ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS.

TÍTULO DE LA INVESTIGACIÓN
Aplicación de machine learning para datos geoquímicos regionales: características geoquímicas de elementos de ocurrencias de trazas de la Cordillera Occidental del Ecuador.

Profesor
Mario Salvador Gonzáles

Autor
Juan Andrés Páez Molineros

2023

RESUMEN

En el presente estudio se materializa la importancia de la interpretación precisa de los datos y la interacción multidisciplinaria para comprender fenómenos geológicos complejos, mediante la aplicación de técnicas de clustering.

El análisis de clusters y la visualización de datos emergieron como herramientas fundamentales para descubrir patrones y relaciones ocultas en conjuntos de datos. Los gráficos Q-Q y de dispersión se destacaron como formas efectivas de evaluar la calidad de los datos y las relaciones entre variables.

La selección del número óptimo de clusters en el análisis de clustering fue abordada a través de técnicas como la curva de deflexión, que permiten identificar un equilibrio entre la explicación de la varianza y la complejidad del modelo.

El conocimiento geológico se demostró crucial para la interpretación de la composición de las rocas. Se exploraron composiciones de basaltos y andesitas, con énfasis en cómo estas composiciones reflejan el origen y la formación de las rocas, así como su potencial para albergar minerales de valor económico.

Es importante perfilar a la investigación a un enfoque holístico y basado en datos para comprender fenómenos científicos y geológicos complejos. La combinación de técnicas analíticas y conocimiento de dominio es esencial para tomar decisiones informadas y extraer información significativa de los datos en diversas aplicaciones.

ABSTRACT

This study materializes the importance of accurate data interpretation and multidisciplinary interaction to understand complex geological phenomena, through the application of clustering techniques.

Cluster analysis and data visualization have emerged as essential tools for discovering hidden patterns and relationships in data sets. Q-Q and scatter plots stood out as effective ways to assess data quality and relationships between variables.

The selection of the optimal number of clusters in the clustering analysis was approached through techniques such as the deflection curve, which allow to identify a balance between the explanation of the variance and the complexity of the model.

Geological knowledge proved crucial for the interpretation of the composition of the rocks. Basalt and andesite compositions were explored, with an emphasis on how these compositions reflect the origin and formation of the rocks, as well as their potential to host minerals of economic value.

It is important to sharpen research to a holistic and data-driven approach to understand complex scientific and geological phenomena. The combination of analytical techniques and domain knowledge is essential to make informed decisions and extract meaningful information from data in various applications.

ÍNDICE DEL CONTENIDO

Tabla de contenido

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Una de las dos principales cadenas montañosas andinas de Ecuador, junto con la Cordillera Central, es la Cordillera Occidental. Está ubicado en la costa del Pacífico y se extiende a lo largo de toda la nación de norte a sur. Con 6.267 metros sobre el nivel del mar, el Chimborazo es el pico más alto de la Cordillera Occidental. La cordillera es baja en comparación con la Cordillera Central y geológicamente joven (Tibaldi, 1993).

Con el propósito de obtener una comprensión más completa de la geología y la composición de la Cordillera Real, se recolectaron meticulosamente un total de diez muestras representativas de diferentes formaciones geológicas en la región. Estas muestras fueron seleccionadas cuidadosamente para abarcar una variedad de características geológicas y contextos de formación. Mediante un enfoque estratégico, se buscó capturar la diversidad de la composición mineralógica y química que caracteriza a esta cordillera. Estas muestras servirán como base fundamental para realizar análisis geoquímicos detallados, lo que permitirá una interpretación más precisa de la historia geológica y los procesos que han dado forma a esta cordillera.
Los datos geoquímicos obtenidos a partir de las muestras recogidas desempeñan un papel fundamental en la determinación del origen y la litología de las rocas en la Cordillera Occidental. Estos datos están siendo sometidos a un análisis exhaustivo para identificar los elementos químicos presentes y sus proporciones relativas. Este proceso permitirá desentrañar la historia geológica de la región, revelando procesos magmáticos y tectónicos que han moldeado su formación. Además, la interpretación de los datos geoquímicos tiene como objetivo establecer una litología precisa, lo que significa categorizar las rocas en términos de su composición mineralógica y química. Esta clasificación es esencial para determinar si la zona tiene el potencial de albergar depósitos minerales valiosos y, por lo tanto, si es un candidato viable para la exploración minera. En última instancia, este análisis detallado de los datos geoquímicos proporcionará información esencial para tomar decisiones informadas sobre futuros proyectos de exploración y aprovechamiento sostenible de los recursos minerales en la Cordillera Occidental.

Se pueden investigar grandes conjuntos de datos mediante la aplicación de algoritmos de machine learning, que es una potente tecnología de ciencia de datos que se emplea con mayor frecuencia en aplicaciones científicas de vanguardia. Dado que los datos geoquímicos en masa con frecuencia consisten en grandes conjuntos de muestras con una variedad de elementos detectados para cada muestra, son excelentes objetivos para el análisis utilizando algoritmos de machine laerning. Gracias a los avances analíticos y las capacidades de intercambio de datos, los conjuntos de datos geoquímicos globales y regionales también se están volviendo menos costosos de producir y más accesibles (Lindsay, 2020).

REVISIÓN DE LITERATURA

Automatización del análisis exploratorio de datos y procesamiento geoquímico univariado empleando Python. (Castillo, 2023)**.**

En diversas áreas de las ciencias geológicas se está utilizando la automatización de procesos, lo que se evidencia en la creación de librerías como Pyrolite, PyGeochemCalc, dh2loop 1.0, NeuralHydrology y GeoPyTools, entre otras. El trabajo presenta una metodología para automatizar el análisis geoquímico univariado utilizando paquetes de código abierto de Python como Pandas, Seaborn, Matplotlib, Statsmodels y Scipy. Estos paquetes se integrarán en un script en un entorno de trabajo local como Jupyter Notebook. El script está diseñado para procesar cualquier tipo de datos geoquímicos, incluyendo cálculos, gráficos y eliminación de excepciones.

A machine learning approach for regional geochemical data: Platinum-group element geochemistry vs geodynamic settings of the North Atlantic Igneous Province. (Lindsay, 2020)**.**

El trabajo permite obtener información sobre procesos magmáticos como la fusión y el fraccionamiento de elementos a partir de métodos geoquímicos tradicionales, pero al analizar objetivamente conjuntos de datos regionales completos utilizando algoritmos de aprendizaje automático (MLA), podemos sacar a la luz nuevos aspectos de la estructura de datos más grande y mejorar en gran medida la geoquímica anterior. interpretaciones. Se ha demostrado que el presupuesto del elemento del grupo del platino (PGE) de las lavas en la Provincia Ígnea del Atlántico Norte (NAIP) varía sistemáticamente con la edad, la ubicación y el entorno geodinámico. Los MLA se utilizaron para investigar los controles magmáticos en estas concentraciones cambiantes porque hay un conjunto considerable de datos geoquímicos de elementos múltiples disponible para el área. La capacidad del aprendizaje automático para agrupar muestras en un espacio multidimensional (es decir, de elementos múltiples) es el principal beneficio de utilizar el aprendizaje automático en el análisis.

A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping.
(Grunsky C. , 2014)**.**
Se utilizaron datos geoquímicos de sedimentos lacustres de lagos de toda la península Melville de Nunavut, Canadá, para probar la eficacia del mapeo geológico predictivo. Se proporciona un método informativo y cuantitativo para el mapeo litológico al tratar los datos geoquímicos de sedimentos lacustres dentro del marco composicional del análisis de logratio y aplicar el análisis de componentes principales, el análisis de varianza, el análisis discriminante lineal y el análisis espacial con kriging convencional de la manera correspondiente. Para comprender las relaciones entre múltiples elementos basadas en el contraste de los tipos de rocas locales, el análisis de componentes principales es una herramienta útil. El análisis espacial arroja luz sobre la dirección y la

continuidad espacial de los elementos conectados a unidades de mapa particulares, mientras que el análisis de varianza detalla qué elementos son los mejores diferenciadores para distinguir entre los tipos de rocas representados por la geoquímica de los sedimentos del lago. La capacidad de discriminar entre las distintas unidades del mapa y validar el poder predictivo del mapeo de las unidades del mapa subyacentes basándose en la geoquímica de los sedimentos lacustres es posible gracias al análisis discriminante lineal. Con el fin de probar y validar mapas geológicos existentes y crear nuevos mapas geológicos en áreas donde no hay datos geológicos suficientes, los métodos estadísticos multivariados sobre datos geoquímicos de sedimentos lacustres forman la base para desarrollar una metodología objetiva para descubrir y clasificar procesos geoquímicos.

