IDENTIFICACIÓN DE LA PRESENCIA DE HIDROCARBUROS EN ARENAS ARCILLOSAS USANDO UN ALGORITMO DE MACHINE LEARNING, REGISTROS DE POZO E IMÁGENES DE FLUORESCENCIA DE CORAZONES

ANGÉLICA YULIETH CASTRO ORTEGA

ANA VALENTINA TEJADA ANGARITA

Proyecto integral de grado para optar por el título de

INGENIERO DE PETRÓLEOS

Director:

José Francisco Peñaloza González

Ingeniero de Petróleos

Codirector:

Sebastián Alejandro Gómez Alba

Ingeniero de Petróleos

Asesor:

Edwin Ortega

Ingeniero de Petróleos, PhD

FUNDACION UNIVERSIDAD DE AMERICA

FACULTAD DE INGENIERIAS

PROGRAMA DE INGENIERÍA DE PETRÓLEOS

BOGOTA, D.C

2021

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

_____

_____
Nombre
Firma del Director

_____
Nombre
Firma del Presidente Jurado

_____
Nombre
Firma del Jurado

_____
Nombre
Firma del Jurado

Bogotá D.C. Julio de 2021.

DIRECTIVOS DE LA UNIVERSIDAD

Presidente Institucional y Rector del Claustro.
Dr. MARIO POSADA GARCÍA PEÑA

Consejero institucional.
Dr. LUIS JAIME POSADA GARCÍA PEÑA

Vicerrectoría Académica y de Investigaciones.
Dra. ALEXANDRA MEJIA GUZMAN

Vicerrectoría Administrativo y financiero.
Dr. RICARDO ALFONSO PEÑARANDA CASTRO

Secretario General.
Dr. JOSE LUIS MACÍAS RODRÍGUEZ

Decano facultad de ingeniería.
ING. JULIO CESAR FUENTES ARISMENDI

Director de Programa Ingeniería de Petróleos.
ING. JUAN CARLOS RODRIGUEZ ESPARZA

DEDICATORIA

Dedico esta nueva etapa culminada de mi vida, a mi familia porque han sido el pilar que me ha sostenido durante todos estos años y porque me ha ayudado a construir las bases de lo que seré. A mi tío, porque sin él no habría podido alcanzar este sueño, porque sus consejos ha forjado en mi mis ganas de seguir adelante y y de convertirme en una gran profesional. A mi abuela materna, por siempre estar para mi y darme las palabras palabras de aliento y amor que tanto he necesitado. A mis padres y hermana, porque son el motivo que necesito para seguir adelante y luchar por mis sueño; nuestros sueños. A mi director de tesis, por haberme dado su tiempo, apoyo incondicional , sus conocimientos durante mi carrera y para culminarla. Finalmente A mi pareja, porque por tantos años me ha dado su apoyo incondicional y me ha impulsado a lograr cada meta que me he trazado.

Angélica Yulieth Castro Ortega

DEDICATORIA

En primera instancia dedico el presente trabajo a Dios por brindarme la oportunidad de estudiar y crecer profesionalmente, a mis padres Iván De Jesús Tejada y Blasina Angarita por su apoyo incondicional, quienes con sus consejos han permitido que yo crezca cada día como persona. También agradezco a mis hermanas Karen Tejada, Laura Tejada y Maximiliano, ellos me han brindado su compañía en la ciudad de Bogotá en el progreso de esta etapa. ¡Son lo mejor que tengo!

Finalmente, a nuestros directores, quienes han sido las personas que nos han guiado para que este proyecto sea una realidad. La consideración de cada uno de ellos en transmitirnos sus conocimientos ha traído consigo el logro de importantes objetivos como culminar el desarrollo de nuestra tesis.

Ana Valentina Tejada Angarita

AGRADECIMIENTOS

Nosotras queremos darle nuestro especial agradecimiento :

A la Universidad América, por abrirnos sus puertas y darnos la posibilidad de conocer excelentes personas. Durante nuestra carrera, los docentes nos brindaron consejos, y nos impartieron conocimientos que sabemos nos van a ayudar en nuestra futura vida profesional.

A nuestro director Ing. José Peñaloza y al Ing. Edwin Ortega , por quienes expresamos nuestro agradecimiento y admiración por los grandes profesionales que tiene la industria. Con su experiencia, tiempo y entrega, nos enseñaron a que debemos esforzarnos y enfocarnos en lo que queremos alcanzar en la vida para llegar al éxito. Gracias, por motivarnos y por impartir nuevos conocimientos que sabemos vamos a poder implementar en nuestra vida profesional.

A nuestro orientador Sebastián Gómez, por brindarnos su tiempo, y por darnos su asesoría durante el desarrollo del proyecto.

A nuestros padres, amigos, compañeros. Por ser parte de nuestro camino, por darnos sus palabras de aliento y por forjar a las futuras de Ingenieras de este país a ser quienes somos para demostrar que somos mujeres valiosas e integras que necesita la industria.

Las directivas de la Fundación Universidad de América, los jurados calificadores y el cuerpo docente no son responsables por los criterios e ideas expuestas en el presente documento, estos corresponden únicamente a los autores.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ABREVIATURAS

%	Porcentaje
Ø	Porosidad
ANS	Alaska North Slope
AT90	Registro de resistividad profunda
BS&W	Basic Sediment and Water (contenido de agua y sedimentos)
CEC	Capacidad de intercambio catiónico
CWLS:	Canadian Well Logging Society
DLIS:	Digital log information system
DTCO	Registro de inverso de velocidad compresional
Ft	Pies
gm/cc	gramos por centímetro cúbico
GR	Gamma Ray
GS	Registro en escala de grises
K	Permeabilidad
LAS	Log ASCII Standard
ML	Machine Learning
MLP	Multilayer Perceptron
MMBO	Millones de barriles de petróleo
MSE	Mean Squared Error (error medio cuadrado)
NPHI	Registro de porosidad neutrón
NPRA	Reserva Nacional de Petróleo de Alaska
PHIT	Porosidad
Psi	Pound per square inch (Libra por pulgada cuadrada)
PTS	Production Testing Services
QFM	Otros minerales, de la tabla XRD
RGB	Rojo, Verde y Azul
RHOZ	Registro de porosidad
RHOG	Densidad de grano
RHOsh	Densidad de la arena
RHOss	Densidad asumida del volumen de arcilla
RMSE	Root Mean Square Error (raíz cuadrática media)
Rsh	Resistividad del shale
RW	Resistividad del agua
So	Saturación de aceite
SVR	Support Vector Regression
Sw	Saturación de agua
SWA	Saturación de agua por el método de Archie
SWp	Saturación de agua del método Poupon
SwWs	Saturación de agua con el método de Waxman-Smits
Sws	Saturación de agua del método Simandoux

RESUMEN

La evaluación de la presencia de hidrocarburos en arenas con intercalaciones de arcilla por debajo de la resolución vertical de los registros de resistividad es una tarea difícil. Las arcillas al ser altamente conductivas generan supresión de los registros de resistividad haciendo que la saturación de agua calculada a través del modelo de Archie sea mayor a la existente en zonas altamente productivas. Otros modelos como Poupon, Waxman-Smits o Simandoux requieren de conocimiento de propiedades eléctricas de las arcillas a través de medidas de laboratorio o calibración en arcillas gruesas y representativas de las laminaciones.

Avances recientes en algoritmos de Machine Learning y la disponibilidad de módulos de procesamiento de imágenes en Python, sugieren el entrenamiento de registros de pozo para reconocer variaciones en imágenes de fluorescencia ultravioleta (UV) como una promisoria forma de reconocer la presencia de hidrocarburos en yacimientos complejos. En este trabajo, se propone un nuevo método basado en Machine Learning, el cual es análogo al uso de algoritmos de reconocimiento facial, para predecir la compleja relación existente entre la presencia de hidrocarburos en arenas laminadas y medidas básicas de registro de pozo sin utilizar expresiones analíticas explícitas.

Mostramos la aplicación del modelo propuesto en dos etapas en la formación Nanushuk en Alaska, famosa por sus prolíficas arenas intercaladas con arcilla resultantes de un ambiente deltaico de depositación. Primero, las imágenes UV de corazones son acondicionadas y procesadas para generar un registro en escala de grises para lograr determinar las zonas con presencia/ausencia de hidrocarburo. Luego, los registros de densidad, neutrón, sónico, gamma ray, y resistividad son utilizados en varios algoritmos de entrenamiento supervisado para predecir la existencia de hidrocarburo con modelos de regresión utilizando subsets de entrenamiento y prueba.

Los resultados de predicciones en un pozo de validación, muestran que el modelo Random Forest logra predecir la presencia de hidrocarburos de forma más rápida y precisa comparado con calibrar un modelo petrofísico estándar. El modelo logró subestimar el Pay con un 18% de error en comparación con el mejor método convencional calculado (Pupon) con un 24% más alto en Pay , en base al valor verdadero de fluorescencia de las imágenes.

Adicionalmente, se logró demostrar con éxito, que el uso de reconocimiento de fluorescencia trae consigo un innovador nivel de simplificación en la evaluación de rocas tipo no-Archie. Este trabajo introduce elementos de una nueva filosofía en petrofísica: el usar las relaciones intrínsecas a los datos para hacer predicciones sin requerir de las expresiones

analíticas que gobiernan la física de las mediciones de pozo en rocas complejas. Aún más, poder predecir rocas complejas con medidas básicas y fotos de fluorescencia significa para las compañías poder tomar decisiones de exploración de forma más rápida y efectiva mientras se reducen recursos humanos.

Palabras Clave: Machine learning, Saturación de agua, Presencia de hidrocarburos, Imágenes de corazones, fotos ultravioleta, registros de pozo, petrofísica convencional, modelos de regresión.

INTRODUCCIÓN

En la cuenca North Slope en Alaska, se encuentra la formación Nanushuk [1][2] famosa por sus prolíficas arenas intercaladas con arcilla que son resultado de ambientes deltaicos de depositación [3] [11]. A pesar de que varias compañías han evidenciado potencial comercial de hidrocarburos, no todos los pozos perforados son exitosos debido a la complejidad geológica y la naturaleza laminada de los yacimientos [19]. Evaluar la presencia de hidrocarburos en intercalaciones de arena-arcilla usando registros de resistividad es una tarea difícil.

Las herramientas utilizadas para tomar registros de resistividad generan una corriente eléctrica que pasa a través de la formación y registran la respuesta de la formación a tal corriente [4]. Hay dos tipos de herramientas principales para medir la resistividad de una roca en el subsuelo: Herramientas de inducción (bobinas que inducen corriente y miden la conductividad de la formación) [12] y herramientas laterolog (electrodos en la herramienta que emiten corriente y miden la resistividad de la formación) [5]. Los registros resistivos se ven afectados por la presencia de arcillas por su alta conductividad modificando los valores de resistividad leídos subestimando el potencial de la zona.

Avances recientes en algoritmos de Machine Learning o también y menos comúnmente llamado Aprendizaje Automático en español (disciplina científica de la inteligencia artificial) [8], demuestran que es viable utilizar fotos de corazones como medida de la presencia de hidrocarburo para descubrir la relación inherente entre esta y la respuesta de los registros de pozo. En este trabajo se utilizaron 8 modelos de Machine Learning , para entrenar los registros de pozo utilizando fotos de corazones [7] tomadas en luz ultravioleta (UV) y obtener resultados más precisos en comparación con los métodos convencionales.

1. MARCO TEORICO

Para desarrollar el presente trabajo, es necesario realizar una descripción de la cuenca de investigación en Alaska. Posteriormente, se presenta la estratigrafía de la cuenca con el fin de asociar la formación de interés y ambientes de depositación. Por último, se muestra como es afectada la medición del registro de resistividad por la presencia de arenas intercaladas con arcilla y el uso de las imágenes ultravioleta en el estudio.

1.1. Generalidades de la cuenca North Slope

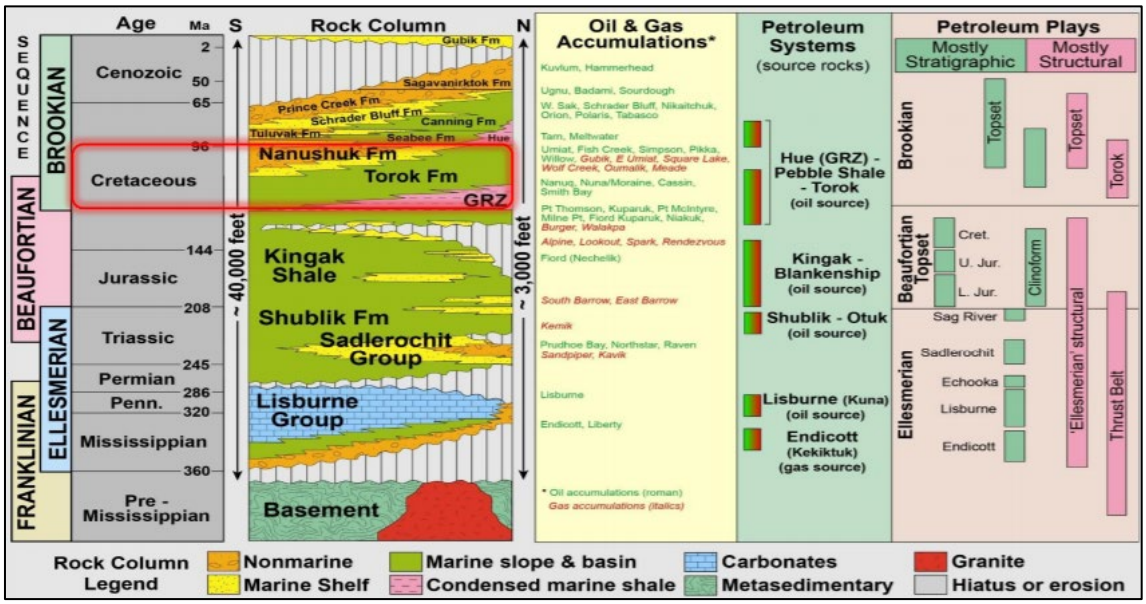
La cuenca North Slope en Alaska (ANS) es una de las cuencas productoras de petróleo más prolíficas de América del Norte. Esta cuenca está situada en el extremo norte de la gama de los arroyos (cadena montañosa) que limita al norte con el mar de Beaufort, al occidente se extiende hacia el mar de Chukchi y la plataforma de Chukchi. Esta cuenca tiene 1.000 km de largo, 600 km en su parte más ancha, y cubre un área total de 240.000Km^2. El North Slope presenta arenas intercaladas con arcilla que son resultado de ambientes deltaicos de depositación[30].

1.1.1. Estratigrafía de la cuenca ANS

La estratigrafía de la cuenca ANS se divide en tres secuencias principales: Ellesmerian que se depositó desde el Mississippian hasta el Triásico y que contiene depósitos marinos poco profundos. La secuencia Beaufortian que se depositó desde el Jurásico hasta el Cretácico Inferior y está dominada por lutitas. Finalmente, la secuencia Brookian que es una secuencia dominada por depósitos siliciclásticos provenientes de depósitos marinos y terrestres poco profundos [37]. Esta última secuencia alberga las formaciones más importantes para la exploración de hidrocarburos como Torok y Nanushuk como se muestra en la Figura 1 y es el objetivo de interés en este estudio.

Figura 1.

Sistema petrolífero y columna estratigráfica North Slope.



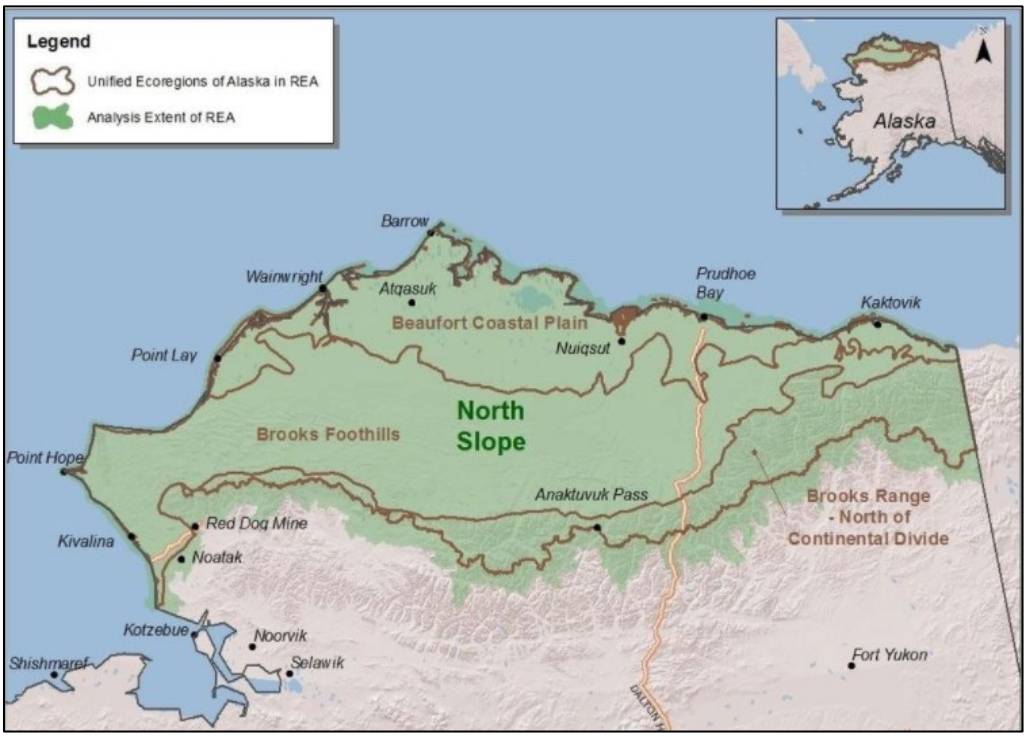
Nota. La figura representa la estratigrafía de la cuenca North Slope, el cuadro rojo enmarca la formación de interés del bajo Brookian de edad cretácea. Tomada de: P. Decker, "Brookian Topset Stratigraphic Play: Petroleum Systems Elements", Alaska Geological Society Meeting, December 2018. [En línea] Disponible en : https://dog.dnr.alaska.gov/Documents/ResourceEvaluation/20181213_BrookianTopsetStratPlay_PetrolSysElements_AGS.PDF

1.2. Formación de interés

La formación Nanushuk tiene varios yacimientos descubiertos según La Reserva Nacional de Petróleo de Alaska (NPR-A) . En una de las zonas geográficas de la NPR-A conocida como Willow, ha sido perforada por la compañía ConocoPhillips. Para el año 2019, se estimó para esta zona una cantidad recuperable de hidrocarburos de 400-750 millones de barriles de petróleo (MMBO) [21] y algo más de 6 MMBO para Nanushuk. Por esta razón, esta formación seguirá siendo considerada como un buen objetivo para la exploración y producción de hidrocarburos.

Figura 2.

Ubicación de la zona geográfica North Slope- Alaska.



Nota. La figura muestra la ubicación de la zona geográfica North Slope- Alaska al noroeste de Estados Unidos. Tomada de: US Bureau of Land Management (Feb 2013).

Basados en la descripción anterior y en la ubicación de las zonas más prospectivas según la NPR-A, se hizo la búsqueda de pozos que cuenten con información de registros eléctricos así como fotos UV realizadas en Corazones en toda la cuenca.

Los pozos T2, T6 y U18 fueron seleccionados al ser parte de la misma formación con un origen sedimentario similar, donde se espera que la composición mineralógica no cambie drásticamente. Los pozos T2 y T6 fueron perforados por la compañía ConocoPhillips en el año 2016 (ubicados en la parte central y occidental de North Slope) [22], mientras que el pozo U18 fue perforado por la compañía Linc Energy, Inc. (parte sur de Umiat) en el año 2013.

