Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemático

Título de la tesis

por

Juan Jesús Torres Solano

como requisito parcial para obtener el grado de

maestría en ciencia de datos

marzo 2025

Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemático

Los miembros del Comité de Tesis recomendamos que la Tesis «Título de la tesis», realizada por el alumno Juan Jesús Torres Solano, con número de matrícula 2173262, sea aceptada para su defensa como requisito parcial para obtener el grado de Maestría en Ciencia de Datos.

El Comité de Tesis

M.C. José Anastacio Hernández Saldaña
Asesor

Nombre del revisor C
Revisor

Nombre del revisor D
Revisor

Vo. Bo.

Dra. Azucena Yoloxóchitl Ríos Mercado
Subdirector de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, marzo 2025

Aquí puedes poner tu dedicatoria
si es que tienes una.

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Aquí puedes poner tus agradecimientos. (No olvides agradecer a tu comité de tesis, a tus profesores, a la facultad).

Capítulo 1
INTRODUCCIÓN

La aplicación de la ciencia de datos en el área de geociencias presenta retos y oportunidades, mayor mente aprovechada en sísmica petrolera, presentando un costo beneficio alto y aprovechando la gran cantidad de datos que se generan en las campañas de exploración e interpretación, otorgando elementos cruciales para el modelado y entrenamiento para reconocimiento de patrones y secuencias geológicas; mientras que la oportunidad de aplicación en otros métodos geofísicos es aun vigente.

La geofísica es la ciencia que se encarga de explorar el medio terrestre, empleando métodos que aprovechan las propiedades físicas del subsuelo como medio de respuesta ante la interacción activa, como la aplicación de un campo eléctrico o una onda mecánica, o pasiva, en la cual se realizan mediciones de las variaciones de los campos naturales de la tierra, como son el campo magnético, gravimétrico o el potencial eléctrico natural del subsuelo.

El método geofísico de prospección geoeléctrica aprovecha las propiedad de resistividad para caracterizar las distintas unidades del medio a partir de su capacidad de resistividad eléctrica, siendo de interés para este trabajo el método de inducción eléctrica en su modalidad de Sondeo Eléctrico Vertical (SEV).

Los SEV se realizan siguiendo un arreglo geométrico de 4 electrodos previamente definido (ver figura 5.2), el cual puedo ejecutarse siguiendo la configuración Wenner, Dipolo-Dipolo o Schlumberger, entre otros, cada arreglo presenta un patrón distinto de dispersión del flujo eléctrico, cambiando su sensibilidad ante variaciones verticales u horizontales.

Durante la adquisición de datos en campo la apertura entre los electrodos

cambia en razón al arreglo geoeléctricos empleado, de manera que obtenemos para cada cambio en apertura un valor de resistividad aparente, Rha o ρa en (Ω m), posterior al proceso de interpretación y modelado, obtenemos una distribución de espesores con una resistividad asociada a cada uno de ellos.

Convencionalmente la prospección eléctrica se aplica, procesa e interpreta siguiendo una metodología claramente establecida, esto no la exime de ser viable la mejora de cada uno de sus aspectos.

Algunas etapas de este proceso an sido optimizadas, como la automatización de la lectura de datos, dicho esto, en el proceso de planeación de adquisición y validación de lecturas en campo pueden mejorar por medio de modelos de aprendizaje, los cuales requieren un amplio numero de datos para su entrenamiento.

Para realizar un entrenamiento efectivo, se requiere de un gran numero de datos de entrenamiento, por lo que se empelaran los espesor por sondeo para definir variaciones de distribuciones de espesores manteniendo la relación entre las distribuciones originales y las variaciones generadas.

Inicialmente se cuenta con 8 sitios, integrados de 2 a 4 modelos SEV's, sumando un total de 25 sondeos, a partir de cada uno se generan 100 variaciones, siendo la variante los espesores de cada unidad resistiva del modelo de interpretación. Establecidas las variaciones, se simular la respuesta eléctrica de cada SEV-sintético, obteniendo como resultado un set compuesto por 75000 registros de resistividad aparente.

Se realiza el entrenamiento empleando y comparando los modelos Support Vector Machines (SVM), Bayesian Compressive Sensing (BCS) y Random Forests (RF), los cuales se seleccionaron a partir de su tolerancia a las características no paramétricas de los datos, el alto nivel de ruido que presente en la respuesta geoeléctrica, si como una relación compleja entre las variables.

A partir de los resultados del entrenamiento se realiza predicciones empleando

como datos de entrada los datos reales de adquisición de un SEV ($AB/2$ y Rha), se realiza una comparación entre las puntuaciones (scorts) de los modelos, los distintos grupos de entrenamiento, identificando la mejor respuesta y el mejor ajuste en la predicción final.

El entrenamiento y predicción representa un gran oportunidad en el proceso de planeación y adquisición de datos geofísicos, permitiendo generar establecer un conjunto de entrenamiento con pocos datos de entrada, procurando que estos sean confiables preferentemente obtenidos por exploración directa mediante reconocimiento geológico, SPT o PCA, de manera que el resultado de predicción de aperturas nos permita definir de mejor manera el muestreo de unidades geológicas.

Capítulo 2
DELIMITACIÓN Y PLANTEAMIENTO DEL PROBLEMA

Durante un trabajo de prospección geofísica, al realizar adquisición de datos geoeléctricos *in situ*, no es posible conocer el resultado del trabajo hasta una vez realizado el procesado de los mismos, por lo que no se tiene certeza de si la adquisición realizada en campo representara con claridad el objeto de prospección, es decir, si el muestreo realizado logra cubrir el espectro de frecuencia, y por consiguiente, representar con claridad las unidades geológicas de un sitio en particular.

El muestreo propuesto en la etapa de planeación de adquisición, en muchas ocasiones incorpora un grado alto de ambigüedad, debido a que no es posible conocer con certeza la distribución y espesores de las unidades geoeléctrica y por lo tanto no es factible un muestreo completamente efectivo, siendo solo parcialmente evidente durante la exploración directa del medio, ya que dicho procedimiento solo permite aprecia una porción ínfima del terreno.

Se plantea como herramienta de análisis y mejora de muestreo la implementación de técnicas de Machine Learning, empleando esta técnica de aprendizaje mediante su entrenamiento con datos procesados y calibrados por sondeo directo, de manera que esto permita identificar oportunidades de mejora en el muestreo y la respuesta en general.

Capítulo 3
JUSTIFICACIÓN

Existen múltiples aplicaciones de ML y DL en el procesamiento, modelado e interpretación geofísica (Li *et al.*, 2024; Liu *et al.*, 2020; El-Qady y Ushijima, 2001; Wrona *et al.*, 2018), algunas de estas aplicaciones corresponden a implementaciones académicas, al igual que comerciales, siendo implementadas en software principalmente de exploración sísmica de hidrocarburos (Diaferia *et al.*, 2024; Panebianco *et al.*, 2024).

La aplicación de herramientas de ML durante la ejecutar de muestreo de Sondeo Eléctrico Vertical (SEV), en particular durante el proceso de adquisición *in situ*, presenta la posibilidad de evaluar y ampliar el intervalo de muestreo original, mejorando la adquisición tradicional, realizando una comparación del muestreo contra la regresiones a fin de establecer nuevos intervalos de apertura de electrodos no considerados previamente.

Durante el muestreo tradicional los intervalos se diseñan considerando el objetivo de exploración (idealmente), este análisis previo es un factor determinante, en esta etapa permitiendo establecer el intervalo de muestreo que permitirá identificar el objeto de exploración, este objeto o anomalía puede asociarse a unidades geológicas, acuíferos, fallas, zonas de fracturas, estructuras antropogénicas, infraestructura moderna, etc.

