



Aplicación del método normativo de CIPW para la denominación de rocas ígneas empleando algoritmos de IA

Zavaleta, J.

Petrología de Rocas Ígneas y Metamórficas (GE433). Sección C. MSc. Ing. César Augusto Mendoza Tarazona

Escuela Profesional de Ingeniería Geológica

Resumen: El presente proyecto forma parte de un trabajo sostenido orientado a integrar la petrología clásica con técnicas modernas de análisis computacional. Se aplica el método normativo CIPW —propuesto por Cross, Iddings, Pirsson y Washington— como herramienta fundamental para caracterizar y comparar composiciones de rocas ígneas mediante minerales normativos derivados de óxidos mayores. Inicialmente se compilieron 3358 muestras de rocas ígneas del Perú provenientes de fuentes reconocidas. Tras un proceso riguroso de QA/QC, estandarización litológica y depuración, el conjunto final utilizado en los algoritmos quedó conformado por 2707 muestras válidas.

Los minerales obtenidos conforme al método normativo CIPW fueron generados mediante la plataforma ShinyNORRRM, obteniéndose parámetros como Q, Cor, Ort, Ab, An, Ne, Le, Ap, ITT y %An en Plg, entre otros, incorporados como features en los algoritmos de IA. Se entrenaron Random Forest, XGBoost y SVM con PCA mediante validación cruzada estratificada.

Los modelos alcanzaron un alto rendimiento en rocas volcánicas ($CA \approx 0.80-0.82$) y un rendimiento moderado en rocas plutónicas ($CA \approx 0.52-0.56$), coherente con su respectiva complejidad mineralógica. La metodología aplicada actúa como una herramienta de apoyo petrológico, integrando razonamiento geoquímico con análisis computacional para la denominación y comparación sistemática de rocas ígneas.

Abstract: The present project is part of an ongoing effort aimed at integrating classical petrology with modern computational analysis techniques. The CIPW normative method—proposed by Cross, Iddings, Pirsson, and Washington—is applied as a fundamental tool to characterize and compare igneous rock compositions through normative minerals derived from major oxides. Initially, 3358 igneous rock samples from Peru were compiled from recognized sources. After a rigorous QA/QC process, lithological standardization, and data cleaning, the final dataset used in the algorithms consisted of 2707 valid samples.

The minerals obtained through the CIPW normative method were generated using the ShinyNORRRM platform, yielding parameters such as Q, Cor, Ort, Ab, An, Ne, Le, Ap, ITT, and %An in Plg, among others, which were incorporated as features in the AI algorithms. Random Forest, XGBoost, and SVM with PCA were trained using stratified cross-validation.

The models achieved high performance for volcanic rocks ($CA \approx 0.80-0.82$) and moderate performance for plutonic rocks ($CA \approx 0.52-0.56$), consistent with their respective mineralogical complexity. The applied methodology serves as a petrological support tool, integrating geochemical reasoning with computational analysis for the systematic naming and comparison of igneous rocks.

1. INTRODUCCIÓN

La denominación de rocas ígneas constituye un proceso que integra observación petrográfica, relaciones de campo y análisis geoquímicos. Entre las herramientas clásicas para relacionar la composición química con la clasificación litológica se encuentra el método normativo CIPW, formulado por Cross, Iddings, Pirsson y Washington. Este procedimiento convierte los óxidos mayores en un conjunto coherente de minerales normativos teóricos, útiles para comparar rocas y series evolutivas.

Los avances recientes en inteligencia artificial permiten investigar cuantitativamente la relación

entre la composición química, los minerales normativos y los nombres litológicos suministrados por repositorios internacionales. En este contexto, el presente estudio forma parte de un trabajo continuado orientado a evaluar la posibilidad de la integración del método CIPW y los algoritmos de IA para asistir en la clasificación litológica de rocas ígneas volcánicas y plutónicas.

2. MATERIALES Y MÉTODOS

2.1. *Softwares utilizados*

El proyecto de investigación hizo uso e integración de diversas herramientas para las etapas de cálculo

normativo, procesamiento geoespacial, administración de bases de datos y análisis predictivo. ShinyNORRRM constituyó la base normativa petroquímica del estudio, pues se empleó para el cálculo de minerales normativos mediante el método CIPW a partir de los óxidos mayores; este aplicó una serie ordenada de reacciones que distribuyen SiO_2 , Al_2O_3 , CaO , Na_2O , K_2O , FeO – Fe_2O_3 , MgO y otros componentes en minerales normativos como Q, Ne, Ab, An, Or, Di, Hy, Ol, Le, Ap, Ilm, Mag, Hem, Cor, Ac, entre muchos otros.

QGIS, como software GIS de código abierto, fue empleado para la revisión espacial del dataset, la validación de ubicaciones y el control de coherencia geológica. Asimismo, permitió integrar operaciones menores de geomática, tales como la reproyección y corrección de coordenadas bajo el sistema de referencia geodésico WGS84 (EPSG:4326) a fin de mantener consistencia en la georreferenciación de las muestras provenientes de USGS, GEOROC e INGEMMET en la zona de estudio. Adicionalmente, QGIS facilitó el manejo estructurado de la base de datos georreferenciada depurada, integrando atributos espaciales usados en este estudio como features dentro del flujo de trabajo analítico; fueron, sin embargo, algunos calificados como metadatos no usados en el procesamiento con machine learning. Dichos atributos serían aplicables y ampliables en un conjunto más robusto de metadatos para investigaciones futuras, lo cual enriquecería la reproducibilidad, precisión regional y capacidad de generalización. Esta proyección contempla la incorporación de features adicionales para su integración en un entorno SIG apoyado por las herramientas que provee el machine learning, permitiendo validaciones regionales más robustas y no solo limitadas a la zona de estudio en cuestión.

La organización y estructuración de la base de datos se realizaron en Microsoft Access donde se establecieron las tablas maestras, las relaciones entre campos y los esquemas de integridad referencial. Mientras que el preprocesamiento y la depuración avanzada (incluyendo estandarización, control de calidad y transformación de variables) se llevaron a cabo sincrónica y exhaustivamente en las plataformas Microsoft Excel y Python, aprovechando su capacidad para operaciones tabulares detalladas y procesamiento automatizado en gran escala.

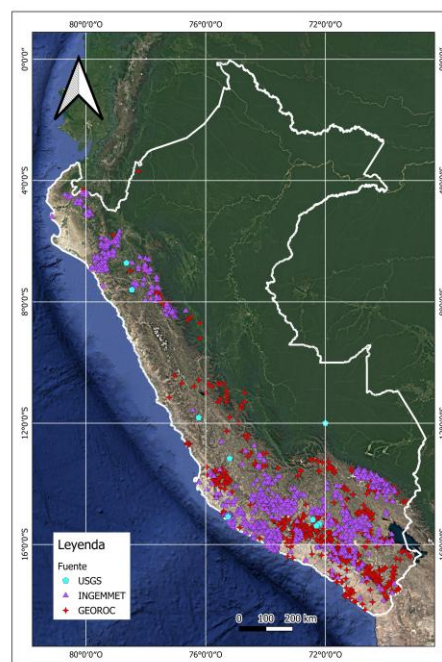
El entrenamiento, validación y evaluación de los algoritmos supervisados se desarrollaron en Orange Data Mining, aplicando Random Forest, XGBoost y SVM con PCA, junto con sus variables de entrada (features) y métricas de desempeño (CA, AUC, F1-Score, Precision, Recall, MCC). En conjunto, estas herramientas permitieron establecer un flujo de trabajo integral, reproducible y orientado al análisis geoquímico y petrológico.

2.2. Obtención de las muestras de estudio

El punto de partida fueron los datasets liberados de tres repositorios reconocidos: el United States Geological Survey (USGS), la base de datos GEOROC (Geochemistry of Rocks of the Oceans and Continents) y el Instituto Geológico, Minero y Metalúrgico del Perú (INGEMMET). Para efectos del estudio, estas bases se consideraron compatibles porque, aunque cada una proviene de instituciones con protocolos analíticos y formatos propios, todas comparten un mismo concepto operativo: son repositorios geoquímicos diseñados para reportar óxidos mayores y metadatos litológicos bajo estándares internacionalmente utilizados. Es decir, no se trata de “sumar elementos inconexos”, sino de integrar distintas fuentes que pertenecen a la misma categoría de información y cuya comparabilidad es alcanzable mediante un proceso previo de estandarización.

Posteriormente, para garantizar la coherencia espacial del estudio, las bases de datos fueron unificadas homogeneizando formatos y nomenclaturas. Finalmente, se aplicó un filtro geográfico para restringir el estudio únicamente al territorio peruano. Esto permitió consolidar un conjunto inicial de 3358 muestras potenciales, representativas de la amplia diversidad de litologías ígneas del país; constituyendo, a su vez, el insumo fundamental sobre la cual las siguientes etapas de QA/QC, depuración geoquímica y estandarización litológica fueron ejecutadas.

Figura 1. Mapa de la distribución geográfica del dataset sin depurar.



