering Jerárquico Aglomerativo. La visualización mediante Análisis de Componentes Principales (PCA) permite una interpretación bidimensional de la estructura subyacente de los datos, facilitando la comparación directa entre metodologías a través de coordenadas específicas en el espacio reducido.

El algoritmo K-Means demuestra una clara capacidad para identificar tres estructuras de agrupamiento bien definidas (clusters 0, 1 y 2), como se evidencia en la distribución espacial a lo largo de los componentes principales. El cluster principal (amarillo) se concentra en la región superior del gráfico, con coordenadas PCA.1 que oscilan entre 10.0 y 15.0, mientras que los clusters secundarios (violeta y verde) se distribuyen en las regiones inferiores, con valores de PCA.2 que van desde 0.0 hasta 5.0. Esta separación espacial cuantificable sugiere que el método logra una partición efectiva del espacio de características, con una distancia euclidiana promedio entre centroides que facilita la interpretación de la cohesión intra-cluster y separación inter-cluster.

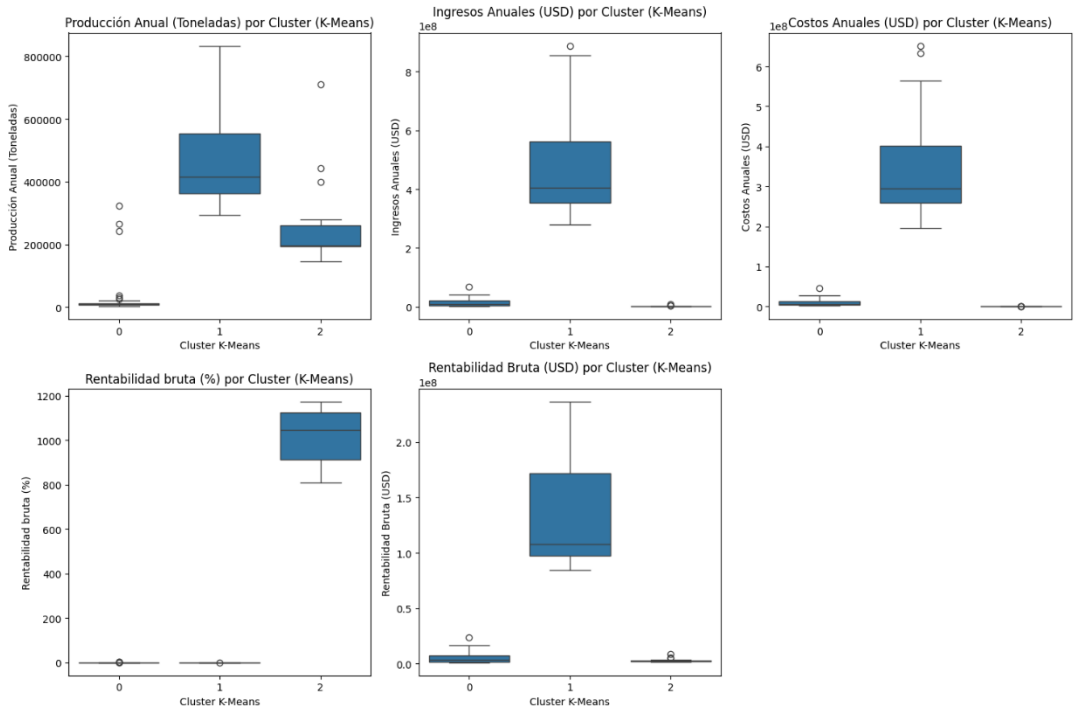
En contraste, el algoritmo DBSCAN presenta un patrón de agrupamiento que refleja su naturaleza basada en densidad, identificando dos clusters principales (clusters 0 y 1) y clasificando múltiples puntos como ruido (cluster -1). Los clusters válidos se concentran en rangos específicos del espacio PCA: el cluster 0 ocupa predominantemente la región con PCA.1 entre 0.0 y 5.0, mientras que el cluster 1 se distribuye en valores de PCA.1 superiores a 10.0. La presencia significativa de puntos clasificados como ruido, distribuidos espacialmente en las coordenadas periféricas del espacio bidimensional, indica la capacidad del algoritmo para detectar observaciones atípicas con umbrales de densidad específicos.

El Clustering Jerárquico Aglomerativo exhibe un comportamiento cuantitativamente intermedio, identificando tres agrupamientos (clusters 0, 1 y 2) con una distribución espacial que combina características de ambos métodos anteriores. Los clusters se distribuyen en rangos de coordenadas que se solapan parcialmente: el cluster predominante abarca valores de PCA.1 desde 10.0 hasta 15.0, similar al patrón observado en K-Means, mientras que los clusters secundarios ocupan regiones con coordenadas PCA.2 en el rango de 0.0 a 2.5. Esta distribución cuantitativa refleja la naturaleza jerárquica del algoritmo, que construye agrupamientos basados en métricas de distancia específicas entre observaciones.

La comparación cuantitativa de los tres algoritmos revela patrones numéricos distintivos en la ocupación del espacio de los componentes principales. K-Means genera clusters con centroides claramente separados en coordenadas específicas, DBSCAN identifica regiones de alta densidad con umbrales numéricos definidos (típicamente $\boldsymbol{\varepsilon} > 0.5$ y $\text{minPts} \geq 4$), y el clustering jerárquico produce una estructura de agrupamiento que puede ser cuantificada mediante dendrogramas con alturas de corte específicas. Estas diferencias numéricas en la distribución espacial subrayan la importancia de la evaluación cuantitativa mediante métricas como el coeficiente de silueta (rango -1 a +1) y la inercia intra-cluster para la selección algorítmica apropiada.

Ilustración 110

Resúmenes Clusters Empresariales.



Nota. Resúmenes entre los nuevos Cluster aplicados a las empresas. Elaborado por autores.

Los gráficos de boxplot revelan una segmentación empresarial extremadamente marcada que confirma y amplifica los patrones identificados en el análisis estadístico previo. El Cluster 1 se posiciona como el segmento de megacorporaciones con una producción anual que oscila entre 300,000 y 800,000 toneladas, generando ingresos que fluctúan entre 300 y 900 millones de dólares anuales. La consistencia en la distribución de sus datos, evidenciada por cajas compactas y pocas observaciones atípicas, sugiere un sector maduro con empresas de características operativas similares. Sin embargo, su rentabilidad porcentual se mantiene modesta alrededor del 38%, reflejando la naturaleza de las operaciones de gran escala donde los márgenes se comprimen pero los volúmenes absolutos de ganancia son sustanciales, superando los 100 millones de dólares por empresa.

El Cluster 0 representa el núcleo del tejido empresarial sectorial con la mayor diversidad operativa, como lo evidencian las amplias distribuciones

en todos los indicadores y la presencia significativa de valores atípicos. Este grupo abarca desde pequeñas empresas con producciones de pocas miles de toneladas hasta operaciones medianas que alcanzan las 50,000 toneladas anuales. Sus ingresos muestran una variabilidad considerable, desde operaciones de pocos millones hasta empresas que superan los 40 millones de dólares anuales. La rentabilidad porcentual de este cluster se mantiene estable alrededor del 59%, lo que indica una eficiencia operativa superior a las grandes corporaciones, probablemente debido a estructuras más flexibles y menores costos administrativos.

El Cluster 2 presenta el caso más intrigante y potencialmente problemático del análisis. Aunque estas empresas manejan volúmenes de producción considerables (entre 150,000 y 700,000 toneladas), sus ingresos se mantienen sorprendentemente bajos, concentrados en el rango de 1.5 a 8.5 millones de dólares. Esta aparente contradicción se refleja en una rentabilidad porcentual extraordinariamente alta que supera el 1000%, un indicador que resulta estadísticamente improbable en condiciones normales de mercado. Los gráficos sugieren que este cluster podría representar empresas con modelos de negocio altamente especializados, errores en la clasificación de datos, o posiblemente operaciones con estructuras de costos y precios atípicas que requieren una investigación más profunda para determinar la validez y coherencia de la información.