De manera general se busca mantener un intervalo de muestreo menor a la frecuencia de ocurrencia del objetivo de estudio, por lo que el éxito de la exploración dependerá en su totalidad de la planeación previa de la adquisición, lo que implicaría conocer previamente la conformación, distribución y espesor de cada unidad, siendo no viable, por lo que la interpretación puede presentar una alta ambigüedad.

De manera que un modelo entrenado permita generar múltiples modelos e identificar regiones de interés no cubiertas por el muestreos a través de predicciones y de esta manera generar puntos adicionales *in situ* optimizando la adquisición de datos e impactando en la calidad del muestreo, mejorando el acotamiento de la respuesta geoeléctrica identificada.

Pese a existir aplicaciones de ML implementadas en geofísica, no se identifica alguna enfocada este esta problemática en particular, sin embargo hay ejemplos en otros campos de estudio con un enfoque en el muestreo y clasificación de datos no paramétricos (Entezami *et al.*, 2022; Bkassiny *et al.*, 2013; Shi y Wang, 2021) empleando técnicas como Dirichlet Process Mixture Model (DPMM), radial basis function network (RBFN), Multiple Point Statistics (MPS) y Bayesian Compressive Sensing (BCS).

Para poder abordar la problemática se requiere de un modelo robusto ante el ruido, que permita trabajar con datos no paramétricos, sea favorable a la distribución de los datos, pueda establecer un modelo mediante entrenamiento supervisado, con capacidad de ejecutar regresiones a partir de entrenamientos previos y permita realizar predicciones de valores de resistividad. Estas condiciones son cubiertas por tres modelos de ML,Random Forests (RF), Support Vector Machines (SVM) y Gradient Boosting Regression (GBR), e identificar mediante la comparación entre los modelos, la puntuación de predicción, la facilidad de implementación y ejecución en pruebas con datos reales, cual presenta el mejor rendimiento y ajuste con respecto a dotos reales, para su implementación en la adquisición geofísica.

Capítulo 4
FORMULACIÓN DE OBJETIVOS

4.1 Generales

- Establecer un modelo de entrenamiento de regresión mediante Machine Learning que genere una respuesta similar a durante adquisición de SEV's permitiendo al usuario identificar intervalos de muestreo en los cuales se pueda mejorar el contraste de respuesta.

4.2 Específicos

- Definir criterios y características para la generación de variantes acotadas para generar datos de entrada para entrenamiento.
- Establecer un modelo de regresión y predicción implementando ML a partir de datos de apertura AB/2 y Rha.
- Realizar un análisis comparativo entre la respuesta de los modelos y datos adquiridos en campo.

Capítulo 5
MARCO TEÓRICO

5.1 Geofísica y Geoeléctrica

5.1.1 Definición de Geofísica

En términos generales la geofísica es la aplicación de los principios físicos de la materia en el estudio del planeta Tierra, o cual quier otro cuerpo celeste, desde el campo magnético, pasando por los fenómenos atmosféricos al medio solido del subsuelo, hasta las profundidades del núcleo interno planetario, ya sea que se aproveche una fuente natural como la propagación de ondas elásticas generadas por sismo, ó bien, la inducción de campo electromagnético de fuente controlada (Parasnis, 2012; Reynolds, 2011; Lay y Wallace, 1995).

El nacimiento de la geofísica es relativamente reciente, la primera prospección geoeléctrica data de 1830 realizados por Fox (1830) en Cornwal, Reino Unido, donde aplico técnicas de Self-Potential en exploración de mineralización de sulfuro en vetas, la medición del potencial natural resulto altamente efectiva para la prospección de este tipo de mineralizaciones ya que su anomalía se caracterizaba por presentar una respuesta muy marcada con respecto al medio (Reynolds, 2011; Revil y Jardani, 2013).

5.1.2 Resistividad de la Tierra

De manera general la materia presenta propiedades físicas definidas a partir de los elementos que la integran, en primer orden la configuración atómica establece

las propiedades físicas, estas se definen a partir de la estructura de electrones, protones y neutrones que presentan los átomos; a su vez, las moléculas pueden estar conformadas por una clase especifica de átomos (moléculas homonucleares) o por conjuntos de diferentes tipos (compuestos), cuya conformación depende de factores físico-químicos (Tiab y Donaldson, 2024).

La configuración molecular inorgánica presente en la materia, definirá el tipo de estructura cristalina (mineral) que formarán, y por consiguiente esta provista de propiedades físicas definidas en base a su composición y estructura; esta configuración cristalina es la que encontramos en el medio geológico conformando los minerales que componen la estructura mineral de una unidad geológica (ver figura 5.1) (Gandhi y Sarkar, 2016; Tiab y Donaldson, 2024).

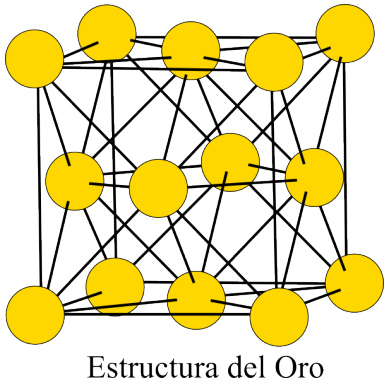

Figura 5.1: Esquema de la estructura atómica de oro que conforma la cristalización octahedral, modificado de Sorrell (1973)

.

Los métodos Geoeléctricos se clasifican en dos grupos, métodos pasivos y de inducción, los primeros corresponden a aquellos en los que se mide el potencial eléctrico natural, usualmente medido en mili volts, en donde se requiere de electrodos no polarizables para tener medidas lo más claras posibles; mientras que los métodos de inducción emplean un arreglo de electrodos, o inductores de campo electromagnéticos, mediante los cuales se induce un campo eléctrico al subsuelo, calculando la diferencia de potencia eléctrica en el medio, o bien, el decaimiento de la polarización

inducida (Revil y Jardani, 2013; Reynolds, 2011; Igboama *et al.*, 2023).

Los métodos de inducción, Sondeo Eléctrico Verticales (VES, por sus siglas en inglés), Tomografía de Resistividad Eléctrica (ERT, por sus siglas en inglés), Polarización Inducida (IP, por sus siglas en inglés), presentan una gran ventaja ya que no dependen del medio para poder realizar una lectura, ademas de poder realizarlos en cualquier momento, manteniendo el equipo en condiciones de operación, y pode diseñar arreglos de adquisición que nos permitan tener un muestreo tan amplio o limitado como sea conveniente, solo limitados por el alcance y potencia de los equipos empleados. Por otro lado su interpretación presenta un alta ambigüedad, solo acotado por la cantidad de referencias que puedan cruzarse para robustecer el modelo geológico y de inversión, y así poder llegar a una interpretación satisfactoria (Reynolds, 2011; Igboama *et al.*, 2023).

El método de prospección geoeléctrica, en especifico el SEV y la TRE, consiste en determinar la distribución de resistividades del subsuelo, de manera que se pueda establecer una correlación entre la resistividad y un modelo ajustado a la realidad geológica-estructural, geotécnica o geohidrológica del objeto de estudio.

5.1.3 Sondeo Eléctrico Vertical

Los SEV corresponden al método de mas rápida ejecución y económicamente mas accesible, por lo que es ampliamente empleado para solucionar problemas de ingeniería, minería, geotecnia, monitoreo e impacto ambiental y abastecimiento de Aguas potable; siendo de gran utilidad en la exploración de hidrogelogica ya que la respuesta resistiva de un medio saturado permite establecer diferencias concisas y discriminar entre agua dulce, salada, rocas fracturadas, arcillas , arenas, conglomerados, etc.

