Análisis y Comportamiento Predictivo Mediante Aprendizaje Automático en Sondeos Eléctricos Verticales, Aplicación en Adquisición de Datos

Ing Torres S. Juan J.[1*]

[1*]Facultad de Ciencias Físico Matemáticas, Maestría Ciencia de Datos, Pedro de Alba, Niños Héroes, San Nicolás de los Garza, 66451, N.L., México.

Corresponding author(s). E-mail(s): juan.j.torres.s@gmail.com;

Resumen

Propósito: Durante el proceso de adquisición de datos geoléctricos, se pueden presentar anomalías, ya sean errores de medición, error de muestreo, ruido ambiental, efectos de sitio, etc.., esta publicación tiene por finalidad generar un algoritmo de Aprendizaje Automático que permita identificar error en etapas tempranas de adquisición así como una pre-interpretación de unidades saturadas, con la finalidad de reducir las incertidumbre, mejorando la calidad de los datos y dar la oportunidad al aperador de corregirlos al momento, así como generar una alerta temprana de unidades saturadas con el objetivo de ampliar la resolución de la unidad en cuestión con el objeto de obtener mayor detalle de estas zonas de interés. **Método:** para lograr el propósito se empleara un conjunto de 647 datos, correspondientes a 29 Sondeos Eléctricos Verticales (SEV), realizados a lo largo de un año de mediciones, en configuraciones geológicas completamente distintas.

Keywords: Resistividad Aparente, Keyword2, Keyword3, Keyword4

1. Introducción

En una campaña de exploración geofísica enfocada en recursos geohídricas se pueden emplear distintos métodos para determinar la existencia, profundidad y distribución del nivel freático de un acuífero, como son método de refracción sísmica,

GPR (Ground Penetrating Radar) y métodos geoeléctricos, en este trabajo nos enfocaremos en la aplicación del método geoeléctrico en la modalidad de Sondeo Eléctrico Vertical, aplicando técnicas de AA en la adquisición geoeléctrica con un Resistivímetro Analógico, el método consiste en inducir una corriente eléctrica en el subsuelo a través de dos electrodos afianzados al terreno, al mismo tiempo se realiza la lectura del potencial inducido por el flujo de la corriente en otro par de electrodos, colocados a una distancia previamente definida, empleando la configuración eléctrica Schlumberger (Referencia año), obteniendo valores de resistividad aparente en cada sondeo, durante la adquisición de datos se recopila la siguiente información:

- Sitio de estudio representado por la clave asignada, numero de sondeo, sitio y localidad o solicitante.
- Personal técnico que realizo el levantamiento de los datos
- Fecha del estudio
- Zona datum UTM
- Coordenada Este
- Coordenada Norte
- Altitud
- Profundidad de muestreo
- K factor geométrico de arreglo Schlumberger
- Distancia Media entre los Electrodos A y B
- Distancia entre los electrodos A y B
- Distancia Media entre los electrodos M y N
- La quinta parte de la Distancia entre los Electrodos M y N
- Distancia entre los Electrodos M y N
- Potencia Natural 1
- Potencial Inducido 1
- Corriente Inyectada 1
- Potencia Natural 2
- Potencial Inducido 2
- Corriente Inyectada 2
- Potencia Natural 3
- Potencial Inducido 3
- Corriente Inyectada 3
- Media del Potencial Natural
- Media Potencial Inducido
- Media de las Diferencias entre Potencial Inducido y Potencial Natural
- Media de la Corriente Inyectada
- Resistividad eléctrica aparente 1
- Resistividad eléctrica aparente 2
- Resistividad eléctrica aparente 3
- Media de las resistividad eléctrica 1 2 y 3
- Resistividad Eléctrica Aparente Final

Normalmente la adquisición de datos, con equipos analógicos, es realizada por personal con experiencia, sin embargo este no siempre es el caso, por lo que es resulta

conveniente tener herramientas que nos permitan tener la oportunidad de mejor la calidad de los datos con alertas tempranas durante el registro de la adquisición geofísica, reduciendo los errores de muestreo y atenuando la incertidumbre de datos adquiridos por un operador inexperto.