Fuente: Elaboración propia

Tras un riguroso QA/QC, que involucró la verificación de la sumatoria de óxidos dentro de intervalos estequiométricamente aceptables para análisis ICP-AES/ICP-MS, depuración de registros redundantes y outliers geoquímicos (valores anómalos), exclusión de muestras afectadas por alteración o meteorización evidente, entre otros, el conjunto de datos para el procesamiento final con machine learning quedó reducido a 2707 registros. En la siguiente sección se detallan los criterios.

2.3. Depuración

2.3.1. Caracterización preliminar del dataset litogeoquímico

La inspección generalizada conjunta de los nombres de roca (Rock_Name) y su clasificación química derivada del contenido de SiO_2 (Clase_ SiO_2) demuestra que el dataset presenta una estructura coherente y representativa de la geología ígnea del Perú. Las rocas intermedias, especialmente las andesitas, constituyen la clase dominante, mientras que las ácidas (granitos, monzogranitos, granodioritas, riolitas) muestran una representación amplia y equilibrada. Las básicas (basaltos, gabros, doleritas, basanitas) aparecen en menor proporción, pero mantienen coherencia geoquímica.

La categoría ultramáfica, en cambio, exhibe una frecuencia prácticamente nula. Debido a esta baja representatividad y con fines de mejorar la estabilidad estadística del modelamiento, se decidió integrar las muestras ultramáficas dentro de la clase “básica”, lo cual es metodológicamente aceptable cuando el objetivo es entrenar modelos supervisados con distribución equilibrada entre clases.

2.3.2. Sobre la unificación lingüística y target

Las bases de datos utilizadas presentaban una fuerte variabilidad en los nombres litológicos, tanto por diferencias idiomáticas (inglés/español) como por enfoques petrográficos distintos. Para evitar inconsistencias terminológicas que afectarían la clasificación, se aplicó una unificación lingüística, traduciendo y estandarizando todos los nombres a un formato coherente en español.

Durante este proceso se separó el nombre en dos campos:

- Rock_Name (target): litología principal, definida exclusivamente según criterios geoquímicos.
- Sec_Rock_Name: descriptor secundario preservado como información contextual litológica.

El target se definió únicamente en función de la coherencia petroquímica y la disponibilidad de datos, manteniendo clases con suficiente soporte estadístico (≥ 10 –15 muestras) y reclasificando litologías ambiguas o texturales —como tobas— hacia su equivalente composicional (p. ej., andesita, dacita, riolita). Esto garantizó una clasificación estable, reproducible y científicamente consistente.

2.3.3. Rango aceptable de óxidos mayores

Se aplicaron umbrales derivados de la literatura geoquímica clásica (Middlemost, 1989; Rollinson, 1993; Le Bas et al., 1986; Le Maitre et al., 2002), ajustados a la realidad de rocas ígneas del arco magmático andino. Fue así como finalmente solo se conservaron muestras cuyos valores se encontraban dentro de los siguientes intervalos:

(a) SiO_2 : 35–82 wt. %

- $< 35 \%$ → posible roca ultramáfica altamente alterada o error analítico.
- 82% → composiciones fuera del rango magmático natural, típicamente producto de alteración hidrotermal ($>$ silica flooding), acumulación de cuarzo o errores de normalización.
- Los valores > 78 –80 % fueron revisados individualmente y descartados si mostraban pérdida significativa de FeOT y CaO.

(b) Na_2O y K_2O : ≥ 0.05 wt. %

Valores extremadamente bajos ($< 0.05 \%$) son geoquímicamente imposibles para rocas ígneas frescas, indicando:

- Alteración hidrotermal avanzada
- Pérdida de álcalis en vidrios volcánicos
- Errores de laboratorio

(c) $(\text{Na}_2\text{O} + \text{K}_2\text{O})$: 0.5–13.5 wt. %

Este es el rango definido por el diagrama TAS (Le Bas et al., 1986). Intervalos no aceptados:

- $< 0.5 \%$ sugiere rocas excesivamente empobrecidas en álcalis (alteración o muestras no representativas).
- $> 13.5 \%$ señala composiciones anómalas, probablemente error o alteración potásica extrema.

(d) $\text{P}_2\text{O}_5 \leq 2.5$ wt. %

Rocas ígneas comunes rara vez superan 1 %. Se permitió hasta 2.5 % para no excluir rocas alcalinas raras o lamproitas, pero valores mayores se descartaron por representar fosforitas o contaminación sedimentaria.

(e) Suma de óxidos: 90–110 %

Control de balance estequiométrico:

- $< 90 \%$ sugiere exceso de volátiles o pérdida de elementos.
- $> 110 \%$ demuestra un claro error de análisis o doble conteo de Fe.

Aunque la normalización CIPW ajusta automáticamente a 100 %, es conveniente verificar consistencia previa del dataset.

2.3.4. Criterio de descarte según LOI (Loss On Ignition)

El metadato LOI se empleó como indicador de alteración. De acuerdo con Rollinson (1993) y Middlemost (1989):

-0.5% $<$ LOI $<$ 4 % constituye el rango normal en rocas ígneas frescas.

LOI > 4–5 % indica alteración hidrotermal, hidratación, arcillificación o carbonatación.

LOI < -0.5 % señala un error analítico.

Estas muestras distorsionan el cálculo CIPW, ya que dicho método asume una composición anhidra ideal sin volátiles.

2.3.5. Exclusión de composiciones químicamente imposibles o contradictorias

Se realizaron revisiones caso por caso en muestras cuyos óxidos sugerían composiciones incompatibles con sus nombres petrográficos o con rocas ígneas reales.

Ejemplos representativos del dataset:

(a) Basaltos con $\text{SiO}_2 = 85\text{--}91$ wt.% son excluidos

Un basalto no puede superar 55 % de SiO_2 . Valores tan altos indican silicificación, vidrio volcánico alterado o error analítico.

(b) Dioritas con $(\text{Na}_2\text{O} + \text{K}_2\text{O}) < 0.1$ wt.% fueron excluidas

Las dioritas naturalmente contienen 3–7 % de álcalis. Valores menores a 1 % son incompatibles con su mineralogía modal.

2.3.6. Casos especiales

Litologías con frecuencias ínfimas (menores a 10) fueron agrupadas dentro de sus familias respectivas. Un algoritmo supervisado no puede aprender una clase que solo tiene 1, 2, 3 o incluso 5 muestras; esto produciría overfitting extremo, matrices de confusión inestables, límites de decisión erráticos. De manera que, si no se procede con este preprocesamiento, el algoritmo no memoriza: generaliza. Distorsiona scores globales y compromete la validación científica del trabajo de investigación, pues no alcanza el umbral de estabilidad estadística.

Se tuvo así, en las rocas plutónicas: Monzogabro reclasificado como Gabro, Dolerita como Gabro, Sienita como Monzonita, Lamproita como Lamprófito (este al no ser claro plutónico o volcánico fue asignado a plutónico); análogamente, para rocas volcánicas: Traquibasalto como Basalto, Basanita como Basalto, Traquita como Riodacita, Traquidacita como Traquita y esta finalmente como Riodacita por no llegar al número suficiente de muestras.

2.3.7. Quality flagging

Se implementó un sistema de control (*Quality_Flag*) para detectar muestras inconsistentes antes del cálculo CIPW y del entrenamiento. La función clasificó los análisis en tres niveles:

- FLAG = 2 (excluida): Muestras con anomalías severas, tales como LOI > 5 %, $\text{SiO}_2 > 82$ %, $\text{Na}_2\text{O} + \text{K}_2\text{O} < 0.5$ %, Sum_Oxs fuera de 90–110 %, coexistencia de Q con Ne o Le, o Q > 90 %. Estas condiciones reflejan composiciones no confiables, por lo que fueron descartadas.
- FLAG = 1 (sospechosa): Composiciones atípicas pero potencialmente

válidas, caracterizadas por relaciones anómalas entre SiO_2 y Q, $\text{P}_2\text{O}_5 > 2$ %, alcalinidad marginal (0.5–1 %) o SiO_2 muy alto (78–82 %). Estas muestras se conservaron con precaución interpretativa.

- FLAG = 0 (normal): Muestras sin las condiciones anteriores, coherentes desde el punto de vista petroquímico y aptas para el modelamiento.

Tras aplicar el sistema al conjunto depurado, todas las muestras resultaron FLAG = 0, confirmando la ausencia de inconsistencias químicas severas en el dataset final utilizado para el análisis supervisado.

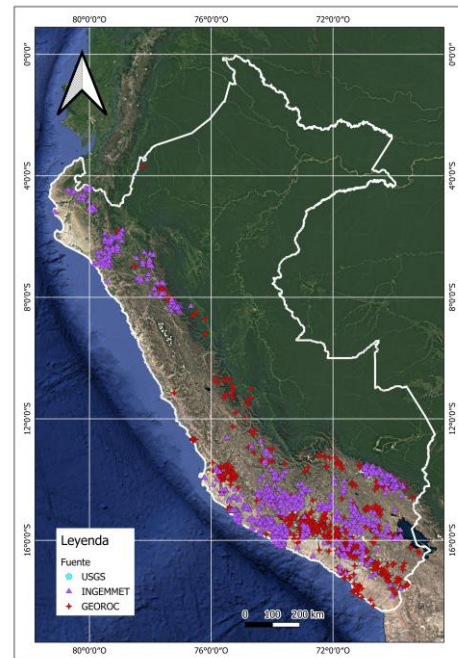
2.3.8. Resultado de la depuración

El dataset depurado retuvo exclusivamente aquellas muestras que cumplieron los requisitos geoquímicos y petrográficos necesarios para asegurar la consistencia del análisis posterior. En particular, se conservaron únicamente las muestras aptas para:

- Cálculo confiable de la norma CIPW.
- Entrenamiento de algoritmos de aprendizaje automático.
- Clasificación geoquímica confiable dentro del rango magmático natural.