1.3. Ambiente de depositación

Los estudios realizados en el Grupo Nanushuk por diferentes compañías, han tenido como principal objetivo la comprensión de los procesos de depositación, distribución y diagénesis de los cuerpos de roca arenisca (estratos potenciales de yacimiento). La formación Nanushuk ubicada en la cuenca North Slope, limita con el mar Beaufort y el mar Chukchi que la expone a procesos de ambientes deltaicos.

Los ambientes deltaicos están formados por la afluencia de agua que transporta una carga de sedimentos a medida que se incorporan en un cuerpo de agua estancada, la velocidad con la que viajan los sedimentos disminuye así como su capacidad de transporte hasta que se depositan. La forma de los deltas varía a causa de los sedimentos que son aportados por los ríos y las fluctuaciones que se generan a nivel del mar y el oleaje[23][25].

En la formación de interés existen deltas antiguos del Holoceno del sistema del río Mississipi que son modificados por el oleaje y tienen un alto contenido de sedimentos. Otros procesos marinos disgregan los sedimentos hacia otras partes de la costa [26]. Los deltas más antiguos, pueden llegar a contener los sistemas petroleros más productivos y que al estar expuestos por diferentes corrientes permiten que existan laminación de arenas-arcillas. En Alaska, las arcillas además de ser laminadas se encuentran en forma estructural, lo que hace difícil poder encontrar aquellas zonas con hidrocarburos.

1.4. Registro de resistividad en formaciones de arenas-arcillosas

La conductividad eléctrica es la capacidad que tiene un fluido en permitir el paso de corriente eléctrica. La resistividad es inversamente proporcional a la conductividad y por eso un valor alto de resistividad indica que el material es mal conductor. Cuando el valor de resistividad es bajo, indica que el fluido es un buen conductor de corriente eléctrica[65].

La complejidad geológica en los yacimientos existentes en la formación de estudio, por los ambientes sedimentarios deltaicos (sección 1.3) se debe a la presencia de minerales de arcilla intercalada en la arena que presentan una alta conductividad. Esta intercalación afecta de diferentes formas la evaluación de la presencia de hidrocarburos en base a las medidas realizadas por los registros de resistividad.

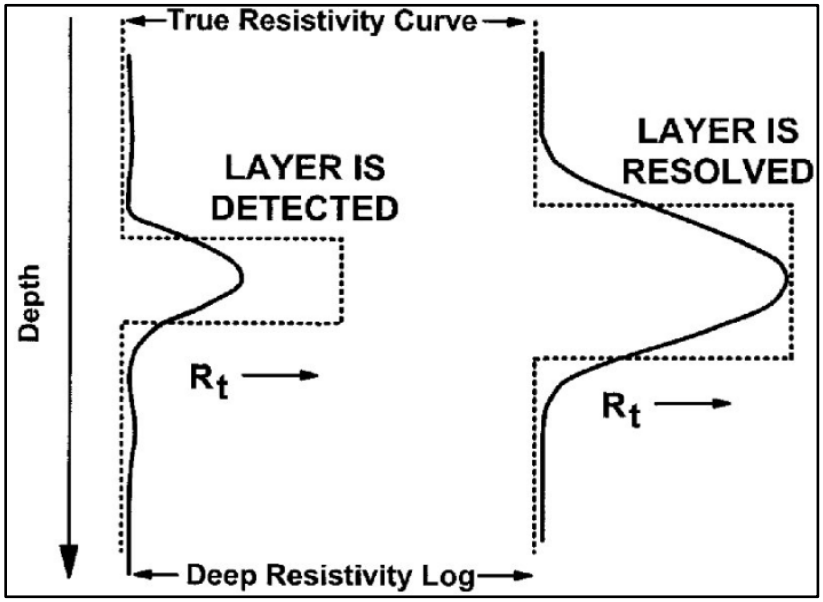
Los registros de resistividad [17] son esenciales para determinar zonas productoras de hidrocarburos. La matriz de las rocas arcillosas no son conductoras, al no serlo la saturación de hidrocarburos de los poros aumenta y la resistividad de la roca debería también aumentar.

Estos registros se ven afectados por la presencia de arcillas, las cuales tienen la capacidad de intercambio catiónico con el agua de formación, es decir, las cargas negativas de las arcillas (aniones) se ligan fuertemente con las cargas positivas (cationes) del agua de formación, aumentando la conductividad total. Debido a este incremento de conductividad, la corriente eléctrica de las herramientas sigue el camino de menor resistencia y por lo tanto la respuesta de la herramienta no corresponde a la resistividad real de una arena gruesa saturada con hidrocarburo [35]. En cambio, se genera una supresión de los registros de resistividad lo

que conduce a que la saturación de agua calculada a través de modelos convencionales como Archie sea mayor a la real, subestimando la saturación de hidrocarburos.

Figura 3.

Resolución de la herramienta de resistividad vs. el tamaño de las arcillas.



Nota. La figura representa la resolución de la herramienta de resistividad profunda vs. el tamaño de las arcillas. La línea continua representa la resistividad paralela donde el pico proporciona el límite inferior a la resistividad verdadera (línea punteada). Tomado de: GEOExPO [En línea]. Disponible: https://www.geoexpro.com/articles/2016/02/an-electrical-rock-physics-framework-for-csem-interpretation

1.5. Imágenes ultravioleta en petrofísica

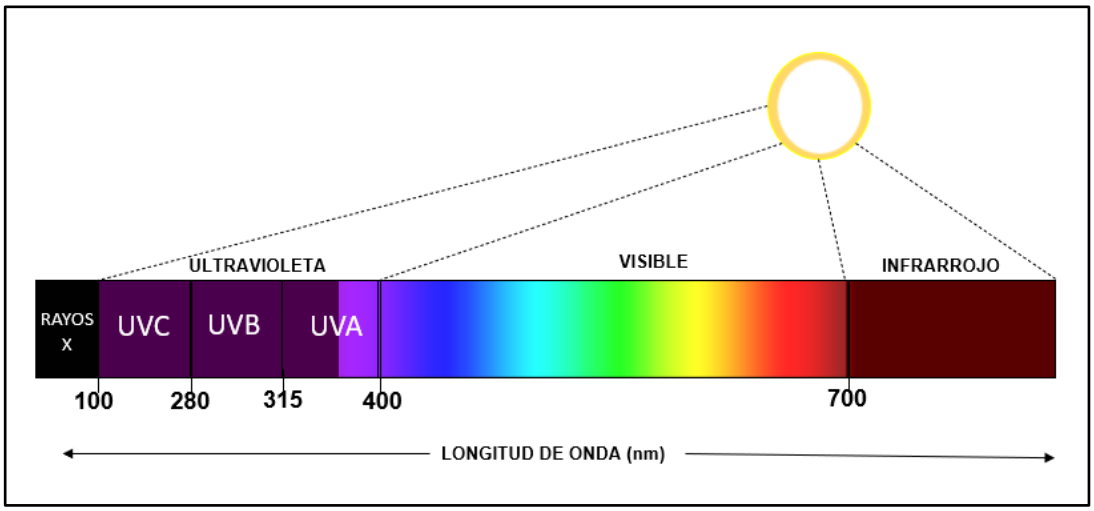
La identificación visual de la presencia o ausencia de hidrocarburos en muestras de roca se hace posible gracias a la propiedad de fluorescencia de los compuestos orgánicos del petróleo. La técnica utilizada para tal identificación tiene el nombre de Fluorescencia UV (Ultravioleta) y consiste en tomar fotos de la roca con luz emitida en el espectro ultravioleta y el uso de equipos de fotografía especiales que permiten capturar el brillo fluorescente de la roca bajo esta exposición.

La fotografía UV es un proceso de captura de imágenes bajo técnicas que utilizan luz ultravioleta. El espectro de luz tiene longitudes de onda de un conjunto de radiaciones electromagnéticas. La luz UV es una porción del espectro de luz que se localiza entre los 100 nm a 400 nm. Este tipo de luz, tiene una longitud de onda más corta que la luz visible (Figura 4.), es decir, no es visible al ojo humano pero sí ante una cámara. En una roca, la luz UV

incidente es absorbida por el material reteniendo su energía y luego la emite en una longitud de onda diferente que se percibe como luminiscencia. Esa luminiscencia es capturada por la cámara como una imagen normal, en un ambiente oscuro con una fuente de luz únicamente UV.

Figura 4.

Espectro de luz

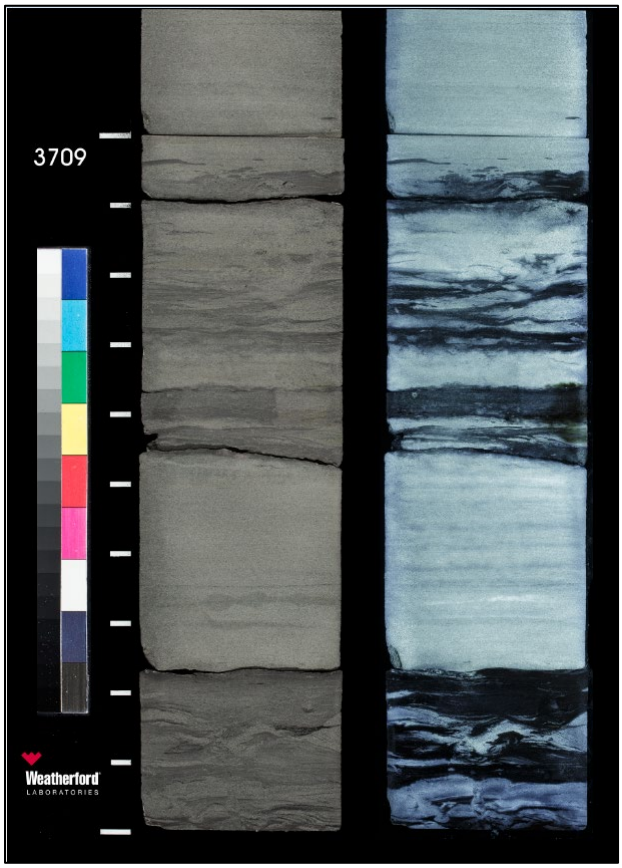


Nota. La figura representa la escala cromática de los espectros de Luz. Se muestran 4 secciones y los rangos de longitud de onda (nanómetros) en la que se encuentra cada espectro. De la sección de luz ultravioleta: UVC es la banda de luz corta de 100-280 nm, UVB banda de luz de onda media de 280-315 nm y UVA banda de luz larga de 315-400 nm.

La fluorescencia representa la capacidad distintiva del aceite de emitir luz en el rango visible cuando se expone a la luz ultravioleta. Cuando los componentes orgánicos de la roca absorben la luz en espectro ultravioleta, los electrones se ubican en un mayor estado de energía de forma temporal y cuando se libera dicha energía es cuando se percibe ese "brillo" al volver a su estado inicial.

Figura 5.

Imagen en luz blanca y en luz UV de una muestra de roca.



Nota. La figura enmarca dos tipos de imagen en una muestra de roca a 3709 ft aproximadamente del pozo T2. Al lado izquierdo se encuentra la imagen capturada bajo luz blanca y al lado derecho a la misma profundidad, la imagen capturada bajo luz UV. Se muestra la diferencia de los espectros de luz incidentes sobre la muestra. Tomada de: Weatherford Laboratories- Imágenes extraídas de las carpetas de pozo encontradas de la información pública.

Desafortunadamente, no solo el hidrocarburo, sino también compuestos minerales orgánicos de la roca pueden ser fluorescentes. Las técnicas explicadas en este documento son desarrolladas para ambientes de siliciclastos en donde la presencia de minerales autigénicos o cementos carbonatos es nula o insignificante. La presencia de reportes de fracciones mineralógicas en muestras de roca apoya nuestra suposición y se explicará en detalle en secciones posteriores del presente documento.

2. METODOLOGÍA Y DATOS

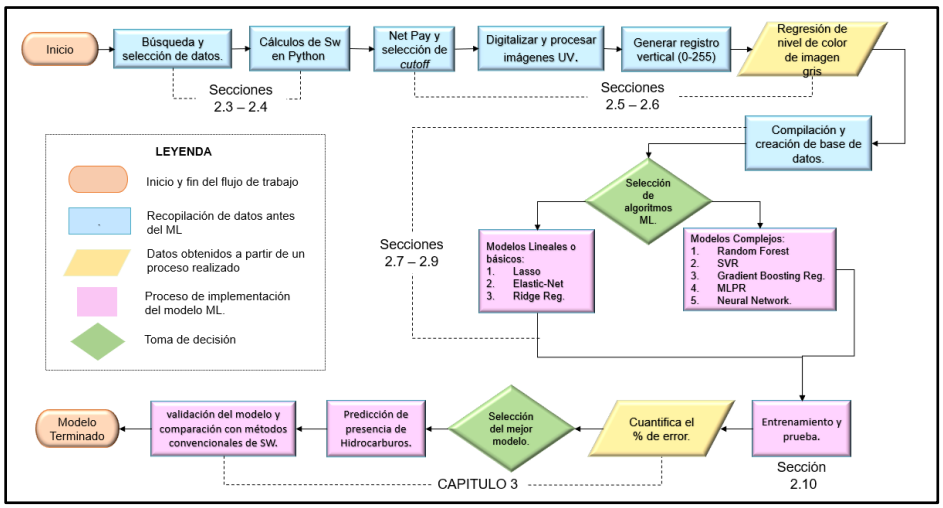
Para esta sección, se muestra la preparación de los datos encontrados para la selección de los pozos de estudio. Una vez son seleccionados los pozos, se identifica la existencia de registros de pozo básicos, análisis en corazones e imágenes ultravioleta disponibles para el desarrollo del proyecto. Finalmente, se presenta la interpretación petrofísica convencional para la obtención de las saturaciones de agua y las zonas prospectivas por los cuatro modelos convencionales.

2.1. Diagrama de flujo de la metodología

En las siguientes secciones se procederá como se relaciona a continuación: En las secciones 2.3-2.4, se hará la búsqueda y selección de los datos de trabajo, así como el cálculo de saturación de agua por métodos convencionales. Para las secciones 2.5-2.6, se desarrollará la selección del *cutoff* de Sw y el procesamiento de las imágenes UV hasta su obtención en escala de grises. Las secciones 2.7-2.9, contendrán la compilación del conjunto de datos para seleccionar los modelos de Machine Learning y clasificarlos según el tipo de modelo. Para la sección 2.10- 2.12 se seccionarán los datos para entrenamiento y prueba hasta la selección de los modelos a predecir. Por último, en el capítulo 3 se mostrarán los resultados obtenidos de las secciones anteriores hasta la selección del mejor modelo de ML para validarlo en el pozo de prueba y compararlo con los modelos convencionales.

Figura 6.

Diagrama de flujo de la metodología - independiente de pozo



Nota. La figura representa la descripción del proceso completo y un paso a paso para la selección- validación del modelo predictivo. La leyenda, indica la acción que se llevará a cabo para cada uno de los capítulos según la figura geométrica.

2.2. Preparación de datos

La preparación de datos consiste en recopilar toda la información de los pozos de estudio. Esta información, permite verificar qué conjunto de datos se tiene disponible y se va a implementar para todo el proyecto. Dentro del conjunto de datos se debe tener: registros eléctricos de pozo, análisis de corazones (porosidad, permeabilidad, salinidad, densidad de grano y saturaciones de los fluidos) y fotos bajo luz ultravioleta para identificar la presencia y ausencia de hidrocarburos.

Luego de tener el conjunto de los datos mencionados anteriormente, se ordena y se anexan los cálculos del "registro de temperatura" y resistividad del agua. Esto se dispone en carpetas y se ubican dentro de Python. Python es un lenguaje de programación gratuito, de código abierto, fácil de entender y utilizar para el análisis de datos. Cuenta con una amplia cantidad de librerías y de herramientas que hacen que este idioma de programación sea único.

La elección de una estrategia apropiada para el análisis de todos los datos depende del plan preliminar del análisis de la información que se estudió desde el plan original. Es indispensable verificar la existencia de problemas y la factibilidad de modificar las estrategias porque de eso depende aplicar una acción de manera oportuna dentro del flujo de trabajo. A continuación, se describen los datos recopilados para los tres (3) pozos seleccionados, tanto para el análisis convencional como para el análisis de los núcleos.

2.3. Selección de pozos

Para el presente trabajo, la selección de pozos se hizo basada en la disponibilidad de imágenes UV. Dado que la adquisición de corazones de roca se hace sólo en ciertos pozos, y que de estos, solo se toman fotografías UV en algunos pozos, su disponibilidad es uno de los mayores limitantes del método propuesto en este estudio. Luego de un análisis detallado de todos los pozos disponibles en(http://dggs.alaska.gov/gmc/seismic well data.php.),encontramos solo 3 pozos con disponibilidad de imágenes.

Los pozos de estudio se referencian en el trabajo como T2, T6, que son dos pozos exploratorios cercanos ubicados al norte de nuestra área de interés y el pozo U18, que es un pozo al sur y más lejano, a 96 kilómetros (detallado en la Figura 8 [77]). La ubicación de los pozos en zonas distantes, pero originadas desde la misma fuente de sedimentos siliciclásticos es conveniente para el entrenamiento de un modelo de *Machine Learning*.

El pozo U18 se encuentra en una zona originalmente llevada a altas profundidades, mayor a la actual, y luego levantada a profundidades someras por procesos geológicos que

resultaron en menor porosidad y menos presencia de arcilla. La inclusión de este pozo en el estudio conlleva a tener tipos de roca que cubren un rango amplio de porosidades y concentración mineralógica. Esto se verá reflejado en mayor aplicabilidad de los modelos de *Machine Learning* a nuevos pozos, ya que el entrenamiento se hace en un amplio rango de posibilidades de roca que posiblemente cubre el tipo de roca presente en futuros pozos a perforar (opuesto a hacer entrenamiento solo en pozos contenidos en una reducida área geográfica).

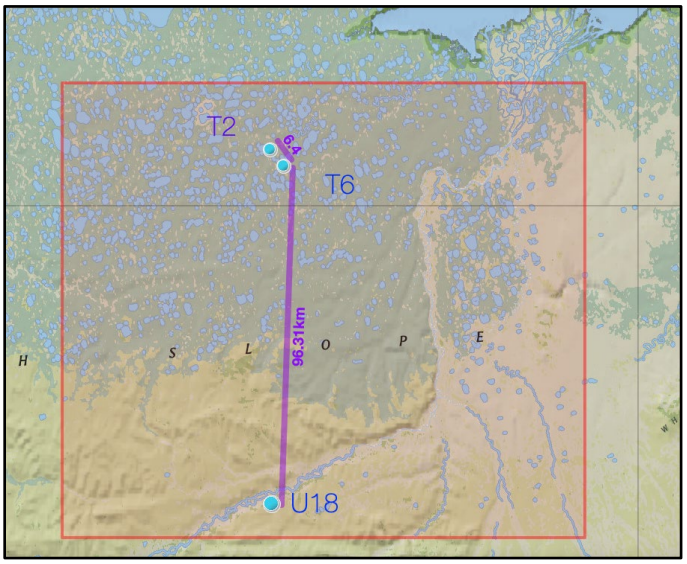
Figura 7.

Ubicación de los tres pozos sobre mapa - Alaska.

Nota. La La figura representa la zona de interés donde se ubican los tres (3) pozos en Alaska. [En línea]. Disponible: https://mapmaker.nationalgeographic.org/ccQJpfu9H6bWim3l85ZN5e//?edit=fZwtreyD4khUC3356wGeA2

Figura 8.