2. Marco Teórico

2.1. Modelo No Supervisado

2.2. Modelo Supervisado: Isolation Forest

3. Metodología

Para este proyecto se trabajaran con datos de Resistividad eléctrica, adquiridos durante campañas de exploración en distintas zonas de México, sumando un total de 647 registros correspondientes a 29 sondeos, por protección de los clientes se cambiaron nombres de personal y ubicaciones geográfica, cambiando la ubicación de los sitios por un cuadrante del municipio San Nicolás de los Garza ubicado en el Estado de Nuevo León, para fines prácticos.

Se aplicaron métodos estadísticos descriptivos básicos, correlación entre variables y entrenamientos de modelos supervisados y no supervisados; en cada paso del proceso se genero información que permitió avanzar en el objetivo del estudio así como otorgar un panorama de lo general a lo particular a fin de alcázar el modelado predictivo.

3.1. Reorganización y Preparación De Los Datos

En esta etapa del trajo se procedió a reagrupar las columnas, descartando aquellas que no se relacionan directamente con el objetivo del trabajo, posteriormente se procese a eliminar las columnas que no contienen información o que presentan valores fuera de los parámetros esperados (Resistividades por a debajo de 0 $ohm * m$), eliminamos los valores nulos, a partir de esta información generamos una nueva base de datos, integrándose de los siguientes elementos: Sitio, Coordenada Este, Coordenada Norte, Altitud, Profundidad de muestreo, Distancia entre los electrodos A y B, Distancia entre los Electrodos M y N, Resistividad Eléctrica Aparente Final (Cuadro 1).

A partir de esta primera etapa obtenemos la siguiente base de datos, denominada *res_geo_copy* (Cuadro 2), sobre este mismo registro agregaremos una columna que contendrá el valor normalizado correspondiente a *Rha* (resistividad aparente), de esta misma dividiremos por sitio y los registros a fin de calcular la estadística básica de cada *sitio*.

La segmentación del DataFrame se realizo con el objetivo de generar una base de datos de entrenamiento, misma que contiene los sitios que presentaron anomalías relacionadas a acuíferos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 647 entries, 0 to 646
Data columns (total 8 columns):
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
0   sitio     647 non-null    object
1   ESTE      647 non-null    float64
2   NORTE     647 non-null    float64
3   ALTITUD   647 non-null    int64
4   Z         647 non-null    float64
5   AB        647 non-null    int64
6   MN        647 non-null    float64
7   Rha       647 non-null    float64
dtypes: float64(5), int64(2), object(1)
memory usage: 40.6+ KB
```

Cuadro 1 Información general de la base de datos resultante, *res_geo_copy.info*()

#	Sitio	Lon Este	Lat Norte	Altitud	Z	AB	MN	Rha
0	S1-TV-VLH	370462.31	2850554.93	489	1.8	6	0.5	17.24
1	S1-TV-VLH	370462.31	2850554.93	489	2.4	8	0.5	25.89
2	S1-TV-VLH	370462.31	2850554.93	489	3.0	10	0.5	47.94
...	...	...	...	...	...	...	...	...
644	S3-HIGA	370645.67	2848071.60	502	90.0	300	20.0	18.98
645	S3-HIGA	370645.67	2848071.60	502	102.0	340	20.0	24.61
646	S3-HIGA	370645.67	2848071.60	502	120.0	400	20.0	22.53

Cuadro 2 Visualización del DataFrame *res_geo_copy*, se observan las principales columnas y la estructura general de los datos que la integran.

3.2. Estadística Descriptiva

Para el conjunto de datos seccionados se obtiene la estadística básica para cada columna, de esta manera podemos observar la el comportamiento general de los datos y a partir de este punto generar nueva estadística que nos proporcione información mas relevante para identificar las áreas de oportunidad de análisis.

Se genero la estadística básica para los datos mediante código en *Phyton*, teniendo los siguientes resultados:

media = 333.4500974025974
media con flotante = 333.4500974025974
media baja de los datos = 108.79
media alta de los datos = 109.09
moda = 12.87
cuartiles = [27.8925, 108.94, 298.90999999999997]
varianza = 304169.9552565759
desviación estándar = 551.5160516762644
valor máximo = 2781.76
valor mínimo = 0.38