Application of principal component analysis and cluster analysis to mineral exploration and mine geology. (Gazley, 2015)**.**
Con frecuencia se recopilan grandes conjuntos de datos mientras se extraen y exploran los depósitos en busca de minerales. Estos conjuntos de datos con frecuencia crecen tanto o se complican tanto que resulta difícil visualizar su organización, y mucho menos convertir esta organización en conocimiento valioso para los geólogos de exploración o minería. Dado que los resultados geoquímicos se presentan frecuentemente como composiciones, un análisis completo debe ser igual al 100%. La varianza de los datos se introduce por esta limitación adicional con correlaciones que no pueden manejarse mediante métodos estadísticos univariados o bivariados estándar. Esto se puede evitar mediante el uso de transformaciones de relación logarítmica para convertir los datos en un espacio de números reales, al que luego se pueden utilizar técnicas estadísticas convencionales. Cuando se utilizan transformaciones de relación logarítmica en un conjunto de datos, es necesario cuantificar todos y cada uno de los aspectos de interés en todas las muestras. Existen rutinas que pueden imputar o reemplazar valores faltantes para evitar correlaciones erróneas. El análisis de componentes principales (PCA), una vez que los datos se han limpiado y transformado, se puede utilizar para descubrir estructuras de datos de alto nivel que no son visibles utilizando métodos univariados o bivariados, simplificar conjuntos de datos geoquímicos y permitir la interpretación de la varianza dentro de los conjuntos de datos. Después de utilizar PCA (u otra ordenación), el análisis de conglomerados se puede utilizar para identificar grupos de muestras sin tener ningún conocimiento previo de sus vínculos geográficos o temporales.

A Cluster Separation Measure. (Davies, 1979)
La similitud de los clusters que se cree que tienen densidades de datos que disminuyen con la distancia desde una característica del cluster se mide utilizando una métrica que se describe. La métrica se puede utilizar para inferir la idoneidad de las particiones de datos y, en consecuencia, para comparar la idoneidad relativa de múltiples divisiones de datos. La métrica se puede utilizar para dirigir un algoritmo de búsqueda de clusters porque es independiente del número de clusters examinados o de cómo se dividieron los datos.

t-SNE Based Visualisation and Clustering of Geological Domain (Balamurali, 2016)
Al estimar los recursos minerales, la definición de dominios geológicos y sus fronteras es crucial. Para identificar problemas con la calidad de los datos o proponer hipótesis novedosas, los geólogos a menudo se interesan por el análisis de datos exploratorios y la visualización de datos geológicos en dos o tres dimensiones. Para la presentación de grandes conjuntos de datos de ensayos geoquímicos, comparamos PCA, algunos enfoques lineales y no lineales adicionales y un método más reciente, la incrustación de vecinos estocásticos t-distribuited (t-SNE). Utilizando registros de exploración y producción, las dimensiones reducidas basadas en t-SNE se pueden combinar con una técnica de agrupación para extraer zonas geológicas bien agrupadas. En espacios objetivo bidimensionales, hubo diferencias discernibles entre los enfoques más modernos y el método t-SNE no lineal.

Treatment of nondetects in multivariate analysis of groundwater geochemistry data (Farnham, 2002)**.**
Para evaluar las similitudes en la química de los oligoelementos de las aguas subterráneas, se realiza un análisis de componentes principales (PCA). Sin embargo, muchos oligoelementos se encuentran en concentraciones inferiores a los límites de detección (DL), lo que complica los cálculos estadísticos. Los enfoques de sustitución simples pueden ser preferibles porque las mejores técnicas para manejar valores 'DL' suponen normalidad o lognormalidad, lo que no es el caso para estos datos.
En este trabajo, se creó un método novedoso para identificar las técnicas de reemplazo más efectivas al tratar con los valores "DL" para una recopilación de datos específica. Se utilizó un modelo multivariado mixto en los estudios de simulación de Monte Carlo para examinar los impactos de reemplazar los valores 'DL' con 0, DL y DL/2 en los hallazgos de PCA. Los resultados generalmente mostraron que la sustitución con DL/2 tuvo mejores resultados que la sustitución con DL o 0. Cuando el porcentaje de valores "DL" superó aproximadamente el 25%, el rendimiento de todos los métodos de sustitución comenzó a verse afectado.

An Investigation of Pb Geochemical Behavior Respect to Those of Fe and Zn Based on k-Means Clustering Method (Ghannadpour, 2013)
La técnica de k-medias, que divide los datos en k clases según un requisito de distancia, es un algoritmo de agrupamiento bien conocido. En el estudio actual, la k óptima se determinó utilizando el método k-medias para clasificar los datos generados a partir de los pozos de exploración en el depósito Parkam. Luego se agruparon los datos y se evaluaron los aspectos relativos del comportamiento geoquímico.
Los criterios utilizados para identificar el mejor k variaron en tamaño de clase desde k=3 hasta k=10, y después de analizar las clasificaciones resultantes, se seleccionó el mejor k. Los resultados mostraron que los agrupamientos con k=3 para Pb y Zn y k=4 para Pb y Fe fueron superiores al número de otras clases en

cada caso. Además, de acuerdo con los casos anteriores y la clasificación derivada, un aumento en la ley de Pb es seguido por un aumento en la ley de Zn, y también hay un aumento seguido de una disminución en la ley de Fe. Así, utilizando el método proporcionado anteriormente, que puede ofrecer una perspectiva muy adecuada frente a otros elementos, es factible analizar la fluctuación de elementos como Cu o Pb con otros componentes existentes en el análisis realizado.

Generalized inferential models for censored data (Cahoon, 2021)**.**
Bajo los marcos tradicionales, se han investigado a fondo las dificultades inferenciales que ocurren cuando se censuran datos. En este estudio, presentamos una técnica de modelo inferencial extendido diferente, cuyo resultado es una función de plausibilidad que depende de los datos. El vínculo entre la distribución de la función de probabilidad relativa en el parámetro de interés y una variable auxiliar no observada es lo que motiva esta construcción. La distribución de un conjunto aleatorio calibrado adecuado destinado a pronosticar una variable auxiliar no observada produce la función de plausibilidad.

The elements of statistical learning: data mining, inference, and prediction (Hastie & Tibshirani, 2009)
Ha habido un enorme aumento en la tecnología de la computación y la información durante los últimos diez años. La dificultad de comprender datos enormes ha impulsado la creación de nuevos métodos estadísticos y ha dado lugar a campos completamente nuevos como la minería de datos, el aprendizaje automático y la bioinformática. Este libro presenta los conceptos clave en estos campos bajo un marco conceptual compartido. Aunque el método es estadístico, la atención se centra en las ideas más que en la aritmética.
Se proporcionan numerosos ejemplos y se utilizan ampliamente imágenes en color. Para los estadísticos y cualquier persona interesada en la minería de datos en la ciencia o los negocios, es un recurso invaluable. El alcance de la cobertura del libro abarca desde el aprendizaje supervisado (predicción) hasta el aprendizaje no supervisado. Las redes neuronales, las SVM, los árboles de clasificación y el impulso se encuentran entre los numerosos temas que se tratan en este libro; es el primero en abordar el tema en su totalidad.