La resistividad es medida mediante la inyección de una corriente en el subsuelo y mientras que se monitorea y captura la diferencia de potencial eléctrico en la

superficie, esta lectura corresponde al valor de la contribución resistiva de todas las capas por donde fluye la corriente.

La inyección de corriente y medición del potencial se realiza a través de un arreglo de dos pares de electrodos, $A, B(C_1, C_2)$ y $M, N(P_1, P_2)$ respectivamente, siendo el electrodo $A(C_1)$ el polo positivo y $B(C_2)$ el polo negativo de inyección, mientras que el electrodo $M(P_1)$ corresponde al polo positivo y $N(P_2)$ al polo negativo de los electrodos de potencial.

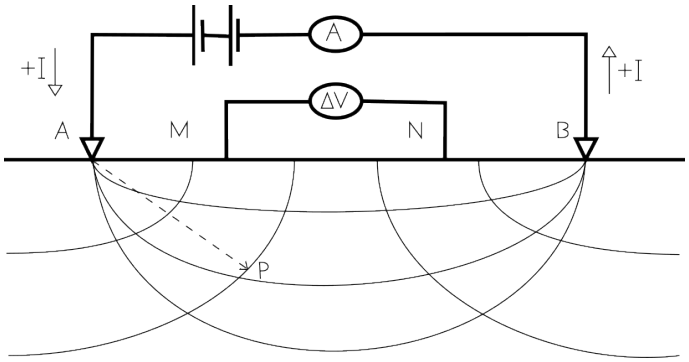

Figura 5.2: Configuración general de arreglo de electrodos, modificado de Reynolds (2011)

.

La resistividad del subsuelo se calcula a partir de la ley de Ohm, considerando el caso general en donde el medios es homogéneo y el arreglo de electrodos presenta una distribución convencional, donde se establece una relación directamente proporcional entre la la resistencia R ,medida en Ohm (Ω), y el cociente entre la diferencia de potencial ΔV y la corriente inducida I, para un valor puntual (Igboama *et al.*, 2023).

$$R = \frac{\Delta V}{I} \tag{5.1}$$

Sabiendo que se puede calcular R para una sección con longitud L y un área A, transversal del material, conociendo la resistividad (ρ) del material (Igboama *et al.*, 2023; Lowrie y Fichtner, 2020), podemos reescribir la ecuación como:

$$R = \rho\frac{L}{A} \rightarrow \rho = R\frac{A}{L} \rightarrow \rho = R \cdot k \tag{5.2}$$

Donde la resistividad (ρ) es una constante de proporcionalidad del medio y k es el factor geométrico de distribución del flujo de corriente en términos de la del arreglo de los electrodos de inducción y potencial (distancias entre los electrodos A-M-N-B) (Igboama *et al.*, 2023; Lowrie y Fichtner, 2020).

$$k = 2\pi\left(\frac{1}{AM} - \frac{1}{AN} - \frac{1}{BM} + \frac{1}{BN}\right) \tag{5.3}$$

Tenemos que la resistividad aparente (ρ_A) de una sección del subsuelo, corresponde a la contribución resistiva de las unidades geológicas en esa sección, en términos de las distancias entre electrodos, la diferencia de potencial y el flujo de corriente en el medio (Igboama *et al.*, 2023; Lowrie y Fichtner, 2020), esta dado por la siguiente ecuación:

$$\rho_A = \frac{\Delta U}{I} \cdot k \tag{5.4}$$

5.1.3.1 Arreglo de Electrodos y Factor Geométrico

Cada arreglo presenta ventas, desventajas, rango de sensibilidad y espacio de ejecución, debido a estas características y se tiene que evaluar e identificar que arreglo cumple con las condiciones adecuadas para ser ejecutado, considerando el espacio disponible en el sitio de estudio, el nivel de ruido (motores, conexiones a tierra mal aterrizadas, antenas, postes metálicos, arboles), la profundidad de objeto de prospección y la resolución vertical alcanzable (ver figura 5.3).

Como se observa en la sección anterior, la resistividad se determina empleando una configuración de los electrodos durante una medición, las distintas configuraciones de electrodos se encuentran ampliamente documentadas, cada una presenta

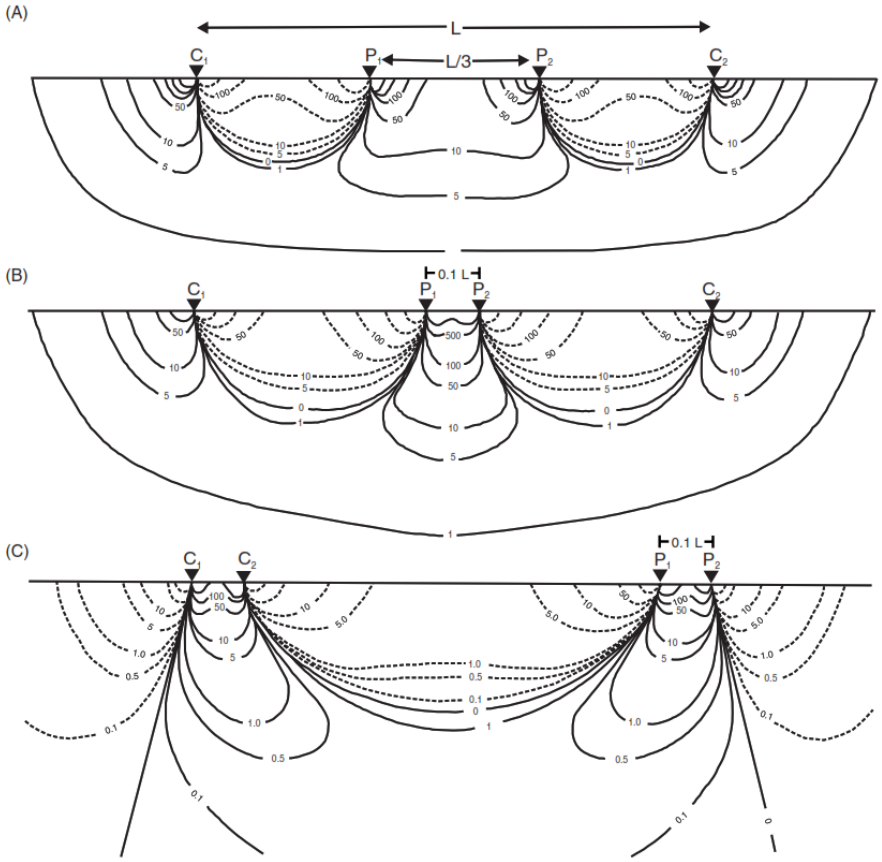

Figura 5.3: Esquema de la contribución de la respuesta de resistividad eléctrica, modificado de Reynolds (2011)

.

un factor geométrico distinto (Igboama *et al.*, 2023; Lowrie y Fichtner, 2020), los principales arreglos geoeléctricos son:

Wenner

$$\rho_A = 2\pi \cdot R \cdot a \tag{5.5}$$

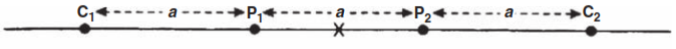

Figura 5.4: Esquema del arreglo Wenner, modificado de Reynolds (2011)

.