```
res_geo_copy.describe()
```

	Z	AB	MN	Rha
count	616.000000	616.000000	616.000000	616.000000
mean	42.870292	142.900974	10.139610	333.450097
std	54.174873	180.582909	11.298715	551.516052
min	0.600000	2.000000	0.500000	0.380000
25%	6.000000	20.000000	4.375000	27.917500
50%	24.000000	80.000000	5.000000	108.940000
75%	57.000000	190.000000	20.000000	297.190000
max	300.000000	1000.000000	100.000000	2781.760000

Cuadro 3 Información general de la base de datos resultante, $res\_geo\_copy.info()$

A partir de la información obtenida se puede concluir que la base de datos una desviación estándar y una varianza bastante elevada,

Subsecuentemente generamos un histograma de estos vales, obteniendo un panorama general de la distribución de las muestras respecto una característica cuantitativa y continua, representando la densidad de frecuencia acumuladas en intervalos iguales,[4] esta información es relevante ya de primera mano conociendo la distribución de las frecuencias podemos intuir el tipo de distribución que presentan los datos, en este caso no presentan una distribución normal, siendo no paramétricos (Ver Figura XX).

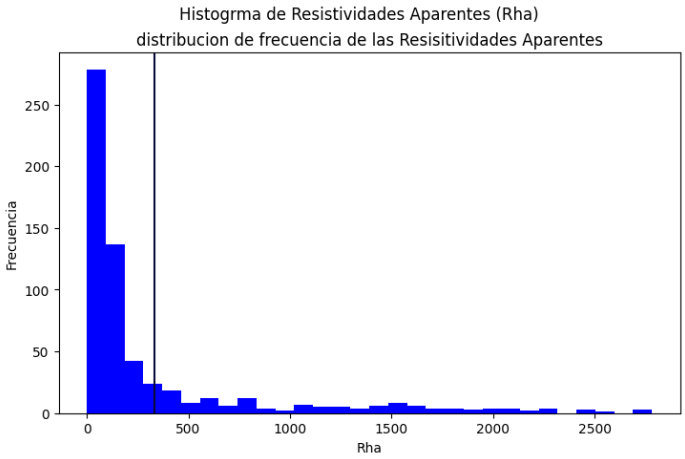

Figura 1 Histograma generado a partir de los datos de Rha, densidad de probabilidad, distribución de las resistividades

De igual manera se obtiene un diagrama de violín el cual se utiliza para visualizar la distribución de los datos y su densidad de probabilidad. Este gráfico es una combinación de un diagrama de cajas, bigotes y un diagrama de densidad, girado y colocado a cada lado, mostrando la distribución de los datos. la linea que divide por el centro

y acotado en su extremo representa el intervalo intercuartil, representando el 95 % de los intervalos de confianza en su parte mas abultada, y la linea vertical representa la posición de la media (Ver Figura XX).

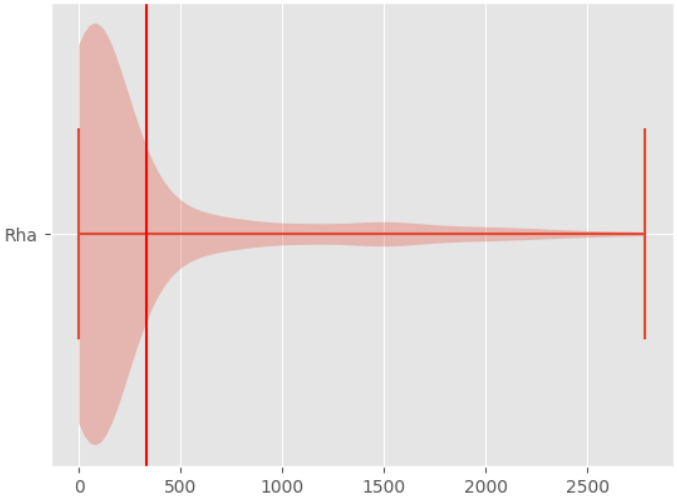

Figura 2 Diagrama de Violín donde se observa la a densidad de probabilidad de la Rha

Estas dos representaciones permiten tener un panorama general de la información contenida en la base de datos, facilita la clasificación de la información, establece los lineamientos para re-clasificación y preparación del *DataFrame* para la aplicación de técnicas de Aprendizaje Automático.

3.3. Correlación y Selección de Variables