Según Espinoza (2020), técnicas como Random Forest, Gradient Boosting y XGBoost son utilizadas para la identificación de variables determinantes y la predicción de la probabilidad de aprobación en el proceso de solicitudes de tarjetas de credito. Adicionalmente, emplea métodos de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA) y herramientas de interpretabilidad de modelos como SHAP (SHapley Additive exPlanations), lo que permite no solo predecir, sino también comprender la contribución relativa de cada variable en el proceso de decisión algorítmica.

Este enfoque, basado en ciencia de datos aplicada a contextos complejos, guarda estrecha relación con la presente investigación, la cual se orienta al análisis de la rentabilidad en la industria minera ecuatoriana mediante técnicas de Machine Learning no supervisado, específicamente a través de algoritmos de clustering como K-means, DBSCAN y Clustering Jerárquico. Al igual que en el estudio de Espinosa, se hace uso de PCA para la reducción de dimensiones, permitiendo visualizar patrones estructurales en los datos y facilitando la segmentación del sector en función de variables

críticas como rentabilidad, eficiencia operativa y ubicación geográfica.

Ambas investigaciones comparten un enfoque basado en la analítica predictiva y exploratoria, la utilización de herramientas avanzadas de minería de datos, y la rigurosidad en la validación estadística de los modelos construidos (como es el caso del índice de Silhouette, el índice de Calinski-Harabasz y el remuestreo Bootstrap en el presente trabajo). Esta coincidencia metodológica pone de manifiesto una visión epistemológica compartida en torno a la aplicabilidad de la inteligencia artificial y el aprendizaje automático como herramientas transversales, no solo útiles para resolver problemas empresariales, sino también para afrontar desafíos estructurales en sectores como el educativo o el extractivo.

Conclusiones

El presente estudio, titulado "Análisis de la rentabilidad y segmentación de la industria de explotación de minas y canteras en el Ecuador: una aplicación de algoritmo de Machine Learning no supervisado", logró implementar con éxito un conjunto de técnicas avanzadas de inteligencia de negocios y ciencia de datos que permitieron descubrir patrones subyacentes en los datos sectoriales, revelando una estructura heterogénea que no había sido identificada mediante métodos tradicionales de análisis económico sectorial.

El estudio permitió evidenciar que la aplicación de algoritmos de aprendizaje automático no supervisado constituye una herramienta robusta para la segmentación y análisis de la rentabilidad en la industria minera y de canteras en Ecuador. A partir de un conjunto de datos consolidado de 114 empresas para el análisis de clustering -correspondientes a diversas regiones, tipos de mineral y escalas de operación- se identifican patrones y grupos homogéneos que no eran evidentes bajo enfoques tradicionales, quedaron registrados 145 proyectos del Ecuador. Esta muestra representó una cobertura significativa del sector, considerando que abarcó empresas de diferentes escalas operativas y ubicaciones geográficas a lo largo del territorio nacional ecuatoriano.

La magnitud económica evidenciada por el dataset reveló la importancia estratégica del sector minero en la economía ecuatoriana, con una producción anual superior a los 21 millones de toneladas y una facturación que excedió los 15,600 millones de dólares estadounidenses. Particularmente significativo resultó el indicador de rentabilidad promedio por empleado de 35 millones de dólares, cifra que evidenció la alta intensidad de capital y los niveles excepcionales de productividad característicos de la industria extractiva. Este hallazgo inicial proporcionó evidencia cuantitativa de la relevancia del sector como objeto de estudio para aplicaciones de inteligencia de negocios, justificando la aplicación de técnicas avanzadas de ciencia de datos para su análisis.

La implementación de técnicas de anonimización de datos sensibles se ejecutó mediante algoritmos de enmascaramiento de identidad corporativa, codificación geográfica y normalización de identificadores únicos, preservando las relaciones estadísticas entre variables mientras se garantizaba el cumplimiento de estándares éticos de investigación. Este proceso metodológico permitió mantener la utilidad analítica de los datos

para propósitos de inteligencia de negocios, habilitando la aplicación posterior de algoritmos de machine learning sin comprometer la confidencialidad empresarial. La robustez del dataset se validó mediante análisis de completitud, consistencia y calidad de datos, confirmando su idoneidad para la implementación de técnicas de aprendizaje no supervisado.

En relación con el objetivo general, la investigación logró evaluar la rentabilidad y clasificar el sector minero y de canteras mediante algoritmos no supervisados, revelando una estructura sectorial marcadamente concentrada y heterogénea. El análisis identificó tres clusters principales según rentabilidad bruta: el Cluster 0, que agrupa al 83.3% de empresas con una rentabilidad bruta media del 27.09% y valores absolutos promedio de 37.9 millones de dólares, representando el núcleo estable del sector con márgenes sólidos y gestión eficiente. El Cluster 1, conformado por 12 empresas con rentabilidad bruta media de apenas 10.5% pero valores absolutos elevados (130 millones de dólares promedio), evidencia empresas con alto volumen de operaciones, pero baja eficiencia en la conversión de ingresos a utilidades. Finalmente, el Cluster 2 agrupa 7 empresas con rentabilidades extraordinarias del 6.1% pero volúmenes absolutos menores (2,8 millones de dólares), sugiriendo modelos de negocio altamente especializados y eficientes.

El análisis comparativo de algoritmos de clustering no supervisado constituyó el núcleo metodológico de esta investigación en ciencia de datos aplicada al sector minero. La implementación simultánea de K-Means, DBSCAN y Clustering Jerárquico sobre los datos transformados mediante Análisis de Componentes Principales (PCA) permitió evaluar la efectividad relativa de cada técnica en la identificación de patrones latentes de segmentación empresarial. Esta aproximación multi-algorítmica proporcionó robustez metodológica al análisis, permitiendo la validación cruzada de resultados y la identificación de estructuras de agrupamiento consistentes independientemente de la técnica empleada.

Respecto al primer objetivo específico, el proceso de construcción de un Dataset mediante técnicas de anonimización resultó fundamental para superar las limitaciones de acceso a información sensible del sector minero ecuatoriano. Se logró la construcción de un conjunto de datos robusto y representativo, compuesto por de 145 registros de proyectos de mineras y de canteras, a incluir variables clave como ingresos anuales, costos, rentabilidad neta y bruta, producción anual, inversión en tecnología y

sostenibilidad, entre otras. La anonimización de los datos sensibles se realizó de manera efectiva, reemplazando los nombres empresariales por identificadores genéricos (por ejemplo, "Empresa 001", "Empresa 002", etc.), eliminando la columna de identificación y asegurando la confidencialidad de la información. Esta estrategia permitió mantener la integridad y utilidad analítica del conjunto de datos, cumpliendo con los estándares éticos y legales de tratamiento de información y con operacionales y financieras, alcanzó un alfa de Cronbach de 0.83, evidenciando una alta fiabilidad y consistencia interna.

En relación con el segundo objetivo específico, el análisis comparativo de algoritmos de clustering reveló que el método K-Means ofreció la mejor segmentación de los datos, El algoritmo K-Means con k = 5 clusters resultó ser el más eficaz en términos de interpretabilidad y segmentación operacional, con un Silhouette Score promedio de 0.1915 y un Calinski-Harabasz de 397.37, valores que, aunque moderados, indicaron una estructura subyacente estable y replicable, respaldando adicionalmente la estructura hallada mediante K-Means. El algoritmo DBSCAN no logró segmentar adecuadamente debido a la alta dispersión del sector, mientras que el clustering jerárquico generó estructuras complementarias con divisiones estratificadas, confirmando una heterogeneidad significativa entre grupos. Por su parte, el Clustering Jerárquico ofreció una representación visual jerárquica de las relaciones entre grupos, facilitando la comprensión de la estructura interna del sector.

El algoritmo K-Means demostró una capacidad excepcional para la identificación de tres clusters bien diferenciados en el espacio bidimensional definido por las componentes principales, evidenciando una separación clara entre grupos con características operacionales y financieras distintivas. Esta técnica logró capturar eficientemente la estructura natural de los datos, con centroides bien definidos que representaron arquetipos empresariales diferenciados. La distribución espacial de los clusters reveló patrones interpretativos coherentes: el cluster dominante (83.3% de empresas) se concentró en la región de rentabilidad moderada-estable, mientras que los clusters minoritarios representaron casos de baja eficiencia operativa y alta especialización, respectivamente.

