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 Anastasio 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

Agradecimientos ix

1. INTRODUCCIÓN 1

2. DELIMITACIÓN Y PLANTEAMIENTO DEL PROBLEMA 2

3. JUSTIFICACIÓN 3

4. FORMULACIÓN DE OBJETIVOS 5

4.1. Generales 5

4.2. Específicos 5

5. MARCO TEÓRICO 6

5.1. Geofísica y Geoeléctrica 6

5.1.1. Definición de Geofísica 6

5.1.2. Resistividad de la Tierra 6

5.1.3. Sondeo Eléctrico Vertical 8

5.2. Adquisición de Datos Geofísicos 12

5.2.1. Intervalo de Muestreo en SEV 13

5.2.2. Proceso de Adquisición In Situ 15

5.3. Machine Learning (ML) en la Geofísica 15

5.4. Random Forests 16

5.4.1. Siembra del bosque 17

5.4.2. Predicciones del bosque 17

5.4.3. Margen y Error de Generalización 18

5.4.4. Robustez y Convergencia 18

5.4.5. Aplicaciones de Random Forests en Geofísica 18

6. METODOLOGÍA 19

6.1. ideas y apuntes 19

6.2. Variables de Entrada 21

6.2.1. Datos de entrada 22

6.2.2. Generación de datos de entrenamiento 24

6.2.3. Clasificación, transformación y escalado de los datos 24

6.3. Diseño del Modelo Random Forest 25

6.3.1. Regresión 25

6.3.2. Configuración inicial del modelo, numero de árboles, profundidad, muestras y muestra mínima 25

6.3.3. Configuración y optimización de hiperparametros 25

6.4. Preparación del dataset para la implementación del modelo 25

6.5. Implementación del modelo 25

6.5.1. Entrenamiento del modelo 25

6.5.2. Mapas de probabilidad y entrenamiento de regresión 25

6.6. Evaluación del modelo 25

6.6.1. Regresión y validación cruzada 25

6.6.2. Análisis de incertidumbre 25

6.7. Reporte estadístico 25

7. RESULTADOS Y CONCLUSIONES **26**

A. Apéndice I **27**

Índice de figuras

5.1. Estructura atómica del oro 7

5.2. Configuración general de electrodos 9

5.3. Esquema de la contribución de la respuesta eléctrica 11

5.4. Esquema del arreglo Wenner 12

5.5. Esquema del arreglo Schlumberger 12

5.6. Esquema del arreglo Dipolo-dipolo 12

Índice de tablas

6.1. Ejemplo de atributos empleados en el calculo de la resistividad aparente. 22

Agradecimientos

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

Capítulo 1
INTRODUCCIÓN

que es la geofisica:

La exploración geofísica consiste en un conjunto de metodologías que a través de la medición de propiedades petrofisicas del subsuelo es

eL geoelectrica

para que sirve

oportunidades de la cuiencia de datos en el análisis en tiempo real de la respuesta

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 es posible identificar aplicaciones comerciales 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, permitiendo evaluar el muestreo contra la regresiones y establecer nuevos intervalos de oportunidad.

Durante el muestreo tradicional los intervalos se diseñan considerando el objetivo de exploración (idealmente), este análisis previo es un factor determinante, en esta etapa se establece el intervalo de muestreo que permitiría identificar el objeto de exploración, este objeto de exploración pueden ser unidades geológicas, acuíferos, fallas, zonas de fracturas, estructuras antropogénicas, 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, lo que resulta irrisorio.

De manera que un modelo entrenado permita generar múltiples regresiones e

identificar regiones de interés no cubiertas por el muestreos 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 las frecuencias identificadas.

Pese a existir aplicaciones de ML implementadas en geofísica, no se identifica alguna enfocada este esta problemática en particular, sin embargo hay aplicaciones 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, establecer un modelo mediante entrenamiento supervisado, dada la complejidad de interpretación, por consiguiente su etiquetado, capas de realizar regresiones a partir de entrenamientos previos y permita realizar predicciones de valores de resistividad. Estas condiciones son cubiertas por dos modelos Bayesian Compressive Sensing (BCS) y Random Forests (RF).

Se Considera ademas de esto la facilidad de implementar e interpretar los resultados, así como el consumo de recursos, por lo que se opta por emplear el modelo Random Forests, así como tolera datos faltantes.

Capítulo 4
FORMULACIÓN DE OBJETIVOS

4.1 Generales

- Establecer un modelo de entrenamiento de regresión mediante Random Forests 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

- Generar modelos geoeléctricos con unidades geológicas de terrenos ya caracterizados mediante SEV's empleando la librería PyGIMLI.
- Establecer un modelo de regresión utilizando Random Forests a partir de variaciones de datos de una sección geológica especifica.
- Realizar un análisis comparativo entre la respuesta del modelo y los resultados pos procesado.

