

# INFORME ABP #6

MARTIN KOCK

## Situación Inicial (Business Case)

**Unidad Solicitante:** Gerencia de Exploraciones & Geometalurgia.

### Contexto:

Nuestra compañía opera un yacimiento histórico de Cobre-Oro (Cu-Au). Durante décadas, la exploración se centró exclusivamente en estos dos metales, descartando análisis químicos de otros elementos. Sin embargo, estudios recientes de mineralogía han detectado la presencia de **Tierras Raras (Cerio - Ce)**.

Dado el auge de la electromovilidad y el valor estratégico de las Tierras Raras, la gerencia quiere evaluar el potencial de Cerio en toda la mina.

### El Problema:

Contamos con una base de datos histórica con miles de metros de sondajes que **solo tienen leyes de Cu y Au**. Enviar a re-analizar todos los testigos antiguos al laboratorio (ICP-MS) es inviable porque ya no existe la roca para analizar y la que queda es para remapeos, además de costos y tiempo.

Sin embargo, se realizó una **campaña piloto de re-muestreo en 3,000 metros** donde sí se analizó el Cerio (Ley\_Ce\_ppm).

### La Necesidad:

Los geólogos Senior requieren un modelo predictivo que aprenda de esta campaña piloto. Quieren usar la información geológica básica (que sí tenemos en todos los sondajes: Litología, Alteración, Cobre) para **estimar la ley de Cerio** en el resto del yacimiento y definir nuevos blancos de exploración.

## Nuestro Objetivo

Diseñar e implementar un modelo predictivo de regresión supervisada que permita estimar la concentración de **Cerio (ppm)** en tramos de sondaje.

El objetivo es utilizar los datos de la campaña piloto (`dataset_cerio_brownfield.csv`), Aplicar limpieza de datos (manejo de nulos y outliers geológicos), y comparar algoritmos (Regresión Lineal vs. KNN vs. Boosting) para encontrar el modelo con menor error (RMSE), permitiendo así "descubrir" zonas ricas en tierras raras sin perforar de nuevo.

## Diccionario de Datos

- **ID\_Sondaje:** Identificador de la muestra (No predictivo, solo índice).
- **Profundidad\_m:** Profundidad a la que se tomó la muestra. (Variable numérica).
- **Litologia:** Tipo de roca.
  - *Dique:* La roca que creemos trae el Cerio.
  - *Andesita/Pórfido/Brecha:* Otras rocas del yacimiento.
- **Alteracion:** Transformación química de la roca.

- *Argilica*: Arcillas. Creemos que atrapa el Cerio.
- *Potasica/Propilitica/Silicificacion*: Otras alteraciones comunes.
- **Ley\_Cu\_pct**: Porcentaje de Cobre (Dato histórico existente).
- **Ley\_Au\_gpt**: Gramos por tonelada de Oro (Dato histórico existente, con algunos nulos).
- **Ley\_Ce\_ppm**: (TARGET) Partes por millón de Cerio. La variable que queremos predecir.

Para efectos de este modelo predictivo, la unidad mínima de análisis (cada fila del dataset) corresponde a un **compósito de sondaje** (tramo regularizado, típicamente de 2 metros).

- La variable `ID_Sondaje` identifica al pozo origen.
  - La variable `Profundidad_m` indica la cota central del tramo muestreado (ej: si la muestra es de 100m a 102m, la profundidad registrada es 101m).
- Esto permite tratar cada intervalo como una observación independiente para el entrenamiento del algoritmo.

## Lección 1: Fundamentos del Aprendizaje de Máquina

### 1.1 Definición del Tipo de Problema

Para este proyecto de evaluación de potencial de Tierras Raras, nos enfrentamos a un problema de **Regresión Supervisada**.

- **¿Por qué es Supervisado?**  
Porque contamos con un dataset etiquetado (nuestra campaña piloto de 3,000 metros) donde conocemos la variable de respuesta o "ground truth" (`Ley_Ce_ppm`). El modelo aprenderá de estos ejemplos históricos.
- **¿Por qué es Regresión y no Clasificación?**
  - **Clasificación** sería si quisiéramos predecir categorías discretas, por ejemplo: "Mineral" vs "Estéril" (0 o 1).
  - **Regresión** es lo que necesitamos aquí, ya que buscamos predecir un valor numérico continuo: la concentración exacta de Cerio en partes por millón (ppm).