Como producto se tuvo un dataset geoquímicamente coherente, representativo de rocas ígneas frescas y estructurado para su uso en análisis estadísticos multivariados, cálculos normativos y procesamientos con inteligencia artificial orientados a la caracterización y denominación litológica.

Figura 2. Mapa de la distribución geográfica del dataset depurado.



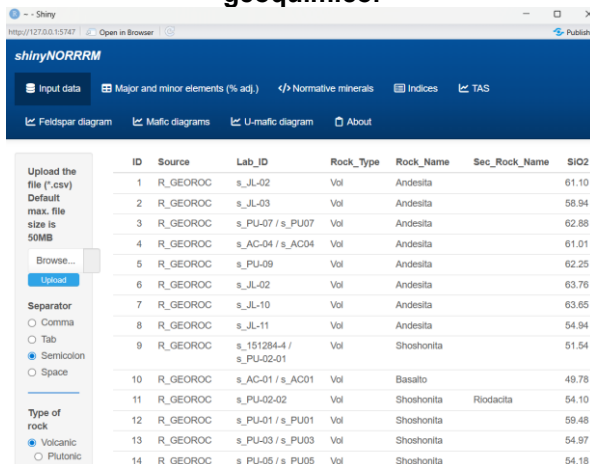
Fuente: Elaboración propia

2.4. Cálculo normativo CIPW con ShinyNORRRM

El cálculo de minerales normativos mediante el método CIPW fue la etapa central de este estudio; de esta manera, las composiciones químicas eran convertidas en supuestos teóricos minerales. Este procedimiento, formulado por Cross, Iddings, Pirsson y Washington, constituye una de las herramientas clásicas de la petrología ígnea para analizar afinidades magmáticas, tendencias evolutivas y consistencia mineralógica.

El cálculo se realizó exclusivamente mediante la plataforma ShinyNORRRM, diseñada para ejecutar la norma CIPW bajo los criterios estandarizados por las recomendaciones de la IUGS; es también considerada en estas las adaptaciones propuestas para rocas modernas en estudios geoquímicos. Esta herramienta permitió automatizar simultáneamente las operaciones de balance estequiométrico, distribución de cationes y construcción sistemática de la mineralogía normativa en un entorno simulado.

Figura 3. Interfaz inicial de ShinyNORRRM durante la carga y visualización del dataset geoquímico.



ID	Source	Lab_ID	Rock_Type	Rock_Name	Sec_Rock_Name	SiO2
1	R_GEOROC	s_JL-02	Vol	Andesita		61.10
2	R_GEOROC	s_JL-03	Vol	Andesita		58.94
3	R_GEOROC	s_PU-07 / s_PU07	Vol	Andesita		62.88
4	R_GEOROC	s_AC-04 / s_AC04	Vol	Andesita		61.01
5	R_GEOROC	s_PU-09	Vol	Andesita		62.25
6	R_GEOROC	s_JL-02	Vol	Andesita		63.76
7	R_GEOROC	s_JL-10	Vol	Andesita		63.65
8	R_GEOROC	s_JL-11	Vol	Andesita		54.94
9	R_GEOROC	s_151284-4 / s_PU-02-01	Vol	Shoshonita		51.54
10	R_GEOROC	s_AC-01 / s_AC01	Vol	Basalto		49.78
11	R_GEOROC	s_PU-02-02	Vol	Shoshonita	Riodacita	54.10
12	R_GEOROC	s_PU-01 / s_PU01	Vol	Shoshonita		59.48
13	R_GEOROC	s_PU-03 / s_PU03	Vol	Shoshonita		54.97
14	R_GEOROC	s_PU-05 / s_PU05	Vol	Shoshonita		54.18

Fuente: Elaboración propia

2.4.1. Reacciones normativas en ShinyNORRRM

El algoritmo CIPW sigue la secuencia clásica:

- I. Precipitación de Ap ($P_2O_5 + CaO$).
- II. Formación secuencial de Ilm, Mag, Hem según TiO_2 y Fe.
- III. Asignación de Al_2O_3 para formar Cor (exceso) o feldespatos (déficit).
- IV. Cálculo de Ort, Ab, An a partir de álcalis y CaO.
- V. Distribución de Fe y Mg en Hy, Ol, Di, Wo según saturación en SiO_2 .
- VI. Asignación de Ne, Le, La, Kp para sistemas subsaturados.
- VII. Formación de fases accesorias: Ru, Ac, Na_ms, Pwk, Cc, Ft, Py cuando se cumplen los criterios geoquímicos necesarios.

2.4.2. Parámetros normativos obtenidos

ShinyNORRRM generó para cada muestra **todas las fases normativas necesarias**, incluyendo minerales principales, accesorios y parámetros derivados.

Estos parámetros constituyeron uno de los aportes más sólidos del pipeline, al capturar relaciones internas (p. ej., saturación en sílice, grado de alcalinidad, equilibrio feldespático) imposibles de interpretar directamente solo mediante óxidos mayores.

2.4.3. Integración a los modelos supervisados

Los parámetros normativos se incorporaron como atributos predictivos en Orange, donde su importancia quedó reflejada en:

- Separación andesita–dacita–riolita explicada por Q, Ab, An, Ort y %An.
- Distinción basalto–basanita–traquibasalto mediante Ne, Le, Kp, Ol, Hy, Di.
- Separación granito–granodiorita–tonalita por Ort–Ab–An–Cor.
- Identificación de gabro, gabrodiorita y diorita a partir de Di, Hy, Ol, Wo.
- Captura de fases accesorias diagnósticas en rocas alcalinas (Ac, La, Na_ms, Pwk).

La combinación CIPW + ML potenció significativamente la capacidad clasificatoria.

2.4.4. Validación de coherencia petroquímica

Se establecieron verificaciones internas:

- $Q \geq 0$ en rocas saturadas; $Q + Ne + Le$ no pueden coexistir.
- Ort–Ab–An coherentes con series calcoalcalinas del arco andino.
- $Ol + Q$ descartado automáticamente.
- $Mag + Hem + Ilm$ en proporciones razonables según Fe – Ti.
- Eliminación de resultados contradictorios o imposibles.

Solo las muestras satisfactoriamente normadas fueron integradas al análisis de machine learning.

2.5. Data mining con algoritmos de machine learning en Orange

El análisis supervisado se realizó mediante la plataforma Orange Data Mining, cuyo entorno modular permite diseñar flujos de trabajo reproducibles para la predicción litológica a partir de variables geoquímicas y normativas. La finalidad del data mining fue evaluar la capacidad de distintos algoritmos para clasificar rocas ígneas peruanas usando como insumo tanto los óxidos mayores como los minerales normativos CIPW generados mediante la plataforma ShinyNORRRM.

2.5.1. Objetivo del análisis supervisado

El propósito central consistió en predecir la litología principal (Rock_Name), definida en este estudio como la categoría objetivo (target), utilizando como variables de entrada (features) los óxidos mayores depurados, los minerales normativos CIPW y otros

atributos calculados durante el preprocesamiento. La separación entre rocas volcánicas y plutónicas —necesaria por la misma naturaleza de génesis de estas— permitió entrenar modelos más coherentes con los procesos petroquímicos propios de cada ambiente.

2.5.2. Selección de features, asignación de metadatos y target

El conjunto final de atributos (*features*) empleados en los modelos supervisados estuvo compuesto por variables organizadas en tres grupos principales:

(1) Óxidos mayores (química elemental)

SiO₂, TiO₂, Al₂O₃, Fe₂O₃, FeO, MgO, MnO, CaO, Na₂O, K₂O, P₂O₅, LOI. Estos constituyen la base geoquímica fundamental sobre la cual se calcula la norma CIPW y se derivan las relaciones petroquímicas del estudio.

(2) Minerales normativos CIPW y fases derivadas

Q, Cor, Ort, Ab, An, Ne, Le, Kp, Hy, Di, Wo, Ol, La, Ac, Na_ms,

Mag, Hem, Ilm, Tit, Ru, Pwk, Ap, Cc, Ft, Py.

Este grupo incluye minerales normativos principales, accesorios, fases diagnósticas en ambientes subsaturados y especies asociadas a óxidos metálicos.

(3) Parámetros petroquímicos derivados

- %An en PI (porcentaje de anortita normativa).
- ITT (Índice Total de Titanio y Fosfatos).
- SUM_QAP (suma normativa del subtriángulo Q–A–P).
- Alc (proporción alcalina normativa)
- Clase_SiO₂ (clasificación por rango de sílice)
- Serie_Mag (clasificación magmática auxiliar).
- Sec_Rock_Name (descriptor litológico contextual secundario).