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 empleé una fuente natural como la propagación de ondas elásticas generadas por sismicidad, 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 características definidas a partir de los elementos que la integran, en primer orden la configuración atómica establece las

propiedades físicas corresponden a 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, en conjunto; 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).

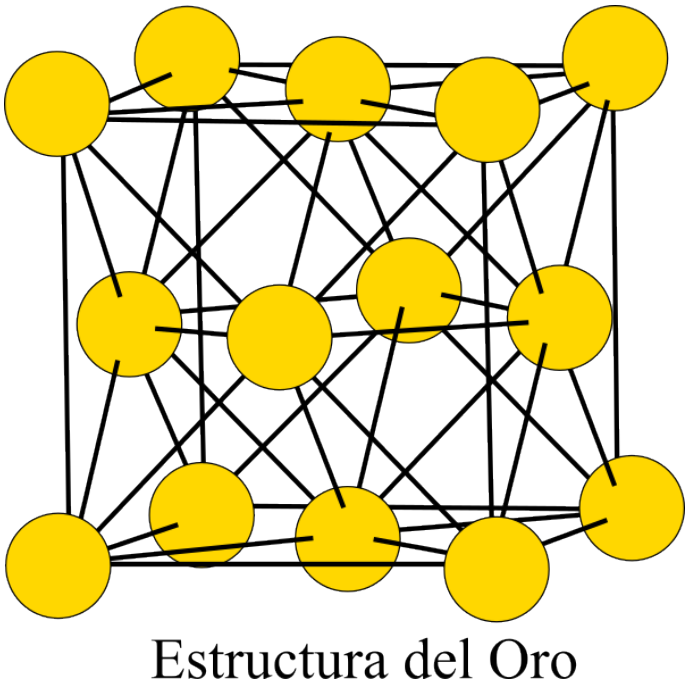

Figura 5.1: Esquema de la estructura atómica de oro que conforma la cristalización octahedral, modificado de Sorrell (1973)

.

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.

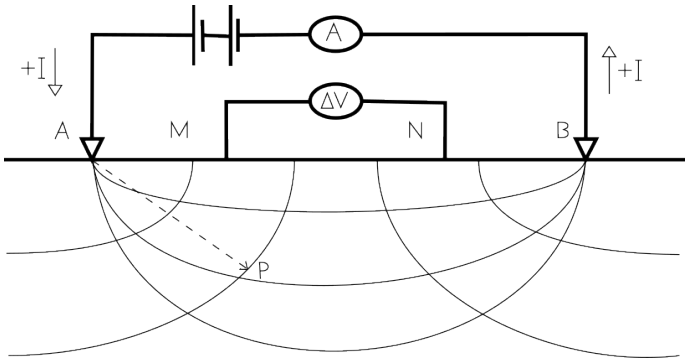

Figura 5.2: Configuración general de arreglo de electrodos, modificado de Reynolds (2011)

.

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).

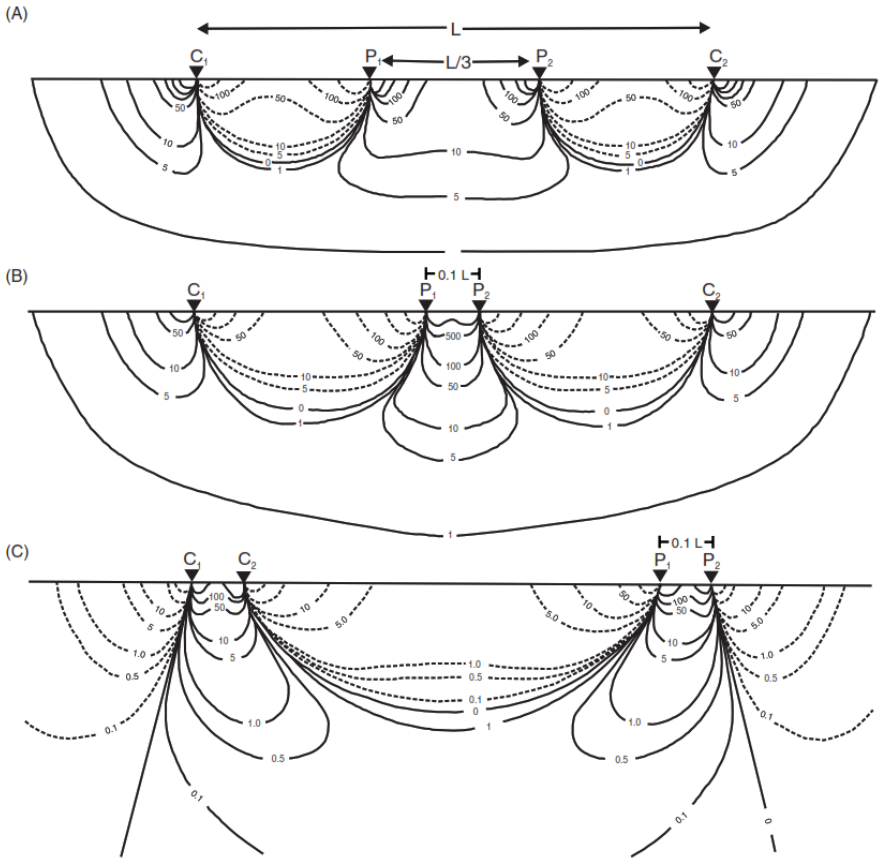

Figura 5.3: Esquema de la contribución de la respuesta de resistividad eléctrica, modificado de Reynolds (2011)

.