Schlumberger

$$\rho_A = \frac{\pi a^2}{b}\left[1 - \frac{b^2}{4a^2}\right] \cdot R, \quad a \geq 5b \tag{5.6}$$

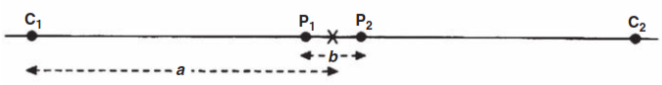

Figura 5.5: Esquema del arreglo Schlumberger, modificado de Reynolds (2011)

.

Dipolo-dipolo

$$\rho_A = \pi n(n+1)(n+2)a \cdot R \tag{5.7}$$

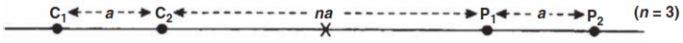

Figura 5.6: Esquema del arreglo Dipolo-dipolo, modificado de Reynolds (2011)

.

5.2 Adquisición de Datos Geofísicos

Previo al trabajo de adquisición se realiza un análisis de entorno, en el cual se verifica la viabilidad del arreglo dadas las condiciones del sitio, considerando lo siguiente: espacio disponible en el sitio de estudio, profundidad de exploración, nivel de ruido eléctrico, interferencias con la estabilidad del potencial natural del subsuelo, profundidad del objeto de exploración y dimensiones aproximadas del mismo.

5.2.1 Intervalo de Muestreo en SEV

El intervalo de muestreo empleado durante la adquisición de un SEV es un parámetro crítico que influye en la calidad y precisión de los datos geofísicos adquiridos, ya que esta estrechamente relacionado con la resolución vertical que deseamos de acuerdo al objeto de estudio. Durante la planeación es necesario considerar distintas condiciones, como son:

- Los espesores de cada unidad.
- La distribución de las distintas unidades.
- Profundidad de investigación
- Ruido en la señal.

Para establecer un intervalo de muestreo apropiado, se deben considerar el Teorema de Muestreo de Nyquist y El teorema de Shannon-Hartley (teorema de codificación de canal ruidoso), al igual que la geometría del arreglo geoeléctrico empleado.

El Teorema de Muestreo de Nyquist, el cual, es un principio fundamental en el procesamiento de señales analógicas y digitales, establece las condiciones mínimas necesarias para la reconstrucción una señal analógica a partir de muestras discretas (Alvarado Reyes y Stern Forgach, 2010).

Nyquist nos garantiza las condiciones necesarias y suficientes para llevar a cabo una adquisición exitosa de muestreo de una señal, llámese distribución de resistividad en un medio heterogéneo y discontinuo (Alvarado Reyes y Stern Forgach, 2010), considerando siempre los espesores como inferidos a partir de muestreo directo.

$$f_s \geq 2 \cdot f_{max} \tag{5.8}$$

Donde la frecuencia de muestreo f_s es por lo menos dos veces mayor a la frecuencia máxima f_{max} conocida, cuando el teorema no se cumple se genera una distorsión en la señal, sumando las frecuencias altas incompletas a la señal natural de baja frecuencia, generando ruido, y problemas de interpretación, se conoce como aliasing (Alvarado Reyes y Stern Forgach, 2010).

Considerando el medio geológico como una región con presencia constante de ruido eléctrico de fuentes tanto naturales como humanas, es imprescindible considerar el teorema de Shannon-Hartley aplicando apilamiento de muestreo como método

de reducción de la relación ruido señal, durante la adquisición de datos; esto quiere decir calcular el promedio de muestreos continuos en un intervalo definido de aperturas entre electrodos.

5.2.1.1 Factores que Determinan el Intervalo de Muestreo

En el contexto de la adquisición de datos mediante SEV, el intervalo de muestreo es equivalente al espaciado entre puntos donde se realizan mediciones de resistividad del subsuelo. Este intervalo de muestreo debe ser lo mas pequeño posible, de modo que permita obtener muestras de resistividad (Telford *et al.*, 1990), esta relación se define de la siguiente manera:

$$f_s = \frac{1}{\Delta x} \tag{5.9}$$

Donde el intervalo de muestreo Δx debe ser menor a la mitad de la longitud de onda (λ_{min}, espesor) asociado al objetivo de exploración.

$$\Delta x \leq \frac{\lambda_{min}}{2} \tag{5.10}$$

5.2.2 Proceso de Adquisición In Situ

La adquisición de datos se realiza mediante la lectura directa en campo, tomando una primera lectura del potencial natural por medio de los electrodos M y N, se continua con la lectura al inducir corriente continua empleando un resistivímetro mediante los electrodos de corriente A (C_1) y B (C_2), mientras se realiza la lectura de potencia en los electrodos M (P_1) y N (P_2), la lectura se realiza en intervalos regulares en instantes de inyección de corriente (Telford *et al.*, 1990).

Durante la toma de datos es importante considerar los modelos previos realizados durante el análisis preliminar, ya que las resistividades esperadas para las unida-

des, permiten tener control en la dispersión de datos, identificando tomas erróneas y corrigiendo al momento con una nueva lectura (Telford *et al.*, 1990).

5.3 Machine Learning en la Geofísica

La aplicación de ML en la geofísica es utilizado en exploración sísmica, abarcando los procesos de adquisición, mejorando los tiempos de procesamiento, clasificación e interpretación, ya que es en este método donde se cuenta con la mayor cantidad de datos para entrenamiento (Wrona *et al.*, 2018); en menor medida se implementan técnicas de ML en la exploración y prospección geoeléctrica, hay algunos ejemplos destacables como son Liu *et al.* (2020); El-Qady y Ushijima (2001); Li *et al.* (2024), sin embargo no es un estándar en la metodología, pese a las ventajas que puede tener su aplicación, como se pretende demostrar en este estudio.

El machine learning se integra por conjunto de técnicas que utilizan algoritmos con los cuales permite a un sistema aprender y generar predicciones, para lo que requiere un conjunto de datos para poder realizar el entrenamiento.

Podemos clasificar los algoritmos de ML por el tipo de entrenamiento que ejecutan, correspondiendo a aprendizaje supervisado, no supervisado y por refuerzo, y por la relación que establecen con los parámetros del conjunto de datos de entrenamiento, es decir, modelos paramétricos y no paramétricos (Li *et al.*, 2024).

De los modelos no paramétricos destacan por su adaptabilidad a la estructura subyacente de los datos, por lo que pueden realizar aprendizaje de relaciones complejas entre datos, así como ausentes de linealidad, teniendo un costo en volumen de datos, requiriendo un número mayor para su entrenamiento, destacan los siguientes algoritmos de ML.

- Random Forests
- Support Vector Machines

- Gradient Boosting Regression

Dada la naturaleza de los datos de SEV's, heterogéneos, discontinuos y no lineales, es conveniente abordar su analistas desde un enfoque no paramétrico, teniendo esto en cuanta, los métodos empleados destaca siendo eficaz en la tarea de clasificación y regresión, teniendo algunos beneficios como son la reducción del sobre ajuste, interpretación de variables, resistencia al aliasing.

5.4 Random Forests

Random Forests es una técnica propuesta por Breiman (2001), la cual emplea múltiples arboles de decisión independientes entre si, donde cada árbol realiza una votación de clases, seleccionando la más popular de la entrada de cada árbol realizando una combinación de salida, permitiendo realizar una clasificación de características complejas o realizar regresiones de datos complejos multivariables (Breiman, 2001; Lan *et al.*, 2020).

La herramienta de Random Forests, de acuerdo con Breiman (2001) emplea tres elementos clave en el proceso de entrenamiento, bagging, selección aleatoria de características y agregación por votación, resultando en la combinación del los resultados en una predicción o clasificación robusta y ajustada (Lan *et al.*, 2020).