Parte del proceso del análisis de los datos se lleva a acabo mediante la correlación de las variables de interés, con el objetivo de esclarecer una relación entre ambas variables, esta parte del trabajo la realizamos mediante una Matriz de Correlación.

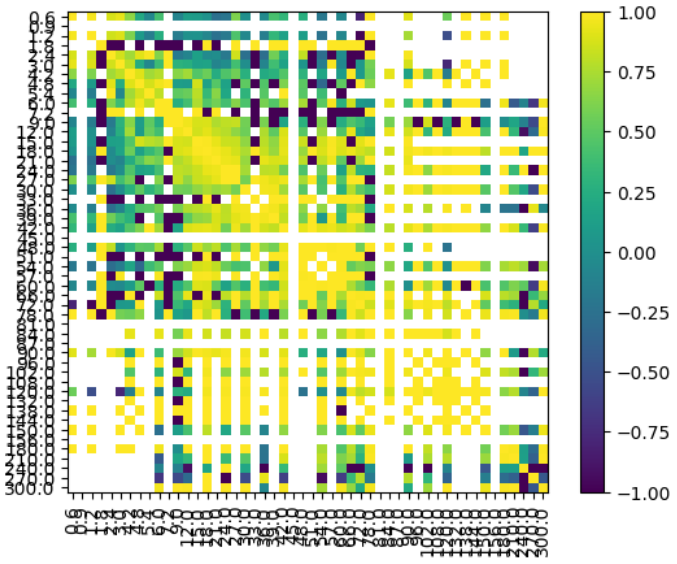

Figura 3 Podemos observar en la escala de color el grado de correlación presente entre la profundidad y el valor de resistividad Aparente.

A partir de la prueba de correlación se puedo identificar el siguiente comportamiento

1.- Tenemos una alta correlación entre los valores de la prueba, entre la profundidad y el aumento de la resistividad, este comportamiento es consistente con los valores.

2.- En cuanto a los valores de baja correlación, nos aportan información valiosa ya que pueden estar representando una anomalía en la distribución esperada, traduciéndose una dispersión de valores de Rha, es decir una baja correlación de los valores, representando valores de interés para la exploración geohídrica.

3.- En cuanto a la distribución se deben considerar como datos no paramétricos, mucho mas similares a una distribución de tipo Gamma.

3.4. Análisis De Anomalías

Como podemos observar los datos presentan un reto adicional debido al tipo de distribución que presenta, por lo que es necesario conocer algunas métricas de la para tener claro la forma de proceder en el uso de la técnicas de Aprendizaje automático, una de estas pruebas es la Homocedasticidad, la cual nos indica que si la varianza de los datos es contante a lo largo del tiempo, esta se obtuvo aplicando el método de "Levene"mediante la librería "Pingouin", la cual es una prueba estadística inferencial utilizada para evaluar la igualdad de las varianzas para una variable calculada de uno o mas grupos [5, Levene, H., 1960],la prueba estadística de Levene se define como:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Zi - Z..)^2}{\sum_{i=1}^{k} \sum_{i=1}^{Ni} (Zi - Zi)^2} \tag{1}$$

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import pingouin as pg

pg.homoscedasticity(data=datos,
                            dv='Rha',
                             group='sitio',
                              method='levene')

         W                  pval              equal_var
levene   14.144112          2.022176e-48      False
```