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El objeto de estudio del machine learning utilizando clustering para datos geoquímicos regionales, en el contexto de las características geoquímicas de elementos de ocurrencias de trazas la cordillera Occidental del Ecuador, es identificar patrones, comportamiento y estructuras subyacentes en los datos geoquímicos para agrupar las ubicaciones con el fin de verificar si el lugar estudiado tiene potencial minero, dependiendo sus características geoquímicas.

El objetivo es utilizar técnicas de clustering para descubrir agrupaciones naturales en los datos geoquímicos y clasificar las ubicaciones en clústeres que compartan características comunes. Esto permite una comprensión más profunda de la variabilidad espacial y la distribución de los elementos de ocurrencias de trazas en la cordillera occidental del Ecuador.

El uso de clustering en datos geoquímicos regionales puede proporcionar información valiosa para diversas aplicaciones, como la identificación de áreas con características similares para la exploración de recursos naturales, el análisis de la calidad del suelo o el agua, la evaluación de riesgos ambientales, y la toma de decisiones en la gestión de recursos y la planificación del desarrollo.

PLANTEAMIENTO DEL PROBLEMA

El objetivo principal es desarrollar un modelo de machine learning que pueda predecir y comprender las características y el comportamiento de los elementos geoquímicos de las ocurrencias de trazas.

Se cuenta con un conjunto de datos geoquímicos que incluye mediciones de múltiples elementos de ocurrencias de trazas tomadas en diferentes ubicaciones de la Cordillera Real del Ecuador. Cada registro de datos contiene información sobre las características geográficas y las mediciones geoquímicas de los elementos.

Se debe realizar un procesamiento y limpieza de los datos para tratar cualquier valor faltante (como NaN), eliminar datos inconsistentes o erróneos, y normalizar las variables si es necesario. Además, podría ser útil explorar y visualizar los datos para comprender mejor su distribución y relaciones.

Es importante realizar técnicas de selección de características para identificar las variables más relevantes y descriptivas que influyen en las características geoquímicas de los elementos de ocurrencias de trazas en la región. Esto ayuda a reducir la dimensionalidad y mejorar la eficiencia del modelo.

Se divide el conjunto de datos en conjuntos de entrenamiento y prueba. Se entrena el modelo utilizando los datos de entrenamiento y se evalúa su rendimiento utilizando métricas apropiadas, como el coeficiente de determinación (R2), precisión, recall, matriz de confusión, entre otros.
Ajustar los hiperparámetros del modelo y aplicar técnicas de validación cruzada para mejorar su rendimiento y generalización.
Se analizan los resultados del modelo para comprender las características geoquímicas más influyentes en las ocurrencias de trazas. Esto puede incluir la identificación de correlaciones, patrones espaciales o variables clave que afectan las mediciones.
Una vez desarrollado y validado el modelo, se puede utilizar para predecir las características geoquímicas de nuevos puntos de muestreo en la región. Esto puede ser útil para la exploración de recursos naturales, la evaluación del medio ambiente o la toma de decisiones relacionadas con la geoquímica regional.
Es importante destacar que este es solo un planteamiento general del problema y que los pasos específicos pueden variar dependiendo de los datos disponibles, la naturaleza del problema y las técnicas de machine learning utilizadas. Se recomienda adaptar este planteamiento según las necesidades y requisitos del proyecto.

OBJETIVO GENERAL

Aplicar clustering en datos geoquímicos regionales para encontrar patrones y agrupaciones que ayuden a comprender mejor la variabilidad espacial y la distribución de los elementos de ocurrencias de trazas en la Cordillera Occidental de Ecuador. Esto proporciona información valiosa para la toma de decisiones y el desarrollo de estrategias en diferentes campos, principalmente la geología, la minería y la conservación ambiental.

OBJETIVOS ESPECÍFICOS

Realizar el procesamiento y limpieza de los datos geoquímicos, incluyendo la eliminación de valores faltantes o atípicos, y normalización de las variables si es necesario.

Identificar las características geoquímicas más relevantes y significativas para el análisis de clustering. Esto podría incluir técnicas de selección de características, como análisis de componentes principales (PCA) o coeficientes de correlación.

Utilizar algoritmos de clustering, como K-means, para agrupar las ubicaciones de la Cordillera Real del Ecuador en clústeres basados en sus características geoquímicas. Estos algoritmos buscan encontrar grupos coherentes y compactos en función de la similitud de las observaciones.

Evaluar la calidad y coherencia de los clústeres obtenidos utilizando métricas de evaluación de clustering, como el índice de Silhouette o el coeficiente de Calinski-Harabasz. Esto ayuda a determinar la calidad y robustez de los grupos identificados.

Analizar e interpretar los clústeres generados para comprender las características geoquímicas que los definen. Esto puede implicar la identificación de patrones geoquímicos comunes en cada clúster y el análisis de las diferencias entre los grupos para determinar zonas mineralizadas.

Utilizar los clústeres obtenidos para tomar decisiones informadas en la exploración de recursos naturales, la gestión ambiental o cualquier otro dominio relevante. Por ejemplo, se pueden identificar áreas con características geoquímicas similares que podrían indicar una mayor probabilidad de ciertos tipos de yacimientos minerales.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

En el campo de la geología y la exploración de recursos naturales, el uso de técnicas de aprendizaje automático en el análisis de datos geoquímicos regionales representa un enfoque innovador y muy prometedor. La combinación de la creciente disponibilidad de datos geoquímicos y las capacidades de procesamiento y análisis de datos proporcionadas por el machine learning se convierte en una herramienta esencial para desentrañar patrones, relaciones y tendencias ocultas en conjuntos de datos de gran envergadura a medida que avanzamos en la era digital (Maimon, 2010). Esta explicación se basa en los siguientes puntos importantes:

Complejidad y Volumen de Datos: Gracias a las tecnologías de muestreo, análisis y cartografía modernas, la recopilación y generación de datos geoquímicos ha alcanzado un nivel sin precedentes. Los conjuntos de datos que se producen con frecuencia son complejos y voluminosos, lo que dificulta su análisis manual y la identificación de patrones significativos. Aquí es donde las técnicas de aprendizaje automático pueden intervenir de manera efectiva para identificar relaciones subyacentes y correlaciones en estos datos.

Descubrimiento de patrones no lineales: las relaciones no lineales regulan los fenómenos geológicos y geoquímicos. Los patrones no lineales en los datos pueden ser detectados por los algoritmos de aprendizaje automático, lo que permite descubrir relaciones sutiles y complejas que podrían pasar desapercibidas en un análisis convencional. Esto es especialmente importante para la exploración de recursos naturales y depósitos minerales, donde las relaciones pueden ser muy contextualmente específicas y altamente no lineales.

Predicciones y Modelización Precisas: la aplicación de algoritmos de aprendizaje automático permite la creación de modelos predictivos precisos basados en datos históricos. Estos modelos pueden predecir la presencia de ciertos minerales o elementos geoquímicos en áreas aún no muestreadas, lo que es esencial para la toma de decisiones sobre la exploración y extracción de recursos naturales. Además, con nuevos datos, estos modelos pueden actualizarse y refinarse, mejorando continuamente su precisión.

Avance Tecnológico y Competitividad: La adopción de tecnologías avanzadas como el aprendizaje automático en geología y la exploración de recursos naturales mejora la eficiencia y la precisión y coloca a las empresas y organizaciones en una posición más competitiva. Los que utilizan estas tecnologías pueden tomar decisiones estratégicas e informadas que les permitan sobresalir en un entorno altamente competitivo.

Seleccionar la base de datos

Diez muestras de roca de diferentes formaciones geológicas representativas de la región de interés se seleccionaron cuidadosamente para este estudio. Estas muestras capturan la variabilidad inherente a la composición geoquímica de la región porque abarcan una amplia gama de contextos geológicos y condiciones

ambientales. La composición de los los óxidos presentes en las rocas se analizó minuciosamente en cada muestra. Esta selección estratégica de muestras tiene como objetivo proporcionar una base sólida y representativa para la aplicación de técnicas de aprendizaje automático, lo que permite identificar patrones y tendencias de los elementos, para entender los procesos geológicos de formación.