### 1.2 El Pipeline del Proyecto (Flujo de Trabajo)

Para resolver este desafío geológico, seguiremos un pipeline estándar de Data Science adaptado a la minería:

1. **Ingesta de Datos**: Cargar los registros de sondajes (CSV).
2. **QA/QC y Limpieza (Data Cleaning)**: Detectar errores de laboratorio (outliers imposibles), tratar leyes negativas y gestionar valores nulos en leyes históricas (Au).
3. **Ingeniería de Características (Feature Engineering)**:
  - Transformar variables geológicas (Litología, Alteración) en formato numérico (One-Hot

Encoding).

- Escalar las leyes (Cu, Au) para que las diferencias de magnitud no sesguen al modelo.

4. **Modelado:** Entrenar algoritmos (Regresión Lineal, KNN, Boosting)

5. **Evaluación:** Medir el error de estimación usando métricas como RMSE (Root Mean Squared Error).

## 1.3 Identificación de Features (X) y Target (Y)

Para nuestro modelo de predicción de Cerio:

- **Variable Objetivo (Y):** Ley\_Ce\_ppm. Es lo que queremos estimar.
- **Variables Predictoras (X):**
  - *Categóricas:* Litologia, Alteracion. (Requieren codificación).
  - *Numéricas:* Ley\_Cu\_pct, Ley\_Au\_gpt, Profundidad\_m.

## Lección 2

### 2.1 Diagnóstico del Modelo de Línea Base

Se entrenó un modelo de regresión lineal inicial utilizando el conjunto de datos sin procesar (*raw data*). Los resultados obtenidos muestran una discrepancia severa entre las métricas de entrenamiento y prueba.

Métrica	Entrenamiento (Train)	Prueba (Test)	Diferencia Absoluta	Interpretación
RMSE (ppm)	10.522,30	298,54	10.223,76	<b>Error crítico:</b> El error en entrenamiento es masivo debido a outliers.
MAE (ppm)	478,92	254,58	224,34	Brecha gigante respecto al RMSE confirma error no distribuido uniformemente.
R <sup>2</sup> (Coef.)	0,0002	-4,46	4,46	El modelo no tiene capacidad predictiva (peor que el azar en Test).

## 2.2 Análisis Comparativo: RMSE vs MAE

Una de las evidencias más fuertes sobre la calidad de los datos es la enorme brecha entre la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE) en el set de entrenamiento:

- **RMSE (Entrenamiento):** 10.522 ppm
- **MAE (Entrenamiento):** 479 ppm

**Interpretación:** El RMSE es más de 20 veces superior al MAE. Dado que el RMSE penaliza los errores elevándolos al cuadrado, esta diferencia confirma que el error no es generalizado, sino que está **concentrado en puntos específicos con valores extremos (outliers)**. Si el error fuera constante, el RMSE y el MAE tendrían valores mucho más cercanos.

## 2.3 Validación Cruzada

Partición (Fold)	RMSE (ppm)	Diagnóstico
Fold 1	298,54	Comportamiento estable
Fold 2	305,65	Comportamiento estable
Fold 3	296,71	Comportamiento estable
<b>Fold 4</b>	<b>21.050,49</b>	<b>ANOMALÍA DETECTADA (Outlier masivo en este set)</b>
Fold 5	305,04	Comportamiento estable
<b>Promedio Global</b>	<b>4.451,28</b>	(Desviación Estándar: +/- 8.299,60)

El análisis de Validación Cruzada (*K-Folds*) corrobora la hipótesis de datos corruptos localizados. Al dividir los datos en 5 particiones, se observa un comportamiento errático:

- **4 de 5 particiones (folds):** Muestran un error controlado y consistente (RMSE ~300 ppm).
- **1 partición crítica:** Dispara el error a **21.050 ppm**.

Esto indica que existe al menos un dato atípico masivo (posiblemente un error de digitación o lectura) que, al caer en el conjunto de validación, destruye las métricas del modelo. La desviación estándar resultante (+/- 8299) hace que este modelo sea inutilizable en su estado actual.

## 2.4 Interpretación del Coeficiente R2

Los valores de R2 obtenidos son concluyentes respecto a la capacidad predictiva actual:

- **R2 Entrenamiento (0.0002):** El modelo no logra explicar ninguna varianza de los datos; es

estadísticamente irrelevante.