Cuadro 4 Prueba de Homocedasticidad mediante el método de Levene

Donde k es el numero de diferentes grupos, N es el número total de casos en todos los grupos, y es el valor de la variable medida para cada sitio de cada localidad; como resultado de la prueba se observa que los datos no presentan Homocedasticidad, por lo que se tendrán se emplearon técnicas apropiadas al tipo de

Debido a la distribución anómala de los datos es importante tener claro que parte de la información es útil para analizar, ya que en cada región de estudio no se presentaron resultados acorde a los esperados para acuíferos, por lo que es importante emplear solo aquellos que presentan la anomalía correcta para implementar las técnicas y generar un modelo que pueda discriminar esta característica del grueso de los datos, tal como se observa en la figura XX, se integraran los datos que presentan un comportamiento acorde a lo esperado, para ello se emplearan los datos correspondientes al área de estudio compuesta por los sitios S1-HIGA, S2-HIGA y S3-HIGA.

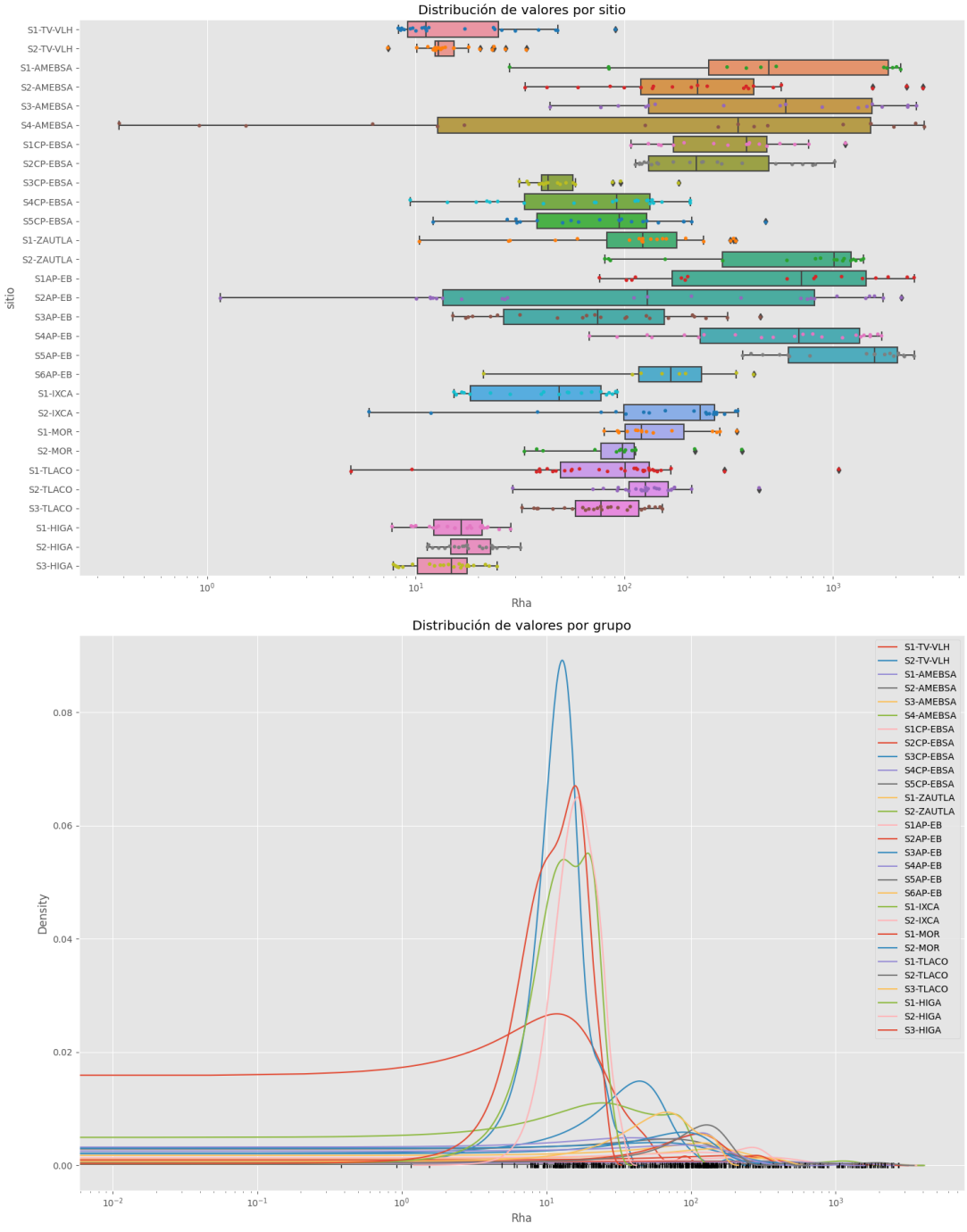

Figura 4 Histograma generado a partir de los datos de Rha, densidad de probabilidad, distribución de las resistividades

4. Resultados

A partir de la información generada, se pueden identificar patrones de agrupamiento, ejemplo de ello es en la distribución de Rha por sitio, en donde se puede observar valores menores a $100 ohm * m$ casi en todos los sitios, el mismo comportamiento se observa en la correlación entre la profundidad y la Rha, en donde los sitios anómalos que se desvían de la relación directa nos aportan información sobre una tendencia subyacente de respuestas de baja resistividad a profundidades en las que no se espera esta correlación, teniendo en cuenta que los sitios anómalos con este comportamiento corresponden a ubicaciones donde la exploración geohídrica resulto positiva, incluyendo la perforación y alumbramiento del pozo, se da pie a buscar una mejor implementación empleando otros métodos para identificar específicamente este comportamiento en los datos.