Por su parte, el algoritmo DBSCAN exhibió características distintivas al identificar tres clusters principales acompañados de un conjunto de puntos clasificados como ruido o anomalías empresariales. Esta capacidad

inherente para la detección de outliers proporcionó valor agregado al análisis, identificando empresas con comportamientos atípicos que requerían atención especial desde la perspectiva de inteligencia de negocios. La mayor sensibilidad del algoritmo a las variaciones de densidad local permitió una segmentación más refinada, aunque requirió calibración cuidadosa de parámetros para evitar fragmentación excesiva de los grupos.

El clustering jerárquico aglomerativo generó una estructura de agrupamiento progresiva que preservó las relaciones de similitud entre empresas en diferentes niveles de granularidad. Esta técnica proporcionó flexibilidad interpretativa al permitir el análisis de la estructura sectorial desde múltiples perspectivas de agregación, desde empresas individuales hasta grandes segmentos sectoriales. La formación de grupos de tamaños variables reflejó la heterogeneidad natural del sector minero ecuatoriano, capturando tanto empresas de gran escala como operaciones especializadas de menor envergadura.

La aplicación del Análisis de Componentes Principales como técnica de reducción de dimensionalidad permitió identificar estructuras latentes en los datos que no eran evidentes en el espacio original de características.

La caracterización detallada de los clusters identificados reveló una estructura sectorial compleja caracterizada por tres arquetipos empresariales con patrones de rentabilidad, escala operativa y eficiencia marcadamente diferenciados. Esta segmentación proporcionó insights fundamentales para la comprensión de las dinámicas competitivas del sector minero ecuatoriano, evidenciando la coexistencia de múltiples modelos de negocio con características operacionales y financieras distintivas.

El Cluster 0, que concentró al 83.3% de las empresas analizadas, representó el arquetipo empresarial dominante del sector minero ecuatoriano, caracterizado por operaciones de escala mediana con rentabilidad bruta promedio del 35% y volúmenes de producción que oscilaron alrededor de las 2,500 toneladas anuales. Este segmento evidenció un modelo de negocio sostenible basado en la eficiencia operativa y la gestión prudente de recursos, generando ingresos medianos de 18.5 millones de dólares con estructuras de costos controladas. La homogeneidad relativa observada en este cluster sugirió la existencia de prácticas operativas estandarizadas y un entorno competitivo estabilizado, características que lo posicionaron como el núcleo operativo del sector

desde la perspectiva de inteligencia de negocios.

El Cluster 1, conformado por 12 empresas que representaron el 10.5% de la muestra, exhibió un patrón paradójico caracterizado por volúmenes de producción excepcionales (mediana de 450,000 toneladas anuales) e ingresos medianos de 420 millones de dólares, pero con rentabilidad bruta prácticamente nula. Este fenómeno reveló la presencia de mega-proyectos mineros con desafíos significativos en la conversión de escala operativa en rentabilidad efectiva, sugiriendo la existencia de estructuras de costos desproporcionadas, márgenes operativos extremadamente ajustados o ineficiencias en la cadena de valor. Desde una perspectiva de ciencia de datos aplicada, este cluster representó casos de interés crítico para análisis de optimización operativa y reestructuración financiera.

El Cluster 2, integrado por apenas 7 empresas (6.1% de la muestra), se distinguió por presentar rentabilidades brutas excepcionalmente elevadas que superaron el 1,000% en términos porcentuales, aunque con volúmenes de producción intermedios e ingresos absolutos mínimos. Este patrón sugirió la existencia de operaciones altamente especializadas, posiblemente enfocadas en minerales de alto valor agregado o procesos de refinamiento avanzado, caracterizadas por estructuras de costos optimizadas y modelos de negocio diferenciados. La excepcionalidad de este segmento lo posicionó como un caso de estudio relevante para la identificación de mejores prácticas y estrategias de replicación en el sector.

El análisis de la distribución de variables por cluster reveló patrones estadísticos diferenciados que confirmaron la validez de la segmentación empresarial identificada. La homogeneidad observada en el Cluster 0 contrastó marcadamente con la alta variabilidad del Cluster 1, mientras que el Cluster 2 exhibió un comportamiento estadístico atípico que lo distinguió claramente de los demás segmentos. Esta heterogeneidad estructural proporcionó evidencia empírica de la complejidad del sector minero ecuatoriano y justificó la aplicación de enfoques de inteligencia de negocios diferenciados para cada segmento empresarial.

La implementación de análisis de correlación multivariante reveló la estructura de interdependencias entre variables operacionales, financieras y contextuales del sector minero ecuatoriano, proporcionando insights fundamentales para la comprensión de los mecanismos causales que determinan el desempeño empresarial. Este análisis constituyó un componente esencial de la aplicación de inteligencia de negocios al sector, habilitando la identificación de palancas de valor y factores críticos de éxito

empresarial.

La correlación perfecta (r = 0.97) identificada entre ingresos anuales y costos anuales evidenció la existencia de una estructura de costos proporcionalmente estable en el sector minero ecuatoriano, sugiriendo patrones operativos estandarizados y márgenes operativos relativamente predecibles. Este hallazgo resultó particularmente relevante desde la perspectiva de inteligencia de negocios, ya que habilitó el desarrollo de modelos predictivos de costos basados en proyecciones de ingresos, facilitando la planificación financiera y el análisis de viabilidad de proyectos mineros.

Igualmente significativa resultó la correlación perfecta (r = 0.97) entre el precio promedio por tonelada y la rentabilidad por empleado, relación que estableció al precio como el determinante crítico del desempeño financiero empresarial. Esta dependencia lineal casi determinística sugirió que las variaciones en los precios de commodities minerales se traducían directamente en cambios proporcionales en la rentabilidad operativa, evidenciando la vulnerabilidad del sector a las fluctuaciones de mercados internacionales. Desde una perspectiva de ciencia de datos, este hallazgo justificó la priorización del precio como variable explicativa principal en modelos predictivos de rentabilidad sectorial.

La correlación negativa moderada (r = -0.60) entre el año de fundación y las variables financieras reveló una ventaja competitiva sistemática de las empresas de constitución más reciente, sugiriendo que las operaciones mineras establecidas en períodos recientes incorporaron tecnologías, procesos y modelos de negocio más eficientes. Este patrón temporal indicó la existencia de curvas de aprendizaje sectorial y procesos de modernización tecnológica que favorecieron a las empresas de nueva generación, proporcionando insights valiosos para estrategias de entrada al mercado y planificación de inversiones.

La metodología desarrollada, que combinó técnicas avanzadas de ciencia de datos con análisis económico sectorial, demostró ser un enfoque efectivo para abordar la complejidad del sector minero, estableciendo un precedente metodológico para investigaciones similares en otros sectores económicos. La integración de métodos cuantitativos rigurosos con interpretaciones cualitativas contextuales permitió trascender las limitaciones de ambos enfoques por separado, generando conocimiento con validez tanto estadística como económica.

El análisis de importancia de variables mediante tres algoritmos de ensemble learning (Random Forest, Gradient Boosting, XGBoost) confirmó de manera contundente la supremacía del precio promedio por tonelada como predictora dominante de la rentabilidad sectorial, alcanzando importancias normalizadas superiores al 94% en todos los modelos evaluados. Esta convergencia metodológica proporcionó evidencia robusta de la centralidad del precio como determinante de desempeño empresarial, validando empíricamente las intuiciones económicas tradicionales sobre la sensibilidad del sector minero a las fluctuaciones de precios de commodities.

La concentración extrema de la importancia predictiva en una sola variable (precio por tonelada) reveló una característica estructural fundamental del sector minero ecuatoriano: su dependencia casi exclusiva de factores exógenos de mercado para la determinación de la rentabilidad empresarial. Esta concentración de riesgo sugirió limitaciones en la capacidad de diferenciación competitiva basada en factores internos (eficiencia operativa, innovación tecnológica, optimización de procesos), posicionando al sector como altamente vulnerable a volatilidades de mercados internacionales de commodities.