5.4.1 Bootstrap aggregating

Esta característica de Random Forest genera una colección de M árboles de decisión no correlacionados, cada uno de ellos se entrena utilizando muestras aleatorias del conjunto de datos originales, identificados como datos de entrenamiento D, este proceso se conoce como bootstrap aggregating (Breiman, 2001).

Deacuerdo con Breiman (2001), Random Forests es un conjunto de clasificadores $H(x, \theta_k)$, x es un vector de entrada y θ_k corresponden a vectores aleatorios

independientes generados a partir de los datos de entrenamiento D, con árboles k (donde $k = 1, 2, ..., M$), esta aleatoriedad puede generar datos replicados en Θ_k, mientras que otros arboles pueden estar faltantes. La dimensión Θ_k es aproximadamente el 80 % del tamaño de D.

5.4.2 Selección Aleatoria de Características

Al construir cada árbol, en cada nodo, en lugar de considerar todas las características para encontrar la mejor división, se selecciona aleatoriamente un subconjunto de K características

En cada nodo de los subconjuntos de entrenamiento se selecciona una características por votación de popularidad, dejando crecer cada árbol sin realizar poda hasta completar los criterios de finalización, es decir un numero de instancias preestablecido(Breiman, 2001).

5.4.3 Predicción por Agregación

La clasificación para una entrada x se basa en las predicciones individuales de los árboles para cada clase $h_k(x)$, se realiza un conteo de cada clase, producto de la predicción de cada árbol, sumando las salidas $I(h_k(x) = c)$, y finalmente se selecciona clase con mayor numero de predicciones, obteniendo la predicción de clasificación $H(x)$,donde x es una función indicadora que vale 1 si $h_k(x) = c$, y 0 en caso contrario (Breiman, 2001).

$$H(x) = \text{argmax}_c \sum_{k=1}^{K} I(h_k(x) = c) \tag{5.11}$$

El proceso de la regresión se obtiene a partir de la media aritmética de cada predicción individual, donde cada árbol produce un valor numérico $h_k(x)$ correspondiente a cada x, al corresponder con promedio de las predicciones se le otorga mas

estabilidad cuando tenemos un numero elevado de arboles y un conjunto de datos grande, entendiéndolo como un modelo central que incorpora información de cada árbol (Breiman, 2001).

$$H(x) = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \tag{5.12}$$

5.4.4 Varianza y Overfitting

Algo notable en el algoritmo Random Forest es su capacidad para reducir la varianza sin que esto afecte significativamente el sesgo, lo que ayuda a evitar el overfitting (sobre ajuste), ya que el bagging, ejecutado durante el proceso de bootstrap, reduce la varianza al promediar las predicciones de múltiples modelos entrenados en diferentes árboles. La selección aleatoria de características descorrelaciona aún más los árboles, lo que contribuye a la reducción de la varianza (Breiman, 2001; Veirana *et al.*, 2021).

La varianza de un Random Forest puede expresarse aproximadamente como:

$$Var(JM_{RF}) = \rho(x)\sigma^2 + \frac{1-\rho(x)}{M}\sigma^2 \tag{5.13}$$

Donde σ^2 es la varianza promedio de cada árbol y $\rho(x)$ es la correlación promedio entre las predicciones de cualquier par de árboles k. A medida que el número de árboles M aumenta, el segundo termino tiende a cero, acercando la varianza a $\rho(x)\sigma^2$ (Breiman, 2001; Veirana *et al.*, 2021).

5.4.5 Bias-Varianza Tradeoff

El objetivo en el aprendizaje automático es encontrar un modelo que minimice el error de prueba esperado ($E[E_{TEST}]$)(Breiman, 2001), el cual puede descompo-

nerse en sesgo al cuadrado ($Bias^2$), varianza (Var) y ruido ($Noise$):

$$E[E_{TEST}] = Bias^2 + Var + Noise$$

Con el método Random Forest se busca un equilibrio entre sesgo y varianza, considerando que los árboles individuales por lo general tienen un bajo sesgo y alta varianza, siendo equilibrado por el proceso bootstrap al reducir la varianza del ensamble, resultando en un mejor rendimiento general en datos no vistos(Breiman, 2001).

5.5 Support Vector Machines

Este algoritmo de aprendizaje supervisado tiene como objetivo encontrar el hiperplano que mejor separa las clases en el espacio de características, maximizando la distancia entre este y los puntos más cercanos de cada clase, conocidos como vectores de soporte. En casos donde los datos no son linealmente separables, permite ciertas violaciones del margen mediante un parámetro de penalización que controla el equilibrio entre precisión y generalización, además, mediante el uso de funciones kernel, es posible proyectar los datos a espacios de mayor dimensión para lograr una separación lineal que no sería posible en el espacio original(James *et al.*, 2013; Veirana *et al.*, 2021).

En problemas de regresión se emplea la variante Support Vector Regression (SVR). En este caso, el modelo busca una función que se mantenga dentro de un margen de tolerancia ϵ respecto a los valores reales, permitiendo cierto grado de error controlado por variables de holgura(James *et al.*, 2013).

Tanto en clasificación como en regresión, SVM es una técnica robusta, especialmente útil en espacios de alta dimensión, aunque su desempeño depende de una adecuada elección de hiperparámetros y de kernel.

5.5.1 Hiperplano Separador

En la clasificación binaria más simple, cuando los datos son linealmente separables, el objetivo es encontrar un hiperplano que divida claramente las dos clases. En un espacio de p dimensiones, dicho hiperplano se define por la ecuación (James *et al.*, 2013):

$$w^T x + b = 0$$

donde:

- w es un vector de pesos que determina la orientación del hiperplano,
- x es el vector de características de una instancia,
- b es un escalar que define la posición del hiperplano (sesgo).

Un hiperplano separador cumple que $w^T x + b > 0$ para una clase y $w^T x + b < 0$ para la otra.

5.5.2 Clasificador de Margen Máximo

El clasificador de margen máximo busca el hiperplano que maximiza la distancia (margen) entre las observaciones más cercanas de cada clase y el propio hiperplano (James *et al.*, 2013). Estas observaciones se denominan *vectores de soporte*.

El problema de optimización se formula como:

$$\underset{w,b}{\text{minimizar}} \quad \frac{1}{2}\|w\|^2 \quad \text{sujeto a} \quad y_i(w^T x_i + b) \geq 1, \quad i = 1, ..., n$$

donde $y_i \in \{-1, 1\}$ representa las clases. Esta formulación garantiza la separación de clases con el mayor margen posible.

5.5.3 Clasificador de Margen Suave (Soft Margin)

Cuando los datos no son perfectamente separables, se introducen variables de holgura $\xi_i \geq 0$ para permitir ciertas violaciones del margen:

$$\underset{w,b,\xi}{\text{minimizar}} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{sujeto a} \quad y_i(w^T x_i + b) \geq 1 - \xi_i$$

Aquí, $C > 0$ es un hiperparámetro que controla el compromiso entre la maximización del margen y la penalización por errores de clasificación (James *et al.*, 2013; Veirana *et al.*, 2021).

5.5.4 Método del Kernel

Para datos no linealmente separables, SVM emplea el *truco del kernel*, que permite proyectar los datos a un espacio de mayor dimensión donde sí pueden ser separados linealmente, sin necesidad de calcular dicha proyección explícitamente (James *et al.*, 2013; Veirana *et al.*, 2021).