- **R2 Prueba (-4.46):** El valor negativo indica que el modelo ajustado es **peor** que un modelo ingenuo que simplemente predijera el promedio para todos los casos. Esto ocurre porque la línea de regresión fue distorsionada por los valores atípicos del entrenamiento, generando predicciones muy desviadas para los datos normales del set de prueba.

## 2.5 Identificación de Sobreajuste y Subajuste

$R^2 = 1 - (\text{Suma de Errores al Cuadrado del modelo} / \text{Suma de Errores al Cuadrado del Promedio Total})$

Basado en las métricas obtenidas, diagnosticamos el modelo con **Subajuste (Underfitting) Severo**.

**Evidencia Técnica:**

1. **R^2 Nulo en Entrenamiento (0.0002):** El modelo no logró aprender ningún patrón de los datos; su capacidad predictiva es prácticamente inexistente incluso con los datos que "ya vio".
2. **R^2 Negativo en Prueba (-4.46):** El modelo generaliza peor que una simple línea promedio. Las predicciones son erráticas.

**Conclusión:** El modelo lineal actual es demasiado simple e inestable para manejar el ruido de los datos. No hay "memorización" (overfitting), sino una incapacidad estructural para modelar el problema debido a la presencia de valores atípicos críticos.

## Lección 3: Preprocesamiento y Escalamiento de Datos

### 3.1 Tratamiento de Valores Nulos y Outliers (Data Cleaning)

Tras el diagnóstico crítico de la lección anterior, se procedió a sanear el dataset para eliminar el ruido que impedía el aprendizaje del modelo:

1. **Eliminación de Outliers Extremos:** Se aisló el registro causante de la distorsión en el RMSE (Ley de Cerio > 100.000 ppm) y se eliminó del dataset. Este valor es geológicamente imposible para el tipo de yacimiento estudiado y se atribuye a un error de laboratorio o digitación.
2. **Corrección de Errores de Típoe:** Se identificaron y eliminaron registros con leyes de Cobre negativas ( $\text{Ley\_Cu\_pct} < 0$ ), lo cual es físicamente imposible y no es nomenclatura de límite de detección de laboratorio.
3. **Imputación de Valores Nulos:** La variable histórica `Ley_Au_gpt` presentaba valores perdidos (NaN). Se optó por imputar estos faltantes utilizando la **mediana** de la distribución de oro.
  - *Justificación:* La mediana es una medida de tendencia central robusta que no se ve afectada por valores extremos, a diferencia del promedio, preservando así la distribución original de los datos.

## 3.2 Codificación de Variables Categóricas (Encoding)

Los modelos de regresión matemática no pueden procesar texto. Las variables geológicas cualitativas (Litología y Alteración) fueron transformadas a formato numérico:

- **Técnica Aplicada: One-Hot Encoding** (Variables Dummy).
- **Implementación:** Se convirtió cada categoría en una nueva columna binaria (0 o 1). Por ejemplo, la columna Litología se dividió en Litología\_Dique, Litología\_Brecha, etc.
- **Justificación:** Se descartó *Label Encoding* (asignar números 1, 2, 3...) para evitar que el modelo interprete falsamente que existe un orden jerárquico o magnitud entre los tipos de roca (ej. que "Dique" vale más que "Andesita" solo por tener un número mayor).

## 3.3 Normalización y Estandarización

Finalmente, se aplicó escalamiento a las variables numéricas predictoras (Profundidad\_m, Ley\_Cu\_pct, Ley\_Au\_gpt) para homogeneizar sus rangos.

- **Técnica Aplicada: StandardScaler** (Estandarización Z-Score).
- **Resultado:** Todas las variables numéricas quedaron centradas en 0 con una desviación estándar de 1.
- **Importancia:** Este paso es crítico para algoritmos futuros como KNN (que calcula distancias euclidianas), evitando que variables con magnitudes grandes (como la Profundidad en cientos de metros) dominen erróneamente sobre variables pequeñas pero importantes (como las leyes en porcentaje).