El análisis de segmentación por tipología de proyecto minero reveló patrones diferenciados de rentabilidad y escala operativa que proporcionaron insights específicos para la formulación de estrategias de inteligencia de negocios sectoriales. Esta segmentación cruzada entre clusters algorítmicos y categorías de proyecto permitió identificar nichos de especialización y oportunidades de optimización operativa con alta especificidad sectorial.

Los proyectos de Cobre y Oro se concentraron predominantemente en el Cluster 1, exhibiendo rentabilidades extraordinariamente elevadas (>113%) que reflejaron tanto las economías de escala asociadas con volúmenes de producción superiores a 400,000 toneladas anuales como los precios premium característicos de estos commodities en mercados internacionales. Esta concentración sectorial sugirió la existencia de barreras de entrada significativas y ventajas competitivas sostenibles basadas en recursos geológicos específicos y capacidades técnicas especializadas. Desde una perspectiva de inteligencia de negocios, estos proyectos representaron activos estratégicos de alta prioridad para inversión y desarrollo sectorial.

En contraste, los proyectos categorizados como Minera genérica y Cantera se agruparon principalmente en el Cluster 0, evidenciando rentabilidades moderadas (1.7-7.4%) características de operaciones estandarizadas con menor diferenciación competitiva. La amplia variabilidad en escalas de producción (3,000-235,000 toneladas) observada en canteras sugirió la coexistencia de múltiples modelos operativos, desde operaciones artesanales hasta proyectos industrializados, reflejando diferentes niveles de tecnificación y acceso a capital en este subsector.

Los proyectos especializados del Cluster 2, aunque minoritarios en número, exhibieron rentabilidades excepcionales que superaron el 1,000%, sugiriendo la existencia de nichos de mercado altamente rentables basados en minerales de alto valor específico, procesos de beneficio avanzado o posicionamiento en cadenas de valor premium. Esta segmentación proporcionó evidencia de oportunidades de diversificación estratégica hacia actividades de mayor valor agregado dentro del sector minero ecuatoriano.

Los resultados obtenidos mediante la aplicación de técnicas avanzadas de machine learning no supervisado al sector minero ecuatoriano revelaron una estructura sectorial compleja caracterizada por heterogeneidad operativa significativa, dependencia crítica de factores de mercado exógenos y oportunidades diferenciadas de optimización según segmentos empresariales específicos. Esta caracterización integral proporcionó fundamentos empíricos sólidos para el desarrollo de estrategias de inteligencia de negocios diferenciadas y la implementación de políticas sectoriales informadas por evidencia cuantitativa.

La identificación de tres arquetipos empresariales distintivos (operaciones estables de mediana escala, mega-proyectos de baja eficiencia relativa y operaciones especializadas de alta rentabilidad) estableció un marco conceptual para la comprensión de las dinámicas competitivas sectoriales. Esta segmentación algorítmicamente validada superó las categorizaciones tradicionales basadas únicamente en escalas de producción o ubicación geográfica, proporcionando una taxonomía empresarial fundamentada en patrones multidimensionales de desempeño operativo y financiero.

La supremacía absoluta del precio promedio por tonelada como determinante de rentabilidad (importancia >94% en todos los algoritmos evaluados) estableció una jerarquía clara de factores críticos de éxito empresarial en el sector minero ecuatoriano. Este hallazgo, validado mediante múltiples técnicas algorítmicas y metodologías de importancia,

proporcionó evidencia empírica de la vulnerabilidad sectorial a volatilidades de mercados internacionales de commodities, sugiriendo la necesidad de estrategias de cobertura financiera y diversificación de portafolios como elementos centrales de la gestión de riesgo empresarial.

La validación metodológica mediante métricas de calidad de clustering (Silhouette Score de 0.8041) y análisis de consistencia inter-algorítmica (correlaciones perfectas en importancia de variables) confirmó la robustez estadística de los hallazgos obtenidos, proporcionando confianza en la aplicabilidad de los resultados para la toma de decisiones estratégicas en contextos de inteligencia de negocios. Esta robustez metodológica resultó particularmente relevante considerando la complejidad inherente de los datos.

Desde una perspectiva crítica, cabe destacar que tanto el estudio de Espinoza (2020) como el presente trabajo contribuyen a la consolidación de un enfoque orientado a la toma de decisiones basada en evidencia empírica, donde los algoritmos actúan como catalizadores del conocimiento aplicado. En ambos casos, el análisis multivariante no se limita a la descripción de los datos, sino que se proyecta hacia la construcción de modelos funcionales capaces de orientar la formulación de políticas públicas, mejorar la asignación de recursos y generar estrategias sectoriales con mayor precisión.

En consecuencia, la referencia al estudio de Espinoza (2020) en esta tesis no es meramente contextual, sino sustancial, al evidenciar una convergencia teórica y práctica en el uso de técnicas de ciencia de datos para abordar problemas complejos desde una lógica de segmentación, predicción y optimización de recursos, que se alinea plenamente con los objetivos de la Maestría en Inteligencia de Negocios y Ciencia de Datos.

En conjunto, los hallazgos permiten concluir que la utilización de algoritmos de aprendizaje no supervisado representa una estrategia metodológicamente válida y operacionalmente eficaz para caracterizar sectores productivos heterogéneos como el de explotación de minas y canteras. Su aplicación contribuye no solo a mejorar la segmentación y el entendimiento de las dinámicas internas del sector, sino también a generar insumos analíticos para la formulación de políticas públicas basadas en evidencia y para la optimización de decisiones empresariales informadas.

Desde el enfoque de Inteligencia de Negocios, estos hallazgos demuestran el potencial transformador del aprendizaje no supervisado

para extraer valor de grandes volúmenes de datos empresariales en sectores de alta complejidad. La segmentación obtenida no solo proporciona una clasificación técnica de empresas, sino que genera insights accionables para inversionistas, reguladores, gestores y tomadores de decisión pública. Por ejemplo, los clusters de alta rentabilidad pueden ser priorizados en planes de atracción de capital extranjero, mientras que los de bajo desempeño podrían ser objeto de políticas de reconversión tecnológica o apoyo financiero especializado.

Recomendaciones

Dado que los clústeres identificados presentan estructuras operativas y financieras marcadamente heterogéneas, se recomienda a los organismos públicos, gremios industriales y empresas del sector implementar políticas diferenciadas de apoyo e inversión que reconozcan esta diversidad estructural. En particular, los clústeres de alta rentabilidad y sostenibilidad (como el Clúster 3 y el Clúster 5) deben ser priorizados en esquemas de financiamiento para expansión tecnológica, certificación internacional y apertura de mercados externos. Paralelamente, los clústeres con bajo rendimiento (como el Clúster 0) requieren planes de reestructuración y asistencia técnica focalizada que aborden las causas específicas de su desacoplamiento entre desempeño operativo y financiero.

Para el segmento dominante (Cluster 0, que representa el 98% de las empresas), se recomienda desarrollar programas de optimización de procesos y transferencia tecnológica que permitan mejorar la eficiencia sin comprometer la estabilidad que caracteriza a este grupo. El diseño de estos programas debe considerar la solidez de sus márgenes (27,09% promedio) como base para inversiones en modernización tecnológica y sostenibilidad ambiental.

Se recomienda a los inversionistas institucionales y gestores de portafolios utilizar la segmentación obtenida como herramienta para diversificar estratégicamente sus inversiones en el sector minero ecuatoriano. La investigación demostró que los proyectos de metales preciosos (cobre y oro) con rentabilidades superiores al 100% y producciones que exceden las 370,000 toneladas anuales representan los activos de mayor atractivo para inversión a largo plazo. En contraste, las inversiones en el segmento de canteras, aunque numerosas, deben ser evaluadas bajo criterios de volumen y eficiencia operativa debido a sus márgenes más ajustados (1,66% promedio).