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

Wenner

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

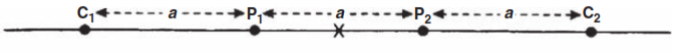

Figura 5.4: Esquema del arreglo Wenner, modificado de Reynolds (2011)

.

Schlumberger

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

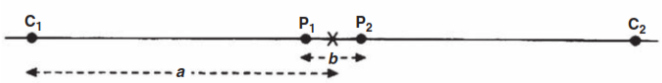

Figura 5.5: Esquema del arreglo Schlumberger, modificado de Reynolds (2011)

.

Dipolo-dipolo

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

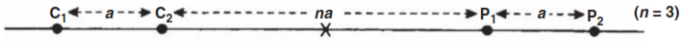

Figura 5.6: Esquema del arreglo Dipolo-dipolo, modificado de Reynolds (2011)

.

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)

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

El teorema de muestreo de 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).

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

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 coanstante de ruido electrico de fuentes tanto naturales como humanas, es impresindible considerar el teorema de Shannon-Hartley aplicando apilamiento de muestreo como metodo de reduccion de la relacion ruido señal, durante la adquisicion de datos; esto quiere decir calcular el promedio de muestreos cointinuos 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, al inducir corriente continua empleando un resistivimetro mediante de 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 distintos (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 unidades, 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 (ML) en la Geofísica

La aplicación de ML y el DL en la geofísica es ampliamente utilizado en exploración sísmica, abarcando los procesos de adquisición, procesado e interpretación, 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 industria, pese a las ventajas que puede tener su aplicación, como es el caso de este estudio

El aprendizaje automático o machine learning, son un 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 de dos maneras, por el tipo de aprendizaje, 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 algoritmos siguientes.

- Árboles de decisión
- Random Forests
- K-Nearest Neighbors (KNN)
- Máquinas de soporte vectorial (kernelizados)

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, la técnica Random Forests 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

La técnica Random Forests emplea múltiples arboles de decisión independientes entre si, donde cada árbol realiza una votación de clases, donde se selecciona 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 Siembra del bosque

Breiman (2001) nos dice que 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.

A partir de los datos de entrada, se generan subconjuntos de datos de entrenamiento, estos se seleccionan con cierta aleatoriedad empleando la técnica bootstrap sampling, en cada nodo de los subconjuntos de entrenamiento se selecciona un subconjunto de 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).

5.4.2 Predicciones del bosque

La salida de un Random Forest 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).

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

5.4.3 Margen y Error de Generalización

5.4.4 Robustez y Convergencia

5.4.5 Aplicaciones de Random Forests en Geofísica

Capítulo 6
METODOLOGÍA

6.1 ideas y apuntes

faltan las referencias....

Esta investigación propone la implementación de un modelo de Random Forests (Bosques Aleatorios) para mejorar la calidad y optimización del proceso de adquisición de datos durante un Sondeo Eléctrico Vertical (SEV). El objetivo es utilizar técnicas de aprendizaje automático para clasificar las lecturas obtenidas y generar una regresión que permita sugerir intervalos de muestreo adicionales, mejorando así la exploración geofísica.

El flujo de trabajo propuesto consta de las siguientes etapas:

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.

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 regresión 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.

En la implementación de técnicas avanzadas como Machine Learning (ML), modelos como Random Forests pueden analizar patrones en los datos y predecir áreas con alta variabilidad de resistividad. Esto permite:

Ajustar dinámicamente el intervalo de muestreo durante la adquisición *in situ*. Generar muestreos adicionales en áreas críticas para aumentar la precisión. Con este enfoque, se asegura que el intervalo de muestreo esté alineado con el teorema de Nyquist, optimizando la calidad de los datos y reduciendo la redundancia.

El uso de Random Forests permite abordar la complejidad y la variabilidad inherentes al proceso de adquisición de datos en el contexto de exploración geofísica. Este enfoque optimiza el proceso de muestreo, mejora la calidad de los datos obtenidos y proporciona una base para generar predicciones más precisas, contribuyendo así al éxito de las campañas de exploración.

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= potencial 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.

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.

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

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 del Modelo Random Forest

6.3.1 Regresión

6.3.2 Configuración inicial del modelo, numero de árboles, profundidad, muestras y muestra mínima

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 modelo

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.

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

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.

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.

Wrona, T., I. Pan, R. L. Gawthorpe y H. Fossen (2018), «Seismic facies analysis using machine learning», *Geophysics*, **83**(5), págs. O83–O95.