Funciones kernel comunes:

- **Lineal**: $K(x_i, x_j) = x_i^T x_j$
- **Polinomial**: $K(x_i, x_j) = (x_i^T x_j + r)^d$
- **RBF (Gaussiano)**: $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$
- **Sigmoidal**: $K(x_i, x_j) = \tanh(\alpha x_i^T x_j + c)$

El problema de optimización dual con kernel se escribe como:

$$\underset{\alpha}{\text{maximizar}} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$

$$\text{sujeto a} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$$

La función de decisión para una nueva instancia x es:

$$f(x) = \text{sgn}\left(\sum_{i \in SV} \alpha_i y_i K(x, x_i) + b\right)$$

5.5.5 Regresión por Vectores de Soporte (SVR)

SVM también puede aplicarse a regresión mediante SVR, cuyo objetivo es encontrar una función $f(x)$ que tenga una desviación máxima de ϵ respecto a los valores verdaderos, penalizando sólo los errores mayores (James *et al.*, 2013; Veirana *et al.*, 2021).

El problema se plantea como:

$$\underset{w,b,\xi,\xi^*}{\text{minimizar}} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{sujeto a} \quad \begin{cases} y_i - (w^T x_i + b) \leq \epsilon + \xi_i \\ (w^T x_i + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Donde ϵ define una región sin penalización y C regula el compromiso entre la planitud del modelo y la tolerancia a errores.

5.6 Gradient Boosting Regression

Gradient Boosting Regression es un método de aprendizaje ensamblado que construye modelos predictivos de manera secuencial. A diferencia de técnicas como Bagging y Random Forest, donde los árboles de decisión se construyen de forma independiente y se combinan al final, en Gradient Boosting cada nuevo árbol se entrena específicamente para corregir los errores cometidos por la suma de árboles anteriores. Esto permite una mejora progresiva en el desempeño del modelo a lo largo de múltiples iteraciones (James *et al.*, 2013; Hastie *et al.*, 2009; Friedman, 2001).

5.6.1 Naturaleza Secuencial

El modelo se construye de manera iterativa. Se inicia con una predicción base (por ejemplo, el promedio de las salidas) y, en cada iteración, se ajusta un nuevo árbol de decisión sobre los errores cometidos por la suma de modelos anteriores. El resultado es un modelo compuesto por muchos árboles débiles, cuya combinación forma un modelo fuerte.

5.6.2 Minimización de la Función de Pérdida

En cada iteración, el algoritmo intenta minimizar una función de pérdida $L(y_i, \hat{f}(x_i))$, que cuantifica la discrepancia entre los valores reales y_i y las predicciones actuales $\hat{f}(x_i)$. La elección de la función de pérdida depende del tipo de problema (por ejemplo, el error cuadrático medio para regresión) (James *et al.*, 2013).

5.6.3 Ajuste a los Residuales

El nuevo árbol se entrena sobre los *pseudo-residuales*, definidos como el gradiente negativo de la función de pérdida con respecto a las predicciones actuales. En el caso de regresión con error cuadrático, estos pseudo-residuales coinciden con los residuales clásicos:

$$r_i^{(b)} = y_i - \hat{f}^{(b-1)}(x_i)$$

Esto permite que cada nuevo árbol aprenda los patrones de error que el modelo anterior no pudo capturar.

5.6.4 Shrinkage (Tasa de Aprendizaje)

Para evitar que cada nuevo árbol domine demasiado el modelo, se introduce un parámetro de *shrinkage* o *tasa de aprendizaje*, denotado como $\lambda \in (0, 1)$ (James

et al., 2013; Veirana *et al.*, 2021). Este factor escala la contribución de cada árbol:

$$\hat{f}^{(b)}(x) = \hat{f}^{(b-1)}(x) + \lambda \cdot f_b(x)$$

Valores pequeños de λ (como 0.01 o 0.001) ralentizan el aprendizaje, lo cual puede mejorar la generalización del modelo.

5.6.5 Regularización

Además del parámetro λ, se pueden emplear otras técnicas de regularización, como la profundidad máxima de los árboles (d), el número mínimo de muestras por hoja, o la fracción de datos usados en cada iteración (submuestreo) (James *et al.*, 2013). Estas estrategias ayudan a reducir la varianza del modelo y mitigar el riesgo de sobreajuste.

5.6.5.1 Algoritmo de Gradient Boosting para Regresión

Antes de implementar este algoritmo es importante entender su dinámica iterativa y cómo se construyen las predicciones dentro del algoritmo, en donde optimiza una función de pérdida específica de manera directa mediante técnicas de gradiente, agregando modelos débiles (árboles de decisión) que corrigen los errores residuales del conjunto de modelos anteriores. El procedimiento se fundamenta en la minimización de una función de pérdida diferenciable, adaptando sucesivamente el modelo a los errores cometidos(James *et al.*, 2013; Hastie *et al.*, 2009; Friedman, 2001).

El siguiente esquema detalla las etapas de inicialización, el cálculo iterativo de los pseudo-residuales, el ajuste de árboles de regresión sobre dichos residuales y la actualización acumulativa del modelo predictivo, incluye la forma general de la salida final, que representa la suma ponderada de los árboles ajustados durante el proceso (James *et al.*, 2013; Hastie *et al.*, 2009; Friedman, 2001).

1. **Inicialización:** Se define una predicción inicial como:

$$\hat{f}^{(0)}(x) = \arg\min_{c} \sum_{i=1}^{n} L(y_i, c)$$

Por ejemplo, si L es el error cuadrático, entonces $\hat{f}^{(0)}(x) = \bar{y}$.

2. **Para cada iteración** $b = 1, 2, ..., B$**:**

a) Cálculo de pseudo-residuales:

$$r_i^{(b)} = -\left[\frac{\partial L(y_i, \hat{f}^{(b-1)}(x_i))}{\partial \hat{f}(x_i)}\right]$$

(En el caso del error cuadrático: $r_i^{(b)} = y_i - \hat{f}^{(b-1)}(x_i)$)

b) Entrenar un árbol de regresión $f_b(x)$ sobre el conjunto $\{x_i, r_i^{(b)}\}$.

c) Actualizar el modelo:

$$\hat{f}^{(b)}(x) = \hat{f}^{(b-1)}(x) + \lambda f_b(x)$$

3. **Salida final:**

$$\hat{f}(x) = \hat{f}^{(B)}(x) = \hat{f}^{(0)}(x) + \lambda \sum_{b=1}^{B} f_b(x)$$

5.6.5.2 Parámetros de Ajuste Importantes

El rendimiento y la capacidad de generalización de un modelo de Gradient Boosting dependende la configuración adecuada de ciertos hiperparámetros, permitiendo controlar tanto la complejidad del modelo como su comportamiento durante el entrenamiento, esto tiene un efecto directo en su precisión, robustez y riesgo de sobreajuste.

- **Número de árboles (B):** Controla la complejidad general del modelo. Más árboles pueden mejorar el ajuste, pero incrementan el riesgo de sobreajuste si no se regula adecuadamente.

- **Tasa de aprendizaje (λ):** Controla cuánto contribuye cada árbol al modelo final. Tiempos de entrenamiento más largos con λ pequeño suelen generar mejores modelos.

- **Profundidad del árbol (d):** Afecta la complejidad de cada árbol individual. Árboles muy profundos pueden capturar relaciones complejas pero también inducir sobreajuste.

5.6.5.3 Compensación Bias-Varianza

Gradient Boosting busca simultáneamente reducir el sesgo (bias) y la varianza del modelo. Al agregar secuencialmente modelos que corrigen errores anteriores, se reduce el sesgo. A su vez, mediante técnicas como el shrinkage, la poda de árboles y el submuestreo, se controla la varianza para evitar que el modelo se ajuste demasiado a los datos de entrenamiento (James *et al.*, 2013).