Los meta-atributos se mantuvieron únicamente para control de calidad, trazabilidad y validación externa, evitando introducir sesgos espaciales o contextuales no relevantes para la clasificación geoquímica.

(4) Target Final

- Rock_Name (litología principal). Variable objetivo del modelo supervisado, definida a partir de criterios petrográficos y petroquímicos unificados.

2.5.3. Preparación y partición del dataset

Debido a que la distribución litológica de las rocas ígneas del Perú es inherentemente desigual, se trabajó con dos subconjuntos independientes. Esta separación evita mezclar dominios petrogenéticos con comportamientos geoquímicos intrínsecamente distintos y permite que los algoritmos aprendan patrones coherentes dentro de cada familia. Así, el dataset se dividió en:

- Litologías volcánicas (n ≈ 1720)
- Litologías plutónicas (n ≈ 987)

Se descartaron categorías con menos de 10 muestras, agrupándose dentro de sus familias petroquímicas para evitar overfitting y asegurar estabilidad estadística en la validación cruzada. La validación se realizó mediante k-fold estratificado, preservando la proporción de clases en cada partición.

2.5.4. Arquitectura del pipeline en Orange

La arquitectura del pipeline implementado en Orange se diseñó para garantizar un flujo transparente, comparable y reproducible al evaluar tres de los algoritmos de machine learning más adecuados para la clasificación litológica usando data geoquímica. El esquema general se compone de seis etapas principales, que integran desde la depuración inicial del dataset hasta la evaluación final mediante matrices de confusión. La estructura del flujo corresponde a la mostrada en el diagrama adjunto.

I. File → Data Table → Select Columns

El proceso inicia con la importación del dataset depurado en formato .csv mediante el widget File, seguido de su visualización y verificación estructural en Data Table. Posteriormente, Select Columns permite:

- Definir la variable objetivo (Rock_Name o su equivalente para cada subconjunto volcan/plutón).
- Seleccionar los atributos predictivos (óxidos mayores, minerales normativos CIPW y derivados).
- Excluir columnas irrelevantes o redundantes identificadas en la fase de preprocesamiento.

Esta etapa asegura consistencia en las variables alimentadas a los modelos.

II. Preprocess (normalización, limpieza y transformación)

El widget Preprocess centraliza las operaciones de preparación de datos requeridas por los modelos:

- Estandarización automática (Z-score): aplicada como requisito para métodos basados en distancias, particularmente SVM y la posterior reducción PCA.
- Imputación automática en caso de valores faltantes residuales (no ocurrió en el dataset final).
- Codificación interna de clases para garantizar compatibilidad con los algoritmos de Orange.

Random Forest y XGBoost son menos sensibles a escalamiento, pero recibieron el mismo tratamiento para mantener comparabilidad entre modelos.

III. PCA (solo para el modelo SVM)

Para el flujo destinado al algoritmo SVM se incorporó un módulo intermedio de PCA, cuyo objetivo fue:

- Reducir la dimensionalidad del espacio original de 44 variables.
- Preservar más del 95 % de la varianza total.

- Suavizar la frontera de decisión para un clasificador que funciona mejor en espacios compactos y ortogonales.

El PCA se alimenta exclusivamente del Preprocess, y su salida es dirigida únicamente al learner SVM.

IV. Learners

El pipeline consideró tres modelos principales, conectados de manera paralela a la salida del preprocesamiento:

a. Random Forest

- Algoritmo basado en ensambles de árboles con capacidad para capturar interacciones no lineales complejas.
- Baja sensibilidad al escalado.
- Rendimiento robusto en datasets geoquímicos heterogéneos.

b. Extreme Gradient Boosting (XGBoost)

- Variante de boosting con optimización secuencial.
- Adecuado para datasets de tamaño medio y patrones altamente no lineales.
- Tiende a obtener métricas más altas cuando existen relaciones complejas entre óxidos-normativos-litoquímica.

c. SVM con PCA

- Clasificador basado en hiperplanos óptimos.
- La combinación SVM + PCA mejora estabilidad y reduce sobreajuste en espacios con alta correlación entre variables (común en geoquímica).

Cada learner recibe como entrada el flujo correspondiente (preprocesado directo o PCA).

V. Test & Score (validación cruzada estratificada)

Las tres ramas convergen en módulos Test & Score, configurados con:

- Validación cruzada estratificada de k-folds (mismo k para todos los modelos).
- Cálculo uniforme de métricas:
 - CA (Classification Accuracy)
 - AUC
 - F1-score
 - MCC (Matthews Correlation Coefficient)

El uso de estratificación asegura una representación balanceada de las litologías en cada partición, lo cual es crítico dada la desigual distribución natural de clases.

VI. Confusion Matrix (diagnóstico detallado por clase)

Finalmente, cada conjunto de resultados se envía al widget Confusion Matrix, que permite:

- Visualizar el número de aciertos y errores por litología.
- Identificar patrones de confusión entre rocas químicamente próximas (p. ej., andesita-dacita-riolita en el caso volcánico).
- Detectar subclases con desempeño inferior y guiar interpretaciones geológicas coherentes.

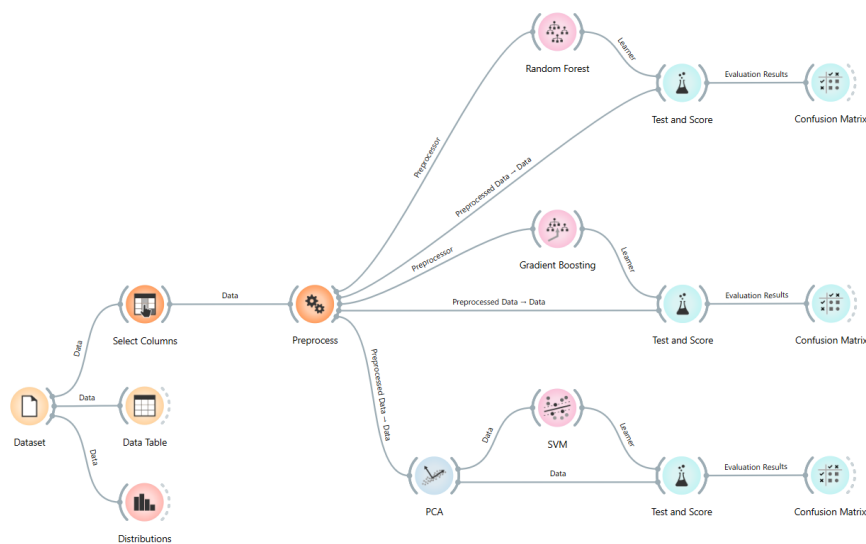
Sobre la replicación del pipeline

Todo el flujo descrito fue replicado de manera independiente para:

- *Dataset de rocas volcánicas*
- *Dataset de rocas plutónicas*

Esto garantiza modelos coherentes con los procesos petrológicos y geoquímicos de cada ambiente, evitando mezclar litologías con génesis, texturas y restricciones químicas distintas.

Figura 4. Pipeline de Procesamiento y Evaluación de Algoritmos.

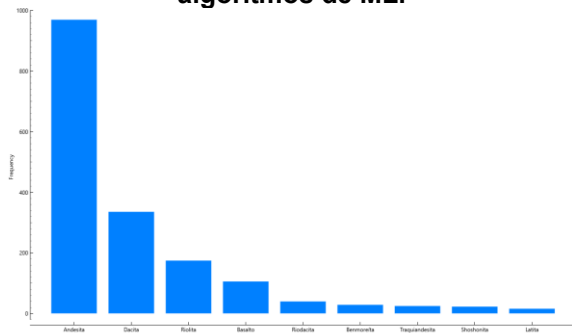


Fuente: Elaboración propia

2.5.5. Frecuencia de clases litológicas

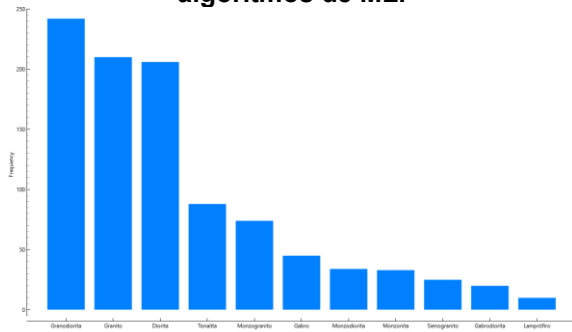
Se generaron histogramas para cuantificar la representatividad de cada litología en ambos conjuntos, lo cual permitió entender la complejidad estadística del problema:

Figura 5. Frecuencia de clases litológicas volcánicas utilizadas como categorías para los algoritmos de ML.



Fuente: Elaboración propia

Figura 6. Frecuencia de clases litológicas plutónicas utilizadas como categorías para los algoritmos de ML.



Fuente: Elaboración propia

Las rocas volcánicas, como se ha comentado, poseen transiciones composicionales más definidas (por ejemplo, de andesita a dacita y esta, a su vez, a riolita), lo cual favorece el aprendizaje.