Zoom de las zonas de interés sobre mapa - Alaska.



Nota. La figura muestra el zoom de la zona de interés que se enmarca en el cuadro rojo. Se representa la ubicación de los tres pozos y la distancia entre ellos (en kilómetros). [En línea]. Disponible: https://mapmaker.nationalgeographic.org/eQzLQJf2nMH7AMoMDrFRnx//?edit=hxaqMVdFYb2soXutWTZQtt

2.3.1. Zonas por pozo

Una vez encontrada la información y datos de cada uno de los pozos, se hizo una revisión de los archivos contenidos para cada uno de ellos, donde se encontraron las zonas de interés y los intervalos correspondientes como se muestra en la siguiente tabla:

Tabla 1.

Campo, Intervalos y zonas de interés por cada pozo.

POZO	CAMPO	ZONAS EXISTENTES	ZONA DE INTERÉS	Intervalo (ft)
T2	NPRA ASP4	Zona 1,2 y 3	Zona 1	3665.65 - 3769.49
T6	NPRA ASP4	K3, Sequence 97 and Sequence 93, Falcon & Willow	Falcón	3370 – 3525
U18	Umiat	Zona 1,2 , 3 y 4	Zona 2	650-1200

Nota. Esta tabla muestra el campo al que pertenecen los pozos, las zonas de interés e intervalos correspondientes para el análisis petrofísico de los pozos.

Estos intervalos serán usados para seleccionar las secciones de registros de pozo a usar en el entrenamiento y prueba, y así delimitar el modelamiento a la zona de interés.

2.3.2. Registros

Los registros de pozos que se usarán en el presente trabajo son: gamma ray (GR), resistividad profunda (AT90), registro de porosidad neutrón (NPHI), registro de inverso de velocidad compresional (DTCO) y el registro de porosidad (RHOZ). Estos registros se encuentran en formato LAS (Log ASCII Standard), que es un formato que almacena y distribuye los datos digitales de registros de pozos. Esta norma (LAS) fue definida y es mantenida por la Canadian Well Logging Society (CWLS) [79].

Otros archivos se encontraron en formato DLIS (digital log information System) que contiene una base de datos de distintas mediciones como resistividad, porosidad, sónicos y resonancia magnética. Estos archivos se guardan en formato estándar de registro digital de intercambio, que se usa para poder intercambiar datos de registros de pozo. Este tipo de formato debe ser abierto por un programa en específico para poder leer, extraer y visualizar toda la información contenida.

2.3.3. Imágenes UV

De toda la información recopilada, se encontraron alrededor de cinco a siete imágenes UV por pozo y los registros mencionados anteriormente (GR, AT90, NPHI, DTCO y RHOZ) en formato LAS o DLIS. Para los pozos T2 y T6, los registros GR , AT90 , NPHI y RHOZ fueron encontrados en un mismo archivo LAS mientras que el registro DTCO fue encontrado en un archivo diferente.

Para el pozo U18, los registros GR, DTCO, AT90 y NPHI se encontraron en un mismo archivo DLIS y el registro faltante (RHOZ) fue tomado y procesado por el software Plotdigitizer para generar una tabla de valores. En Python, los valores de RHOZ fueron interpolados a la misma profundidad de los demás registros encontrados para el mismo pozo.

A continuación, se muestra una tabla con el listado de validación de la información encontrada para cada uno de los pozos:

Tabla 2.

Información disponible por pozo.

POZO	PROFUNDIDAD (ft)	Fotos UV	GR	DTCO	RHOZ	AT90	NPHI
T2	3660 - 3895	✓	Si	Si	Si	Si	Si
T6	3370 - 3525	✓	Si	Si	Si	Si	Si
U18	650 - 1200	✓	Si	Si	Si	Si	Si

Nota. Esta tabla muestra la profundidad de interés, disponibilidad de fotos UV y disponibilidad de registros eléctricos que serán usados durante todo el trabajo para los pozos T2, T6 y U18.

2.3.4. Datos de corazones

Los núcleos o corazones son muestras de roca tomadas del pozo a una profundidad específica. Estos fragmentos son tomados por herramientas especiales, que permiten preservar la estructura geológica y sus características fisicoquímicas de la mejor manera posible. Los análisis realizados en los corazones permiten evidenciar la presencia de hidrocarburos, capacidad de producción del yacimiento e incluso poder detectar rasgos del yacimiento que no hayan sido captados por las mediciones hechas por los registros de fondo de pozo [7].

A pesar de que nuestro objetivo es desarrollar un modelo de *Machine Learning*, desarrollar modelos petrofísicos que representan los datos de núcleo es importante. La solidez del modelo de Machine Learning será evaluada respecto a los resultados de espesor de hidrocarburo de estos modelos convencionales. Debido a que varios parámetros de un modelo petrofísico se deben asumir, los datos del núcleo, siendo medidas más directas de las propiedades de la roca que un registro de pozo, ayudan a validar los resultados de porosidad y saturación. A continuación presentamos los tipos de datos de núcleo a usar y su disponibilidad en los pozos de estudio.

2.3.4.a.Difracción de rayos x. La difracción de rayos X (XRD por su sigla en inglés X-Ray Diffraction) es una de las técnicas más eficaces para identificar de forma cualitativa y cuantitativa de arcillas y otros minerales. Esta técnica permite obtener información detallada sobre la estructura cristalográfica presentes en las muestras, que puede usarse para identificar las fases presentes [20].

Este tipo de técnica también determina los porcentajes normativos o de peso de las fases presentes así como la fracción de cada fase mineral contenida en sus muestras [20]. Para esta

sección, se utilizó el XRD en porcentaje en peso para cada uno de los pozos. Los archivos utilizados para el XRD fueron tomados de las diferentes bases de datos de los pozos realizados por la compañía Weatherford Laboratories. Los XRD encontrados por Pozo fueron:

- T2: Se encontró un único archivo XRD que contenía los valores para el tipo de muestra de núcleo lateral rotativo.
- T6: Se encontraron dos tipos de archivos XRD. El primer archivo (HH-91264ccXRD), contenía los valores asociados a la muestra de núcleo convencional mientras que el segundo (HH-91264rXRD) fue realizado para el tipo de muestra de núcleo lateral rotativo. Para tener el mismo tipo de análisis se tomó el segundo archivo encontrado para este pozo.
- U18: Se encontraron dos archivos XRD del mismo tipo de muestra (núcleo convencional), se utilizó el que tenía mayor cantidad de datos.

Las tablas XRD encontradas para cada pozo, contenían los datos tanto en porcentaje en peso como en volumen de cada uno de los minerales encontrados en la formación. Los porcentajes que se emplearon fueron los del porcentaje en volumen para los dos primeros pozos y para el último porcentaje en peso. El laboratorio muestra los valores obtenidos para cada uno de los minerales que contiene el núcleo extraído de la formación. Dentro de estos minerales se encuentran:

- *Clays:* Los grupos de arcillas presentes en el estudio de laboratorio son la clorita, caolinita, la illita/mica e illita/esméctica.
- *Carbonatos:* Se encuentran la calcita, dolomita ,dolomita (Fe/Ca+)2 y la siderita. Otros minerales: cuarzo, feldespato (K-spar), plagioclasa, pitita, dióxido de titanio, halita y pirita.

Para este estudio, se tomaron los valores totales del grupo de arcillas (*Clays*) contenidos en los diferentes archivos por pozo. Con esta información, se elaboró una tabla con los valores de los cuatro minerales del grupo de las arcillas por profundidad, el total de las arcillas (*total Clays*), el total de los carbonatos y de otros minerales para cada uno de los pozos como se muestra en la siguiente tabla:

Tabla 3.

Visualización parcial de la tabla original de XRD (Excel)

Depth	Chlorite	Kaolinite	Ilite Mica	Ilite Smectite	Total Clays	Tota Carb.	QFM	Well
3311	0.05	0.02	0.20	0.05	0.32	0.04	0.64	T2
3318	0.04	0.03	0.11	0.03	0.21	0.05	0.74	T2
3331	0.04	0.03	0.11	0.03	0.21	0.01	0.78	T2
3347	0.04	0.02	0.12	0.02	0.20	0.05	0.75	T2
3391	0.05	0.03	0.19	0.08	0.35	0.04	0.61	T2

Nota. En esta tabla se muestra la visualización parcial de la tabla original realizada en Excel del listado de valores obtenidos de las tablas XRD encontradas (laboratorio Weatherford). Se tomaron los valores de los minerales arcillosos como la clorita, caolinita e ilitas y el total del conjunto de los minerales. El valor total de arcillas para cada pozo por profundidad fue tomado para diferentes cálculos en el presente trabajo.

2.3.4.b.Porosidad y permeabilidad. Del conjunto de datos recopilados, se hallaron las muestras de corazón analizadas por el laboratorio Weatherford en archivos denominados "*Routine Core*". El *Routine Core*, proporciona una medición directa de la formación e información crítica sobre el yacimiento.

Para los tres pozos, se encontraron dos tipos de archivos que tenían en común tablas de valores de porosidad Klinkenberg (md), porosidad (%) a 1010 psi de presión, densidad de grano (gm/cc). Uno de los dos archivos, contiene además de los valores antes mencionados también valores de saturación de fluidos (agua (Sw) y aceite (So)). Todos estos valores fueron tomados de las tablas de "*Routine Core- SUMMARY OF ROUTINE CORE ANALYSES RESULTS*" presentadas por el laboratorio Weatherford.

Es clave para el proceso, la implementación de la porosidad como factor principal en conjunto con la permeabilidad para definir el corte de agua para cada uno de los pozos. Esto permite poder obtener las zonas prospectivas por métodos convencionales y poderlas diferenciar de las calculadas por el procesamiento de las imágenes.

Tabla 4.

Visualización parcial de la tabla original de valores de Routine Core (Excel)

DEPTH	K	PHIT	RHOG	SW_O	SW	SO	WELL
3675.5	0.094	0.1746	2.71	0.5815	0.5395	0.2773	T2
3677.15	0.048	0.1600	2.70	0.5904	0.5656	0.3314	T2
3681.50	3.24	0.2016	2.70	0.4049	0.3872	0.4348	T2
3683.30	0.013	0.0918	2.69	0.7202	0.6710	0.1364	T2
3686.40	3.49	0.1989	2.70	0.4492	0.4285	0.4006	T2
3689.20	6.01	0.2052	2.70	0.3974	0.3810	0.4484	T2

Nota. La tabla muestra los datos por profundidad y pozo de permeabilidad (K), porosidad (PHIT), densidad de grano (RHOG), saturación de agua ajustada (SW_O), saturación de agua (SW) y saturación de aceite (SO). Esta tabla es una visualización parcial que muestra algunos de los datos del pozo T2.

2.3.4.c.Salinidad. Con el fin de obtener el valor de la salinidad, fue necesario buscar en la base de datos de los pozos el reporte final de las pruebas de fluidos. Se encontró que la compañía de servicios de pruebas de producción (PTS por sus siglas en ingles), fue quien realizó las pruebas de fluidos para los tres pozos. En el archivo *FINAL_PTS_Final_Main_Flow_Test_Report*, se hallaron lecturas y resultados del campo, durante diferentes periodos de tiempo.

En los comentarios realizados por la compañía sobre los cloruros y el ph se encontraron los siguientes datos: BS&W = 2% salmuera, 98% Crudo, Trazas de Sólidos. 0% H2S, 0% CO2, Cl = 18,000ppm, pH = 7. La interpretación petrofísica básica del presente trabajo asume una salinidad de agua 18 ppm NaCl equivalentes dada la precisión de una prueba real de producción y muestreo en cabezal de pozo. Ese valor de salinidad se utilizó para el posterior análisis y cálculo de Rw (resistividad del agua).

Se sabe que los pozos seleccionados (T2, T6 y U18) se encuentran en la misma formación, el valor de salinidad (cloruros) se asume constante debido a que el origen sedimentario y ambientes de depositación son similares.

2.4. Interpretación petrofísica convencional

2.4.1. Zonas de interés

Los valores e información recopilada de las zonas de interés para cada uno de los pozos se mencionan en la sección 2.3. Para los pozos T2 y T6, se analizaron dos para-secuencias, es

decir, dos tipos de secuencias estratigráficas que de base a tope son más pequeñas en escala. Estas consecuencias , son resultantes de las oscilaciones en un periodo de tiempo corto entre el aporte de sedimentos y cómo se estructura. El paquete de estratos es genéticamente relacional si se depositan en el mismo ciclo según los cambios por el nivel del mar [32].

Para el pozo U18, se encontró que la formación es más gruesa y por tanto se escogieron las zonas dentro de la formación que contaban con análisis en corazones. Una de las zonas que contaba con estas imágenes, oscila entre los 650 a 1200 ft de profundidad como se muestra en la Tabla 1.

2.4.2. Registros de pozo

Los registros en formato LAS contienen diferentes secciones para identificar toda la información de los pozos. Dentro de esas secciones se encuentran:

- *VERSIÓN* - es el delimitador para las secciones de datos e información de envoltura.
- *WELL* - (obligatorio) nombre del pozo, ubicación e información sobre las profundidades de inicio y parada, intervalo de paso entre las profundidades y definiciones de valores nulos.
- *CURVE* - (obligatorio) lista de curvas de registro de pozo incluidas en el archivo.
- *PARAMETER* - datos sobre la elevación del pozo, la profundidad del revestimiento, etc.
- ASCII - (obligatorio) datos de la curva de registro digital.

En la programación realizada en Python se importó la librería Lasio para leer los archivos LAS de los registros. Luego de leerlos, se extraen del formato a un *Data Frame* (estructura de datos heterogénea más usada como una versión más flexible de una matriz en el análisis de datos) para ser graficados en función de la profundidad como se muestra en la Figura 32.

En el caso del U18 fue necesario realizar el siguiente procedimiento para la lectura de los registros, en la programación así:

- Los formatos ".las" del pozo no contenían los 5 registros básicos (GR, DTCO, AT90, NPHI, RHOZ) así que se buscaron los formatos .dlis
- Se realizó la conversión del archivo de .dlis a .las , generando 20 archivos con los registros GR, DTCO, AT90 y NPHI.
- Para el registro RHOZ se utilizaron dos *softwares* de uso libre *(Java SE Runtime Environment 8 y Plotdigitizer*) para la generación de una tabla de valores del registro.

- Por último, se extrajo la tabla de datos a formato .xlsx para poder ser graficados en Python junto con los otros registros.

2.4.3. Cálculo de temperatura de formación (Tf) y resistividad de agua de formación (Rw)

Para los cálculos se tomaron diferentes fórmulas del manual Crain [39], para convertir una salinidad medida en laboratorio en una resistividad del agua de formación (Rw) a cualquier temperatura específica (FT) en grados Fahrenheit como se muestra en la Tabla 5. Para la construcción total de la tabla se realizaron los siguientes pasos:

2.4.3.a.Cálculo de Rw a temperatura ambiente y salinidad específica

- La temperatura promedio anual en North Slope es de 25 ° F (-3.9°C).
- A superficie se toma la temperatura promedio y profundidad MD a 0 ft.

Para convertir la salinidad a cualquier temperatura arbitraria y encontrar la resistividad en condiciones de laboratorio, se utilizó la siguiente fórmula así:

Ecuación 1

$$RW@FT = \left(\frac{\frac{400000}{FT_1}}{WS}\right)^{0.88}$$

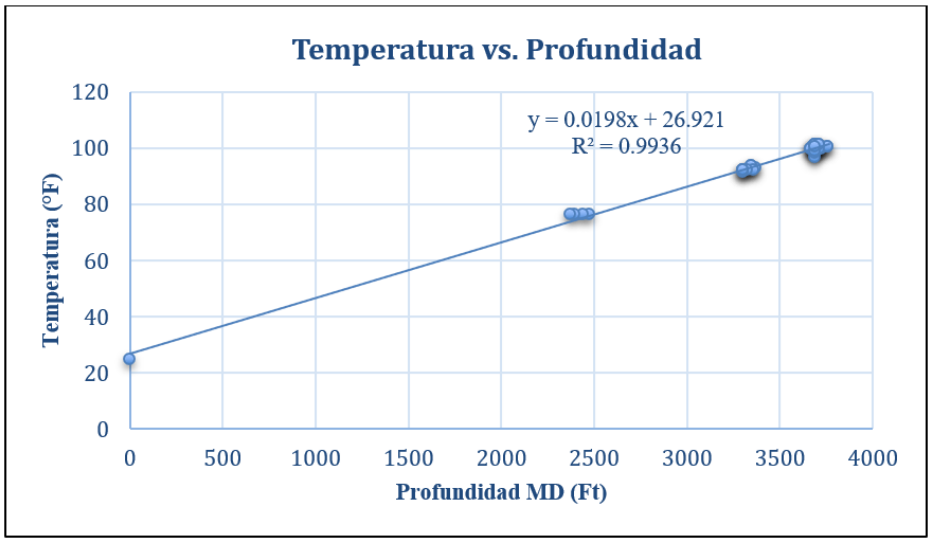
$$RW@25°F = \left(\frac{\frac{400000}{25}}{18000}\right)^{0.88} = 0.90\ ohm.m$$

Donde RW@FT es la resistividad del agua a temperaturas de formación (ohm-m), WS es la salinidad del agua (ppm NaCl) y FT1 es la temperatura de formación (grados Fahrenheit)[39].

2.4.3.b.Cálculo temperatura como función de profundidad. Se graficaron los valores de la profundidad MD en pies (ft) y la temperatura "after" en grados Fahrenheit para obtener la fórmula de regresión lineal que se muestra en la Figura 9. La fórmula de y, se utiliza para poder hallar los valores de cada una de las temperaturas (x) en función de la profundidad Y(FT). Al hacer todos los cálculos, se obtiene un "registro de temperatura" que varía para cada pozo.

Figura 9.

Gráfica de temperatura vs. profundidad - pozo representativo.



Nota. La figura muestra la gráfica de temperatura en grados Fahrenheit en función de la profundidad MD en pies (ft) con regresión lineal del pozo representativo (T2).

Con el registro calculado de temperatura, se hicieron los cálculos para convertir Rw a otras temperaturas con la siguiente ecuación del manual de Crain [39]:

Ecuación 2

$$RW\ @\ FT = \frac{RW@TRW\ (TRW+KT1)}{FT+KT1}$$

Donde RW@FT es la resistividad a la temperatura de formación (ohm-m), RW@TRW resistividad a una temperatura distinta a la de la formación (*Rw_sup*), TRW es la temperatura a la que se midió la resistividad (normalmente Fahrenheit para la profundidad en pies, Celsius para la profundidad en metros), FT es la temperatura de formación (grados Fahrenheit o Celsius) y KT1= 6.77 es un factor de conversión [39].

Tabla 5.

Visualización parcial de la tabla original de valores de resistividad del agua.