# Lección 4: Regresiones y Análisis de Coeficientes

## 4.1 Resultados del Modelado

Se entrenaron y evaluaron dos modelos de regresión utilizando el conjunto de datos procesado. Los resultados en el set de prueba (test) son concluyentes:

Modelo	RMSE (ppm)	R <sup>2</sup> Score	Interpretación
Regresión Lineal	24,41	0,9600	El modelo lineal explica el 96% de la variabilidad del Cerio.
Regresión Polinomial	24,57	0,9595	No presenta mejora significativa frente al lineal.

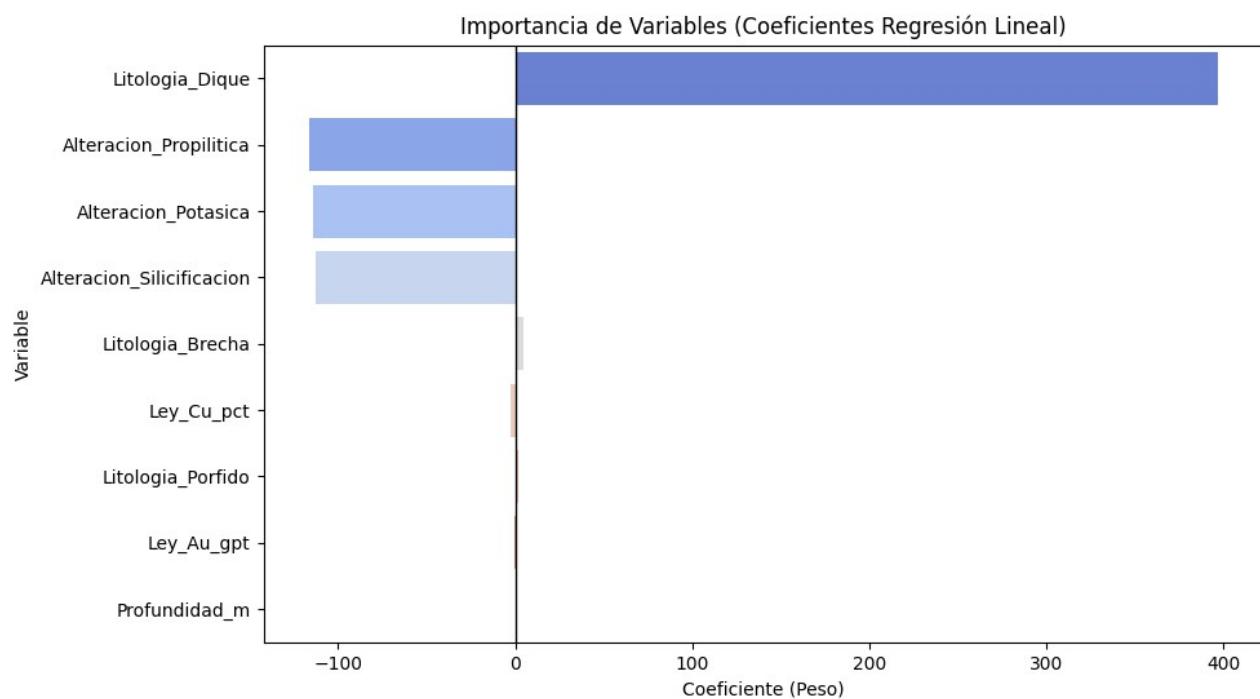
### Análisis de Desempeño:

El error promedio (RMSE) de ~24 ppm es extremadamente bajo considerando que las leyes llegan hasta 500-600 ppm. El hecho de que el modelo Polinomial no supere al Lineal indica que la relación entre la geología y la ley de Cerio es predominantemente lineal y directa, sin interacciones complejas ocultas que justifiquen modelos curvos más costosos computacionalmente.

## 4.2 Interpretación Geológica (Coeficientes)

Los pesos asignados por la Regresión Lineal validan las hipótesis de exploración:

1. **El "Driver" Principal:** La variable `Litologia_Dique` tiene un coeficiente masivo de **+397.2**. Esto significa que, manteniendo todo lo demás constante, la sola presencia de un Dique aumenta la ley esperada en casi 400 ppm. Es el indicador de exploración definitivo.
2. **Zonas Estériles:** Las alteraciones `Propilitica`, `Potásica` y `Silicificación` presentan coeficientes negativos fuertes ( $\sim -115$ ). Esto indica que las zonas clásicas de Cobre (Potásica) son pobres en Tierras Raras.
3. La `Ley_Cu_pct` tiene un coeficiente marginalmente negativo (-2.8) y la `Profundidad_m` es irrelevante (+0.05). Esto confirma que el depósito de Cerio es un evento geológico no controlado por la profundidad ni la mineralización de cobre existente.