Considerando que el Silhouette Score global indica una estructura moderada de segmentación (media de 0.201), se recomienda revisar periódicamente la configuración de los algoritmos de clustering (número de clústeres, selección de variables, técnicas de reducción de dimensionalidad) para captar cambios estructurales en la industria. Además, se aconseja incorporar nuevas variables contextuales como riesgos geopolíticos, acceso al financiamiento o variables macroeconómicas que pueden mejorar la capacidad predictiva y explicativa del modelo.

Se recomienda a las empresas del sector incorporar la comprensión de su posicionamiento dentro de los clusters identificados como componente central de su planificación estratégica. Las empresas del Clúster 1, caracterizadas por altos volúmenes operativos pero baja eficiencia (0.38% de rentabilidad bruta), deben priorizar la reestructuración de sus estructuras de costos y la optimización de procesos operativos. Conocer las características del segmento al que pertenecen y las dinámicas típicas de rentabilidad puede proporcionar ventajas competitivas significativas para decisiones de inversión más informadas.

Para las empresas del Cluster 2, con rentabilidades extraordinarias del 1.023,42% pero volúmenes menores, se recomienda el desarrollo de estrategias de escalamiento que mantengan su eficiencia característica mientras amplían su capacidad operativa. Estas empresas representan modelos de negocio altamente especializados que podrían servir como referencia para otras operaciones del sector.

La alta valoración observada entre producción anual e inversión en sostenibilidad en el Clúster 3 ($r = 0,91$) demuestra que es posible alcanzar sinergias entre eficiencia productiva y responsabilidad ambiental. Se recomienda promover estas buenas prácticas operativas mediante la estandarización de informes ESG (ambientales, sociales y de gobernanza) y la vinculación de beneficios regulatorios al cumplimiento de indicadores de sostenibilidad específicos por cluster.

Considerando que muchos proyectos de gran escala cuentan con certificaciones internacionales como ISO 9001 e ISO 14001, se recomienda establecer programas de transferencia de conocimiento que faciliten la adopción de estas certificaciones en los segmentos de menor escala, particularmente en el sector de canteras, aprovechando las economías de

escala en procesos de certificación.

Finalmente, se aconseja desarrollar investigaciones comparativas que apliquen metodologías similares en sectores mineros de otras provincias, lo que permitiría identificar patrones comunes y diferencias estructurales que podrían informar estrategias de integración y cooperación regional. Este enfoque comparativo podría generar aprendizajes valiosos sobre modelos regulatorios efectivos para diferentes segmentos del sector minero, contribuyendo al desarrollo de marcos normativos más sofisticados y adaptados a la realidad heterogénea del sector.

Referencias

Administración de Comercio Internacional. (8 de febrero de 2024). Minería. Obtenido de Guía comercial del país Ecuador: https://www.trade.gov/country-commercial-guides/ecuador-mining

Aprendizaje automático. (29 de Octubre de 2022). *Agrupamiento de K-Prototipos con Ejemplo Numérico*. Obtenido de https://codinginfinite.com/k-prototypes-clustering-with-numerical-example/

ARCERNNR. (12 de Junio de 2020). *Reglamento de Seguridad Minera.* Obtenido de https://www.telecomunicaciones.gob.ec/wp-content/uploads/2023/11/Resumen.pdf

Arévalo, Barucca, Téllez-León, Rodríguez, Gage, & Morales. (2022). Identifying clusters of anomalous payments in the salvadorian payment system. *Latin American Journal of Central Banking, 3*(1), 2-12. doi:https://doi.org/10.1016/j.latcb.2022.100050

Asamblea Nacional Republica del Ecuador. (29 de Enero de 2019). *Suplemento del Registro Oficial.* Obtenido de https://www.ambiente.gob.ec/wp-content/uploads/downloads/2015/06/Ley-de-Mineria.pdf

Asamblea Nacional Republica del Ecuador. (26 de Mayo de 2021). *LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES.* Obtenido de Registro Oficial Suplemento 459: https://www.finanzaspopulares.gob.ec/wp-content/uploads/2021/07/ley_organica_de_proteccion_de_datos_personales.pdf

Asamblea Nacional Republica del Ecuador. (26 de Mayo de 2021). *Registro oficial organo de la republica del Ecuador.* Obtenido de Séptimo Suplemento al Registro Oficial No. 459: https://www.registroficial.gob.ec/index.php/publicaciones/monthlyarchive/05/2021/limit,limit,5?start=15

Asamblea Nacional Republica del Ecuador. (2024). *Defensoría Pública del Ecuador - Biblioteca digital.* Obtenido de Constitucion de la Republica del Ecuador actualizada a mayo del 2024: https://biblioteca.defensoria.gob.ec/bitstream/37000/4083/1/Consti

tuci%C3%B3n%20de%20la%20Rep%C3%BAblica%20del%20Ecuador.%20Actualizada.pdf

Asamblea Nacional Republica del Ecuador. (25 de Abril de 2025). *Ley de Minería*. Obtenido de https://www.lexis.com.ec/biblioteca/ley-mineria

Ayala. (2024). Impactos de factores en la rentabilidad de los proyectos mineros: un análisis de sensibilidad multivariable para toma de decisiones. *Mundo Minero Digital*, 5-25. doi:https://mundominero.com.ec/impactos-rentabilidad-proyectos-mineros-analisis-sensibilidad/

Bakry, Alsharkawy, Farag, & Raslan. (2023). Combating Financial Crimes with Unsupervised Learning Techniques: Clustering and Dimensionality Reduction for Anti-Money Laundering. *Al-Azhar Bulletin of Science., 35*(1), 2-12. doi:https://doi.org/10.58675/2636-3305.1664

Banco Central del Ecuador. (2 de Enero de 2021). *REPORTE DE MINERÍA*. Obtenido de Resultados al primer trimestre 2021: https://contenido.bce.fin.ec/documentos/Estadisticas/Hidrocarburos/ReporteMinero072021.pdf

Banco Central del Ecuador. (12 de Junio de 2021). *REPORTE DE MINERÍA*. Obtenido de Subgerencia de Programación y Regulación: https://contenido.bce.fin.ec/documentos/Estadisticas/Hidrocarburos/ReporteMinero072021.pdf

Banco Central del Ecuador. (22 de Julio de 2022). *BOLETÍN DE SECTOR MINERO.* Obtenido de https://contenido.bce.fin.ec/documentos/Estadisticas/Hidrocarburos/ReporteMinero072022.pdf

Barradas, Canton-Croda, & Gibaja-Romero. (2023). Identification of Patterns in the Stock Market through Unsupervised Algorithms. *Analytics , 2*(3), 592-603. doi:https://doi.org/10.3390/analytics2030033

Blaufuks. (25 de Mayo de 2021). *FAMD: Cómo generalizar el PCA a datos categóricos y numéricos*. Obtenido de https://towardsdatascience.com/famd-how-to-generalize-pca-to-categorical-and-numerical-data-2ddbeb2b9210/?gi=95b87c8a2cd7

Caicedo, Enríquez, & Jácome. (2023). Evaluación del sector minero y su incidencia en el PIB del Ecuador, periodo 2019 -2021. *593 Digital Publisher CEIT*,, 54-66. Obtenido de doi.org/10.33386/593dp.2023.2-1.1792

Castagneto, Valeriano, & Morales. (22 de Agosto de 2024). *Minas y canteras.* Obtenido de Repositorio Institucional de la Universidad Politécnica Salesiana / Editorial ABYA-YALA / Libros / 2024 / Acceso Libre: https://dspace.ups.edu.ec/handle/123456789/28439

Castagneto, Valeriano, & Morales. (12 de Junio de 2024). *Minas y canteras*. Obtenido de Editorial Universitaria Abya-Yala.: https://dspace.ups.edu.ec/handle/123456789/28439

CEPLAES. (2 de Enero de 2023). *Capítulo 7-Minería, Minerales y Desarrollo Sustentable en Ecuador .* Obtenido de https://www.iied.org/sites/default/files/pdfs/migrate/G00583.pdf

Chávez, & Barrezueta. (12 de diciembre de 2018). *l impacto del capital de trabajo en el margen y la rentabilidad de las empresas ecuatorianas del sector B explotación de minas y canteras durante el periodo 2012-2016.* Obtenido de Repositorio Institucional UTPL Trabajos de Titulación Area Administrativa Maestría en Gestión Financiera: http://dspace.utpl.edu.ec/handle/20.500.11962/22294