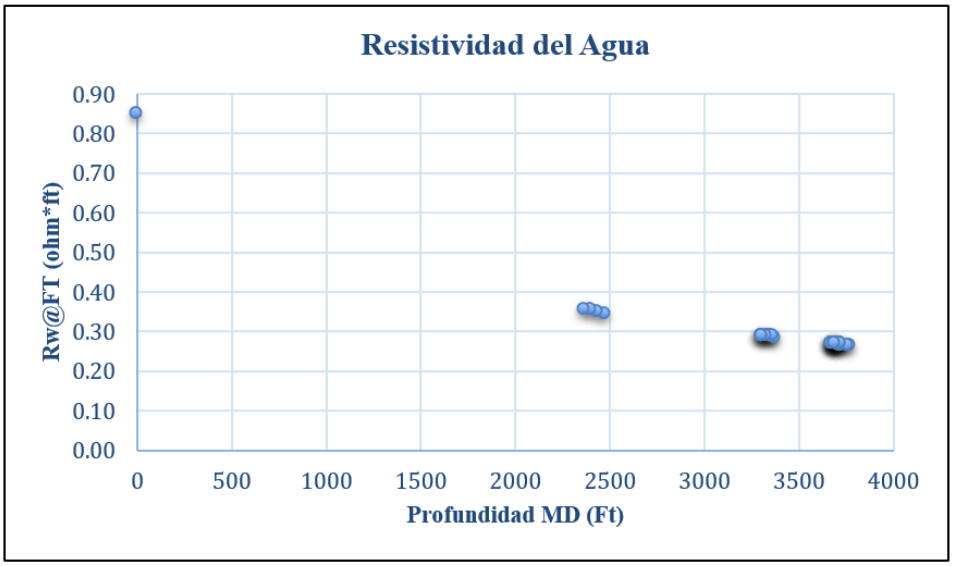
Profundidad MD (ft)	Temp. after (°F)	FT	Rw @ ft
0	25	26.921	0.8501
3693.13	97.03	100.04	0.2681
3696.56	98.38	100.11	0.2680
3699.63	98.75	100.17	0.2678
3719.47	99.51	100.57	0.2668
3726.22	99.92	100.70	0.2665

Nota. La tabla muestra los valores obtenidos por profundidad de la resistividad del agua con los valores calculados de cada uno de los parámetros de la ecuación.2.

Para complementar este análisis, la totalidad de las resistividades calculadas se graficaron en función de la profundidad como se muestra en la siguiente figura :

Figura 10.

Gráfica de resistividad del agua - pozo representativo.



Nota. La figura muestra la gráfica de resistividad del agua (ohm-ft) en función de la profundidad MD (ft) del pozo representativo T2. En la gráfica se visualiza el comportamiento descendente de la resistividad a medida que aumenta la profundidad.

2.4.4. Cálculo del volumen de arcilla (Vsh)

El volumen de arcilla (Vsh) se define como el porcentaje neto de arcilla que está presente en la formación. Para el cálculo del Vsh, se tomaron los valores de la tabla de datos reportados en el registro gamma ray, el valor máximo y mínimo para hacer el cálculo con la siguiente ecuación:

Ecuación 3

$$Vsh = \frac{GR_{LOG} - GR_{MIN}}{GR_{MAX} - GR_{MIN}}$$

Donde Vsh es el volumen de arcilla, GR_{LOG} es el valor tomado de la tabla de valores del registro a cada profundidad, GR_{MIN} es el valor mínimo obtenido del registro y GR_{MAX} es el valor máximo obtenido de los valores del registro [40].

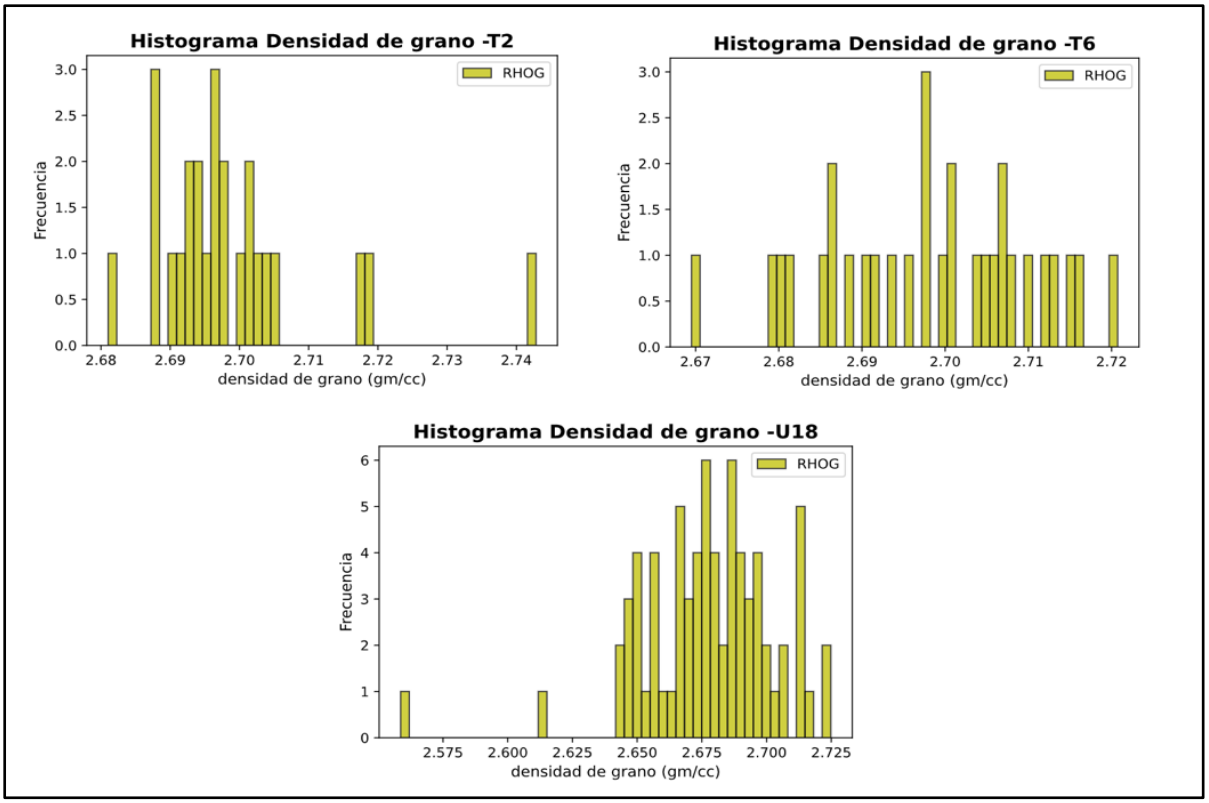
Los valores tomados tanto mínimos como máximos del registro para el cálculo del Vsh, varían para cada uno de los cálculos realizados en los tres pozos.

2.4.5. Cálculo de la densidad de grano variable ("Grain_Density")

El cálculo de densidad de grano en las formaciones de interés requiere del análisis de densidad de granos en las muestras de corazones. El siguiente histograma muestra la distribución de densidad de grano en las muestras que reportaban algún valor (no todas las muestras de núcleo tienen el valor disponible).

Figura 11.

Histogramas de densidad de grano por pozo



Nota. La figura muestra los histogramas de la densidad de grano medida en gramos por centímetro cúbico (gm/cc). Las barras de color amarillo en cada uno representan la frecuencia de los valores representados para cada uno de los pozos.

El histograma junto con el conocimiento de que las muestras fueron adquiridas en un amplio espectro con concentraciones de arcilla bajas (arenas) y altas (lutitas) sugiere usar valores límite de densidad de grano así: 2.65 para las arenas con menos arcilla y 2.75 para las lutitas.

La densidad de grano variable, para cada pozo fue calculada para los métodos convencionales mediante la siguiente ecuación :

Ecuación 4

$$RHOG = (Vsh \times RHO_{sh}) + \big((1 - Vsh) \times RHO_{ss}\big)$$

Donde RHOG es la densidad de la matriz (grano variable) , Vsh es el volumen de arcilla encontrado en el paso anterior, RHO_{sh} es la densidad de la arena y RHO_{ss} es la densidad asumida del volumen de arcilla .

2.4.6. Cálculo de porosidad

La porosidad es el porcentaje de volumen de poros o espacio poroso de la roca que puede contener fluidos. Para determinar los valores de porosidad se tuvieron en cuenta los valores del registro RHOZ a cada una de las profundidades en cada uno de los pozos. Para el cálculo se empleó la siguiente ecuación:

Ecuación 5

$$Ød = \frac{(RHOG - RHO_b)}{(RHOG - RHO_f)}$$

Donde $Ød$ es la porosidad a calcular, $RHOG$ es la densidad de la matriz (grano variable) ,RHO_b es la densidad tomada del registro de densidad (RHOZ) a cada profundidad y RHO_f es la densidad del lodo de perforación[42].

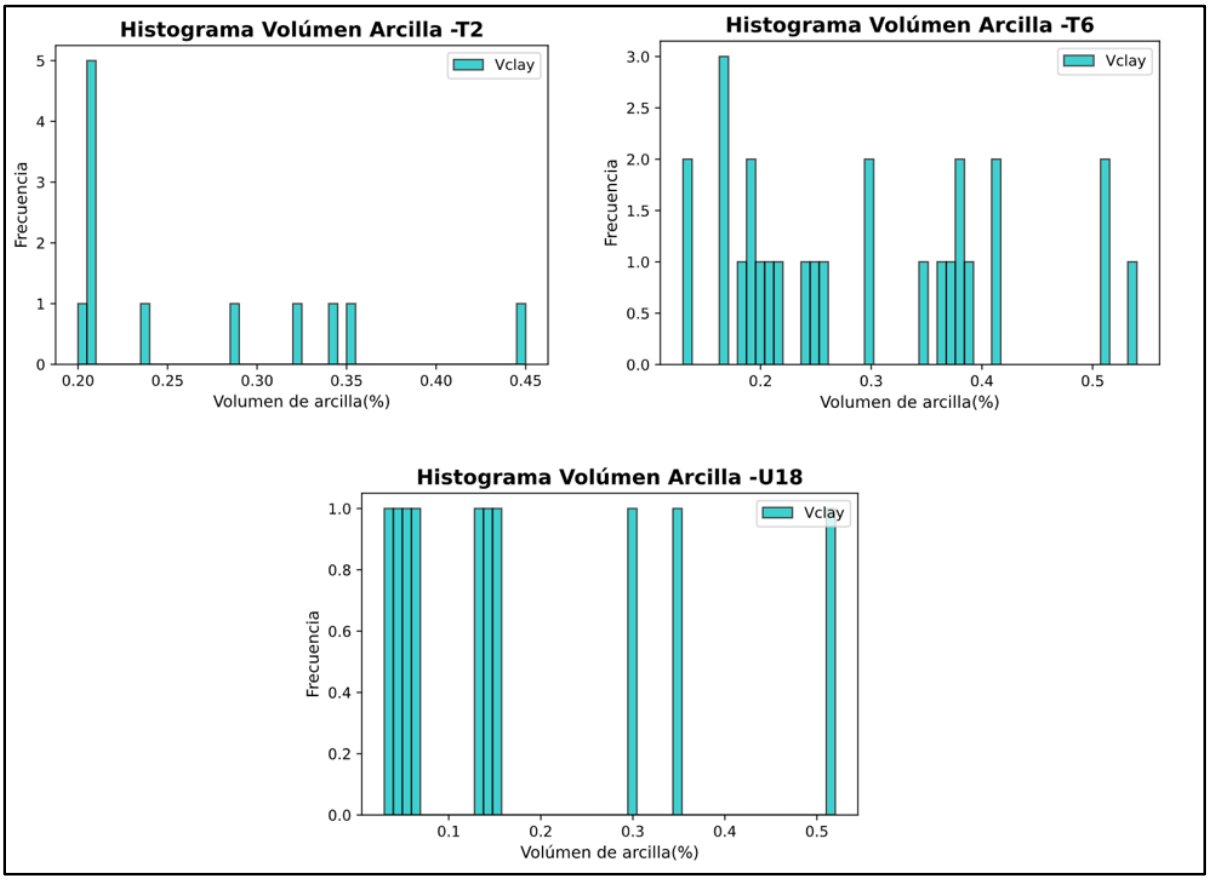
2.4.7. Calibración Vsh y RHOG con datos de corazón

A partir los datos encontrados de porosidad, permeabilidad Klinkenberg, saturaciones de los fluidos (agua y aceite) y densidad de grano, se realizó el siguiente procedimiento:

- Se realizó el ajuste de la densidad de grano del *Core* con la densidad obtenida mediante ecuaciones convencionales basado en la normalización de los puntos.
- Se hizo el cálculo del volumen de arcilla (*Vclay*) en función del volumen de arcilla (*Vsh*) basado en los datos arrojados por el laboratorio Weatherford de la Difracción de Rayos X, en volumen calculado de "*total Clays*" para cada uno de los pozos (*XRD %Vol.*).
- Con los datos obtenidos del paso anterior, se realizó un histograma del porcentaje de *Vclay* respecto a la frecuencia de repetición de los datos para cada uno de los pozos como se muestra en la Figura 12.

Figura 12.

Histogramas de volumen de arcilla por pozo



Nota. La figura muestra las gráficas del Histograma del volumen total de arcillas *(Vclay*) en porcentaje en función, de la frecuencia de los datos obtenidos del laboratorio. Las barras de color cian en cada histograma, representa la frecuencia de los valores representados para cada uno de los pozos.

2.4.8. Cálculo de la saturación de agua por métodos convencionales

El valor del "registro de saturación" se ubicó en el mismo *track* que corresponde a cada una de las saturaciones calculadas con los métodos convencionales como se muestran en las Figuras 34 , 35 y 36 respectivamente.

Para el cálculo de la saturación de agua, se implementaron los siguientes métodos convencionales: Archie, Poupon-Leveaux, Waxman-Smits y Simandoux para cada uno de los pozos [8]. A continuación explicamos los métodos de saturación y las ecuaciones utilizadas:

2.4.8.a.Método para cálculo de Sw de Archie (1942).El método de Archie es uno de los más conocidos y utilizados para el cálculo de la saturación de agua. Archie incorpora propiedades físicas de la roca y mediciones de registros de pozo como resistividad del agua y de la

formación, tortuosidad y porosidad. La presencia de otros materiales conductivos como la arcilla, implica que la ecuación de Archie sea modificada [8].

Para el cálculo de la saturación de agua, se hizo uso de los siguientes parámetros: los exponentes de cementación (m=2), el factor de tortuosidad (a=1), el exponente de saturación (n=2), los valores de las curvas generadas de porosidad (Ø), resistividad del agua (Rw) y el registro de resistividad (AT90) encontrado en la base datos de cada pozo [44].

Ecuación 6

$$SwA = \left(\frac{a}{Ø^m} * \frac{Rw}{Rt}\right)^{\left(\frac{1}{n}\right)}$$

Donde SwA es la saturación de agua por el método de Archie (fracción), Rw es la resistividad de agua de formación (ohm-m), Rt es la Resistividad de formación (AT90)(ohm-m), Ø es la porosidad (fracción), a es el factor de tortuosidad, m es el exponente de cementación que tiene un valor entre 1.8 y 2.0 para areniscas consolidadas y n es el exponente de la saturación.

2.4.8.b.Método de Poupon-Leveaux. En 1954, Poupon y Leveaux incorporaron el término 1-Vsh por primera vez para lograr crear un balance volumétrico entre el volumen de lutita y el volumen de arcilla presente en el yacimiento. Este método, se utiliza para determinar la saturación de agua de areniscas y lutitas finas asumiendo que la conductividad de un medio en particular se basa en su tamaño y material conductor en los espacios porosos. La ecuación de Poupon se expresa de la siguiente forma [44]:

Ecuación 7

$$F = \frac{a}{Ø^m}$$

Ecuación 8

$$Swp = \left(\left(\frac{1}{Rt} - \frac{Vsh}{Rsh}\right) * \left(\frac{F*Rw}{1 - Vsh}\right)\right)^{\left(\frac{1}{n}\right)}$$

Donde F es el factor de formación (Ec.7), Swp es la saturación de agua del método Poupon (fraccional), Rt es la Resistividad de formación (AT90)(ohm-m), Vsh es el Volumen de *shale*, Rsh es la resistividad del *shale* (ohm-m), Rw es la resistividad de agua de formación (ohm-m), m es el exponente de cementación, n es el exponente de saturación, a es el factor de tortuosidad y Ø es la porosidad (fracción).

Para el valor de resistividad del *shale* (Rsh) a partir del registro de GR, se calcula el Vsh (sección 2.4.4) . Con el uso de los valores del registro de resistividad (AT90), se calcula un valor de Rsh promedio como se muestra en la Figura 13.

Figura 13.

Previsualización de código - funciones de almacenamiento.

```
dfSh = df_1[df_1['Vsh']>0.5]
Rsh = np. percentile(dfSh['AT90'],20)
```

Nota. La figura muestra la previsualización de la sección de código en Python para generar un valor promedio de Rsh en el cálculo de la Swp (Ec.8) y Sws (Ec.18) para cada uno de los 3 pozos.

Como se muestra en la Figura 13, se define tomar los valores mayores a 0.5 de Vsh para el cálculo de Rsh. Luego, se calcularon los valores del registro de AT90 con un percentil (indica la posición de medida respecto a todos los datos) de 20. Para eso, se utilizó la función *np.percentile* en Python que permite tomar un arreglo de datos y sacarle la medida estadística del percentil que se está buscando.

2.4.8.c.Método de Waxman-Smits. El modelo de Waxman-Smits (1968), se basa en el resultado de mediciones hechas en laboratorio por el efecto que tiene la arcillosidad en la conductividad. Este método sirve para calcular la saturación de agua en formaciones con presencia de arcilla a partir de información aportada por los registros de resistividad [44][8]. Para el cálculo de la saturación de agua, se realizaron los siguientes 4 pasos con la ecuación del manual de Crain [80].

- Cálculo de la concentración de iones presentes en la formación.

Ecuación 9

$$Qv = \frac{RHOG*(1-Ø)*CEC}{Ø*100}$$

Donde Qv es el contador de la concentración de iones presentes (meq/gm), $RHOG$ es la densidad de la matriz (gm/cc) , CEC es la capacidad de intercambio catiónico y Ø es la porosidad (fracción) [83].

- Cálculo de la conductividad equivalente de intercambio de cationes de arcilla.

Ecuación 10

$$B = 3.83 * (1 - 0.83^{(\frac{-0.5}{Rw})})$$

Ecuación 11

$$BQv = Qv * B$$

Donde B es la conductividad equivalente de intercambio de cationes de la arcilla (mS/m) y Rw es la resistividad de agua de formación (ohm-m) (Ec.10). BQv es la multiplicación entre la conductividad equivalente de cationes y el contador de la concentración de iones [83].

- Cálculo de conductividades.

Ecuación 12

$$E = \frac{\emptyset^m}{a}$$

Ecuación 13

$$Ct = \frac{1}{Rt}$$

Ecuación 14

$$Cw = \frac{1}{Rw}$$

Donde Ct es la conductividad de la matriz, Rt es la resistividad de formación (AT90)(ohm-m), Cw es la conductividad del agua de formación y E es el factor de formación pero inverso (adim.) [83].

- Cálculo del Sw por el método de Waxman-Smits.

Ecuación 15

$$Error = E * Cw * (Sw\_a1^{(n)}) + E * BQv * (Sw\_a1^{(n-1)}) - Ct$$

Ecuación 16

$$gp = n * E * Cw * (Sw\_a1^{(n-1)}) + (n-1) * E * BQv * (Sw\_a1^{(n-2)})$$

Ecuación 17

$$SwWs = \frac{Sw\_a - Error}{gp}$$

Donde SwWs es la saturación de agua con el método de Waxman-Smits (fracción), Sw_a1 es la saturación de agua método de Archie (fracción), gp es la derivada del Error y "n" es el exponente de la saturación (adim) [83].

2.4.8.c.Método de Simandoux. El modelo de Simandoux (1963), es útil en el cálculo de la saturación de agua para formaciones conductoras con presencia de arenas arcillosas [44]. En el manual de Crain [81] la ecuación del método Simandoux es la siguiente :

Ecuación 18

$$Sws = \left(\frac{a*Rw}{2*(\emptyset)^m} * \left(\frac{Vsh}{Rsh}\right)^2 + \frac{4*(\emptyset)^m}{a*Rw*Rt}\right)^{\left(\frac{1}{2}\right)} - \frac{Vsh}{Rsh}$$

Donde Sws es la saturación de agua del método Simandoux (fraccional), a es el factor de tortuosidad, m es el exponente de cementación, n es el exponente de saturación, $\emptyset$ es la porosidad (fracción), Rt es la Resistividad de formación (AT90)(ohm-m), Rsh es la resistividad del *shale* (ohm-m), Rw es la resistividad de agua de formación (ohm-m) y Vsh es el Volumen de *shale*.