Las plutónicas, sin embargo, se distribuyen de manera continua en el diagrama QAPF, generando mayores confusiones (granito, granodiorita, tonalita). Estas distribuciones explican en gran medida la diferencia de desempeño entre ambos grupos: las litologías volcánicas presentan transiciones composicionales más definidas, mientras que las plutónicas se organizan de manera más continua en el espacio QAPF, dificultando la clasificación.

2.5.6. Modelos de machine learning aplicados

Una vez definida la arquitectura del pipeline, y tomando en cuenta las particularidades por clase de litología, se justifica a continuación la selección de los tres modelos supervisados empleados, considerando su desempeño frente a datos geoquímicos de alta dimensionalidad y su compatibilidad con los patrones petroquímicos del

estudio. Estos algoritmos son ampliamente utilizados en geociencias debido a su capacidad para capturar relaciones no lineales y estructuras complejas en los datos:

- **Random Forest:** Algoritmo basado en ensamblajes de múltiples árboles de decisión, caracterizado por inmutabilidad y alta robustez frente al ruido, valores atípicos y presencia de variables poco informativas son irrelevantes. En un dataset geoquímico de composición mixta —que integra óxidos mayores, minerales normativos CIPW, parámetros petroquímicos derivados y descriptores auxiliares— RF resulta especialmente adecuado debido a que no requiere escalamiento, tolera de manera eficiente la multicolinealidad natural entre óxidos y minerales normativos, y modela interacciones no lineales típicas de sistemas magmáticos. Además, proporciona métricas de importancia de variables, lo cual facilita interpretar qué componentes geoquímicos y normativos influyen más en la clasificación litológica.
- **XGBoost:** Técnica avanzada de *gradient boosting* optimizada para rendimiento y estabilidad en datasets de tamaño mediano. Su arquitectura iterativa permite capturar patrones sutiles en distribuciones composicionales complejas, como aquellas que vinculan óxidos mayores con fases normativas en rocas plutónicas y volcánicas. XGBoost maneja eficientemente relaciones no lineales, interacciones de segundo orden y efectos combinados entre variables químicas y normativas. Asimismo, su regularización interna contribuye a controlar el sobreajuste, un aspecto relevante en datasets donde ciertas litologías presentan baja frecuencia.
- **SVM + PCA:** El esquema basado en Máquinas de Vectores de Soporte combinado con Análisis de Componentes Principales es conveniente en escenarios donde el espacio de variables es numeroso —como ocurre al integrar 44 *features* geoquímicos y normativos— y donde los límites de decisión entre litologías tienden a ser suaves y continuos. El PCA reduce dimensionalidad preservando la varianza química dominante, mitigando la colinealidad entre óxidos y minerales normativos; sobre esta base reducida, SVM define fronteras de decisión flexibles en espacios transformados. Sin embargo, su desempeño depende del escalamiento previo de las variables, debido a la sensibilidad del algoritmo a diferencias de rango entre

óxidos, minerales normativos y parámetros derivados.

Los modelos fueron evaluados utilizando un esquema de validación cruzada estratificada de 5 pliegues, configurado para mantener la proporción real de clases litológicas en cada subconjunto. Este enfoque garantiza que cada partición del dataset preserve la distribución original de las litologías, condición esencial en conjuntos de datos donde algunas unidades presentan menor representación. El proceso se repitió de manera estándar para que cada pliegue actuara una vez como conjunto de prueba y cuatro veces como conjunto de entrenamiento. Esta estrategia reduce la varianza inducida por el muestreo aleatorio, genera métricas más estables y permite comparar directamente el desempeño de los modelos bajo un mismo criterio experimental.

La estratificación asegura que la evaluación no esté sesgada hacia clases mayoritarias y que todas las litologías sean evaluadas en condiciones equivalentes. Bajo este mismo esquema se entrenaron Random Forest, XGBoost y SVM + PCA, lo que permite interpretar las diferencias de rendimiento como inherentes a los algoritmos y no como producto de variaciones en el método de validación.

3. RESULTADOS

3.1. Desempeño general de los algoritmos

Las métricas globales permiten identificar las diferencias claras del rendimiento de los modelos al actual sobre los conjuntos volcánicos y plutónicos. A continuación, se presentan los valores de AUC, CA, F1, precisión, recall y MCC para cada algoritmo, junto con las matrices de confusión correspondientes y las gráficas comparativas. Estos resultados sintetizan el desempeño general de los clasificadores y sirven como base para el análisis detallado que se desarrolla más adelante.

3.1.1 Métricas globales por modelo

Tabla 1. Comparación de métricas de desempeño de los algoritmos de clasificación en rocas volcánicas.

Modelo	AUC	CA	F1
RF	0.945	0.824	0.818
XGB	0.944	0.825	0.822
SVM+PCA	0.922	0.799	0.783
Modelo	Prec	Recall	MCC
RF	0.823	0.824	0.712
XGB	0.825	0.825	0.714
SVM+PCA	0.806	0.799	0.664

Fuente: Elaboración propia

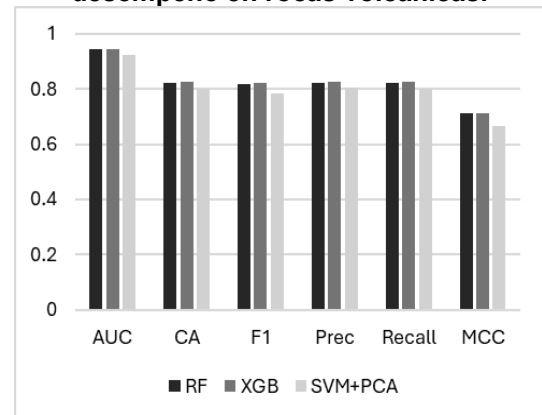
Tabla 2. Comparación de métricas de desempeño de los algoritmos de clasificación en rocas plutónicas.

Modelo	AUC	CA	F1
RF	0.862	0.565	0.538
XGB	0.855	0.564	0.541
SVM+PCA	0.842	0.520	0.453
Modelo	Prec	Recall	MCC
RF	0.538	0.565	0.466
XGB	0.541	0.564	0.465
SVM+PCA	0.468	0.520	0.402

Fuente: Elaboración propia

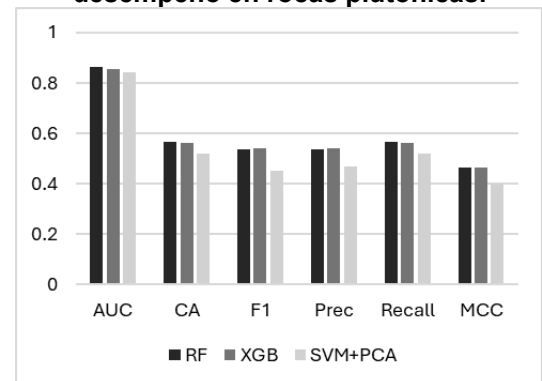
3.1.2. Comparación visual del desempeño de los algoritmos

Figura 7. Gráfico comparativo de métricas de desempeño en rocas volcánicas.



Fuente: Elaboración propia

Figura 8. Gráfico comparativo de métricas de desempeño en rocas plutónicas.



Fuente: Elaboración propia

En el conjunto volcánico, las métricas muestran un rendimiento alto y consistente en los tres modelos. Random Forest y XGBoost presentan AUC > 0.94 y valores prácticamente idénticos de CA (0.824–0.825) y F1 (0.818–0.822). SVM+PCA obtiene valores menores en todas las métricas (AUC = 0.922, CA = 0.799, F1 = 0.783), lo que indica un rendimiento inferior frente a los modelos basados en árboles.

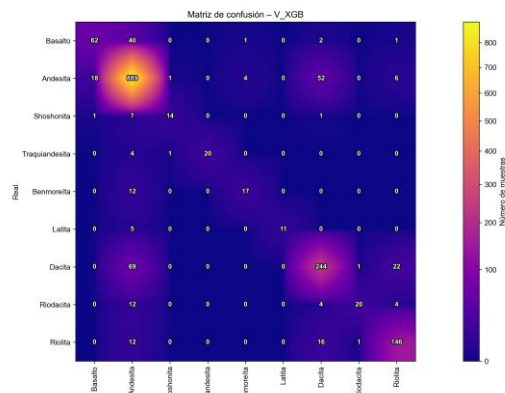
Para el conjunto plutónico, el rendimiento disminuye en los tres modelos. Random Forest alcanza el MCC más alto (0.466), seguido de XGB (0.465), mientras que SVM+PCA presenta el desempeño más bajo (F1 = 0.453, MCC = 0.402). Las diferencias estrechas en CA (0.565 en RF vs. 0.564 en XGB) muestran que la separación entre clases plutónicas es limitada.