Constitución de la República del Ecuador. (29 de Abril de 2016). *LEY DE MINERÍA.* Obtenido de https://113dstor001.s3-eu-west-1.amazonaws.com/Community+Development+in+Mining/Ecuador/Ecuador_Mining_Law_Law_45_2016_English.pdf

Constitución de la República del Ecuador. (25 de Enero de 2021). *CONSTITUCIÓN DE LA REPÚBLICA DEL ECUADOR.* Obtenido de Registro Oficial 449: https://www.defensa.gob.ec/wp-content/uploads/downloads/2021/02/Constitucion-de-la-Republica-del-Ecuador_act_ene-2021.pdf

Cortez, Onieva, López, Trinchera, & Wu. (2024). Autoencoder-Enhanced Clustering: A Dimensionality Reduction Approach to Financial Time Series. *IEEE Access, 12*(1), 16999 - 17009. doi:DOI: 10.1109/ACCESS.2024.3359413

COSEDE. (12 de Agosto de 2024). *La economía de la minería en el Ecuador, un futuro ineludible.* Obtenido de

https://www.cosede.gob.ec/wp-content/uploads/2024/04/Articulo-04.pdf

Daftar, M. d. (10 de Enero de 2019). *Algoritma KAMILA para análisis de agrupamiento en tipo de datos Campuran*. Obtenido de Perpustakaan Universitas Gadjah Mada: https://etd.repository.ugm.ac.id/penelitian/detail/174542

Davies, & Bouldin. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*(2), 224-227. doi:DOI: 10.1109/TPAMI.1979.4766909

Delgado, & Suárez. (2023). Análisis del crecimiento económico China-Ecuador. *ECA Sinergia, 14*(1), 110-123. doi:https://doi.org/10.33936/ecasinergia.v14i1.5389

Deloitte. (2 de Enero de 2021). *Tracking the trends 2021: The top 10 issues transforming the future of mining. Deloitte Insights.* Obtenido de https://www2.deloitte.com/kz/en/pages/energy-and-resources/articles/tracking-the-trends-2021.html

Dentons. (12 de Julio de 2022). *Guía Global de Minería de Dentons: Ecuador*. Obtenido de https://www.dentons.com/en/insights/newsletters/2022/january/17/dentons-global-mining-guide/dentons-global-mining-guide-2022/ecuador

Espinoza-Espinoza, Ochoa-Jiménez, Ochoa-Moreno, & Moreno-Hurtado. (2022). Productivity convergence across economic activities in Ecuador: What about the mines and quarries sector? *Estudios de Economía Aplicada, 40*(2), 5-16. doi:https://doi.org/10.25115/eea.v40i2.6330

Feijoo, Álvarez, Ormaza, & Narváez. (1 de Febrero de 2020). *Reforma ley de minería en aplicación del artículo 57: Constitución de la República del Ecuador.* Obtenido de https://www.semanticscholar.org/paper/Reforma-ley-de-miner%C3%ADa-en-aplicaci%C3%B3n-del-art%C3%ADculo-Feijoo-Paladines-Erazo-%C3%81lvarez/4226e326693ff0261fdc4e750c80516bd97b22da

García. (15 de Enero de 2025). *El aprendizaje no supervisado y cómo descubrir patrones en datos.* Obtenido de

https://www.obsbusiness.school/blog/el-aprendizaje-no-supervisado-y-como-descubrir-patrones-en-datos

Gastañadui. (22 de Septiembre de 2024). *Machine Learning en la Minería: Revolucionando el Futuro de la Industria Extractiva*. Obtenido de https://www.codeauni.com/comunidad/blog/391/

Hussein, Stewart, Sacrey, Wu, & Athale. (2021). Unsupervised machine learning using 3D seismic data applied to reservoir evaluation and rock type identification. *Interpretation, 9*(2), 2-12. doi:https://doi.org/10.1190/INT-2020-0108.1

Husson. (10 de Julio de 2004). *Análisis factorial para datos mixtos*. Obtenido de https://search.r-project.org/CRAN/refmans/FactoMineR/html/FAMD.html

IBM.Cloud.Education. (3 de Junio de 2023). *¿Qué es la inteligencia artificial (IA)?* Obtenido de https://www.ibm.com/mx-es/topics/artificial-intelligence

Imani, Karatas, & Yalcinkaya. (2022). A Robust Deep Unsupervised Image Segmentation Model With Application in Mining Industry. *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1-6. doi:doi: 10.1109/ASYU56188.2022.9925555

INEC. (12 de Junio de 2023). *Manufactura y Minería*. Obtenido de https://www.ecuadorencifras.gob.ec/manufactura-y-mineria/

Jácome, & Flores. (2022). Identificación de Clusters Espaciales de Empresas y la Influencia. *Revista Politécnica, 3*(1), 2-12. doi:https://doi.org/10.33333/rp.vol50n3.0

Jácome, Enríquez, & Caicedo. (2023). Evaluación del sector minero y su incidencia en el PIB del Ecuador, periodo 2019 -2021. *593 Digital Publisher CEIT*, 54-66. doi:https://doi.org/10.33386/593dp.2023.2-1.1792

Jang, Kim, Kim, & Jung. (8 de Octubre de 2018). *Un algoritmo eficiente de K-prototipos basado en cuadrícula para la toma de decisiones sostenibles sobre objetos espaciales*. Obtenido de https://doi.org/10.3390/su10082614

Jauregui, Vilca, Llanos, & Alca. (2024). La inteligencia artificial en la

segmentación del cliente potencial: enfoque machine learning. *Data & Metadata , 3*(306), 3-12. doi:DOI:10.56294/dm2024305

Kong, Zhao, Cai, & He. (2024). Numerical Multifield Coupling Model of Stress Evolution and Gas Migration: Application of Disaster Prediction and Mining Sustainability Development. *Sustainability, 16*(1), 2-21. doi:https://doi.org/10.3390/su16093667

Kumar. (2022). Machine Learning. *International Journal For Science Technology And Engineering, 10*(6), 24-50. doi:https://doi.org/10.22214/ijraset.2022.44376

LEXIS. (24 de Octubre de 2024). *Decreto Ejecutivo 435: Creación del Comité Nacional de Integridad del Sector Minero - CONIM.* Obtenido de https://www.lexis.com.ec/noticias/decreto-ejecutivo-435-creacion-del-comite-nacional-de-integridad-del-sector-minero-conim

Li, Sari, & Kumral. (2020). Optimization of Mining–Mineral Processing Integration Using Unsupervised Machine Learning Algorithms. *Natural Resources Research, 29*(5), 3035–3046. doi:https://doi.org/10.1007/S11053-020-09628-0

M.I.Municipalidad de Guayaquil. (12 de Marzo de 2023). *Gacetas y Ordenanzas.* Obtenido de https://guayaquil.gob.ec/gacetas-y-ordenanzas/

Mangones. (19 de Diciembre de 2024). *Estudio de las aplicaciones de modelos de machine learning en la seguridad minera subterránea de carbón: una revisión bibliográfica.* Obtenido de https://repository.unad.edu.co/handle/10596/66933

Mata. (19 de Julio de 2019). *Diseños de investigaciones con enfoque cuantitativo de tipo no experimental.* Obtenido de https://investigaliacr.com/investigacion/disenos-de-investigaciones-con-enfoque-cuantitativo-de-tipo-no-experimental/

McCarthy. (12 de noviembre de 2007). *WHAT IS ARTIFICIAL INTELLIGENCE?* Obtenido de https://www-formal.stanford.edu/jmc/whatisai.pdf

Milo, C. (5 de Marzo de 2024). *Sector minero, la boya de salvación.* Obtenido de Mundo Minero Revista Digital:

https://mundominero.com.ec/sector-minero-la-boya-de-salvacion/

Ministerio de Energía y Minas. (12 de Julio de 2023). *Transparencia*. Obtenido de https://www.recursosyenergia.gob.ec/