En las Figuras 34 ,35 y 36 (sección 3.1.3) , se muestra el resultado de la saturación de agua calculada por los métodos convencionales para cada uno de los tres (3) pozos. En las figuras también se ilustra la saturación de agua del *set* de valores obtenidos de los corazones en los mismos *tracks.*

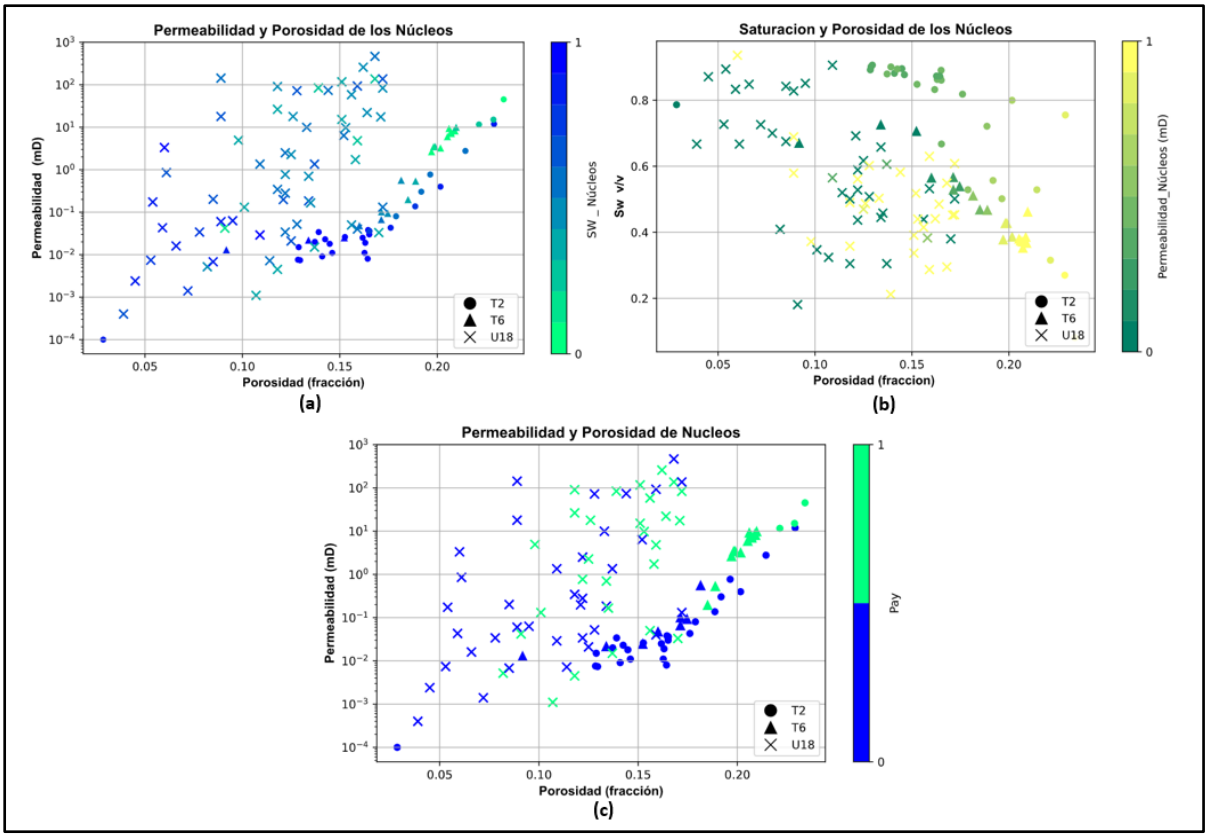
2.5. Cálculo del net Pay y corte del Sw.

El *Net Pay* es una porción del reservorio que contiene hidrocarburo recuperable (zona productora) y que se establece por criterios de corte o *cutoff* según las propiedades de permeabilidad, saturación y porosidad de la roca [45]. A partir de estas propiedades, se realizan los *crossplots* (gráficos de dispersión) como se presentan en la Figura 14. Lo anterior con el fin de identificar las regiones con mejor potencial de hidrocarburos que corresponderían a arenas de alta permeabilidad con baja saturación de agua, dado que existe una correlación directa entre ambas variables donde se aprecia que el *cutoff* de Sw se puede definir en un 50% asegurando permeabilidad suficiente para fluir.

El corte de agua aplicado, permite la clasificación de las zonas Pay de las no Pay justificando el valor del 50%. Los valores de saturación por encima del 0.5 representan la existencia de más agua que de hidrocarburos (PAY =0) mientras que los valores por debajo del *cutoff* indican que hay mayor presencia de hidrocarburos que de agua (PAY =1) (Figura 37).

Figura 14.

Gráficos de dispersión para valor del cutoff.



Nota. La figura representa los *crossplots* (gráficos de dispersión) en base a los datos recopilados del análisis de núcleos para determinar el cutoff. En la leyenda de cada crossplot se ve representado cada uno de los pozos con diferente forma así: círculos - pozo T2, triángulos - pozo T6 y X - pozo U18. **(a)** *crossplot* de porosidad (fracción) en función de la permeabilidad (md) en base a los valores de saturación de agua de los núcleos. **(b)** *crossplot* de porosidad (fracción) en función de los valores de saturación de agua en base a los valores de la permeabilidad (md). **(c)** *crossplot* de porosidad (fracción) en función de la permeabilidad (md) en base a los valores de Pay.

2.6. Procesamiento de imágenes

Para procesar una imagen, se requiere de un conjunto de técnicas que se aplican para lograr mejorar la calidad y facilitar la búsqueda de información. Las imágenes UV, son extraídas, colapsadas y apiladas para obtener una imagen final en escala de grises.

2.6.1. Imágenes de corazones bajo luz ultravioleta

Los núcleos proveen información esencial tanto para la exploración como para la producción de un yacimiento. Estos núcleos son analizados para determinar varias de las propiedades de la roca en laboratorio como saturación, porosidad, permeabilidad, etc. Los núcleos son

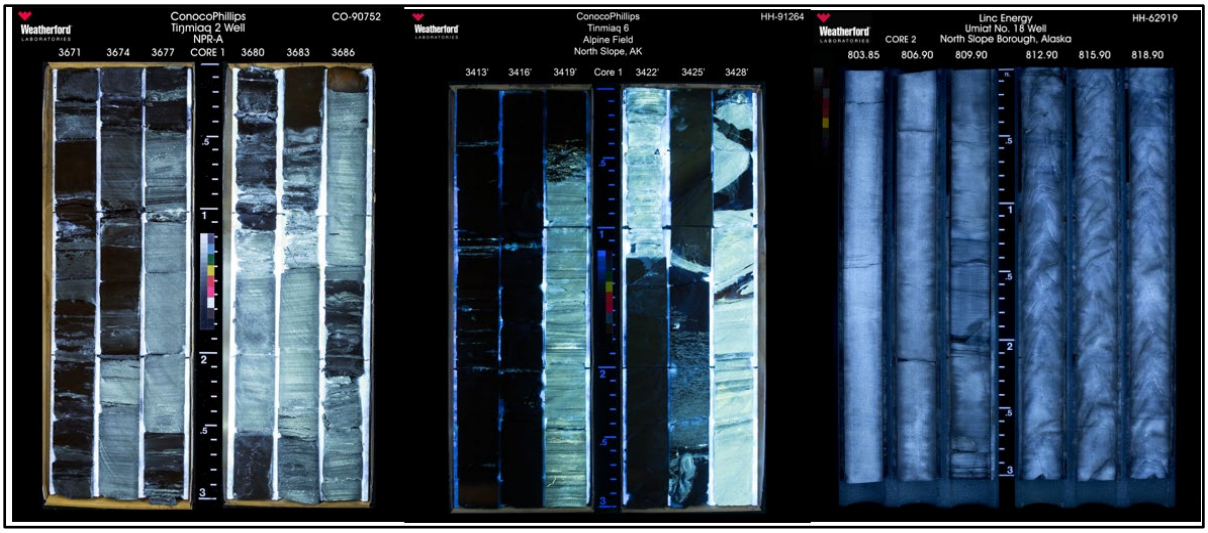
fotografiados con cámara y luego la foto es transmitida a un ordenador para digitalizarla. Las fotos de los núcleos son tomadas con luz blanca y luz ultravioleta [7].

La luz ultravioleta permite resaltar el contraste que hay entre las zonas con hidrocarburos y sin hidrocarburos. Algunos de los compuestos de los hidrocarburos, tienen la propiedad de emitir luz, la cual es conocida como "fluorescencia". En el momento de capturar las imágenes, la fluorescencia es fácil de identificar porque tiene un brillo y color característico. A los geólogos, les es posible filtrar o manipular los colores de la imagen para resaltar características importantes de la muestra [51].

Las fotos UV para cada uno de los pozos fueron encontradas en formato de cajas. Es decir, los núcleos se encuentran uno al lado del otro por profundidad. Cada una de las imágenes contienen seis (6) piezas de núcleos como se muestra en la siguiente figura:

Figura 15.

Imágenes capturadas bajo luz UV de núcleos en formato de cajas



Nota. La figura muestra tres fotografías digitalizadas de corazones tomadas bajo luz ultravioleta (representativas del conjunto de fotos UV) de los pozos T2, T6 y U18 en formato de cajas a diferentes profundidades.

El lenguaje de programación Python posee diferentes tipos de librerías para el procesamiento de imágenes. Esto permite que se pueda cargar la fotografía en formato digital para su análisis y manipulación. Para lograr el procesamiento completo de las imágenes para cada uno de los pozos, se hizo el siguiente paso a paso:

2.6.2. Importación de librerías

El primer paso es la importación de las librerías que se requieren. *Glob* para ordenar la ruta y un patrón especificado a uno arbitrario [53], *cv2* para cargar una imagen desde un archivo en

específico, *Lasio* para leer archivos .las que contienen los valores de los registros [34], *Numpy* para arrays [52], *pandas* para el almacenamiento de datos, *matplotlib* para poder visualizar los datos [54], *os* para crear y eliminar carpetas obteniendo su contenido [62] y *SciPy* para operaciones matemáticas (interpolación)[60].

Figura 16.

Previsualización de código - Importación de librerías.

```
import glob
import cv2
import Lasio
import os. path
import NumPy as np
import matplotlib. pyplot as plt
import pandas as pd
import os
from scipy import interpolate
```

Nota. La figura representa la previsualización de la sección de código de la importación de las librerías seleccionadas en Python para el procesamiento de las imágenes.

Una vez importadas las librerías, se toman las fotos UV de las carpetas por pozo y se ubican en una carpeta independiente en Python. Todas las fotos tomadas de los pozos deben tener el mismo formato .jpg para que no exista error en el momento de correr el algoritmo (Figura 25 (a)).

Al encontrarse las fotos UV en formato de caja, es necesario indicarle al algoritmo la cantidad de piezas de núcleos que contienen. Con el largo de cada pieza se puede dar una profundidad a cada imagen. En el algoritmo, la librería *Glob* permite buscar cada una de las fotos UV guardadas en las carpetas para identificar el formato y cargarlo al programa. Esto se muestra en el siguiente fragmento del código realizado:

Figura 17.

Previsualización de código - localización de imágenes en Python.

```
Well = 'T2'
cores_per_image = 6
uvFiles = glob. glob('./Photos/T2/*.jpg')
filedos = []

core_length = 3
```

Nota. La figura representa la previsualización de la sección de código inicial de la ubicación de las imágenes (glob.glob('./Photos/T2/*.jpg'), cantidad de núcleos por imagen (cores_per_image) , creación de matriz para almacenamiento de datos (filedos=[]) y longitud por pieza de núcleo (*core_length*) para cada pozo.

Luego de extraer la primera profundidad, se ordena de menor profundidad a mayor profundidad para que en el proceso queden las imágenes con un orden específico. Se crea una función con la librería *cv2* para mostrar siempre la primera foto del *set* que se tiene en cada carpeta generada y pueda implementarse sin importar la cantidad de fotos que se tengan (Figura 25 (b)).

2.6.3. Extracción de las piezas de núcleo por profundidad

Para extraer cada una de las piezas del núcleo de forma independiente, fue necesario generar dos bucles. Los bucles o *loops,* son funciones que ejecutan secuencialmente sentencias hasta que se cumpla una condición dada [36]. El primer *loop*, ejecuta la secuencia para ir de foto en foto y el segundo *loop* se ejecuta una vez esté dentro de cada imagen para hacer el recorte de la siguiente manera:

Cuando la primera imagen del set se muestra, el algoritmo indica que se deben hacer 3 puntos: en la primera pieza, se hace un punto en la esquina superior izquierda y en la esquina inferior derecha (indicando al código el alto y ancho de la imagen). Una vez se tienen esos dos primeros, se ubica el otro punto al inicio de la siguiente pieza (esquina superior izquierda). El algoritmo de forma interna realiza el recorte de las 6 piezas secuencialmente y ejecuta la extraccion de cada una de ellas (Figura 25 (c)).

Para obtener la información de cada pieza extraída, se toma solo la sección central de todo el ancho de la pieza, ya que cada una no tiene cortes simétricos de núcleo. El corte de cada pieza es guardada en la variable *crop_img* donde toma los puntos generados y los convierte en cantidad de pixeles que darán la forma a cada corte. Dentro de la variable *vc* se implementa la

librería *cv2* que permite concatenar cada pieza cortada, es decir, las apila una sobre la otra (Figura 18).

Figura 18.

Previsualización de código - recorte de piezas de núcleo en Python.

```
crop_img = img [y: y + dy, x:x + dx]

    if i == 0: # and k == 0:
        vc = crop_img
    else:
        vc = cv2.vconcat([vc, crop_img])
```

Nota. La figura representa la previsualización de la sección de código del recorte de la imagen en las 6 piezas mediante la selección de puntos en el eje x, eje y de cada tramo para almacenarlas en la variable *crop_img* . Luego de hacer el corte, cada tramo es concatenado (apilado) en secuencia vertical.

2.6.4. Color principal de la imagen y escala de grises

Las imágenes están compuestas por pixeles e información de color, saturación y brillo. Cada uno de los píxeles puede llegar a tener millones de combinaciones posibles de colores para definir una imagen; cuantos más píxeles haya, mayor será su resolución. Los colores son determinados a partir de tres (3) colores básicos: Rojo, Verde y Azul (RGB) y que al aumentar el espectro de luz se puede llegar a diferenciar muchos más tonos como pasa en un ordenador [63].

En la programación existe un código que está formado por un conjunto de números que van desde 0 (negro) hasta 255 (blanco). El código, está representado por los colores verde, rojo y azul de la siguiente forma: [R][G][B]. Dentro de cada corchete, va un número independiente que representa que tan iluminado está o no el color [63].

Cada una de las piezas cortadas de núcleo se convierte en un array (matriz) en 3 dimensiones con mallas de color RGB. Para hacer el procesamiento es necesario colapsar las tres mallas de color (RGB) de 3 dimensiones a una sola dimensión. La matriz colapsada, se convierte en una imagen a escala de grises (imagen plana) (Figura 25(d y e)). y para lograrlo se muestra un fragmento de código realizado en la Figura 19.

Figura 19.

Previsualización de código - Transformación del color de RGB a escala de grises.

```
vc_gray = cv2.cvtColor(vc, cv2.COLOR_BGR2GRAY)
      crop_name = str (int(fname [0:4]) + (core_length * i)) + ".jpg"
```

Nota. La figura representa la previsualización de la sección de código del cambio de escala de color de RGB a escala de grises (vc_gray = cv2.cvtColor(vc, cv2.COLOR_BGR2GRAY)). Se genera el nombre por pieza cortada, en formato .jpg según la profundidad y largo de la pieza (crop_name = str(int(fname[0:4]) + (core_length * i)) + ".jpg").

2.6.5. Apilado (stacking) de las imágenes por profundidad

En el algoritmo, cuando termina el proceso en la última pieza (extraccion por profundidad) sigue con el proceso externo (loop 1) para apilar (stack) cada una de las piezas extraídas. Se apila cada pieza, una sobre la otra a medida que se mueve el algoritmo hacia la derecha (Figura 25(f)).

Figura 20.

Previsualización de código - Apilado de las 6 piezas de núcleo cortadas.

```
if k == 0:
      ImgStack = vc_gray
  else:
      ImgStack = np. concatenate((ImgStack, vc_gray), axis =0)
```

Nota. La figura representa la previsualización de la sección de código de la asignación de las imágenes en escala de grises a la variable *Image Stack* (apilado de imagen). Luego se le pide que cada imagen se concatena en una sola columna en vertical (ImgStack = np.concatenate((ImgStack, vc_gray), axis =0)).

En el código mostrado, se le asigna a la variable *ImgStack* cada una de las imágenes completas (6 piezas cortadas) en escala de grises para luego concatenarlas con las restantes del *set* de fotos cargadas por pozo. Todas estas imágenes son agregadas a una carpeta, de donde se selecciona únicamente el tramo resultante de cada concatenación.

A medida que el código pasa por cada imagen, se guardan las profundidades en una misma variable para obtener el mínimo y el máximo valor de profundidad que se van a emplear para graficar.

2.6.6. Creación del registro de la imagen

Un *stack* completo por imagen tiene en el eje x una escala de 0 a 120, en nuestro caso se tomó de 20 a 100 para recortar los bordes de cada pieza como se muestra en la siguiente variable**:**

Figura 21.

Previsualización de código – Promedio de corte por pieza.

```
img_log = np. average(vc_gray[:, 20:100], axis=1)
  depths = np.arange(do, dn, (dn - do) / len(img_log))
```

Nota. La figura representa la previsualización de la sección de código del recorte de los extremos por pieza dejando un *stack* promedio de cada una (img_log = np.average(vc_gray[:, 20:100], axis=1). La variable *Depth*, *do* y *dn* son las profundidades por pieza y len(img_log) es la longitud del *stack* completo para obtener la misma cantidad de puntos en el eje y para poder graficar.

Como una imagen depende de la resolución y no tiene profundidades como un registro, se tuvo que crear un registro de profundidad para cada una utilizando las funciones representadas en la Figura 22. En esas variables se almacenan los datos de las profundidades (una encima de la otra), los valores de las imágenes en escala de grises a cada profundidad apiladas y el número de foto que se procesa.

Figura 22.

Previsualización de código - funciones de almacenamiento.

```
DEPTH.extend(depths. tolist())
GRAY.extend(img_log.tolist())
PHOTO.extend(photo_number.tolist())
```

Nota. La figura muestra la previsualización de la sección de código de las funciones de almacenamiento para crear el registro de profundidad que contiene todas las imágenes concatenadas. DEPTH.extend(depths.tolist()), almacena todas las profundidades, GRAY.extend(img_log.tolist()), almacena todas las imágenes concatenadas en escala de grises y PHOTO.extend(photo_number.tolist()), guarda el número de foto que se procesó.

2.6.7. Gráficas del registro y el procesamiento de las imágenes

Las imágenes al estar en escala de grises quedan en una sola dimensión (xy) lo que dificulta controlar que tanto de la imagen se va a graficar. Estas imágenes, no están directamente asociadas a las profundidades calculadas en el paso anterior sino que contienen miles de pixeles en el eje y .