3.2. Resultados por clases litológicas

3.2.1. Mapas de calor de matrices de confusión principales

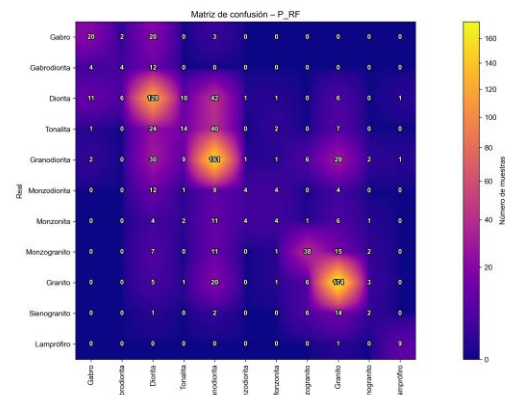
Se presentan a continuación las matrices de confusión de los modelos con mayor capacidad predictiva. Han sido ordenadas siguiendo series petroquímicas continuas a fin de visualizar de manera coherente las confusiones entre litologías genéticamente cercanas. En esta sección se muestran únicamente las matrices del mejor modelo para cada grupo (XGB para volcánicas y RF para plutónicas).

Figura 9. Mapa de calor de la matriz de confusión de rocas volcánicas mediante el algoritmo XGBoost.



Fuente: Elaboración propia

Figura 10. Mapa de calor de la matriz de confusión de rocas plutónicas mediante el algoritmo Random Forest.



Fuente: Elaboración propia

La matriz de confusión volcánica (V_XGB) muestra una diagonal fuerte en andesita, dacita y riolita, con confusiones entre andesita–basalto y dacita–riolita. En la matriz plutónica (P_RF) se observan aciertos en granito, granodiorita y diorita, junto con confusiones en los pares monzogranito–granito, diorita–tonalita y la secuencia gabro–gabrodiorita–diorita.

3.2.2. Interpretación de las confusiones: continuidad petroquímica y patrones esperables

En el conjunto volcánico, la matriz de confusión del modelo XGBoost muestra una diagonal bien definida, con altas tasas de acierto en andesita, dacita y riolita. Estas litologías presentan los porcentajes de clasificación correcta más elevados del grupo volcánico. Asimismo, se observa que andesita presenta leves cruces hacia basalto, mientras que dacita muestra intersecciones con riolita. Aunque estos valores son menores que los aciertos de la diagonal, representan los principales focos de redistribución de etiquetas dentro del conjunto. Las demás litologías volcánicas presentan niveles de confusión considerablemente menores o más dispersos.

En el conjunto plutónico, la matriz de Random Forest evidencia un patrón distinto, aunque igualmente concentrado en familias litológicas específicas. Granito, granodiorita y diorita muestran porcentajes de clasificación correcta relativamente altos, mientras que las redistribuciones más notorias aparecen entre granito–monzogranito y entre granodiorita–tonalita. En las rocas máficas, gabro presenta los principales cruces hacia gabrodiorita y diorita, siguiendo un gradiente claro dentro de este subconjunto. Las demás litologías plutónicas exhiben errores menores en frecuencia, distribuidos principalmente en categorías adyacentes dentro de la matriz.

En conjunto, ambas matrices evidencian que los modelos concentran los errores en subconjuntos litológicos específicos, mientras que el resto de las clases mantiene tasas de acierto relativamente estables. Esto permite identificar, dentro del comportamiento general del modelo, cuáles son las clases que presentan mayor consistencia interna y cuáles tienden a redistribuirse hacia clases cercanas dentro de la tabla.

3.3. Patrones de error y consistencia geológica

El análisis conjunto de ambas matrices permite identificar patrones de error recurrentes. En rocas volcánicas, los principales cruces se concentran en el grupo andesita–dacita–riolita, seguido por una menor contribución de andesita hacia basalto. El resto de las clases volcánicas mantiene distribuciones más compactas, lo que sugiere que los errores se concentran en regiones bien delimitadas de la matriz.

En las rocas plutónicas, los errores se distribuyen de manera más extensa, aunque también con núcleos bien definidos. Las secuencias granito–granodiorita–

tonalita y gabro–gabrodiorita–diorita son las que presentan mayor concentración de cruces, mientras que litologías intermedias menos representadas muestran errores dispersos. Este comportamiento es más notorio que en el caso volcánico y se refleja en un mayor ensanchamiento de las columnas y filas asociadas a estas litologías.

Estos resultados permiten delimitar con claridad cuáles son las categorías que el modelo distingue con mayor estabilidad y cuáles representan desafíos para la clasificación estrictamente basada en composiciones químicas y minerales normativos.

3.4. Síntesis comparativa y modelo óptimo

La comparación integral entre las métricas globales, las gráficas de desempeño y los patrones observados en las matrices de confusión permite establecer con claridad cuáles modelos presentan el comportamiento más estable en cada conjunto litológico. En el grupo volcánico, XGBoost muestra la combinación más equilibrada entre precisión, recall y F1, acompañado de los valores más altos de MCC y un mapa de calor con una diagonal más definida y menor dispersión hacia clases no relacionadas. Random Forest presenta un rendimiento muy cercano, pero con ligeras variaciones en algunas litologías específicas que lo sitúan ligeramente por debajo de XGBoost en términos de consistencia general. SVM+PCA, aunque adecuado en la clasificación de clases bien definidas, muestra una caída más evidente en las métricas y una mayor dispersión en las matrices de confusión.

En el conjunto plutónico, los resultados indican un panorama distinto. Si bien las métricas globales muestran valores más bajos para todos los modelos, Random Forest mantiene los indicadores más estables, especialmente en CA, F1 y MCC. Su matriz de confusión exhibe los patrones de redistribución menos extensos y los aciertos más consistentes dentro de las litologías principales, lo que lo posiciona como el modelo con mejor desempeño dentro de este subconjunto. XGBoost le sigue de cerca con valores prácticamente equivalentes en CA y F1, aunque con una ligera tendencia a redistribuir ciertas litologías en mayor magnitud. Por su parte, SVM+PCA muestra la mayor pérdida de desempeño en este grupo, tanto en métricas como en su patrón de confusión.

En síntesis, los resultados permiten identificar claramente un modelo óptimo por cada grupo litológico:

- **XGBoost** para el conjunto volcánico, debido a su mayor estabilidad en las métricas y en la distribución de aciertos por clase.
- **Random Forest** para el conjunto plutónico, gracias a su consistencia en las litologías principales y al patrón de redistribución más controlado.

Estas observaciones proporcionan una base objetiva para el análisis interpretativo desarrollado en la sección de Discusión, donde se examinan las razones petroquímicas y metodológicas que explican el comportamiento diferencial de los modelos.

4. DISCUSIÓN

Los resultados obtenidos muestran tendencias coherentes con lo descrito en la literatura petroquímica clásica. Diversos autores señalan que las rocas volcánicas presentan trayectorias geoquímicas bien definidas, con patrones evolutivos claros; por dar ejemplos: en diagramas TAS, AFM y Q–Ab–An (Rollinson, 1993; Winter, 2010). Esto se refleja claramente en los valores altos de AUC y CA obtenidos en el presente estudio, donde XGBoost y Random Forest superan el 82 % de acierto en rocas volcánicas. De manera que, el buen rendimiento de los algoritmos coincide con lo reportado para rocas con límites composicionales más marcados y gradientes químicos consistentes.

En contraste, los resultados más bajos obtenidos en rocas plutónicas concuerdan con los trabajos de Le Maitre (2002) y Streckeisen (1976), quienes documentan que las rocas faneríticas presentan campos amplios y superpuestos dentro del QAPF. La química total, por sí sola, suele ser insuficiente para separar de manera robusta litologías como granito, granodiorita y tonalita, lo cual explica las métricas inferiores observadas (CA ~0.56) y la redistribución frecuente entre estas clases. Estudios recientes que aplican algoritmos de machine learning a datos de litogeoquímica muestran que los modelos basados únicamente en geoquímica de roca total tienden a presentar un desempeño relativamente menor o una mayor confusión en subconjuntos plutónicos, en comparación con enfoques que integran información modal, textural o mineralógica adicional (p. ej., Antonini et al., 2024; Nathwani et al., 2022; Petrelli, 2024; Caté et al., 2018).

Al comparar los resultados con lo esperado según estas fuentes, las rocas volcánicas muestran un comportamiento predecible. Sin embargo, algunas desviaciones destacan. La confusión ligeramente mayor entre andesita y dacita sugiere que parte de las muestras podría situarse cerca de límites geoquímicos o presentar signos débiles de alteración, lo cual afecta el cálculo normativo CIPW. Este comportamiento ha sido documentado en rocas volcánicas calcoalcalinas afectadas por procesos de hidrotermalización o redistribución de volátiles (Gill, 1981).

En las rocas plutónicas, la magnitud de ciertos cruces —como monzogranito–granito y gabro–diorita— es mayor a la reportada en clasificaciones estrictamente modales. Esto es atribuible a dos factores: (i) las clases amplias del dataset, donde el rango composicional de “granito” excede los límites

ideales sugeridos por Le Maitre (2002), y (ii) la falta de información modal, crucial para diferenciar rocas con composiciones totales similares, pero proporciones mineralógicas distintas (Streckeisen, 1976).

Entre las posibles fuentes de error, destacan el desbalance de clases, la variabilidad analítica (especialmente en muestras con LOI elevado), y la aplicación de parámetros normativos CIPW —que asumen minerales ideales y equilibrio perfecto— a rocas que pueden haber sufrido procesos de alteración o meteorización (Best, 2003). Estas limitaciones no invalidan los resultados, pero explican parte de la dispersión observada en las matrices de confusión, especialmente en litologías faneríticas.