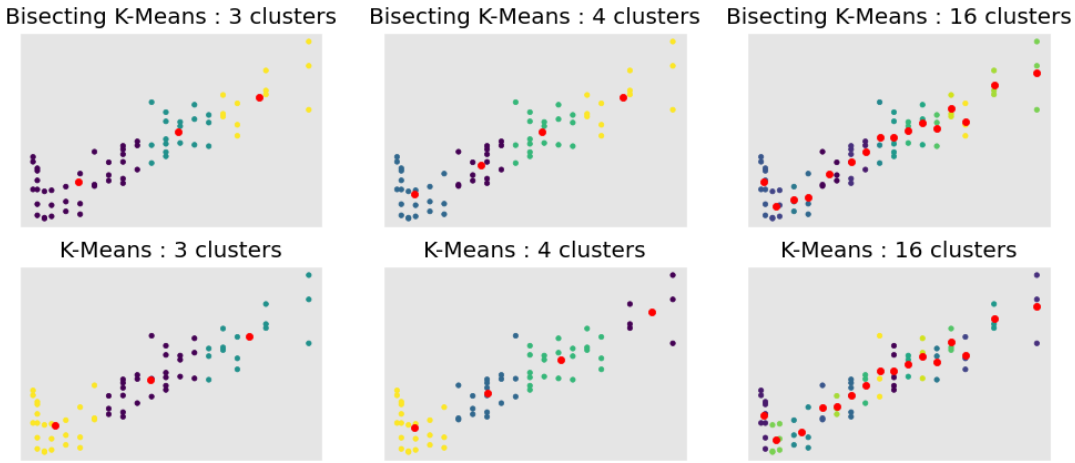

Figura 5 Comparación entre K-Means y Bisecting K-Means, es posible identificar las tendencias de agrupamiento por secciones.

Como podemos observar en los resultados de la prueba de ambos algoritmos, K-Means y Bisecting K-Means presentan gran similitud, sin embargo K-Means se enfoca unicamente en los puntos mas cercanos al centro, mientras que su símil logra tener mas sensibilidad entre los puntos vecinos, en general presenta una mejor respuesta Bisecting K-Means, ya que logra identificar de forma muy precisa en el cluster 4 los puntos que engloban a el acuífero en cuestión, sin embargo hay que destacar que el ajuste necesario forzar el agrupamiento a presentar este comportamiento.

5. Conclusión

De acuerdo con los resultados así como los objetivos del trabajo, se puede establecer un análisis cualitativo y cuantitativo de las variables dependientes, su comportamiento

así como la relación entre estos, siendo positivo lograr comprender mejor la distribución y estadística de la información que aportan los sondeos, siendo de utilidad para la identificación de anomalías en datos, sin embargo al no lograrse el objetivo de identificar claramente las anomalías y lograr predecirlas, se requerirá de un trabajo extra para lograr este objetivo, así como del entrenamiento del algoritmo IsolationForest, así como implementarlo de forma directa durante una prueba de adquisición para verificar su uso practico.

Agradecimientos. Agradezco a chuchita por no darse por vencida después de que la bolsearon

Referencias

[1] Joaquín Amat, Rodrigo *Machine learning con Python y Scikit-learn*, available under a Attribution 4.0 International (CC BY 4.0), Fecha de consulta: 26/03/23, Url: https://www.cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn.html.

[2] Cesano, D., Olofsson, B. y Bagtzoglou, A. C.(2000), *Parameters regulating groundwater inflows into hard rock tunnels—a statistical study of the Bolmen tunnel in southern Sweden*, Tunnelling and Underground Space Technology, 15(2), 153-165.

[3] Parsekian, A. D., Claes, N., Singha, K., Minsley, B. J., Carr, B., Voytek, E. y Flinchum, B.(2017), *Comparing measurement response and inverted results of electrical resistivity tomography instruments*, Journal of Environmental and Engineering Geophysics, 22(3), 249-266.

[4] Gutierrez, R. B., y i Cintas, P. G.(2013), *El histograma como un instrumento para la comprensión de las funciones de densidad de probabilidad.*, robabilidad Condicionada: Revista de didáctica de la Estadística, 22(3), P, (2), 229-235.

[5] Levene, H.(1960), *Robust tests for equality of variances. Contributions to probability and statistics*, r278-292.