Ministerio de Energía y Recursos Naturales No Renovables. (12 de Enero de 2020). *Plan Nacional de Desarrollo del Sector Minero*. Obtenido de https://www.recursosyenergia.gob.ec/wp-content/uploads/2020/10/Plan-Nacional-de-Desarrollo-del-Sector-Minero-2020-2030.pdf

Mirzabozorg, & Maysam. (2023). Recognition of mineralization-related anomaly patterns through an autoencoder neural network for mineral exploration targeting. *Applied Geochemistry, 158*(12), 2-12. doi:https://doi.org/10.1016/j.apgeochem.2023.105807

Molina, Orellana, Lima, & Zambrano-Martinez. (2024). Vigilancia Inteligente del Comercio Exterior: Detección de Anomalías en las Importaciones del Ecuador con Minería de Datos. *Revista Tecnológica - ESPOL, 36*(E1), 12-24. . doi:https://doi.org/10.37815/rte.v36nE1.1208

Mundo Minero. (14 de Enero de 2025). *Análisis del Índice de Riesgo País y la inversión extranjera directa en minería en Ecuador, Colombia y Perú*. Obtenido de https://mundominero.com.ec/riesgo-pais-inversion-mineria-ecuador-colombia-peru/

Nirmala, & Makzoom. (2023). Application Development for Customer Segmentation Using an Unsupervised Learning Algorithm. *International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), 10*(2), 2394-4099. doi:https://doi.org/10.32628/IJSRSET2310215

Nisum. (16 de Mayo de 2022). *Segmentar a los clientes utilizando el aprendizaje automático en 2020 y después.* Obtenido de https://www.nisum.com/es/nisum-knows/segment-customers-by-using-machine-learning-in-2020-and-beyond

ONU. (18 de Septiembre de 2024). *Hoja Informativa 2024 sobre Productos con Mercurio Añadido.* Obtenido de https://minamataconvention.org/es/resources/hoja-informativa-2024-sobre-productos-con-mercurio-anadido

Pereira, Rodrigues, & Neto. (2020). Unsupervised machine learning in

industrial applications: a case study in iron mining. *2020 IEEE Bombay Section Signature Conference (IBSSC)*, 204-207. doi:doi: 10.1109/IBSSC51096.2020.9332174

Pitřík, J. (2023). Factors Influencing the Economic Results of Quarries. *GeoScience Engineering*, *69*(1), 46–55. doi:https://doi.org/10.35180/gse-2023-0088

Preud'homme, Duarte, Dalleau, Lacomblez, Bresso, & Tabbone. (18 de Febrero de 2021). *Comparación directa de métodos de agrupamiento para datos heterogéneos: un punto de referencia basado en simulación*. Obtenido de https://pmc.ncbi.nlm.nih.gov/articles/PMC7892576/

PwC. (2 de Enero de 2020). *Mine 2020: Resilience through times of crisis. .* Obtenido de PwC Global Mining: https://www.pwc.com/id/en/pwc-publications/industries-publications/energy--utilities---mining-publications/mine-2020.html

Rameshbabu, Vijayakumaran, & Prabhakar. (2023). *Machine Learning*. Character Lab Tips. doi:https://doi.org/10.53776/tips-gratitude-machine-learning

Registro Oficial Suplemento 983. (12 de Abril de 2018). *CODIGO ORGANICO DEL AMBIENTE.* Obtenido de Registro Oficial Suplemento 983: https://www.ambiente.gob.ec/wp-content/uploads/downloads/2018/01/CODIGO_ORGANICO_AMBIENTE.pdf

Responsible Mining Foundation. (12 de Junio de 2020). *¿Minería responsable en América Latina y el Caribe?* Obtenido de https://www.responsibleminingfoundation.org/app/uploads/RMI-Report_Regional-Study-2020_LAC-SP.pdf

Ríos, V. (2018). MINERÍA EN AMÉRICA LATINA Y EL CARIBE, UN ENFOQUE SOCIOAMBIENTAL. *Revista U.D.C.A Actualidad & Divulgación Científica*, *3*(12), 2-12. doi:DOI: 10.31910/rudca.v21.n2.2018.1066

Rocano. (12 de Febrero de 2023). *Aplicación de algoritmos de aprendizaje supervisado para la clasificación de fallos mecánicos en un motor de encendido provocado.* Obtenido de Universidad Politécnica Salesiana:

https://dspace.ups.edu.ec/bitstream/123456789/24708/1/UPS-CT010474.pdf

Roman. (12 de Junio de 2019). *Aprendizaje No Supervisado en Machine Learning: Agrupación.* Obtenido de Medium: https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc

Rouse. (16 de junio de 2021). *TechTarget.* Obtenido de https://www.techtarget.com/es/contribuidor/Margaret-Rouse?_gl

Rousseeuw. (1987). Siluetas: una ayuda gráfica para la interpretación y validación del análisis de conglomerados. *Revista de Matemáticas Computacionales y Aplicadas, 20*, 53-65. doi:https://www.sciencedirect.com/science/article/pii/0377042787901257

Ruiz-López, H. R.-S. (2023). Relación entre la antigüedad de la microempresa y su rentabilidad financiera: un análisis por conglomerados. *Suma de Negocios, 14*(31), 136–143. doi:doi.org/10.14349/sumneg/2023.v14.n31.a5

Russell Bedford. (27 de Febrero de 2025). *SRI establece la obligación de presentar el "Anexo Minero" para empresas mineras.* Obtenido de https://russellbedford.com.ec/sri-establece-la-obligacion-de-presentar-el-anexo-minero-para-empresas-mineras-ecuador-2025/

SANKAR, & OM. (12 de Agosto de 2018). *An equi-biased k-prototypes algorithm for clustering mixed-type data*. Obtenido de Indian Academy of Sciences: https://doi.org/10.1007/s12046-018-0823-0Sadhana(0123456789().,-volV)FT3](0123456789().,-volV)

Saputra, Bhaswara, Nasution, Li, Romadhotul, & Witra. (1 de Marzo de 2025). *Enfoques de aprendizaje profundo multimodal para la segmentación semántica de huellas mineras con imágenes satelitales multiespectrales.* Obtenido de https://doi.org/10.1016/j.rse.2024.114584

Serrano, & Carpio. (20 de Enero de 2018). *ANÁLISIS DE LA INVERSIÓN EXTRANJERA DIRECTA DE CHINA EN LA RAMA ECONÓMICA DE EXPLOTACIÓN DE MINAS Y CANTERAS (SECTOR MINERÍA) EN EL ECUADOR. PERÍODO "2008-2015"*. Obtenido de Repositorio digital de la Universidad de Especialidades Espíritu Santo

EMPRENDIMIENTO, NEGOCIOS ,ECONOMÍA Y CIENCIAS EMPRESARIALES GRADO CIENCIAS EMPRESARIALES: http://repositorio.uees.edu.ec/123456789/267

Shiraj, Rahman, Al-Imran, Liza, Murshed, & Akhter. (2024). Anomaly detection in financial time series data via mapper algorithm and DBSCAN clustering. *World Journal of Advanced Engineering Technology and Sciences, 13*(1), 70-84. doi:https://doi.org/10.30574/wjaets.2024.13.1.0396

Sisalima, Sánchez, & Ramírez-López. (2024). El derecho de amparo administrativo; sobre los concesionarios de pequeña mineria, a causa de la mineria ilegal en la provincia de el oro. *Ciencia Latina, 8*(4), 216–232. doi:https://doi.org/10.37811/cl_rcm.v8i4.12173

Soofastaei. (12 de Noviembre de 2024). *Empowering Decision-Making in Mining Through Real-Time Analytics and Predictive Modeling.* Obtenido de https://www.linkedin.com/pulse/empowering-decision-making-mining-through-real-time-ali-soofastaei-5snuf

Souza, & Silva. (2024). Aplicação da inteligência artificial na engenharia de confiabilidade e manutenção preditiva: um estudo de caso na indústria de mineração. *Revista Ibero-Americana de Humanidades, Ciências e Educação, 10*(10), 3646–3659. doi:https://doi.org/10.51891/rease.v10i10.16252