Durante el análisis también se identificaron dificultades específicas. La ausencia de información modal fue la más relevante, ya que la clasificación QAPF se basa en proporciones modales reales, no en química total. La literatura señala que la composición modal puede cambiar el nombre de la roca incluso cuando la química permanece constante (Le Maitre, 2002), lo que introduce inevitablemente errores cuando se usan solo óxidos mayores y normativos como features. De manera adicional, el desbalance en el número de muestras por clase —particularmente en litologías minoritarias como lamprófito al ser casi ínfimos— afectó la capacidad de aprendizaje del algoritmo, un efecto ampliamente documentado en estudios de ML aplicados a geociencias (Fernández-Delgado et al., 2014).

Aun así, las aproximaciones metodológicas empleadas se justifican plenamente. La separación del dataset en rocas volcánicas y plutónicas sigue las recomendaciones de Winter (2010), quien enfatiza que ambas familias siguen trayectorias evolutivas distintas y deben analizarse por separado. Asimismo, el uso de minerales normativos CIPW ha sido validado en numerosos estudios petrogenéticos para capturar tendencias químicas fundamentales cuando no se dispone de mineralogía modal directa (Rollinson, 1993). El hecho de que los patrones observados en las matrices de confusión correspondan a secuencias petroquímicas reales demuestra la coherencia del enfoque aplicado.

5. CONCLUSIONES

- a. **Los algoritmos basados en árboles mostraron el mejor desempeño general**, con XGBoost como modelo óptimo para el conjunto volcánico y Random Forest para el conjunto plutónico. Esto se reflejó en las métricas globales (AUC, CA, F1 y MCC), donde ambos modelos superaron consistentemente a SVM+PCA.
- b. **Las rocas volcánicas presentan una separabilidad química mucho mayor que**

las rocas plutónicas, lo que permitió obtener precisión y estabilidad significativamente superiores. Los altos valores de CA (>0.82) y MCC (>0.71) en rocas volcánicas corroboran que los patrones evolutivos de estas rocas se ajustan bien a enfoques supervisados de clasificación litogeoquímica.

- c. **Las rocas plutónicas mostraron un rendimiento reducido**, con confusiones más frecuentes entre granito–granodiorita–tonalita y gabro–gabrodiorita–diorita. Este comportamiento es coherente con la literatura, ya que estas rocas presentan campos composicionales superpuestos en el QAPF y una variabilidad modal que no queda capturada por la química total.
- d. **Las confusiones observadas en las matrices de calor reproducen patrones petroquímicos reales**, lo que valida la coherencia geológica del enfoque aplicado. Las transiciones andesita–dacita–riolita y granito–granodiorita son esperables dentro de sus respectivas series ígneas.
- e. **Las limitaciones principales del estudio derivan del dataset**, incluyendo clases amplias, clases minoritarias con bajo número de muestras, variaciones analíticas (LOI, sumas de óxidos), y la ausencia de información modal y textural, especialmente relevante para las rocas plutónicas.
- f. **El uso de minerales normativos CIPW como features resultó ser una aproximación útil y coherente**, capaz de capturar tendencias evolutivas clave, aunque insuficiente para resolver con claridad todas las fronteras litológicas en rocas faneríticas.

Así, en conjunto, los resultados demuestran que la geoquímica total combinada con técnicas de machine learning permite una clasificación litológica robusta, especialmente en rocas volcánicas, y proporciona una base sólida para sistemas automatizados de reconocimiento litológico.

6. RECOMENDACIONES

- a. **Incrementar la representatividad del dataset**, ampliando el número de muestras en litologías minoritarias (p. ej., lamprófitos, monzodioritas) y reduciendo la amplitud de clases excesivamente heterogéneas como andesita o granito.
- b. **Incorporar información modal y textural**, especialmente en rocas plutónicas. La integración de imágenes de lámina delgada mediante procesamiento computacional o segmentación mineralógica puede mejorar notablemente la separabilidad entre rocas con composiciones químicas similares.

- c. **Aplicar técnicas de balanceo de clases** (SMOTE, oversampling controlado o pesos ajustados en el entrenamiento) para reducir el sesgo hacia clases mayoritarias.
- d. **Refinar la etapa de preprocesamiento**, incluyendo correcciones específicas para muestras con LOI elevado, imputación geoquímicamente coherente y normalizaciones ajustadas al comportamiento de los óxidos mayores.
- e. **Explorar modelos adicionales** como LightGBM, CatBoost o ensembles híbridos, que suelen mostrar mejor rendimiento en datasets con alta colinealidad y solapamiento composicional.
- f. **Separar rocas volcánicas en subgrupos más específicos** (basálticas, intermedias, ácidas) para capturar con mayor detalle las tendencias internas de cada familia y reducir el ruido en las transiciones químicas.
- g. **Integrar información contextual adicional**, como edad, entorno tectónico o serie magmática, para construir modelos multimodales que superen la dependencia exclusiva de la química total.
- h. **Validar los modelos con nuevos conjuntos externos**, provenientes de otras bases de datos o campañas de muestreo, para evaluar su capacidad de generalización y robustez.
- i. **Desarrollar versiones futuras del clasificador orientadas a aplicaciones de campo**, facilitando la identificación preliminar de litologías mediante análisis químicos portátiles (p. ej., XRF móvil) combinados con modelos entrenados.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] E. A. K. Middlemost, "Iron oxidation ratios, norms and the classification of volcanic rocks," *Chemical Geology*, vol. 77, no. 1, pp. 19–26, Jan. 1989.

[2] H. Rollinson, *Using Geochemical Data: Evaluation, Presentation, Interpretation*. London, U.K.: Longman, 1993.

[3] M. J. Le Bas, R. W. Le Maitre, A. Streckeisen, and B. Zanettin, "A chemical classification of volcanic rocks based on the total alkali–silica diagram," *Journal of Petrology*, vol. 27, no. 3, pp. 745–750, Jun. 1986.

[4] R. W. Le Maitre et al., *Igneous Rocks: A Classification and Glossary of Terms*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[5] W. Cross, J. P. Iddings, L. V. Pirsson, and H. S. Washington, "A quantitative chemico-mineralogical classification and nomenclature of igneous rocks," *Journal of Geology*, vol. 10, no. 6, pp. 555–690, Sep. 1902.

[6] R. González-Guzmán, L. A. Elizondo-Pacheco, A. González-Roque, C. E. Sánchez-Torres, and K. S. Cárdenas-Muñoz, "shinyNORRRM: A cross-platform software to calculate the CIPW norm," *Mathematical Geosciences*, vol. 55, no. 3, pp. 563–577, Mar. 2023.

[7] USGS, "USGS Rock Geochemistry Database," U.S. Geological Survey, Reston, VA, USA. Available: <https://www.usgs.gov>

[8] GEOROC, "Geochemistry of Rocks of the Oceans and Continents (GEOROC Database)," Max Planck Inst. Chem., Mainz, Germany. Available: <https://georoc.eu>

[9] INGEMMET, "Base de datos geoquímica institucional," Instituto Geológico Minero y Metalúrgico del Perú, Lima, Perú. Available: <https://www.ingemmet.gob.pe>

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, 1995.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, pp. 785–794, Aug. 2016.

[13] C. Winter, *Principles of Igneous and Metamorphic Petrology*, 2nd ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2010.

[14] A. Streckeisen, "To each plutonic rock its proper name," *Earth-Science Reviews*, vol. 12, no. 1, 1976.

[15] J. B. Gill, *Orogenic Andesites and Plate Tectonics*. Berlin, Germany: Springer-Verlag, 1981.

[16] M. G. Best, *Igneous and Metamorphic Petrology*, 2nd ed. Hoboken, NJ, USA: Wiley-Blackwell, 2003.

[17] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[18] A. S. Antonini, J. Tanzola, L. Asiain, and G. Johnson, "Machine learning model interpretability using SHAP values: application to igneous rock classification task," *Applied Computing and Geosciences*, vol. 23, art. 100178, 2024.

[19] C. L. Nathwani, J. J. Wilkinson, G. Fry, R. N. Armstrong, D. J. Smith, and C. Ihlenfeld, "Machine learning for geochemical exploration: classifying metallogenic fertility in arc magmas and insights into porphyry copper deposit formation," *Mineralium Deposita*, vol. 57, pp. 1143–1166, 2022.

[20] M. Petrelli, "Machine Learning in Petrology: State-of-the-Art and Future Perspectives," *Journal of Petrology*, vol. 65, no. 5, art. egae036, 2024.

[21] A. Caté, E. Schetselaar, P. Mercier-Langevin, and P.-S. Ross, "Classification of lithostratigraphic and alteration units from drillhole lithogeochemical data using machine learning: a case study from the Lalor volcanogenic massive sulphide deposit, Snow Lake, Manitoba, Canada," *Journal of Geochemical Exploration*, vol. 188, pp. 216–228, 2018.