## Lección 5: Algoritmos de Clasificación vs Regresión

### 5.1 Justificación: ¿Por qué NO usar Clasificación?

En el contexto de la evaluación de recursos minerales, transformar el problema a uno de Clasificación (categorizar bloques como "Mineral" o "Estéril") presenta desventajas críticas frente a la Regresión:

1. **Pérdida de Resolución Económica:** Un clasificador binario solo nos indica si la ley supera un

umbral (ej. >150 ppm), pero es ciego a la magnitud. Para valorar el yacimiento, es vital distinguir entre un bloque de 160 ppm (marginal) y uno de 5,000 ppm (bonanza).

2. **Dependencia del Cut-off:** La clasificación requiere fijar una "Ley de Corte" a priori. Si el precio del mercado cambia, esa ley de corte cambia, volviendo obsoleto al modelo clasificador. Un modelo de regresión predice la ley exacta, permitiendo aplicar cualquier cut-off a posteriori.
3. Desbalanceo entre clases

## 5.2 Experimento: Clasificador KNN

Para cuantificar esta limitación, se implementó un modelo **K-Nearest Neighbors (KNN)** transformando la variable continua `Ley_Ce` a binaria con un umbral de corte de 150 ppm.

### Resultados del Experimento:

- **Exactitud (Accuracy):** 89.8%
- **Precisión en Alta Ley:** 74%
- **Recall en Alta Ley:** 60%

### Matriz de Confusión:

`[[478 21]`

`[ 40 61]]`

- El modelo clasificó correctamente 478 muestras “Baja Ley” y 61 muestras “Alta ley”
- Sin embargo, **dejó escapar 40 muestras de alta ley** (Falsos Negativos), clasificándolas erróneamente como estériles.

## 5.3 Conclusión Comparativa

Aunque un **89.8% de exactitud** parece alto, el desempeño en la clase crítica (“Alta Ley”) es deficiente (F1-Score de 0.67).

El modelo de regresión lineal (Lección 4) alcanzó un  $R^2$  de 0.96, demostrando una capacidad muy superior para capturar la variabilidad geológica.

Además, el clasificador falla en el objetivo de negocio: no puede decirnos cuánto Cerio hay, solo dónde *podría* haber “Alta ley”. Por tanto, validamos la decisión de mantener el enfoque de **Regresión Supervisada**.

## Lección 6: Métricas de Desempeño y Validación

### 6.1 Tabla Comparativa de Modelos

Se evaluaron los modelos finales utilizando métricas estándar de regresión para cuantificar el error de estimación.