Limpieza, pre-procesamiento de datos

Identificación y descripción de variables
La identificación y la descripción de variables son pasos esenciales en el análisis de datos geoquímicos regionales mediante técnicas de machine learning. En este caso, las variables representan las diversas características y partes de las muestras de roca que se utilizarán como entradas para los algoritmos de aprendizaje automático. Algunas de las variables relevantes para el análisis son:

- Composición de Clinopiroxenos: Estas variables representan las proporciones relativas de diversos elementos químicos presentes en los clinopiroxenos, que son minerales comunes en rocas ígneas y metamórficas. Las composiciones de clinopiroxenos pueden incluir elementos como wollostonita (wo), Estantita (En), Ferricilita (Fe), Acmita (Ac) y otros elementos traza. Estas variables capturan las características específicas de la mineralogía de las rocas y pueden indicar la génesis y evolución geológica.
- Composición de óxidos: Las proporciones de elementos químicos en forma de óxidos presentes en las muestras de roca son descritas por estas variables. Los óxidos de hierro (FeO), el óxido de titanio (TiO_2), el óxido de aluminio (Al_2O_3) y otros son óxidos comunes. Para comprender las condiciones de formación y la historia geológica de las rocas, es crucial comprender la composición de los óxidos.

Visualización de variables
Al proporcionar una comprensión visual de cómo se agrupan los datos y cómo las variables contribuyen a la formación de grupos, la visualización de variables es un componente esencial del análisis de agrupamiento. Los patrones, tendencias y relaciones que pueden no ser evidentes en los datos brutos se pueden identificar mediante la representación gráfica de las variables. El uso de gráficos de dispersión, donde cada punto representa una instancia y se colorea o etiqueta según el grupo al que pertenece, es una de las técnicas más comunes (Grunsky E. , 2012). Esta visualización permite observar la distribución de los grupos en relación con dos variables a la vez, así como clusters naturales y solapamientos entre ellos. Los gráficos q-q, además de los gráficos de dispersión, son útiles para comprender la variabilidad de cada variable en los diferentes grupos. Estos gráficos muestran la mediana, los cuartiles y los valores potenciales atípicos de cada grupo, lo que ayuda an encontrar diferencias en la distribución de las variables entre los grupos. Estos gráficos ayudan an encontrar patrones de similitud y diferencias en los valores de las variables entre los grupos (Grunsky C. , 2014).

Finalmente, la visualización de variables en el análisis de grupos es esencial para identificar patrones ocultos, verificar la coherencia de los grupos y guiar el proceso de toma de decisiones. Los analistas pueden obtener una representación más completa y significativa de la estructura de los datos y cómo los grupos se relacionan con las características subyacentes al combinar varias técnicas de visualización.

Selección del modelo estadístico

El objetivo del análisis de clustering, también conocido como segmentación, es identificar y describir grupos en una colección de datos de modo que los elementos que pertenecen al mismo grupo sean similares entre sí y los que pertenecen a diferentes grupos sean diferentes entre sí. En general, la medición de distancia es la que con mayor frecuencia se utiliza para realizar comparaciones entre los atributos de las distintas instancias. Cuando se completa el proceso de agrupamiento, es posible etiquetar cada instancia del conjunto de datos asociándolo con uno de los grupos creados (Maimon, 2010).

Análisis exploratorio de datos

El análisis de datos exploratorio (EDA) es un enfoque estadístico que se centra en el análisis de datos para resumir sus características principales, con frecuencia utilizando métodos visuales. El objetivo del EDA es comprender los datos y descubrir patrones, relaciones y anomalías que puedan ser útiles para la toma de decisiones. La investigación científica, la ingeniería, la medicina y otras áreas utilizan con frecuencia el EDA para explorar y analizar grandes conjuntos de datos (Grunsky C. , 2014).

El análisis de poblaciones univariadas sirve como base para la descripción de la exploración de datos. Los gráficos EDA suelen ser útiles cuando se agrupan porque ofrecen una variedad de métodos para resumir datos, que consisten en una combinación de texto y gráficos, brindan una base para el contexto y la comparación de varios tipos de datos. (Grunsky E. , 2012)

En el proceso de análisis exploratorio de datos, se emplearon dos herramientas visuales fundamentales para examinar y comprender la naturaleza de los datos. Los gráficos Q-Q plot se utilizaron para evaluar la conformidad de las distribuciones observadas con las distribuciones teóricas, lo que permitió identificar posibles desviaciones y patrones inusuales en los datos. Además, se emplearon los gráficos de dispersión para visualizar la relación entre pares de variables, brindando información sobre patrones de agrupación, dispersión y posibles correlaciones. Estas dos herramientas complementarias facilitaron la detección temprana de tendencias significativas y la identificación de posibles anomalías en el conjunto de datos, lo que a su vez sentó las bases para análisis más profundos y específicos en etapas posteriores del estudio.

Q-Q plot

Una distribución de frecuencia se puede comparar gráficamente con una distribución de frecuencia esperada, que suele ser la distribución normal, utilizando un gráfico Q-Q. Se crea comparando los valores cuantiles de la distribución de frecuencia normal con los datos observados ordenados. Si una

distribución de frecuencia tiene una distribución normal, se dibujará una línea recta cuando los valores de los cuantiles se representen frente a los valores ordenados de la población. Los gráficos Q-Q son útiles para detectar valores extremos en las colas de la distribución y para examinar los rasgos específicos de conjuntos de observaciones. Además, pueden revelar aspectos de la naturaleza de los datos que otras visualizaciones no pueden revelar (Grunsky E. , 2012).

Se observó que los gráficos Q-Q plots generados figura 1 para los datos de clinopiroxenos exhiben un patrón irregular y significativas desviaciones de la línea diagonal esperada. Dada la falta de linealidad y coherencia en estos gráficos, se ha decidido descartar su utilización en el análisis subsiguiente. Es importante destacar que, para los objetivos específicos de este estudio, los datos de clinopiroxenos no aportan información relevante ni contribuyen de manera significativa al análisis general. Por lo tanto, se optó por enfocar la atención en otras variables y aspectos más pertinentes para la investigación en curso.

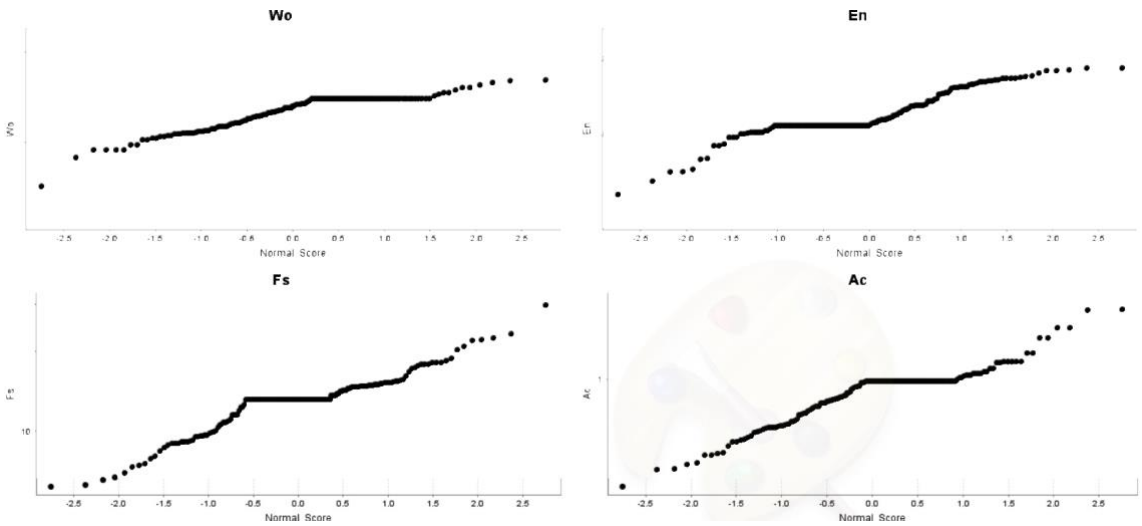

Ilustración 1. Q-Q plot clinopiroxenos

El análisis exploratorio de datos aplicado a los óxidos reveló ciertas curvaturas y variaciones en los gráficos Q-Q plots generados. Aunque estas desviaciones podrían indicar ciertas discrepancias con las distribuciones teóricas asumidas, es importante destacar que, en el contexto del propósito de este estudio, estas curvaturas no invalidan la utilidad del análisis. Estas variaciones podrían atribuirse a características intrínsecas de los datos de óxidos y no necesariamente representan incongruencias sustanciales para los objetivos de la investigación.