Para asociar las profundidades con los píxeles, se utilizó la variable *istr* e *iend* (variables que permiten graficar toda la imagen completa) para normalizar el número de píxel respecto a la profundidad. Para esto se utilizaron las siguientes variables : *ImgStack. shape* (valor de la cantidad de píxeles contenidos) , *dplot_o* y *dplot_o* (profundidad inicial y final a la cual se quiere graficar respectivamente) , *doo = 0* y *dnn=1* como se muestra en el siguiente fragmento de código:

Figura 23.

Previsualización de código - Variables para graficar imagen procesada.

```
istr = int (ImgStack.shape[0]*(dplot_o - doo)/(dnn-doo))
iend = int(ImgStack.shape[0]*(dplot_n - doo)/(dnn-doo))
```

Nota. La figura muestra la previsualización de la sección de código de las variables que permiten graficar la imagen completa en un mismo *stack*. La variable *istr* calcula el inicio e *iend* calcula el final del tamaño que tendrá la imagen completa procesada.

2.6.8. Imágenes procesadas

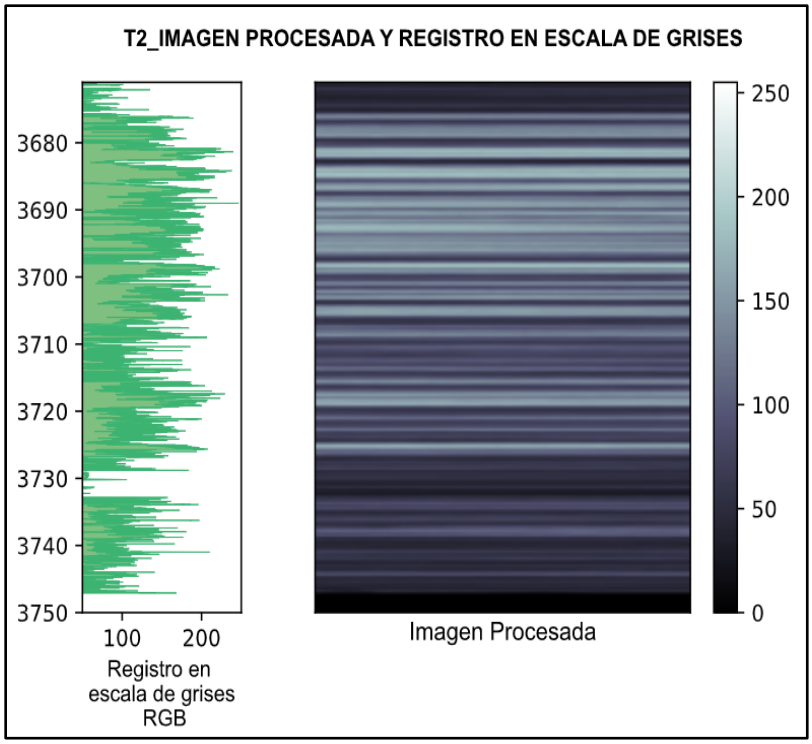
Para graficar el procesamiento completo de las imágenes, se toman las variables *doo* y *dnn* para el eje x mientras que para el eje y se toman las variables *dplot_o* y *dplot_n* para graficar como se muestra en la Figura 23.

2.6.8.a.Registro de imágenes en escala de grises en función de la profundidad. Para graficar el registro de las imágenes en escala de grises en función de la profundidad, se toman las variables *dplot_o*, *dplot_n* para el eje y (profundidades) y un rango de 50 a 250 para el eje x. Para poder graficar este registro, se toman las variables de almacenamiento *GRAY* y *DEPTH* como se ve representado en la Figura 22.

Tanto el procesamiento completo de la imagen en escala de grises como el registro de las imágenes en función de la profundidad fueron graficadas una al lado de la otra (Figura 25 (g y h)). Esto se hizo con el fin de mostrar que el registro muestra las zonas con y sin hidrocarburos en paralelo como se muestra en *Imagen procesada*: zonas con hidrocarburo (brillante) y sin hidrocarburo (oscuras) como se representa en la siguiente figura:

Figura 24.

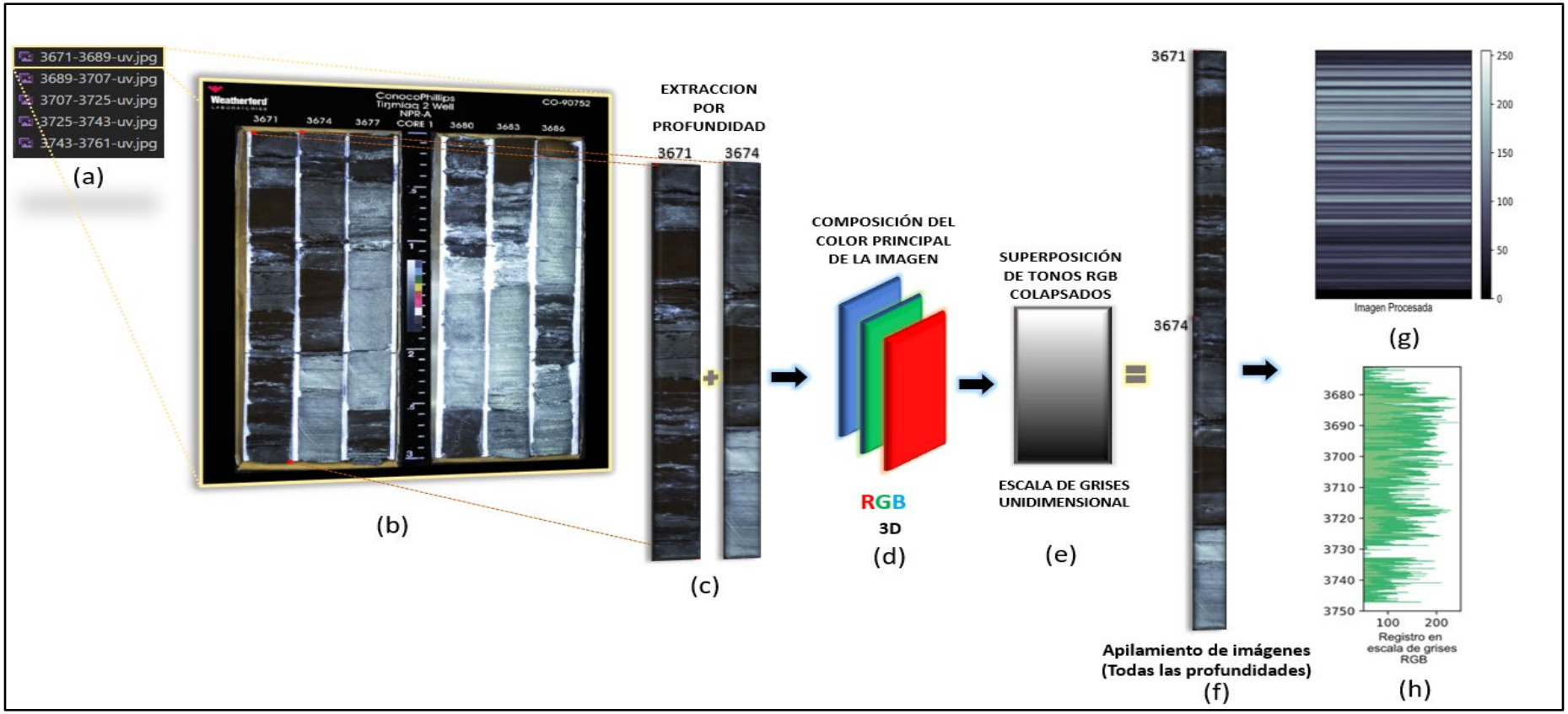
Imagen procesada y registro en escala de grises.



Nota. La figura muestra la representación gráfica del procesamiento de imágenes UV y el registro en escala de grises del pozo T2. A la derecha de la figura, está el procesamiento final del conjunto de imágenes UV en una escala de grises entre 0 y 255. A la izquierda de la figura, está el registro en escala de grises-RGB que toma en el eje x los valores del rango de las imágenes y en el eje y las profundidades para cada valor.

Figura 25.

Paso a paso del procesamiento de imágenes en Python.



Nota. La figura representa el paso a paso del procesamiento de imágenes en Python. (a) carga de fotos UV en formato de profundidad con extensión .jpg , (b) Imagen principal del set de fotos UV, (c) Extracción de las piezas de núcleo por profundidad , (d) Composición principal de la imagen en escala RGB en tres dimensiones, (e) representación visual en escala de grises luego de superponer y colapsar la escala RGB a una sola dimensión,(f) Apilamiento de las piezas luego de ser extraídas por el algoritmo, (g) resultado final del procesamiento de las imágenes en escala de grises y (h) Registro de las imágenes en función de la profundidad.

2.7. Machine Learning

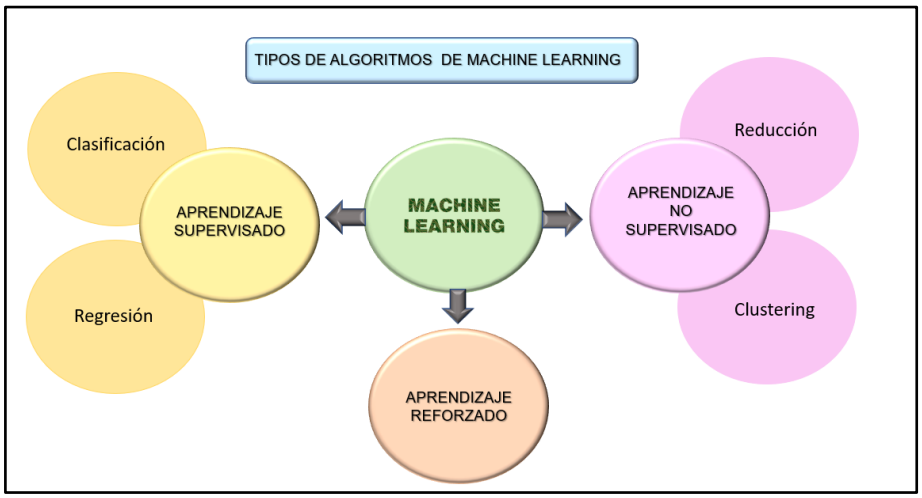
La técnica de aprendizaje automático es una parte fundamental del *Big Data* (datos a gran escala), ya que maneja un conjunto de métodos capaces de identificar patrones de datos masivos por medio de algoritmos para hacer predicciones [66]. Es una ciencia de inteligencia artificial capaz de generar sistemas de aprendizajes que mejoran de manera automática cualquier proceso que lo requiera.

2.7.1. Tipos de algoritmos de Machine Learning

Los algoritmos de Machine Learning se clasifican según el tipo de variables suministradas y el objetivo final de la implementación. Estos algoritmos se dividen en tres tipos: aprendizaje supervisado, no supervisado y reforzado; siendo las dos primeras las más comunes.

Figura 26.

Tipos de algoritmos de Machine Learning



Nota. La figura representa la clasificación de los algoritmos de Machine Learning en tres divisiones. La primera es el aprendizaje supervisado (color amarillo) que se subdivide en regresión y clasificación de los datos. La segunda división es el aprendizaje no supervisado (rosado) que se subdivide en reducción y *clustering* (agrupamiento). Por último, el aprendizaje reforzado (color naranja).

2.7.1.a.Aprendizaje supervisado. El aprendizaje supervisado encuentra patrones en los datos que se pueden aplicar a un proceso analítico. Estos tipos de algoritmos utilizan técnicas para aprender de forma automática de un modelo basado en un sistema de etiquetas, las cuales están asociadas a datos para poder tomar decisiones o hacer predicciones [38]. Este tipo de

aprendizaje es muy útil cuando la variable a predecir hace parte del conjunto de datos y se pueden implementar dentro de un modelo.

Se divide en dos tipos de aprendizaje: clasificación y regresión. Ambos tipos se diferencian por la variable objetivo, para clasificación su objetivo es una variable categórica mientras que el de regresión tiene como objetivo una variable numérica. Un ejemplo claro es el detector de spam en los correos electrónicos, el algoritmo de Machine Learning debe definir entre dos clases: spam o no-spam (categórico). Un ejemplo de variable numérica es el caso del valor en que se vende una propiedad o estimar cuánto inventario queda de una empresa, etc.

2.7.1.b.Aprendizaje no supervisado. Este tipo de aprendizaje dirige un proceso iterativo (de repetición) cuando no existen "datos etiquetados" para el entrenamiento. Los datos sin etiquetas tienen una estructura desconocida u oculta que son el principal objetivo de este tipo de aprendizaje. Cuando no conoce los datos, hace que exista un caos y se requiera de un algoritmo que permita que el modelo genere clasificaciones de un grupo de información [38].

El aprendizaje no supervisado se divide en dos tipos de problema que trata de resolver: agrupamiento (*clustering*) y reducción de dimensión. Los algoritmos tipo *clustering* encuentran una estructura en la que los datos de este *clúster* se diferencien de otro tipo de grupos *(clústeres)* [85]. Un ejemplo es cuando se tiene un conjunto de frutas (bananos y peras) y se quiere clasificar. Dentro del conjunto, solo hay dos tipos de frutas que son fáciles de identificar ya sea por su color o su forma. Los algoritmos de reducción de dimensión reducen el número de variables que se requieren para encontrar una información en específico [64].

2.7.1.c.Aprendizaje reforzado. Estos algoritmos encajan perfectamente en las características que no son predecibles por los anteriores tipos de aprendizaje. Este aprendizaje se basa en el proceso de retroalimentación para mejorar las respuestas del modelo implementado y que logra aprender en base a su propia experiencia. Es capaz de tomar la mejor decisión en diferentes situaciones, de acuerdo con un proceso de prueba y error en el que se recompensan las decisiones correctas [38]. Este tipo de modelo es muy útil para herramientas donde se requiera reconocimiento facial, también para hacer diagnósticos médicos, o robots para que aprendan a realizar una tarea en específico.

2.7.2. Lineamiento para la selección del modelo

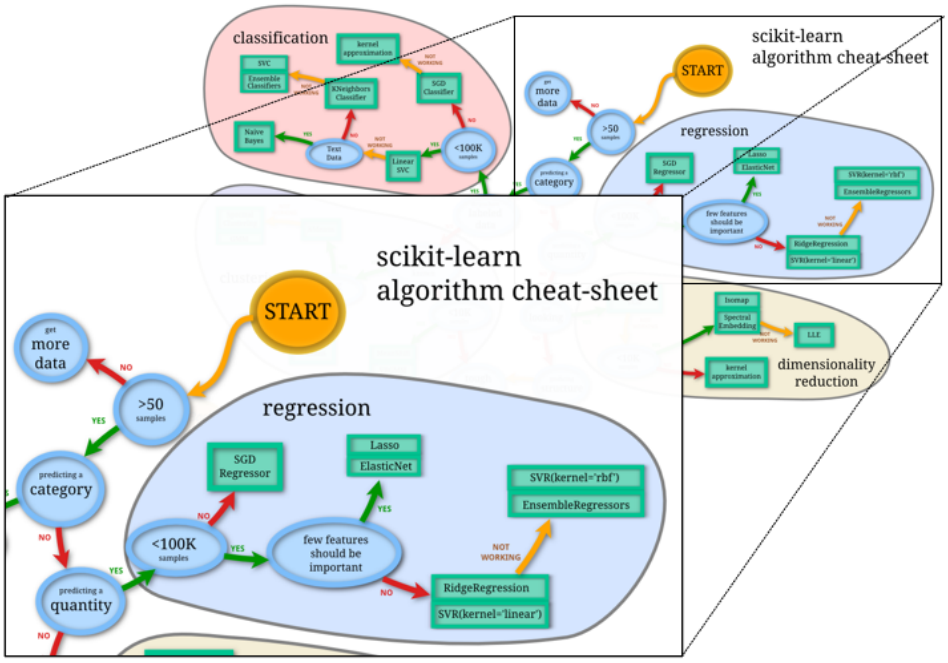
Una vez explicado los modelos de aprendizaje, el modelo que mejor se ajusta a nuestra base de datos y objetivos es el aprendizaje supervisado.

Inicialmente se buscaba predecir el clasificador binario (presencia (1) y ausencia (0) de hidrocarburo) derivado de fotos UV. Cuando se procesaron las imágenes, al tener cierta cantidad limitada de datos decidimos hacer las predicciones bajo un modelo de regresión en vez de un modelo de clasificación.

La librería Scikit-learn cuenta con un diagrama de flujo en base a los diferentes estimadores (regresión, clasificación, clustering y reducción de la dimensionalidad) que permite tomar el mejor camino para resolver cualquier problema en Machine Learning. Los estimadores son usados para diferentes tipos de datos y problemas [67], por ello en la Figura 27 se muestra la dirección que se tomó (en base a los datos de entrada que se tenían) para poder hacer las diferentes predicciones en los diferentes modelos.

Figura 27.

Diagrama de flujo de trabajo para Machine Learning.



Nota. La figura representa el diagrama de flujo diseñado por la librería Scikit learn que será usado como guía de trabajo para la selección de los métodos predictivos de Machine Learning. A la izquierda de la imagen se realizó el zoom del diagrama que se llevará a cabo para el proceso como se explica en la sección 2.7.3. Tomado de: Scikit learn [En línea] Disponible en: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html. [Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.].

2.7.3. Descripción del flujo de trabajo - regresión

Dentro del flujo de trabajo seleccionado (regresión) existen diferentes modelos que se pueden implementar en el aprendizaje automatizado. Para poder evaluar los diferentes modelos se identificó que:

- La cantidad de muestras (datos) disponibles es mayor a 50 (> 50 *samples*) para la predicción.
- Como los valores son limitados, no es una predicción por categoría (*predicting category*), sino una predicción por cantidad (*predicting a quintity*).
- En nuestro caso, como la base de datos tiene una cantidad menor de 100.000 muestras (<100k *samples*), no es posible aplicar la regresión por descenso de gradiente estocástico (SGD por sus siglas en inglés).
- Como el nivel de importancia que tienen las entradas es el mismo, los modelos que se implementan son Lasso y ElasticNet.

2.8. Tipos de modelos de Machine Learning seleccionados

Los modelos de aprendizaje de regresión utilizados en este proyecto se clasifican en dos grupos teniendo en cuenta su complejidad matemática y costo computacional [46]. El primer grupo, son los modelos lineales o básicos (Lasso, ElasticNet y Ridge Regression) y en el segundo los modelos complejos (Random Forest, Support Vector Regression (SVR), GradientBoostingRegressor, MLPRegressor y Neural Network) que se explican a continuación:

2.8.1. Modelos Lineales (Linear models) o básicos

Los modelos lineales de regresión ayudan a comprender y predecir el comportamiento de sistemas complejos o analizar datos experimentales. Las estrategias de regularización empleadas por los modelos como: Ridge Regression, Lasso o ElasticNet, es la implementación de una ecuación lineal a la hora de hacer el entrenamiento [46].

2.8.1.a.Lasso. Es un modelo lineal que estima coeficientes escasos y es útil en ciertas situaciones ya que tiende a elegir soluciones con menos coeficientes distintos de cero. Al ocurrir esto, se reduce efectivamente el número de características de las que depende una solución dada [47].

Lasso, es la base del campo de la detección comprimida ya que no solo ayuda a reducir el ajuste excesivo sino que también puede ayudar en la selección de funciones. En determinadas condiciones, puede recuperar un conjunto preciso de coeficientes distintos de cero[43][46]. La función de costo para la regresión de Lasso es:

Ecuación 19

$$\sum_{i=1}^{n}(yi - \beta_0 - \sum_{j=1}^{p}\beta jXij)^2 + \lambda\sum_{j=1}^{p}|\beta j| = suma\ residuos\ cuadrados + \lambda\sum_{j=1}^{p}|\beta j|$$

Donde β_0 es la ordenada en el origen, y corresponde al valor promedio de la variable de respuesta cuando todos los predictores son cero. βj es el coeficiente parcial de regresión, *xj* es la variable predictora, α es una constante y λ es el parámetro de regularización.