En resumen, el algoritmo Gradient Boosting Regression construye un modelo fuerte combinando múltiples árboles débiles entrenados de forma secuencial. Cada nuevo árbol se enfoca en aprender los errores del modelo anterior, y su contribución se regula mediante una tasa de aprendizaje. Gracias a su capacidad para capturar patrones complejos y su flexibilidad, se ha convertido en una de las técnicas más potentes y ampliamente utilizadas en tareas de regresión y clasificación (James *et al.*, 2013; Veirana *et al.*, 2021; Hastie *et al.*, 2009; Friedman, 2001).

Capítulo 6
METODOLOGÍA

6.1 ideas y apuntes

Adquisición de datos y preprocesamiento: En primer lugar, se realiza la adquisición de datos in situ utilizando el método de Sondeo Eléctrico Vertical (SEV), el cual consiste en medir la resistividad del terreno a diferentes profundidades con un intervalo de muestreo predefinido. Este intervalo se determina de acuerdo con el objetivo de exploración, como la identificación de unidades geológicas, acuíferos, fallas, fracturas o estructuras antropogénicas. Para garantizar que los datos sean adecuados para el análisis, se lleva a cabo un preprocesamiento de los datos, que incluye la limpieza de valores atípicos, normalización de las lecturas y manejo de valores faltantes.

6.2 Variables de Entrada

Al observar las ecuaciones 5.1, 5.3 y 5.4, es posible identificar las variables involucradas en el calculo de la resistividad aparente, estos valores son medidos por un equipo automático o bien, de forma manual a través de la lectura directa en un resistivímetro, para lo cual se requiere de comprobaciones durante la adquisición.

Como observamos en la tabla 6.1, se integran datos generados durante la planeación y valores medidos en la etapa de adquisición; Z= profundidad aparente de exploración; K= factor geométrico; AB/2= apertura total de muestreo entre dos; MN= distancia entre electrodos de potencial; Pn= potencial natural; Pi= poten-

Z	K	AB/2	*AB/5 >*	*MN*	*>AB/20*	Pn	Pi	I	Pn	Pi	I	Pn	Pi	I	PPn	PPi	U	PI	Rha
1.8	**56.549**	**3**	***1.2***	***0.5***	***0.3***	1	6	15	0	6	15	0	5	15	0.3	5.7	5.3	15.0	20.11
2.4	**100.531**	**4**	***1.6***	***0.5***	***0.4***	12	19	3	0	25	17	5	24	13	5.7	22.7	17.0	11.0	176.45
3	**157.080**	**5**	***2***	***0.5***	***0.5***	8	54	73	8	52	85	8	40	83	8.0	48.7	40.7	80.3	80.28
3	**31.416**	**5**	***2***	***2.5***	***0.5***	52	508	42	52	648	72	52	606	21	52.0	587.3	535.3	45.0	476.64
4.8	**80.425**	**8**	***3.2***	***2.5***	***0.8***	48	60	6	48	62	7	51	59	4	49.0	60.3	11.3	5.7	160.85
6	**125.664**	**10**	***4***	***2.5***	***1***	52	60	2	52	60	5	52	60	5	52.0	60.0	8.0	4.0	301.59
9	**282.743**	**15**	***6***	***2.5***	***1.5***	52	76	54	52	78	87	50	64	44	51.3	72.7	21.3	61.7	100.04

Tabla 6.1: Ejemplo de atributos empleados en el calculo de la resistividad aparente.

cial inducido; I= corriente Inducida; PPn= promedio del potencial natural; PPi= promedio de potencial inducido; U= diferencia entre PPn y PPi; PI= promedio de corriente inducida; Rha= resistividad aparente ponderada.

En términos generales, el proceso de adquisición consiste en la inducción de una corriente eléctrica a través del medio geológico, dicha intensidad de corriente es registrada, junto con el valor del potencial natural (Self-Potential) y el potencial inducido, generado por la inyección de corriente a tierra, obteniendo así los elementos necesarios para calcular el valor de la resistividad, habiendo previamente planeado la configuración geométrica del arreglo.

6.2.1 Datos de entrada

Para los datos de entrada se emplearon levantamientos de SEV, empleando la configuración geométrica Shlumberger, previamente procesados e interpretados, ya sea mediante correlación geológica o con muestreo directo por sondeo de penetración estándar (SPT por sus siglas en inglés), correspondientes a ambiente de deposito sedimentario y flujos volcánico, en ambos caso subyaciendo a unidades sedimentarias recientes.

A partir de estos resultados etiquetados, valores de resistividad aparente, es como de modela variaciones, modificando el espesor de las unidades, ya que cada muestreo de resistividad aparente integra la respuesta conjunta de las unidades que

la preceden, es decir las capas geoeléctricas por arriba de la profundidad aparente de exploración, para iguar las condiciones en los modelos que integran los datos de entrenamiento, se emplea la Librería PyGIMLI, la cual esta preparada para realizar esta tarea.

Los datos corresponden a trabajos realizados en en distintas condiciones geológicas, correspondientes a proyectos de exploración hidrológica y minera, se integran un total de 99999 SEV's, procesados, interpretados y validados, a partir de los cuales se generaran las variantes para generar la base de entrenamiento.

La información particular de nombres de proyectos, localidad, ubicación geográfica o cualquier información que pueda relacionar directamente al propietario del proyecto, son omitidos.

6.2.1.1 Limpieza de datos

En esta etapa se consideran los siguientes criterios para la selección y limpieza de datos, permitiendo detectar y corregir errores, identificar valores atípicos en el muestreo así como datos inconsistentes.

1.- Interpretacion geologica de los perfiles Deberán incluir interpretación geológica de los resultados de inversión, es decir, es necesario conocer la unidad geológica correspondiente al modelo de resistividad, correspondiendo a etiquetas de datos.

2.- Muestreo continuos a intervalos regularres En caso de que no se cuente con un muestreo adecuado, se integraran los intervalos faltantes, de manera que se modele la señal completa en un muestreo extenso, considerando como mínimo 30 datos de muestreo por sondeo.

3.- profundidad de exploracion calculada Poder identificar durante la adquisición, la profundidad de exploración y los valores de resistividad aparente en

el medio, permite detectar variaciones o puntos de inflexión en la curva de la señal.

4.- Valores de resitividad atipicos Estos errores por inconsistencia pueden surgir por una mala lectura en campo, al no identificar un cambio de polaridad en el medio.

5.- Valores duplicados identificación de modelos duplicados en los registros.

6.2.2 Generación de datos de entrenamiento

a partir d

6.2.2.1 Resistividad aparente

6.2.2.2 Atributos cualitativos asociados a la curva de resistividad

etiquetas de acuerdo a unidades geológicas especificas

Clasificación de las lecturas con Random Forests: Se emplea el algoritmo Random Forests, una técnica de aprendizaje automático no paramétrica que consiste en crear múltiples árboles de decisión que luego se combinan para mejorar la precisión del modelo. El modelo se entrena utilizando las lecturas de resistividad obtenidas de las distintas profundidades y las características asociadas (como el tipo de terreno, las propiedades geológicas conocidas o las variables del sondeo). Los árboles se entrenan para clasificar los datos en diferentes categorías, tales como las unidades geológicas presentes, los acuíferos, las fracturas, entre otros. La clasificación permitirá identificar patrones en las lecturas de resistividad que corresponden a diferentes tipos de formaciones geológicas.