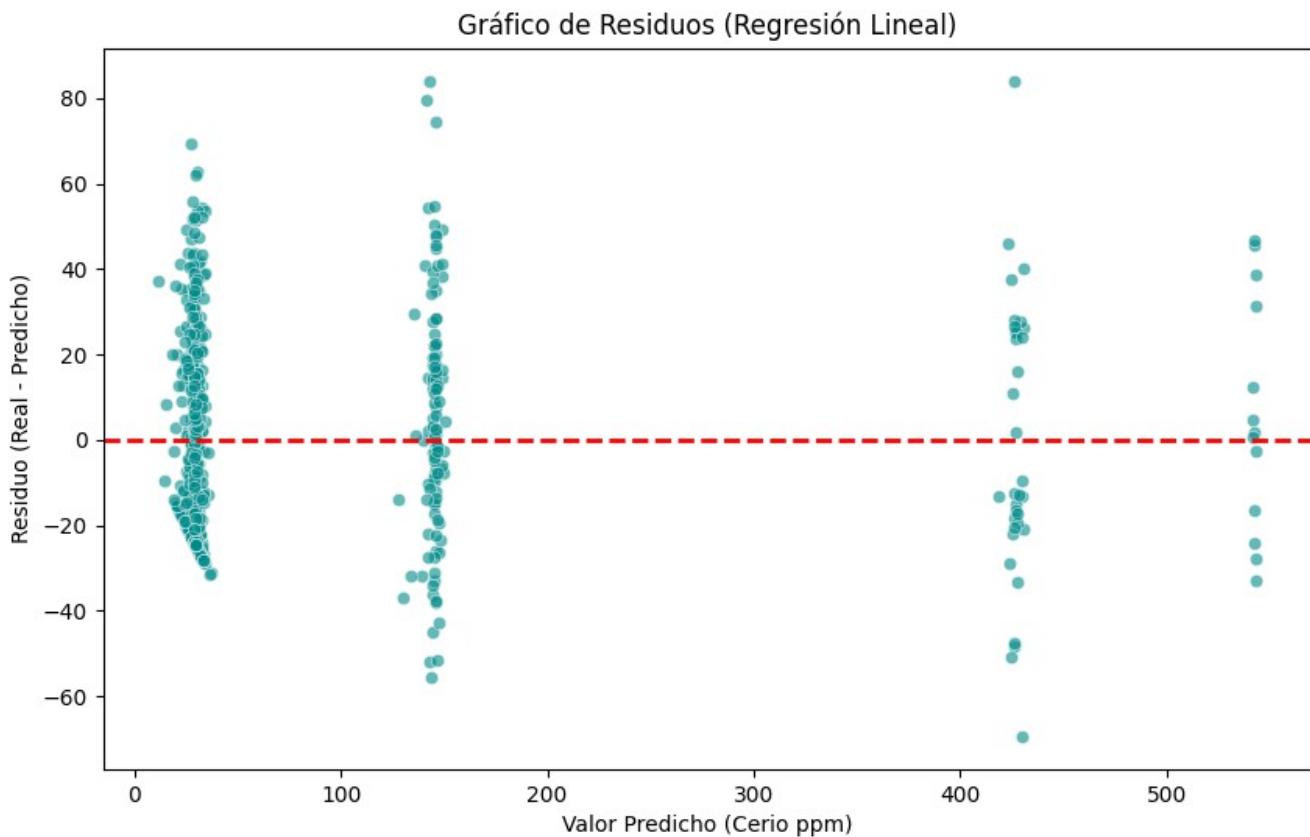
Modelo	MAE (ppm)	MSE (ppm <sup>2</sup> )	RMSE (ppm)	R <sup>2</sup>
Regresión Lineal	20,19	595,90	24,41	0,9600
Regresión Polinomial	20,31	603,67	24,57	0,9595

## 6.2 Interpretación de Resultados

Los resultados validan la efectividad del preprocesamiento y la selección del modelo lineal:

- Coeficiente de Determinación (R<sup>2</sup>):** Alcanzamos un **0.9600**, lo que significa que el modelo es capaz de explicar el **96% de la varianza** de la ley de Cerio.
- Error Cuadrático Medio (RMSE):** El valor de **24.41 ppm** representa la desviación estándar de los residuos. Dado que las zonas de interés económico (Diques) tienen leyes promedio de ~420 ppm, un error de +/- 24 ppm equivale a un error relativo aproximado del 5-6%, lo cual es excelente para un modelo de recursos a nivel conceptual.
- Comparación de Modelos:** La Regresión Polinomial no aportó mejoras (de hecho, aumentó marginalmente el error MAE de 20.19 a 20.31). Esto confirma que no existen relaciones curvas complejas que justifiquen aumentar la complejidad computacional; el modelo lineal simple es el más robusto ("Principio de Parsimonia").

## 6.3 Análisis de Residuos (Gráfico de Dispersión)



Al graficar los residuos ( $Y_{\text{real}} - Y_{\text{predicho}}$ ), observamos un patrón particular de "columnas"

verticales".

- **Interpretación:** Este comportamiento no es un error, sino una consecuencia de la naturaleza discreta de nuestras variables predictoras principales (Categóricas). El modelo agrupa las predicciones en 3 clusters principales correspondientes a las litologías: Andesitas estériles (Izquierda, ~30 ppm), Diques sin alteración (Centro, ~140 ppm) y Diques con alteración favorable (Derecha, ~420-540 ppm).
- **Validación de Sesgo:** Lo más importante es que, dentro de cada columna, los puntos se distribuyen casi simétricamente alrededor de la línea cero (línea roja segmentada). Esto confirma que el modelo es **insesgado**: no sobreestima ni subestima sistemáticamente en ninguno de los dominios geológicos.