En el contexto de las metas de este estudio, las curvaturas observadas en los gráficos Q-Q plots para los óxidos no presentan un obstáculo insuperable. Dado que el análisis exploratorio de datos tiene como objetivo principal establecer tendencias y relaciones generales, las desviaciones identificadas no comprometen la validez del enfoque. Estas curvaturas podrían incluso revelar peculiaridades interesantes en la distribución de

óxidos, lo que podría enriquecer la comprensión de las relaciones investigadas. En resumen, a pesar de las curvaturas presentes en los gráficos Q-Q, el análisis exploratorio de datos para los óxidos se considera aceptable y viable para el propósito central de este estudio.

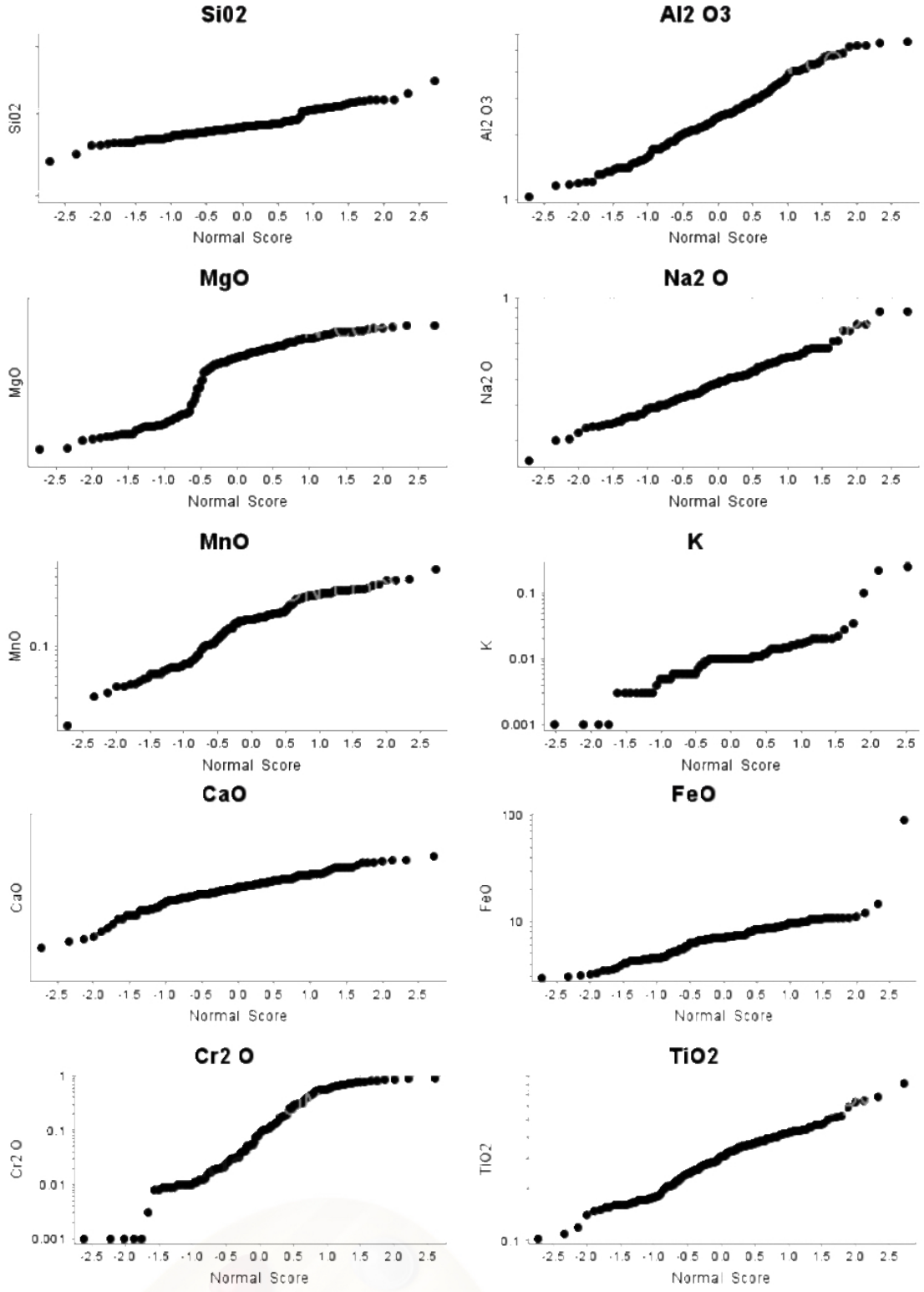

Ilustración 2. Q-Q plot Oxidos

Graficos de Dispersión

Cuando existe muchos puntos de datos diferentes y se desea destacar las similitudes en el conjunto de datos, se utiliza una gráfica de dispersión. Esto es útil para encontrar valores atípicos o comprender la distribución de datos.