Generación de la regresión para optimizar el intervalo de muestreo: Una vez que se haya realizado la clasificación, se utiliza el modelo de re-

gresión basado en Random Forests para predecir la resistividad eléctrica en profundidades no muestreadas. Esto permitirá estimar la resistividad del terreno en puntos específicos que no han sido cubiertos por el muestreo inicial. A partir de estas predicciones, se podrán proponer intervalos de muestreo adicionales in situ que mejoren la representación de las formaciones geológicas de interés. La regresión también proporcionará un modelo predictivo que puede ajustarse dinámicamente para adaptar los intervalos de muestreo según las características del terreno y los objetivos de exploración.

Evaluación del modelo y ajuste de parámetros: Se realiza una evaluación exhaustiva del modelo mediante técnicas de validación cruzada para asegurarse de que el modelo esté bien entrenado y sea capaz de generalizar correctamente a nuevos datos. Además, se compara el rendimiento del modelo de Random Forests con otras técnicas de clasificación y regresión para determinar su eficacia en comparación con otros enfoques. Se analizan métricas como la precisión en la clasificación, el error cuadrático medio (RMSE) en la regresión y la capacidad de predicción en términos de muestreos adicionales.

Optimización y mejora continua: Finalmente, se optimizan los parámetros del modelo (como el número de árboles y la profundidad de los mismos) para mejorar la precisión y eficiencia del modelo. A medida que se incorporan nuevos datos de exploración y se obtienen más lecturas de resistividad, el modelo puede ser recalibrado y ajustado para mantener su efectividad en la identificación de objetivos geofísicos y en la optimización del intervalo de muestreo.

6.2.3 Clasificación, transformación y escalado de los datos

6.2.3.1 Normalización o estandarización de resistividades si se observan grandes variaciones

Normalización y escalado: Al tener datos con escalas diferentes, es decir, sistemas de unidades de medición muy distintas como en este caso donde encontramos valores de Resistividad en Ohm·m y distancias en metros.

6.2.3.2 Transformación logarítmica de resistividad para reducir el sesgo de valores extremos

6.2.3.3 Codificación de categorías litológicas si se incluyen como variable adicional

6.3 Diseño de los Modelos ML

6.3.1 Regresión

6.3.2 Configuración inicial del modelo

6.3.3 Configuración y optimización de hiperparametros

6.4 Preparación del dataset para la implementación del modelo

6.5 Implementación del modelo

6.5.1 Entrenamiento del modelo

6.5.2 Mapas de probabilidad y entrenamiento de regresión

6.6 Evaluación del modelos

6.6.1 Regresión y validación cruzada

6.6.2 Análisis de incertidumbre

6.7 Reporte estadístico

Capítulo 7
RESULTADOS Y CONCLUSIONES

Apéndice A
Apéndice I

Bibliografía

Alvarado Reyes, J. y C. Stern Forgach (2010), «Un complemento al teorema de Nyquist», *Revista mexicana de física E*, **56**(2), págs. 165–171.

Bkassiny, M., S. K. Jayaweera y Y. Li (2013), «Multidimensional dirichlet process-based non-parametric signal classification for autonomous self-learning cognitive radios», *IEEE Transactions on Wireless Communications*, **12**(11), págs. 5413–5423.

Breiman, L. (2001), «Random forests», *Machine learning*, **45**, págs. 5–32.

Diaferia, G., L. Valoroso, L. Improta y D. Piccinini (2024), «A high-resolution seismic catalog for the Southern Apennines (Italy) built through template-matching», *Geochemistry, Geophysics, Geosystems*, **25**(3), pág. e2023GC011 160.

El-Qady, G. y K. Ushijima (2001), «Inversion of DC resistivity data using neural networks», *Geophysical Prospecting*, **49**(4), págs. 417–430.

Entezami, A., H. Shariatmadar y C. De Michele (2022), «Non-parametric empirical machine learning for short-term and long-term structural health monitoring», *Structural Health Monitoring*, **21**(6), págs. 2700–2718.

Fox, R. W. (1830), «On the electro-magnetic properties of metalliferous veins in the mines of Cornwall», *Philosophical Transactions of the Royal Society of London*, págs. 399–414.

Friedman, J. H. (2001), «Greedy Function Approximation: A Gradient Boosting Machine», *Annals of Statistics*, **29**(5), págs. 1189–1232.

GANDHI, S. y B. SARKAR (2016), *Essentials of mineral exploration and evaluation*, Elsevier.

HASTIE, T., R. TIBSHIRANI y J. FRIEDMAN (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, segunda edición, Springer, New York.

IGBOAMA, W. N., M. AROYEHUN, J. AMOSUN, O. AYANDA, O. HAMMED y J. OLOWOFELA (2023), «Review of geoelectrical methods in geophysical exploration», *Nigerian Journal of Physics*, **32**(3), págs. 141–158.

JAMES, G., D. WITTEN, T. HASTIE, R. TIBSHIRANI *et al.* (2013), *An introduction to statistical learning*, tomo 112, Springer.

LAN, T., H. HU, C. JIANG, G. YANG y Z. ZHAO (2020), «A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification», *Advances in Space Research*, **65**(8), págs. 2052–2061.

LAY, T. y T. C. WALLACE (1995), *Modern global seismology*, Elsevier.

LI, M., S. YIN, Z. LIU y H. ZHANG (2024), «Machine learning enables electrical resistivity modeling of printed lines in aerosol jet 3D printing», *Scientific Reports*, **14**(1), pág. 14 614.

LIU, B., Q. GUO, S. LI, B. LIU, Y. REN, Y. PANG, X. GUO, L. LIU y P. JIANG (2020), «Deep learning inversion of electrical resistivity data», *IEEE Transactions on Geoscience and Remote Sensing*, **58**(8), págs. 5715–5728.

LOWRIE, W. y A. FICHTNER (2020), *Fundamentals of geophysics*, Cambridge university press.

PANEBIANCO, S., C. SATRIANO, G. VIVONE, M. PICOZZI, A. STROLLO y T. A. STABILE (2024), «Automated detection and machine learning-based classification of seismic tremors associated with a non-volcanic gas emission (Mefite d'Ansanto, Southern Italy)», *Geochemistry, Geophysics, Geosystems*, **25**(2), pág. e2023GC011 286.

PARASNIS, D. S. (2012), *Principles of applied geophysics*, Springer Science & Business Media.

REVIL, A. y A. JARDANI (2013), *The self-potential method: Theory and applications in environmental geosciences*, Cambridge University Press.

REYNOLDS, J. M. (2011), *An introduction to applied and environmental geophysics*, John Wiley & Sons.

SHI, C. y Y. WANG (2021), «Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties», *Geoscience Frontiers*, **12**(1), págs. 339–350.

SORRELL, C. A. (1973), *Rocks and minerals: A guide to field identification*, Macmillan.

TELFORD, W. M., L. P. GELDART y R. E. SHERIFF (1990), *Applied geophysics*, Cambridge university press.

TIAB, D. y E. C. DONALDSON (2024), *Petrophysics: theory and practice of measuring reservoir rock and fluid transport properties*, Elsevier.

VEIRANA, G. M. M., S. PERDOMO y J. AINCHIL (2021), «Three-dimensional modelling using spatial regression machine learning and hydrogeological basement VES», *Computers & Geosciences*, **156**, pág. 104 907.

WRONA, T., I. PAN, R. L. GAWTHORPE y H. FOSSEN (2018), «Seismic facies analysis using machine learning», *Geophysics*, **83**(5), págs. O83–O95.