ANEXOS

A. Features usados en ML

Óxidos mayores (química elemental)

- SiO_2 – Sílice
- TiO_2 – Óxido de titanio
- Al_2O_3 – Alúmina
- Fe_2O_3 – Óxido férrico (Fe^{3+})
- FeO – Óxido ferroso (Fe^{2+})
- MgO – Óxido de magnesio
- MnO – Óxido de manganeso
- CaO – Óxido de calcio
- Na_2O – Óxido de sodio
- K_2O – Óxido de potasio
- P_2O_5 – Óxido de fósforo
- LOI – Pérdida por ignición (volátiles)

Minerales normativos principales

- Q – Cuarzo normativo
- Cor – Corindón
- Ort – Ortoclasa
- Ab – Albita
- An – Anortita
- Ne – Nefelina
- Le – Leucita
- Kp – Kaliophilita
- Hy – Hiperstena
- Di – Diópsido
- Wo – Wollastonita
- Ol – Olivino
- La – Larnita
- Ac – Acmita

Fases accesorias / óxidos normativos

- Na_ms – Metasilicato de sodio
- Mag – Magnetita
- Hem – Hematita
- Ilm – Ilmenita
- Tit – Titanita
- Ru – Rutilo
- Pwk – Perowskita
- Ap – Apatita
- Cc – Calcita normativa
- Ft – Ferritos
- Py – Pirita normativa o sulfuros

Parámetros derivados

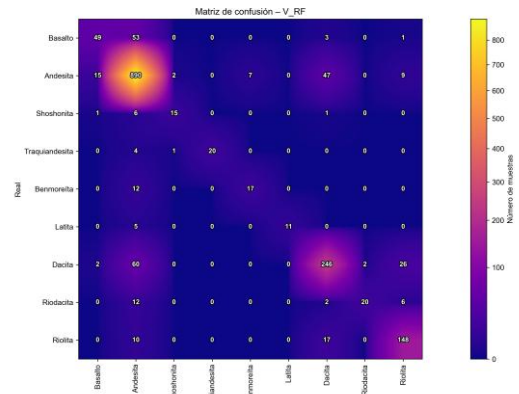
- %An en Pl – Porcentaje de anortita en la plagioclasa normativa
- ITT – Índice Total de Titanio y Fosfatos, indicador de fases accesorias Ti-P
- SUM_QAP – suma normativa del subtriángulo Q-A-P.
- Alc – proporción alcalina normativa
- Clase_SiO₂ – clasificación por rango de sílice)
- Serie_Mag – clasificación magmática auxiliar.
- Sec_Rock_Name – descriptor litológico contextual secundario.

B. Matrices de confusión

B.1 Matrices de rocas volcánicas

B.1.1. Matriz de confusión – Rocas Volcánicas (Random Forest)

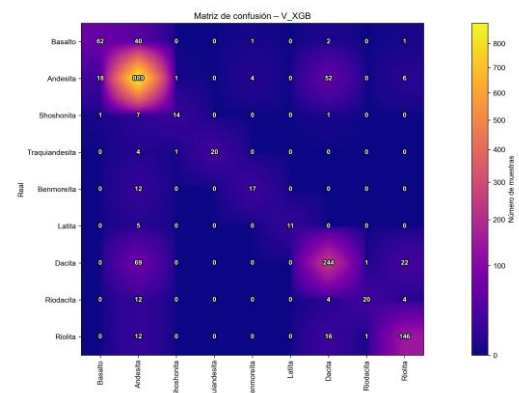
Figura B-1. Matriz de confusión del modelo Random Forest para rocas volcánicas.



Fuente: Elaboración propia

B.1.2. Matriz de confusión – Rocas Volcánicas (XGBoost)

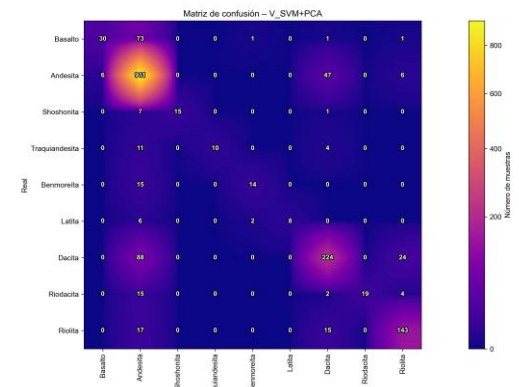
Figura B-2. Matriz de confusión del modelo XGBoost para rocas volcánicas.



Fuente: Elaboración propia

B.1.3. Matriz de confusión – Rocas Volcánicas (SVM+PCA)

Figura B-3. Matriz de confusión del modelo SVM+PCA para rocas volcánicas.

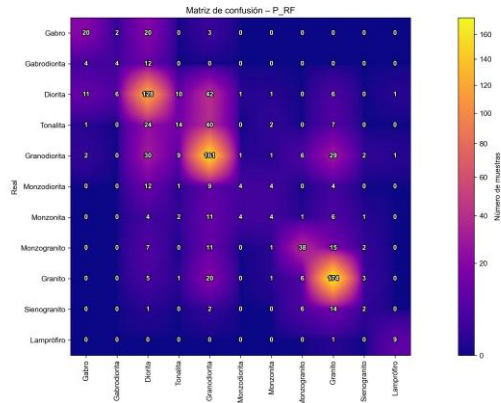


Fuente: Elaboración propia

B.2. Matrices de rocas plutónicas

B.2.1. Matriz de confusión – Rocas Plutónicas (Random Forest)

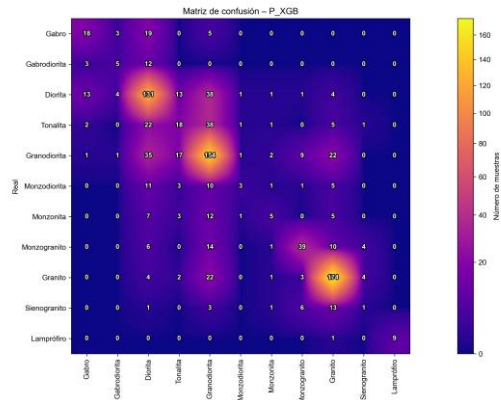
Figura B-4. Matriz de confusión del modelo Random Forest para rocas plutónicas.



Fuente: Elaboración propia

B.2.2. Matriz de confusión – Rocas Plutónicas (XGBoost)

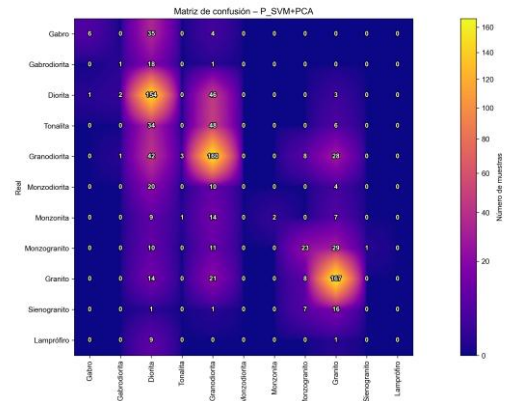
Figura B-5. Matriz de confusión del modelo XGBoost para rocas plutónicas.



Fuente: Elaboración propia

B.2.3. Matriz de confusión – Rocas Plutónicas (SVM+PCA)

Figura B-6. Matriz de confusión del modelo SVM+PCA para rocas plutónicas.



Fuente: Elaboración propia

C. Extracto del dataset final depurado

ID	RockType	Rock_Name	SiO ₂	Al ₂ O ₃	FeO	MgO	Na ₂ O	K ₂ O	Q	Cor	Ort	Ab	An
1	Vol	Andesita	61.1	17.62	4.76	2.42	3.2	3.81	12.51	0.94	22.62	27.21	19.98
2	Vol	Andesita	58.9	16.59	6.08	2.61	3.38	3.54	8.54	0	21.05	28.77	19.76
3	Vol	Andesita	62.88	17.17	5.092868	1.36	3.28	3.29	16.52	0.28	19.55	27.91	21.75
4	Vol	Andesita	61.01	16.4	5.12886	2.86	3.4	3.95	11.1	0	23.46	28.92	17.91
5	Vol	Andesita	62.25	17.09	5.119862	2.25	2.83	3.33	16.79	0.71	19.78	24.08	22.28
6	Vol	Andesita	63.76	16.08	4.328038	2.17	2.58	5.07	17.04	0.84	30.09	21.93	15.09
7	Vol	Andesita	63.65	16.11	4.463008	2.58	3.18	3.81	16.65	0.58	22.62	27.03	16.92
8	Vol	Andesita	54.94	15.48	6.52355	6.05	3.2	2.92	0.7	0	17.37	27.26	19.38
9	Vol	Shoshonita	51.54	17.58	7.234392	7.47	2.8	0.67	0	0	3.98	23.87	33.67
10	Vol	Basalto	49.78	13.31	6.982448	10.44	2.87	4.11	0	0	24.46	11.73	11.38

Nota: Debido al gran número de columnas y registros del dataset final, no es posible representarlo en una tabla legible dentro del formato A4. Por ello, se muestra únicamente un extracto ilustrativo. El archivo completo se adjunta en el siguiente enlace: <https://tinyurl.com/Jz-PRIM-ABET20205>