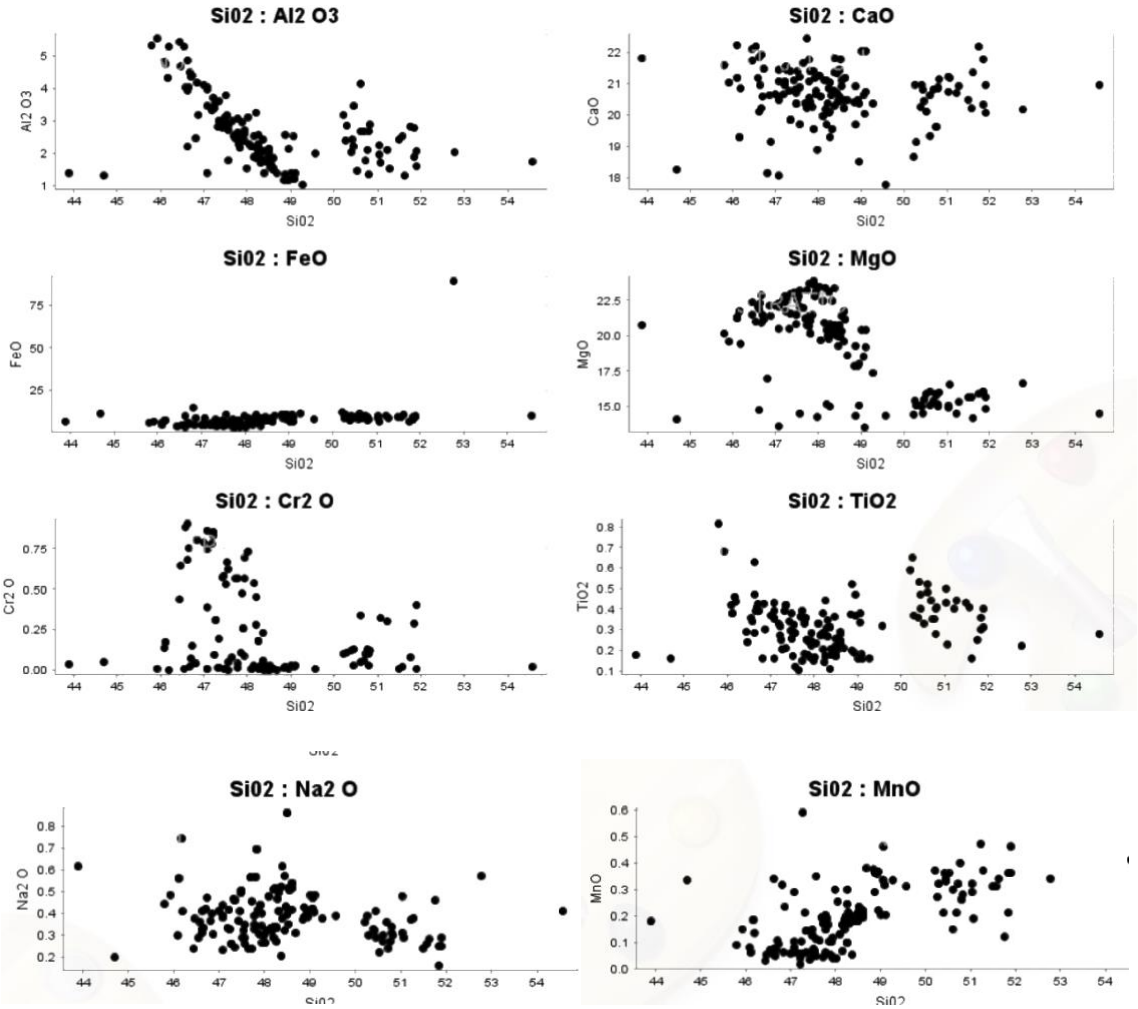

Ilustración 3. Gráfico de dispersión para oxidos

Analisis de componentes principales

La análisis de componente principal (PCA) es una técnica estadística que reduce la cantidad de variables necesarias para describir la variación observada en un conjunto de datos. Esto se logra mediante la creación de combinaciones lineales de las variables (componentes) que describen la distribución de la información. Estas combinaciones lineales provienen de alguna medida de asociación (es decir, matriz de correlación o covarianza). La creación de indicadores empiricos para la atención específica de un elemento es uno de los muchos usos de PCA en la evaluación de datos geoquímicos. Una técnica de PCA conocida como análisis de componentes principales de modo simultáneo RQ tiene la ventaja de presentar los puntajes de componentes principales de las

observaciones y las variables (elementos) en la misma escala, lo que permite plotar las observaciones y las variables en el mismo diagrama (Grunsky E. , 2012)
La análisis de componentes fundamentales, o PCA, es una técnica estadística que se utiliza para reducir la dimensionalidad de un conjunto de datos con múltiples variaciones. El objetivo de la PCA es encontrar un conjunto de variables no correlacionadas, llamadas componentes principales, que puedan explicar la mayor cantidad posible de varianza en los datos iniciales. Los componentes principales están organizados en función de la cantidad de varianza que explican, por lo que el primer componente principal explica la mayor cantidad de varianza, el segundo componente principal explica la siguiente mayor cantidad de varianza, etc. En el análisis de datos geoquímicos, la PCA se utiliza con frecuencia para identificar patrones y relaciones entre elementos químicos (Grunsky C. , 2014).

Tabla 1. PCA de las varibales utilizadas

Scaled Coordinates	PC1	PC2	PC3	PC4
Si02	0.5423	0.5376	0.08229	0.3715
Al2 03	-0.6624	0.3627	0.4882	-0.04774
CaO	-0.5169	0.1736	0.02543	0.6908
K	0.164	0.1874	-0.4165	0.6214
FeO	0.9418	-0.08167	0.1058	-0.07025
MgO	-0.8372	-0.4111	-0.02018	0.01364
Na2 O	-0.1252	-0.681	0.464	0.3904
Cr2 O	-0.5298	0.6695	-0.2611	-0.2264
TiO2	0.1167	0.4139	0.8364	-7.1847E-4
MnO	0.9504	-0.04045	0.06932	0.03024

A través del análisis de componentes principales (PCA), se lograron identificar cuatro componentes principales que capturan de manera efectiva la variabilidad en los datos. Estos componentes revelan patrones significativos en la relación entre los elementos analizados, indicando tendencias de crecimiento conjunto en múltiples variables. La identificación de estas cuatro componentes principales destaca la existencia de relaciones intrínsecas entre los elementos estudiados, proporcionando una visión más profunda de cómo estos elementos tienden a aumentar en conjunto a lo largo de los datos analizados.

RESULTADOS

Se lograron identificar con éxito cuatro componentes principales que arrojan luz sobre las interrelaciones presentes entre los elementos examinados. Estos cuatro componentes revelan patrones notables en la relación entre los elementos, lo que se refleja en la tendencia compartida de crecimiento. Esta relación se evidencia de manera más comprensible a través de gráficos que representan la variación de los elementos a lo largo de los componentes principales, la reación se la ralizó con respecto a PCA1. Estos gráficos ilustran cómo los elementos convergen y se expanden conjuntamente en función de los valores de los componentes, brindando una representación visual clara y

reveladora de las relaciones subyacentes identificadas mediante el análisis de componentes principales.

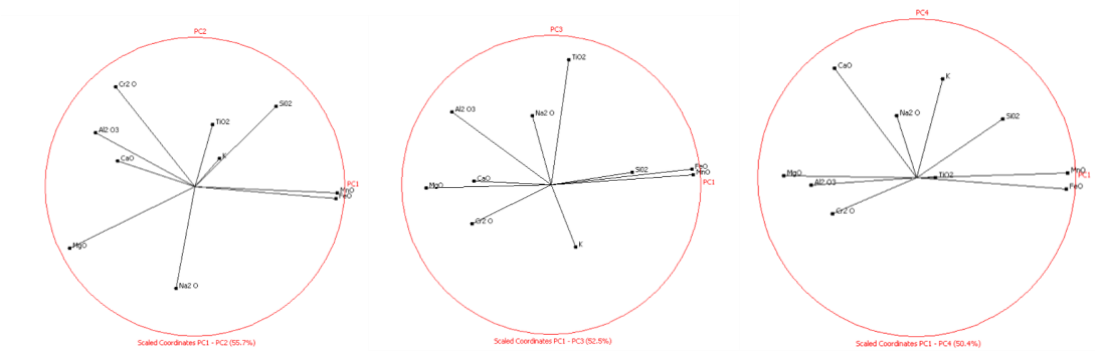

Ilustración 4. Diagrama PCA con respecto al PCA1

Los valores de los óxidos principales, como el dióxido de silicio (SiO2) y el dióxido de manganeso (MnO), en la composición de un balsalto, han demostrado tener una relación directamente proporcional con el conjunto de otros óxidos presentes en la misma. Esto significa que a medida que aumentan las concentraciones de SiO2 y MnO en la roca, también tienden a aumentar las concentraciones de otros óxidos como el óxido de hierro (FeO), el óxido de magnesio (MgO), el óxido de calcio (CaO), el óxido de aluminio (Al2O3) y otros componentes.

K-means

k-means es una técnica de agrupación que divide un conjunto de puntos de datos en k grupos para el aprendizaje automático y la minería de datos, cada punto de datos se asigna iterativamente al centroide del grupo más cercano como parte de la operación del algoritmo, y el centroide del grupo para cada grupo luego se recalcula utilizando las nuevas asignaciones, hasta que los centroides dejen de desplazarse o se alcance un número predeterminado de iteraciones, se repite el proceso. en el análisis de datos exploratorios, k-means se utiliza con frecuencia para encontrar patrones y agrupaciones en grandes conjuntos de datos (Lindsay, 2020). La agrupación de k-medias se aplicó en el contexto de este trabajo para distinguir diferentes unidades litológicas en función de sus firmas geoquímicas.