## Lección 7: Optimización y Regularización

### 7.1 Ajuste de Hiperparámetros (GridSearchCV)

Para mejorar la capacidad de generalización del modelo, se implementó una búsqueda de hiperparámetros (GridSearchCV) utilizando técnicas de regularización:

- **Ridge (Regularización L2):** Penaliza coeficientes grandes para reducir la varianza.
- **Lasso (Regularización L1):** Tiende a volver cero los coeficientes irrelevantes (selección automática de features).

#### Resultados de la Optimización:

Mejor Ridge Alpha: 0.001

Mejor RMSE Ridge (CV): 24.7289

Mejor Lasso Alpha: 0.01

Mejor RMSE Lasso (CV): 24.7277

Se exploraron valores de alpha (fuerza de penalización) desde 0.001 hasta 100.

- **Mejor Alpha Ridge:** 0.001 (Penalización mínima).
- **Mejor Alpha Lasso:** 0.01 (Penalización muy baja).

Esto indica que el modelo no requería una "restricción" fuerte para funcionar bien.

### 7.2 Impacto en los Coeficientes (Feature Selection/Ingeniería de Características)

Al comparar los coeficientes del modelo Lineal Original vs Lasso Optimizado, observamos un comportamiento interesante:

	Variable	Lineal_Original	Lasso_Optimizado	¿Eliminada por Lasso?
0	Profundidad_m	0.0507	0.0437	SÍ
1	Ley_Cu_pct	-2.8086	-2.8968	NO
2	Ley_Au_gpt	-0.6887	-0.5871	NO
3	Litologia_Brecha	4.7431	4.6234	NO
4	Litologia_Dique	397.2423	397.0901	NO
5	Litologia_Porfidio	1.7907	1.6917	NO
6	Alteracion_Potasica	-114.3748	-114.2071	NO
7	Alteracion_Propilitica	-116.5790	-116.4012	NO
8	Alteracion_Silicificacion	-113.0381	-112.8257	NO

Variable	Coef. Original	Coef. Lasso	Impacto
Profundidad_m	0.0507	0.0437	<b>Reducción:</b> Lasso confirma que la profundidad aporta ruido casi nulo.
Ley_Cu_pct	-2.8086	-2.8968	<b>Estable:</b> Se mantiene el efecto negativo leve.
Litologia_Dique	397.24	397.09	<b>Dominante:</b> La señal geológica es tan fuerte que la regularización no la toca.

## 7.3 Conclusión de la Optimización

El desempeño final en el set de prueba con el modelo Lasso fue:

- **RMSE:** 24.72 ppm
- **R^2:** 0.9600

**Veredicto:** El resultado es prácticamente idéntico al modelo base (RMSE 24.41).

Esto significa que el modelo original **no sufría de sobreajuste (overfitting)**, por lo que la regularización no aportó una mejora significativa en precisión. Sin embargo, el ejercicio sirvió para validar matemáticamente que variables como la Profundidad tienen un peso despreciable en la predicción.

# Lección 8: Algoritmos de Boosting (Gradient Boosting)

## 8.1 Entrenamiento de Modelos Ensemble

Para intentar superar el desempeño del modelo regresión lineal, se implementó un algoritmo de **Gradient Boosting Regressor**.