En la regresión Lasso, lambda (λ) es el mismo parámetro *alpha* usado en la programación y que se implementa por conveniencia para poderlo controlar.

2.8.1.b.ElasticNet. Es un modelo de regresión lineal, de regularización donde combina las penalizaciones *L1*y *L2* (suma absoluta de los coeficientes y suma de coeficientes al cuadrado respectivamente) de los modelos de regresion Lasso y Ridge Regression. ElasticNet, combina ambas penalizaciones convexas para dar lugar a buenos resultados [43] y poder aprender de un modelo disperso (valores distintos de cero o nulos) comprendido en los intervalos de [0,1]. *Alpha* es una constante que multiplica los términos de penalización , cuando es igual a 0 se aplica el modelo Ridge Regression y es igual a 1 se aplica el modelo Lasso. ElasticNet es útil cuando hay varias características que están correlacionadas entre sí [49].

Ecuación 20

$$\frac{\sum_{i=1}^{n}(yi - \beta_0 - \sum_{j=1}^{P}\beta jXij)^2}{2n} + \lambda(\alpha\sum_{j=1}^{P}|\beta j| + \frac{1-\alpha}{2} + \sum_{j=1}^{P}\beta j^2)$$

Donde β_0 es la ordenada en el origen, y corresponde al valor promedio de la variable de respuesta cuando todos los predictores son cero. βj es el coeficiente parcial de regresión, *xj* es la variable predictora, α es una constante y λ es el parámetro de regularización.

Los parámetros principales utilizados en este proyecto para el modelo ElasticNet fueron: *L1_ratio* y *alpha*(λ). El parámetro *L1_ratio* corresponde a *alpha* (λ) en la librería *glmnet* (trabajar con modelos ridge, Lasso y ElasticNet). El parámetro *L1_ratio* con valores menores o iguales a 0.01, proporcionaría un resultado engañoso sino implementa su propia

secuencia de *alpha*. Si *alpha* es igual a 0 , corresponde a un mínimo cuadrado ordinario y por ende no se recomienda usar valores de de *alpha* con este valor [49].

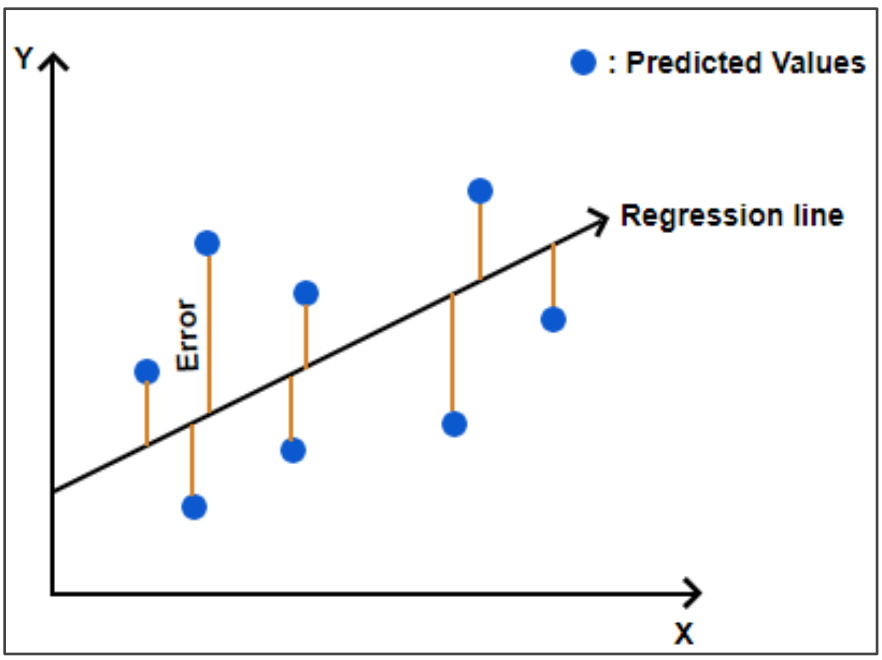
2.8.1.c.Ridge Regression. Es un modelo de regresión, que analiza datos de de regresion multilineal donde la función de pérdida es la función lineal de mínimos cuadrados y la regularización viene dada por la norma *L2*. La principal ventaja del modelo, depende de un valor adecuado de λ para permitir la reducción de varianza y conseguir un valor mínimo de error total [43][46].

Cuando el parámetro λ es mayor a 0, controla la cantidad de penalización mientras que cuando es igual a 0 la penalización es nula y el resultado es equivalente al de un modelo lineal. A medida que λ aumenta, mayor es la penalización y menor es el valor de los predictores haciendo que los coeficientes sean más resistentes a la colinealidad.

El parámetro principal utilizado en el modelo Ridge Regression, fue *alpha (*fuerza de regularización) para mejorar el condicionamiento del problema y reducir la varianza de las estimaciones.

Figura 28.

Ridge Regression.



Nota. La figura representa el modelo de Ridge Regression con el objetivo de encontrar la línea de mejor ajuste (línea de regresión) y minimizar la diferencia entre los valores reales y los valores predichos**.** Tomado de: Robofied [En línea]. Disponible: https://blog.robofied.com/ridge-regression/

2.8.2. Modelos de regresión complejos

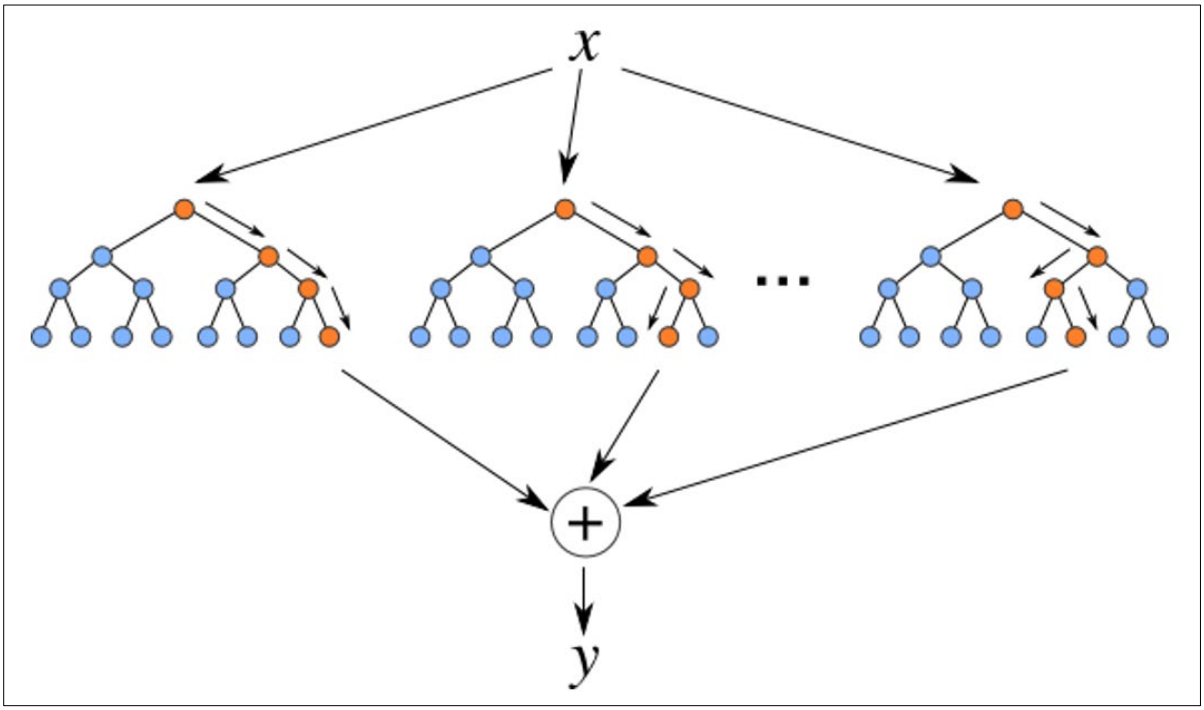
Existen diferentes modelos de aprendizaje supervisado, que son más robustos y pueden aplicarse para lograr mejores predicciones con pocos datos. Los siguientes modelos, fueron los utilizados en este proyecto:

2.8.2.a.Random Forest. El modelo de aprendizaje supervizado de bosques aleatorios (Random Forest), puede ser implementado tanto para regresión como para clasificación [41]. Esta técnica de aprendizaje automático, ensambla y promedia varios árboles de decisión obteniendo por cada uno un puntaje o voto a la salida de este. El árbol más votado dará como resultado la mejor opción para el modelo [55].

Una de las grandes ventajas de este tipo de algoritmo es que funciona de manera eficiente en grandes volúmenes de datos y aporta estimaciones de qué variables son relevantes en la clasificación. Por otra parte, es un modelo eficaz para la estimación de los datos faltantes y mantiene la precisión cuando una gran parte de los datos faltan. Una de las desventajas del algoritmo, es que un solo árbol de decisión tiende a variar mucho y hace que exista un sobre ajuste por la aleatoriedad inyectada en los bosques produciendo árboles de decisión con errores de predicción. La variable más importante para el modelo es el número de árboles seleccionados (*n_estimators*), que indica la cantidad de árboles en el bosque [48]. Cuantos más árboles haya, mejor será el rendimiento del bosque aleatorio.

Figura 29.

Proceso interno del modelo Random Forest.



Nota. La figura muestra la estructura del modelo de Random Forest durante su periodo de ejecución. La X en la figura representa el conjunto de datos de entrada para que sean procesados. El símbolo (+) representado en la figura, es la acumulación de todas las predicciones ejecutadas. Mientras que (Y) es la predicción final del modelo. Tomado de: Gitconnected [En línea]. Disponible: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

2.8.2.b.Support Vector Regression (SVR). La regresión de vectores de soporte (SVR por sus siglas en inglés), contiene herramientas capaces de resolver problemas de regresión de manera análoga. El modelo depende solamente de un subconjunto de datos de entrenamiento ya que la función de costo ignora las muestras cercanas a su objetivo. Los requisitos de computación y almacenamiento aumentan rápidamente con la cantidad de vectores de entrenamiento.

La líbreria *libsvm* , permite la implementación de los parámetros *C*, *épsilon* y *kernel* que fueron utilizados en nuestro código de programación. C es el parámetro de regularización donde la fuerza de la regularización es inversamente proporcional a C. El parámetro *épsilon* no se asocia con ninguna penalización en la función de pérdida de entrenamiento con puntos predichos dentro de una distancia del valor real. Por último el parámetro *kernel* especifica el tipo de *kernel* que se utilizará en el algoritmo [56].

2.8.2.c.Gradient Boosting Regressor. Este meta-algoritmo de aprendizaje supervisado, se usa para la predicción de las características de salida mediante la potenciación de los árboles de decisión que están óptimamente construidos. El modelo, se utiliza tanto para problemas de regresión como de clasificación (GradientBoostingClassifier y GradientBoostingRegressor) y es usado de forma simultánea para reducir el error y la varianza del conjunto de árboles de decisión construidos secuencialmente.

Gradient Boosting, construye un modelo aditivo de manera progresiva por etapas que pueden ser preferiblemente usados para tamaños de muestras pequeñas. La agrupación de datos puede generar puntos de división que son demasiado aproximados en esta configuración. Además, tiene un soporte incorporado para valores perdidos, es decir, en cada etapa los árboles de regresión se ajustan al gradiente negativo de la función de pérdida de desviación binomial o multinomial.

Los parámetros utilizados para este modelo fueron: la tasa de aprendizaje (*learning_rate*) que permite reducir la contribución de cada árbol mientras que *n_estimator* es el número de etapas de impulso a realizar y *alpha* corresponde al parámetro de regulación [57].

2.8.2.d.MLPRegressor (MLP).Es un modelo de optimización de funciones, con un gran número de parámetros que optimiza la pérdida cuadrática. Este modelo hace el entrenamiento, usando la propagación hacia atrás sin función de activación en la capa de salida (conjunto de valores continuos). MLP puede manejar multiples salidas de regresión donde cada una de ellas puede tener más de un objetivo.

Los parámetros usados que fueron implementados son: tasa de aprendizaje que controla el tamaño del paso para la actualización de peso (*learning_rate_init)* , el parámetro de regularización-*L2* (*alpha)* y el parámetro del número de neuronas en la capa oculta (*hidden_layer_sizes*) [58].

2.8.2.e.Neural Network. El modelo computacional de redes neuronales (Neural Network), se basa en un gran conjunto de unidades neuronales simples (neuronas artificiales) cuya estructura de capas se asemeja a la estructura interconectada de las neuronas en el cerebro (capas de nodos conectados). La información de entrada atraviesa la red neuronal donde se somete a diversas operaciones produciendo diferentes valores de salida.

Su comportamiento está definido por la forma en que se conectan sus elementos individuales, esto es posible porque cada neurona se conecta a otras por medio de enlaces por el grado de importancia (o ponderación) de dichas conexiones. En estos enlaces, el valor de

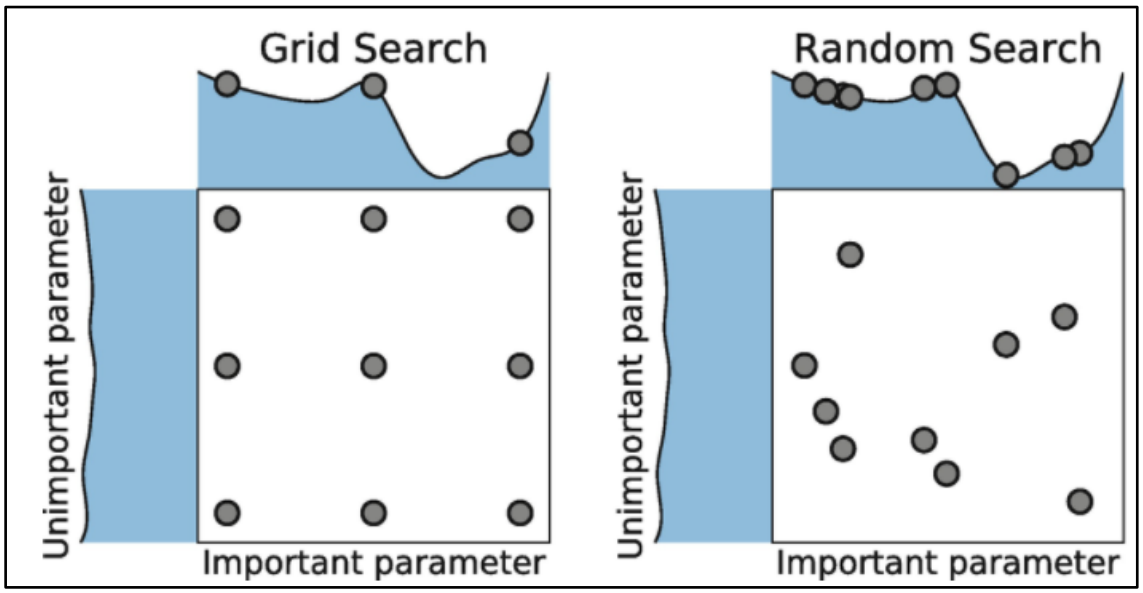
salida de la neurona anterior es multiplicado por un valor de peso que puede incrementar o inhibir el estado de activación de las neuronas adyacentes. Las ponderaciones son ajustadas automáticamente durante el entrenamiento hasta que la red neuronal lleva a cabo la tarea deseada correctamente [59].

2.9. Descripción general de los parámetros implementados

Cada modelo anteriormente descrito maneja una arquitectura fija de parámetros. Los métodos son capaces de optimizar ciertos coeficientes, pero una inadecuada selección de estos genera que el modelo no se lleve a cabo de forma efectiva. Para evitar esto, se utilizó una técnica de ajuste llamada búsqueda de cuadrícula (Grid search) que calcula los valores óptimos de los hiper parámetros y que facilita la búsqueda de los factores específicos de los modelos [82].

Figura 30.

Grid Search.



Nota. La figura muestra representa la búsqueda aleatoria de optimización de hiper parámetros, comparando el método de Grid Search vs Random Search donde las distribuciones que se muestran en cada eje representan la puntuación del modelo. Tomado de: Elyse Lee [En línea]. Disponible en: https://medium.com/@cjl2fv/an-intro-to-hyper-parameter-optimization-using-grid-search-and-random-search-d73b9834ca0a

Los resultados que se evidencian en la sección 3.5.4. han sido basados en el desarrollo de la arquitectura de los modelos y los parámetros elegidos para ser variados a la hora de entrenar. Los parámetros elegidos y sus rangos son los siguientes:

Tabla 6.

Parámetros modificados por modelo predictivo.

Modelo predictivo	Parámetro que sintonizar	Rangos
Lasso	alpha	(0.01,0.99, 10)
ElasticNet	alpha	(0.01,0.99, 10)
	l1_ratio	(0.01,0.99, 10)
Ridge Regressor	alpha	(0.01,0.99, 10)
Support Vector Regression(SVR)	C	(0.01,3, 10)
	épsilon	(0.01,0.3, 4)
	kernel	['linear', 'poly', 'rbf', 'sigmoid']
RandomForest	n_estimators	[10, 50, 100, 500, 1000]
GradientBoostingRegressor	alpha	(0.01,0.99, 10)
	learning_rate	(0.01,3, 10)
	n_estimators	[10, 50, 100, 500, 1000]
MLPRegressor	learning_rate_init	[0.0001, 0.001, 0.1]
	alpha	(0.01,0.99, 10)
	hidden_layer_sizes	(10,50,100)

Nota. Esta tabla muestra los diferentes parámetros que fueron modificados con su respectivo rango para cada uno de los siete modelos predictivos evaluados.

2.10. División del set de entrenamiento y prueba (Train & Test Splitting)

Para el entrenamiento y validación, los modelos predictivos segmentan en dos partes las variables. La función *Train_test_split* (función en la selección de modelos de Sklearn), permite seleccionar un porcentaje de datos para entrenamiento (*x_train, y_train*) y el porcentaje restante para probar el modelo (*x_test, y_test*).

El porcentaje de entrenamiento y prueba debe ser ajustado según el tamaño del conjunto de datos y la complejidad de los parámetros [28]. Inicialmente, se pensó tomar los datos de un

pozo modelo para entrenamiento pero se decidió tomar el conjunto completo de datos de los 3 pozos (T2, T6, U18). El conjunto completo de datos es muy limitado por lo que se decidió ser muy conservadores y entrenar solo el 30% de los datos como se muestra en Figura 31[84].

Figura 31.

Previsualización de código - división de datos para entrenamiento y prueba.

```
train, test = train_test_split (data, test_size=0.7, random_state=42)
```

Nota. La figura representa la previsualización de la sección de código de la división del conjunto de datos para entrenamiento y prueba. La variable *train,test*, almacena la división de los datos tomando un tamaño de prueba del 70% y una ejecución (*random_state*) de 42 veces que no afecta el resultado del conjunto de datos de entrenamiento y prueba.

Para la figura anterior, el parámetro *random_state* fija una semilla para el generador de números aleatorios. El generador es útil porque le permite reproducir la aleatoriedad para sus propósitos de desarrollo y prueba. El parámetro *test_size*, establece el tamaño del conjunto de datos de prueba que en este caso se tomó como el 70% de los datos totales para que el modelo sea válido.