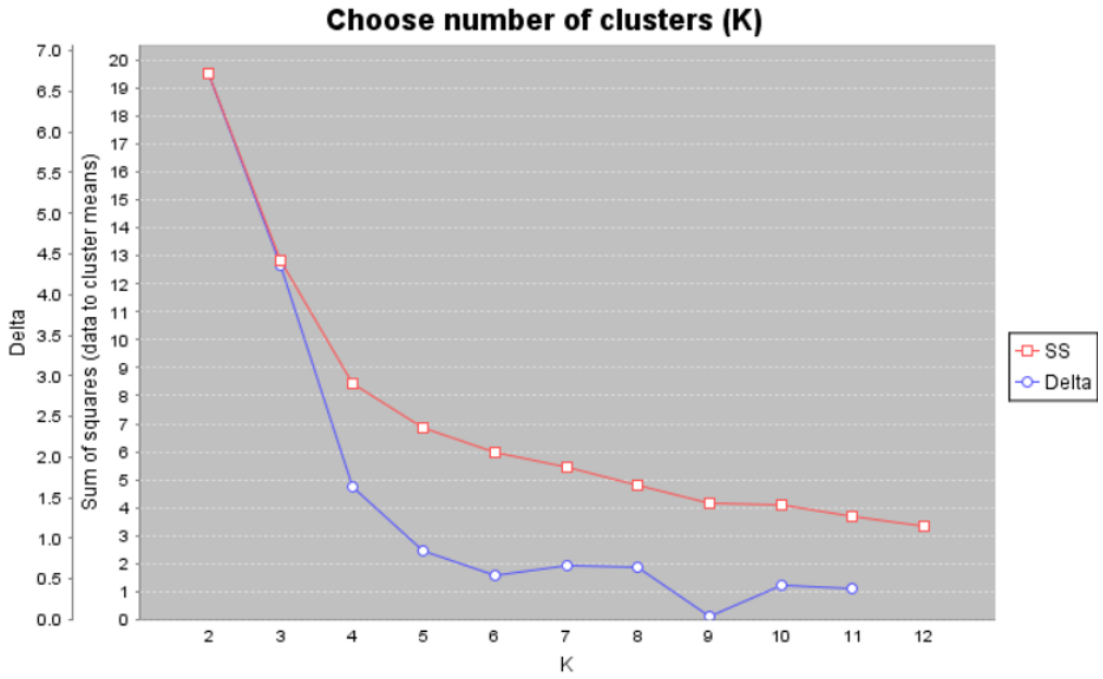

Ilustración 5. Determinación del numero de clusters.

El valor de k, que en este caso se estableció en k=3, fue determinado a través de un análisis de la curva de deflexión en el proceso de selección de clusters. Al examinar el gráfico de la curva de deflexión, se identificó un punto de inflexión donde la curva comienza a cambiar después de un descenso inicial pronunciado. Este punto de inflexión marcó la elección del número óptimo de clusters, en este caso, tres clusters. Esta decisión se basó en la noción de que, a medida que se agregan más clusters, las ganancias en la varianza explicada se vuelven menos significativas en comparación con el aumento en la complejidad. por lo tanto, la deflexión en la curva indicó un equilibrio donde se obtiene una cantidad razonable de clusters que maximizan la capacidad del modelo para explicar la variabilidad de los datos, al mismo tiempo que se evita una complejidad innecesaria.

Clusters Analys

El análisis de clusters, una técnica estadística multivariable, se puede utilizar en cualquier conjunto de datos para identificar agrupaciones de muestras sin conocer sus relaciones espaciales o temporales entre sí. Este analsisi se utiliza para encontrar patrones y estructuras en los datos que a simple vista pueden ser invisibles. En algunas situaciones, el análisis de componentes principales solo puede revelar una estructura de datos limitada, y el análisis de clusters puede ser necesario para encontrar patrones más sutiles (Gazley, 2015).

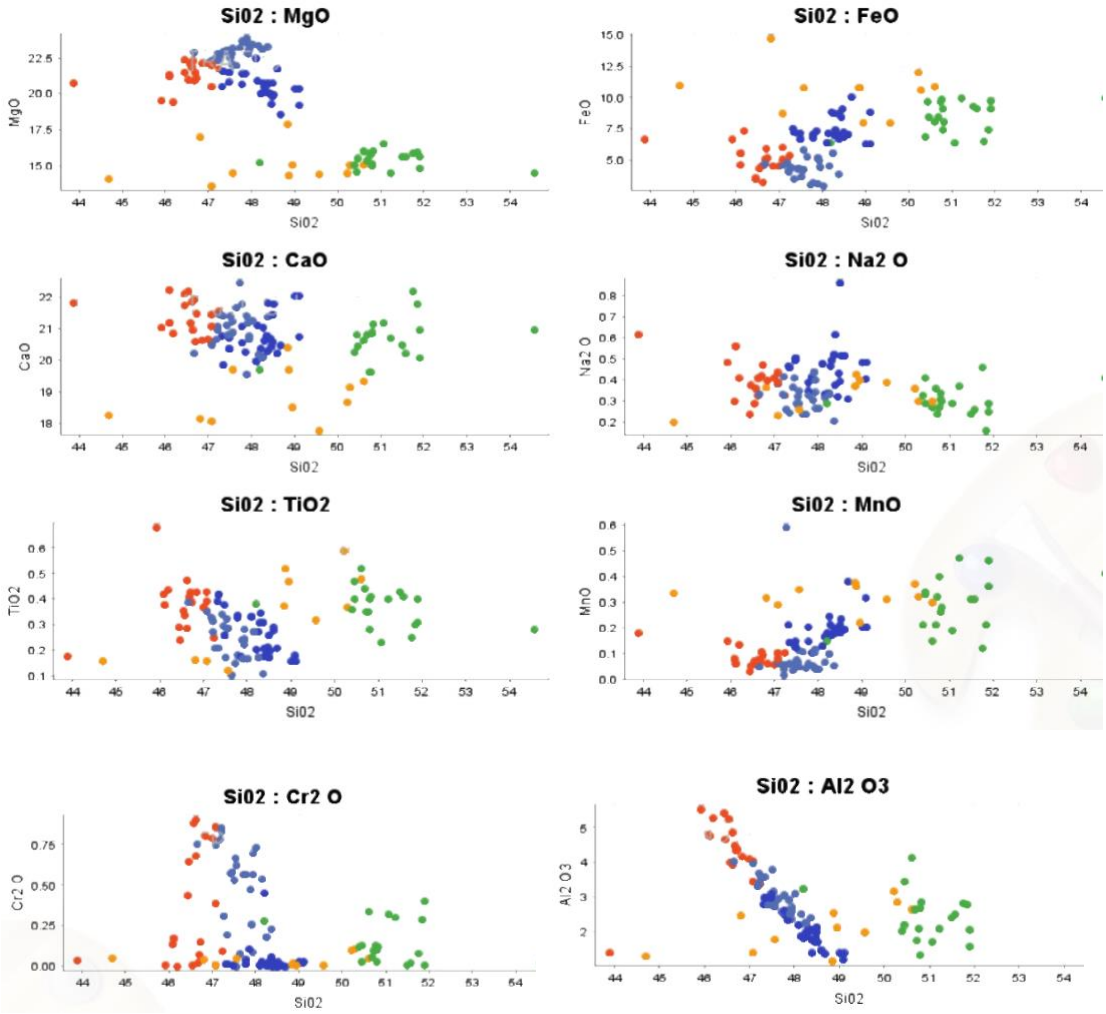

Ilustración 6. Clusters con respecto a SiO2

Los análisis de clusters realizados en relación con la variable SiO2 han proporcionado resultados esclarecedores que permiten interpretar el tipo de roca en cuestión como predominantemente básica en naturaleza. Estos resultados sugieren una característica básica en las muestras, lo cual apunta hacia una composición con rasgos claramente basálticos. Sin embargo, se observa cierta variabilidad dentro de este grupo que sugiere la presencia de una composición andesítica. La combinación de estas características nos conduce a una interpretación más matizada de las rocas estudiadas, que se alinean con una naturaleza basal basáltica, pero con indicios de una influencia andesítica en su composición general.

Esta revelación es significativa en el contexto geológico, ya que nos brinda una comprensión más completa de la diversidad de las rocas en la región estudiada. La coexistencia de características basálticas y andesíticas en las muestras sugiere una historia geológica rica y compleja, con procesos magmáticos variados que han dado forma a la composición de las rocas en la zona. Esta conclusión subraya la importancia de llevar a cabo análisis detallados de las

propiedades químicas para comprender de manera más precisa la génesis y la evolución geológica de esta región.

DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

El analisis llevado a cabo muestra resultados concluyentes que permiten caracterizar la composición de las rocas en cuestión como de origen volcánico intrusivo. Tras un análisis minucioso de las propiedades geoquímicas, se evidenció de manera consistente una serie de características que respaldan esta clasificación. La presencia de determinados minerales y la disposición de los componentes en la matriz de las rocas revelan una intrusión magmática en la corteza terrestre. Estos resultados se alinean de manera coherente con la clasificación de basaltos, sugiriendo que las rocas estudiadas presentan las características distintivas típicas de esta formación geológica.

El reconocimiento de estas características es un hito significativo en el entendimiento de la geología de la región, ya que contribuye a una interpretación más precisa de la historia geológica local. La identificación de la composición basáltica refleja la actividad magmática que ocurrió en profundidad y que posteriormente emergió en forma de rocas volcánicas. Este hallazgo proporciona una base sólida para futuras investigaciones sobre la evolución geológica de la zona y abre puertas a un análisis más profundo de los procesos geodinámicos que dieron forma a esta región volcánica.

La identificación de rocas basálticas con cierta composición andesítica es un hallazgo importante en el estudio geológico. Sin embargo, es crucial destacar que este tipo de rocas no es el ambiente propicio para la formación de minerales económicamente rentables. Las características químicas y mineralógicas de las rocas influyen en la posibilidad de que se desarrollen depósitos minerales valiosos. En el caso de las rocas basálticas, la composición no es adecuada para la concentración y acumulación de minerales de interés económico (Mahmut, 2023). Por lo tanto, aunque estas rocas pueden tener un valor geológico y científico, no presentan el potencial necesario para albergar depósitos minerales significativos desde una perspectiva económica.

Además, es esencial considerar el ambiente de formación de estas rocas. Los procesos geológicos que dieron origen a estas rocas no implicaron condiciones adecuadas para la concentración y acumulación de minerales en niveles rentables. Por lo tanto, aunque puedan presentar ciertas características de interés, no satisfacen los requisitos geológicos necesarios para albergar minerales con potencial comercial. Aunque el estudio de estas rocas contribuye a nuestro entendimiento geológico, no son el tipo de roca adecuado para la formación de depósitos minerales de valor económico (Mahmut, 2023).

La identificación de rocas basálticas con composición andesítica al no ser adecuadas para la formación de minerales económicamente rentables, es fundamental considerar otras áreas geológicas que puedan albergar depósitos minerales valiosos. Una solución efectiva podría ser llevar a cabo un análisis más

exhaustivo y detallado de diferentes regiones geológicas con características químicas y mineralógicas propicias para la formación de minerales rentables. Esto implicaría una investigación geológica más amplia que identifique áreas con las condiciones adecuadas para la concentración y acumulación de minerales valiosos.

Además, sería recomendable utilizar técnicas geofísicas y geoquímicas avanzadas para identificar zonas donde la probabilidad de encontrar depósitos minerales rentables sea más alta. La combinación de datos geológicos, geofísicos y geoquímicos puede proporcionar información valiosa sobre la distribución y concentración de minerales en las diferentes áreas estudiadas. Esto permitiría enfocar los esfuerzos de exploración en lugares con un mayor potencial para el descubrimiento de minerales de valor económico.

La solución al desafío planteado sería ampliar el alcance de la investigación geológica, utilizando técnicas avanzadas de análisis y exploración para identificar áreas con las condiciones geológicas adecuadas para la formación de depósitos minerales valiosos. Esto garantizaría una búsqueda más eficiente y efectiva de recursos minerales que puedan tener un impacto económico positivo.

CONCLUSIONES Y RECOMENDACIONES

El análisis de clusters es una técnica valiosa para identificar patrones y agrupaciones en los datos geoquímicos. La visualización de estos datos, a partir de gráficos de dispersión y Q-Q plots, es esencial para comprender la estructura de los datos y detectar anomalías, relaciones y tendencias. Estas herramientas proporcionan una comprensión más profunda de las relaciones entre variables y grupos de datos.

El análisis realizado muestra resultados imporantes que facilitan la caracterización de la composición de las rocas estudiadas, originadas por procesos volcánicos intrusivos. Estos hallazgos están en línea con la clasificación de basaltos y sugieren que las rocas analizadas tienen características únicas pertenecientes a la formación geológica.

La combinación de propiedades químicas y mineralógicas presentes en los basaltos andesíticos no coincide con las condiciones ideales para la acumulación y concentración de minerales de valor económico. Aunque estas rocas pueden tener relevancia en contextos geológicos y académicos, no son adecuadas como huéspedes para la formación de depósitos minerales explotables económicamente en la zona estudiada.

Se recomienda encarecidamente la realización de estudios geológicos complementarios, como investigaciones geofísicas y exploración detallada, con el propósito de obtener una evaluación integral de la viabilidad minera en el área de estudio. Aunque los análisis geoquímicos han proporcionado una valiosa comprensión de la composición de las rocas y sus características, es crucial contar con datos adicionales para tomar decisiones informadas sobre la rentabilidad de la minería en esta región. La aplicación de técnicas geofísicas permitiría mapear las características subsuperficiales y revelar posibles estructuras geológicas, lo que podría arrojar luz sobre la presencia de depósitos minerales de valor económico. Asimismo, una exploración más detallada en el terreno proporcionaría información concreta sobre la concentración y distribución de minerales en la zona.

Se propone la implementación de una política pública que priorice el uso del análisis geoquímico de rocas como factor clave en la toma de decisiones sobre exploración minera. El objetivo principal de esta política es evaluar la idoneidad de una región para la actividad minera de una manera científica, garantizando al mismo tiempo un enfoque responsable y sostenible para la explotación de recursos. El análisis geoquímico de rocas se utiliza para identificar la presencia de minerales económicamente valiosos y evaluar su valor económico potencial determinando la composición química y mineralógica de los depósitos presentes en una región. Esta información permitirá a las agencias gubernamentales y a la industria minera tomar decisiones informadas sobre la viabilidad económica y ambiental de los proyectos mineros, asegurando la protección ambiental y maximizando los beneficios económicos para las comunidades locales y la nación en su conjunto. La implementación de esta política fomentaría la gestión responsable de los recursos minerales y apoyaría el crecimiento sostenible de la

industria minera en armonía con el medio ambiente y las comunidades afectadas.

REFERENCIAS

Castillo, B. (2023). *Automatización del análisis exploratorio de datos y procesamiento geoquímico univariado empleando Python.*

Lindsay, J. (2020). *A machine learning approach for regional geochemical data: Platinum-group element geochemistry vs geodynamic settings of the North Atlantic Igneous Province.*

Grunsky, C. (2014). *A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping.*

Gazley, C. (2015). *Application of principal component analysis and cluster analysis to mineral exploration and mine geology .*

Maimon, O. a. (2010). *Data Mining and Knowledge Discovery Handbook.* Israel: Oded Maimon and Lior Rokach Tel-Aviv University.

Pérez, L. (2022). *Análisis de clústers K-means.*

Grunsky, E. (2012). The interpretation of geochemical survery data. *Geochemistry: Exploration, Enviroment, Analysis 2012*, 27-74.

Davies, D. (1979). A Cluster Separation Measure(Article). En *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vols. PAMI-1, págs. 224-227).

Balamurali, M. (2016). t-SNE Based Visualisation and Clustering of Geological Domain. *Neural Information Processing*, 565-572.

Farnham, I. (2002). Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics and Intelligent Laboratory Systems*, 265-281.

Ghannadpour, S. (2013). An Investigation of Pb Geochemical Behavior Respect to Those of Fe and Zn Based on k-Means Clustering Method. En *Quartery Journal of Tethys* (Vol. 1, págs. 291-2013).

Cahoon, J. (2021). Generalized inferential models for censored data. En I. J. Reasoning.

Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer.

Mahmut, M. (2023). *Basalt.*