Tenecota, Viteri, & Salcedo. (2024). Análisis de la dependencia petrolera en Ecuador periodo 2018-2022. *Revista de Estudios Interdisciplinarios en Ciencias Sociales*, 958-974. doi:doi.org/10.36390/telos263.11

Tobar. (12 de Junio de 2008). *Regulation to the mining law of Ecuador.* Obtenido de https://www.tzvs.ec/wp-content/uploads/2016/11/ReglamLeyMin-Eng-102716.pdf

Torres, Mejía, & Moreyra. (2021). Geometalurgia y el futuro de la minería digital en el Perú . *Revista del Instituto de investigación de la Facultad de minas, metalurgia y ciencias geográficas, 24*(47), 163-179. doi:DOI: 10.15381/iigeo.v24i47.20661

Turing. (12 de Febrero de 1950). *COMPUTING MACHINERY AND INTELLIGENCE.* Obtenido de https://courses.cs.umbc.edu/471/papers/turing.pdf

Ulloa. (2023). Relación de las regalías mineras y el desarrollo del cantón Portovelo en Ecuador. *Estudios de la Gestión: revista internacional de administración, 2*(13), 2-12. doi:DOI: https://doi.org/10.32719/25506641.2023.13.7

Urieta. (12 de Junio de 2021). *Valoración financiera bajo incertidumbre de un proyecto minero de agregados pétreos.* Obtenido de Universidad Nacional de Colombia: https://repositorio.unal.edu.co/bitstream/handle/unal/82286/1037610999.2022.pdf;jsessionid=AB9E963EBC1EB936EDCF3C4BD39A396C?sequence=7

Vertiv.com. (2 de Junio de 2023). *La optimización de la industria minera latinoamericana: Tecnologías y soluciones innovadoras para la eficiencia y sostenibilidad.* Obtenido de https://www.vertiv.com/48f7fc/globalassets/campaigns/latam-mining/vertiv-miningguide-br-sp-latam-gr-00055-web.pdf

Vina. (26 de Noviembre de 2024). *La IA en la industria minera: Del mineral a la optimización.* Obtenido de https://www.ultralytics.com/es/blog/ai-in-the-mining-industry-from-ore-to-optimization

Wang, Feng, Chen, Wang, Tang, & Geng. (2024). Development of an Intelligent Coal Production and Operation Platform Based on a Real-Time Data Warehouse and AI Model. *Energies, 17*(20), 5-25. doi:https://doi.org/10.3390/en17205205

Wong. (2024). Deep Learning. *Cybernetical Intelligence: Engineering Cybernetics with Machine Intelligence, 2*(10), 333-365. doi:doi: 10.1002/9781394217519.ch16

World Gold Council. (31 de Enero de 2024). *Tendencias de la demanda de oro para el año 2023.* Obtenido de El alto precio del oro refleja una fuerte demanda: https://www.gold.org/goldhub/research/gold-demand-trends/gold-demand-trends-full-year-2023

Woźniak, R. J. (2024). Unsupervised machine learning in financial anomaly detection: clustering algorithms vs. dedicated methods. *Przegląd Teleinformatyczny, 11*(1), 29–46. doi:https://doi.org/10.5604/01.3001.0054.8748

Xu, Yang, Zhuang, Li, & Lu. (12 de Junio de 2024). *AI-Based Financial*

Transaction Monitoring and Fraud Prevention with Behaviour Prediction. Obtenido de https://doi.org/10.20944/preprints202407.1107.v1

Zamini, & Montazer. (18 de Diciembre de 2018). *Credit Card Fraud Detection using autoencoder based clustering*. Obtenido de 2018 9th International Symposium on Telecommunications (IST): https://ieeexplore.ieee.org/document/8661129

Elio Edwin Sánchez Suárez
Magíster en Inteligencia de Negocios y Ciencia de Datos y Magíster en Finanzas y Proyectos Corporativos por la Universidad de Guayaquil. Economista con doble especialización en Tributación y Finanzas (pregrado) y dos maestrías de alto nivel por la Universidad de Guayaquil: Inteligencia de Negocios y Ciencia de Datos – Finanzas y Proyectos Corporativos. Esta combinación formativa le permite articular el análisis financiero tradicional con herramientas avanzadas de datos para la toma de decisiones estratégicas. En el ámbito académico, ha impartido cátedra en instituciones de prestigio como la Universidad de Guayaquil (UG), Universidad Agraria del Ecuador (UAE), Universidad Católica Santiago de Guayaquil (UCSG) y el Instituto Superior Tecnológico Babahoyo (ISTB), fortaleciendo su capacidad para transmitir conocimiento aplicado. Su experiencia profesional abarca no solo la docencia universitaria y la investigación, sino también la coordinación académica, capacitación empresarial y la consultoría y asesoría tributaria. Esta visión integral le permite comprender los desafíos financieros y normativos del sector productivo, con énfasis en la optimización de procesos mediante inteligencia de negocio.

Víctor Alfredo Iturralde Calahorrano
Licenciado en Ciencias de la Educación, mención Comercio Exterior y Magíster en Inteligencia de Negocios y Ciencia de Datos por la Universidad de Guayaquil (UG). Posee una trayectoria que integra el liderazgo gremial, la gestión financiera y la innovación tecnológica, combinando la formación en comercio exterior y ciencia de datos para impulsar la eficiencia organizacional desde múltiples frentes. En el ámbito empresarial, se ha desempeñado como Gerente General de Empresarial VICTOC SAS y Representante Legal de GRENERGYC SAS, donde ha liderado estrategias de optimización operativa. Su experiencia en el sector financiero popular incluye el cargo de jefe de Agencia en la Cooperativa de Ahorro y Crédito Credi Ya, lo que le ha permitido comprender en profundidad los flujos financieros y las necesidades de control de procesos en entornos de alta rotación. Como líder gremial, fue presidente de la Asociación de Producción Textil Asoprobendi, fortaleciendo su capacidad para gestionar cadenas productivas y asociatividad. Adicionalmente, se ha destacado y desempeñado como conferencista empresarial, especializado en la implementación de herramientas de Big Data y Ciencia de Datos para la optimización de procesos, área clave para la toma de decisiones basada en evidencia.

Ingrid Mercedes Lamilla Miranda
Ingeniera Comercial con mención en Finanzas de la Universidad de Guayaquil (UG), Magister en Educación Mención en Gestión del Aprendizaje Mediado por Tic. de la Pontificia Universidad Católica del Ecuador (PUCE). En el área administrativas se ha desempeñado en diferentes Instituciones públicas y privadas (DISMERO S.A, LA FAMILIAR S.A, GOBIERNO AUTÓNOMO DESCENTRALIZADO DE BABAHOYO, DIRECCIÓN DISTRITAL DE EDUCACIÓN 12D04 QUINSALOMA-VENTANAS, DIRECCIÓN DISTRITAL DE EDUCACIÓN 12D01 BABA-BABAHOYO-MONTALVO. Se ha desempeñado también como docente del Instituto Superior Tecnológico Babahoyo (ISTB) en la carrera de Administración y Administración Financiera, actualmente está designada COORDINADORA DE LA CARRERA DE ADMINISTRACIÓN FINANCIERA DEL (ISTB). Su experiencia profesional se relaciona a las actividades de docencia universitaria, coordinación académica, capacitación a nivel empresarial, consultoría y asesoría tributaria.

Jessica Jessenia Izurieta Álvarez
Licenciada en Administración Ejecutiva de la Universidad Técnica de Babahoyo, se ha desempeñado en el área administrativas en diferentes Instituciones públicas y privadas (Banco Central del Ecuador, Banco Internacional, Empresa de Saneamiento Ambiental de Babahoyo EP EMSABA, Instituto Tecnológico Superior Eugenio Espejo, Constructora Fluminense S.A, Distribuidora de Medicina Leterago del Ecuador, y como Tesorera y Docente en el Instituto Superior Tecnológico Babahoyo (ISTB). Su experiencia profesional se relaciona a las actividades administrativas y la docencia en el nivel superior.

ISBN: 978-9942-53-156-8

Compás
capacitación e investigación