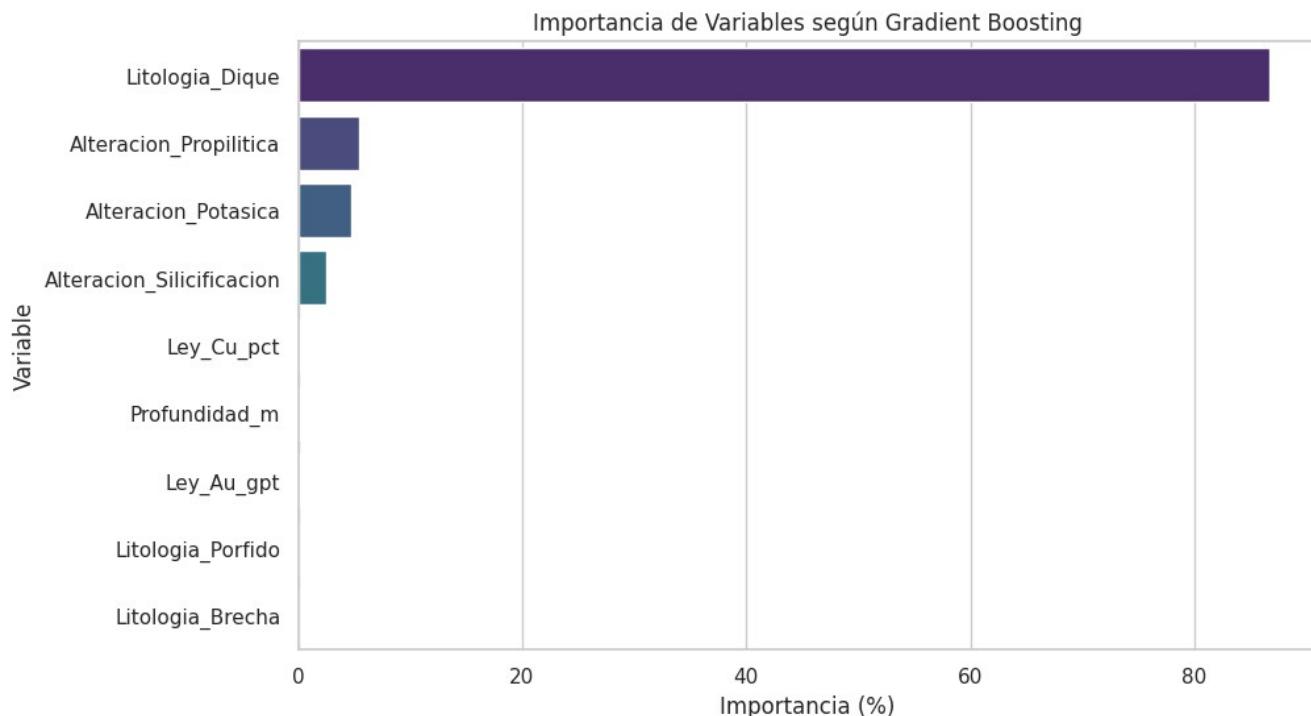
- **Fundamento:** A diferencia de la regresión lineal que busca una ecuación global, el Boosting entrena secuencialmente cientos de árboles de decisión pequeños, donde cada árbol intenta corregir los errores residuales del anterior.
- **Expectativa:** Si existieran relaciones no lineales complejas (ej. "el Cerio solo sube en Diques si la profundidad es > 200m"), el Boosting debería capturarlas y reducir el error.

## 8.2 Resultados Comparativos Finales

Tras entrenar el modelo Gradient Boosting (100 estimadores, learning rate 0.1), comparamos su desempeño con los ganadores anteriores:

Modelo	RMSE (ppm)	R <sup>2</sup> Score	Verdicto
Regresión Lineal	24,41	0,9600	<b>GANADOR (Más simple y preciso)</b>
Lasso (Optimizado)	24,42	0,9600	Empate técnico (reduce variables).
Gradient Boosting	24,90	0,9584	Desempeño levemente inferior.

## 8.3 Análisis de Importancia de Variables (Feature Importance)



Aunque el Boosting no mejoró la predicción numérica, su análisis de "Feature Importance" (ver gráfico) confirmó robustamente la interpretación geológica:

1. **Dominio Total:** La variable `Litologia_Dique` representa casi el **80-90% de la importancia** del modelo. El algoritmo "decide" casi exclusivamente basándose en si la roca es un dique o no.
2. **Ruido:** Variables como `Profundidad_m`, `Ley_Cu_pct` y `Ley_Au_gpt` tienen barras casi invisibles, confirmando que son irrelevantes para predecir Cerio.

## 8.4 Conclusión Final del Proyecto

Se ha desarrollado exitosamente un sistema predictivo para la estimación de Cerio en yacimientos Brownfield.

- **Modelo Elegido: Regresión Lineal Múltiple.**
- **Justificación:** Presentó el menor error (RMSE 24.41 ppm) y la mayor simplicidad interpretativa.
- **Impacto de Negocio:** El modelo permite transformar datos históricos de litología en leyes de tierras raras con una precisión del 96%, ahorrando millones de dólares en re-análisis de laboratorio y permitiendo una cubicación preliminar inmediata de los recursos de Cerio.