Se tomó el tamaño de prueba y entrenamiento de 70/30 respectivamente, para disminuir la probabilidad de caer en el efecto de sobreajuste (*overfitting*) de los datos [29]. Este efecto genera que el algoritmo solo reconozca los datos de entrenamiento y no los nuevos datos de entrada, afectando la eficiencia de la predicción.

2.11. Análisis descriptivo

Un componente importante de hacer modelamiento basado en datos es entender las variables que hacen parte del modelo. La inspección de valores como el promedio, los límites de las variables y su distribución nos ayudan a entender que tanto procesamiento hay que hacer de los datos antes de alimentarlos a un modelo de Machine Learning. En esta sección presentamos las estadísticas de datos de la media, valores máximos, mínimos y la cantidad de datos utilizados en cada una de las variables independientes para los 3 pozos. Esto se hace con el fin de comprender las relaciones entre las variables, identificar valores atípicos, problemas con los datos y las relaciones aparentes entre los *features* y el *target*.

2.11.1. Datos de media

La media aritmética o promedio, indica el valor central del total de los datos para cada una de las variables mostrando una distribucion de los datos en partes iguales. La media es utilizada para distribuciones normales de números, con una cantidad baja de valores atípicos [72]. La tabla 7 presenta los valores promedio de las variables independientes y dependientes (GRAY) de nuestro problema.

Tabla 7.

Datos de media de las variables por pozo.

Well	DEPT	GR	RHOB	logRT (AT90)	DTCO	NPHI	GRAY
T2	3717.00	76.09420	2.429600	1.060860	92.14902	0.3234	139.858875
T6	3431.50	84.12250	2.397800	0.941019	97.33488	0.3437	17.590574
U18	848.25	45.35793	2.448586	1.969598	84.00787	0.2332	33.389167

Nota. Esta tabla muestra los datos de media para cada uno de los pozos en función de las variables relevantes como los cinco registros base, registro GS y Pay.

2.11.2. Datos totales registrados

Los datos totales, son la cantidad de datos obtenidos para cada uno de los pozos. Las cantidades presentadas en la tabla 8, muestran un resumen del número de datos trabajados para el entrenamiento y prueba de los modelos.

Tabla 8.

Datos totales registrados.

Well	DEPT	GR	RHOB	logRT (AT90)	DTCO	NPHI	GRAY
T2	341	341	341	341	341	341	341
T6	417	417	417	417	417	417	417
U18	600	600	600	600	600	600	600

Nota. Esta tabla muestra los valores de la cantidad de muestras totales registradas en cada variable por pozo.

2.11.3. Datos de mínimo y máximo

Saber el tamaño de la muestra es importante para aumentar la rapidez del estudio. Una de las maneras más sencillas de evaluar la dispersión de los datos que consiste en comparar el mínimo y el máximo.

El mínimo es el valor más pequeño de la muestra o del conjunto de datos y se utiliza para identificar un valor anormal o un error de entrada. La Tabla 9 muestra los valores mínimos registrados en cada una de las variables para cada pozo [73].

Tabla 9.

Valores mínimos registrados por pozo.

WELL	DEPT	GR	RHOB	logRT (AT90)	DTCO	NPHI	GRAY	PAY
T2	3672.0	57.41320	2.348700	0.784567	67.45719	0.2089	0.0	0
T6	3378.0	57.53930	2.193000	0.616108	67.50480	0.1729	0.0	0
U18	773.0	23.98474	2.310451	1.438275	72.97357	0.1570	0.0	0

Nota. La tabla muestra los valores de los datos mínimos registrados para cada variable relevante por pozo.

El número de datos máximo es el valor de datos más grande que se registró en cada una de las variables características (*features*) por cada pozo como se muestra en la Tabla 10. Estos valores, permiten identificar un posible valor anormal o un error de entrada de datos [73].

Tabla 10.

Valores máximos registrados por pozo.

Well	DEPT	GR	RHOB	logRT (AT90)	DTCO	NPHI	GRAY	PAY
T2	3759.5	100.13100	2.629600	1.408438	100.85811	0.3749	255.0	1
T6	3483.5	108.09000	2.618000	1.556353	117.59015	0.4704	255.0	1
U18	923.0	113.31291	2.642403	2.476211	93.68692	0.4132	255.0	1

Nota. La tabla muestra los valores de los datos máximos registrados para cada variable relevante por pozo.

2.12. Modelamiento

Los modelos de Machine Learning proporcionan la salida de la información que es generada una vez que se entrene un algoritmo con diferentes cantidades de datos. Luego de hacer el

entrenamiento con un conjunto de datos de entrada, el modelo dará como resultado uno de salida . Si es un algoritmo predictivo, creará un modelo predictivo.

2.12.1. Selección de Modelos

Una vez se conoce el concepto de modelamiento, se debe buscar que modelos de predicción por regresión son los que mejor se ajustan a los datos disponibles. Para esto, se hace un análisis basado en las métricas de regresión que darán como resultado los porcentajes de error y distribución de los datos para cada uno de los modelos.

2.12.1.a.Métricas de regresión. En la evaluación y selección del mejor modelo, se tuvo en cuenta la métrica de regresión disponible en la librería Sklearn: error cuadrático medio (*mean squared error -MSE*) y el Error de la media cuadrática (*Root Mean Square Error - RMSE*) para medir el rendimiento de la regresión.

El error cuadrático medio (*mean_squared_error*) mide el riesgo del valor esperado de error o pérdida cuadrática. La ecuacion utilizada para el calculo de error se representa de la siguiente forma:

Ecuación 21

$$MSE(y, \widehat{y}) = \frac{1}{nsamples} \sum_{i=0}^{nsamples-1} (yi - \widehat{yi})^2$$

Donde $\widehat{yi}$ es el valor predicho de la muestra, yi es el valor verdadero, *MSE* es el error cuadrático medio estimado sobre n muestras (*n_samples*).

El error de la media cuadrática (*Root_Mean_Square_Error)* es la raíz cuadrada de la media del cuadrado de todos los errores [27]. Es una buena medida de precisión, que permite comparar errores de predicción de diferentes modelos para una variable en particular que se puede calcular por la siguiente ecuación:

Ecuación 22

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Si - Oi)^2}$$

Donde Oi son las observaciones, Si son los valores predichos de de una variable, y n el número de observaciones disponibles para el análisis.

3. ANÁLISIS Y RESULTADOS

Este capítulo tiene como objetivo presentar los resultados de los datos obtenidos e interpretados de los métodos convencionales, procesamiento de imágenes y la realización del Machine Learning con el mejor modelo de predicción. En este proceso, se logra contrarrestar los resultados obtenidos con la predicción en Machine Learning con los resultados por métodos convencionales evidenciando la existencia de mayores zonas con hidrocarburos.

3.1. Resultados de la interpretación petrofísica convencional

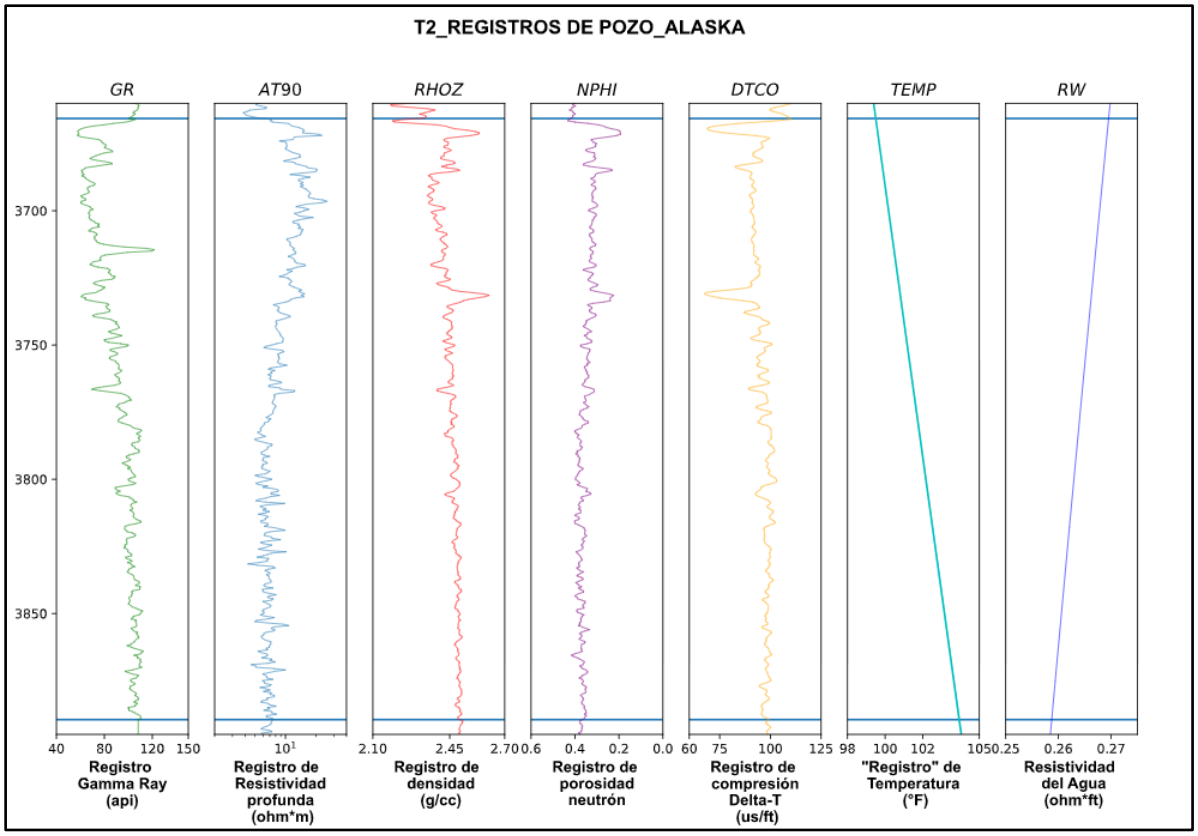
En esta sección se muestran los resultados obtenidos en Python de los registros de pozo (Figura 32), datos convencionales y de corazones (Figura 33) y los valores obtenidos de la saturación de agua calculada por métodos convencionales para cada uno de los pozos (Figuras 34, 35 y 36).

3.1.1. Registros basicos de pozo

Los diferentes *tracks* (conjunto de datos ordenados en función de dos variables) que se muestran a continuación, recopilan la información obtenida de los registros de pozo, los cálculos del "registro de temperatura" y del "registro de resistividad del agua" en función de la profundidad en la zona de interés del pozo T2.

Figura 32.

Tracks de los registros de pozo en función de la profundidad y zona de interés.



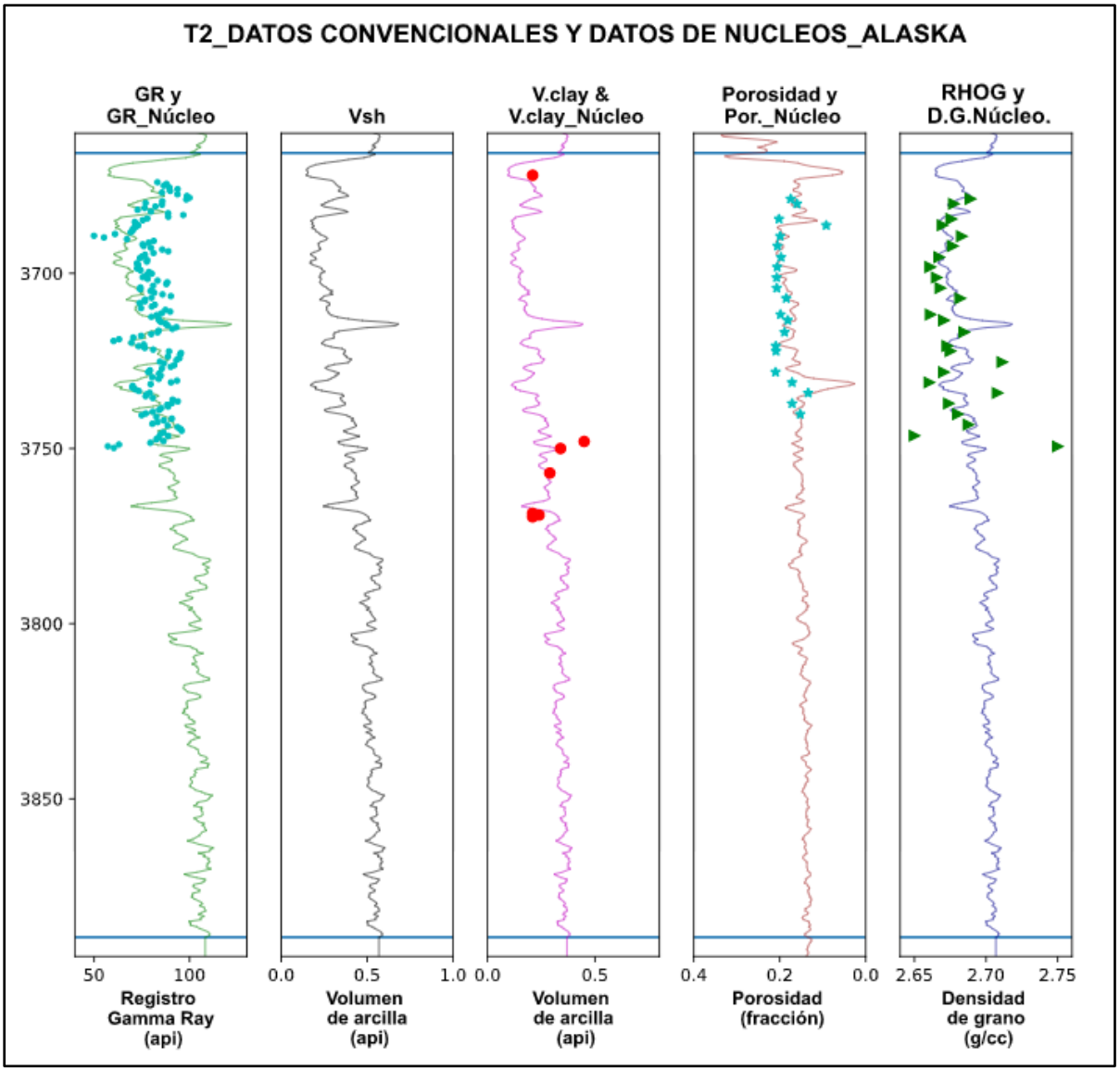
Nota. La figura muestra siete *tracks* del pozo representativo T2. De izquierda a derecha, está el registro GR (línea verde), registro AT90 (línea azul), registro RHOZ (línea roja), registro NPHI (línea morada), registro DTCO (línea amarilla). Seguidos se encuentran el "registro de temperatura" (línea cyan) y por último los valores calculados a modo de registro de la resistividad del agua (línea azul oscura) (sección 2.4.3).

3.1.2. Datos convencionales y de núcleos.

Luego de haber realizado los cálculos correspondientes de Vsh, Vclay y RHOG , se graficaron en *tracks* las curvas de los resultados obtenidos (datos continuos). En los mismos *tracks*, se muestra el registro de gamma ray y los valores obtenidos de núcleos que estan representados como datos discretos (puntos) en un color diferente para ver las variaciones de lo calculado respecto a lo real. Ademas, se realizó un desplazamiento y ajuste a los datos discretos de los núcleos a la misma profundidad de los datos continuos para los tres pozos.

Figura 33.

Tracks de los datos convencionales y de núcleos- pozo representativo.



Nota. La figura muestra los datos gráficados de: 1. registro gamma ray (línea verde) y los valores calculados de Vsh (línea gris), *Vclay* (línea fucsia), porosidad (línea café) y densidad de grano (línea azul oscura) por ecuaciones convencionales como datos continuos. 2. En los mismos *tracks* de los datos continuos se encuentran los datos discretos de núcleo graficados: GR (puntos cian), *Vclay* (puntos rojos), porosidad (estrellas cian) y densidad de grano (triángulos verdes) en función de la profundidad (ft) para pozo representativo T2.

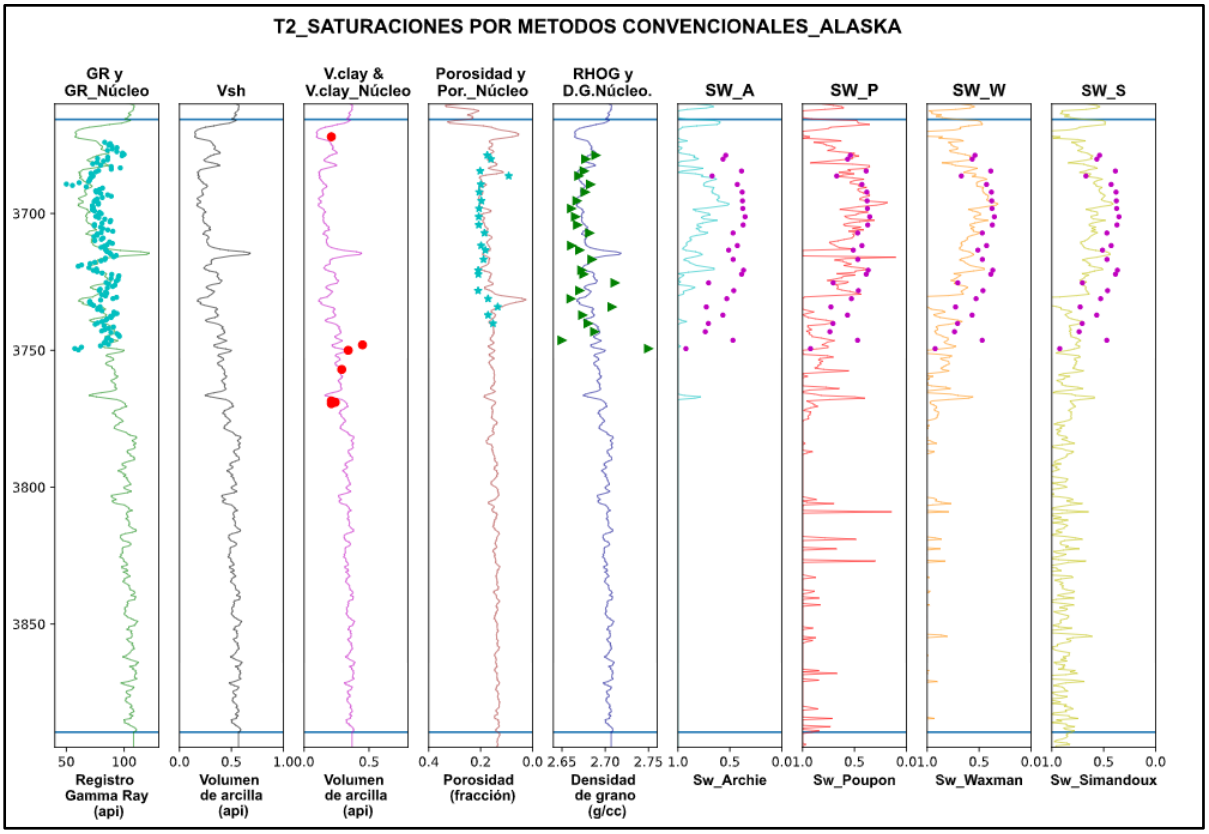
3.1.3. Datos convencionales, de núcleos y de saturación de agua por pozo

Una vez obtenidos los valores del cálculo de saturación de agua por los métodos de Archie, Poupon, Waxman-Smits y Simandoux, se procede a graficar en Python cada uno de los *tracks* en conjunto con los datos obtenidos del paso anterior. Ademas, se representan en datos

discretos los valores de la saturación de agua (SW_O) de los núcleos en los mismos *tracks* de la saturación calculada en la zona de interés de cada uno de los pozos.

Figura 34.

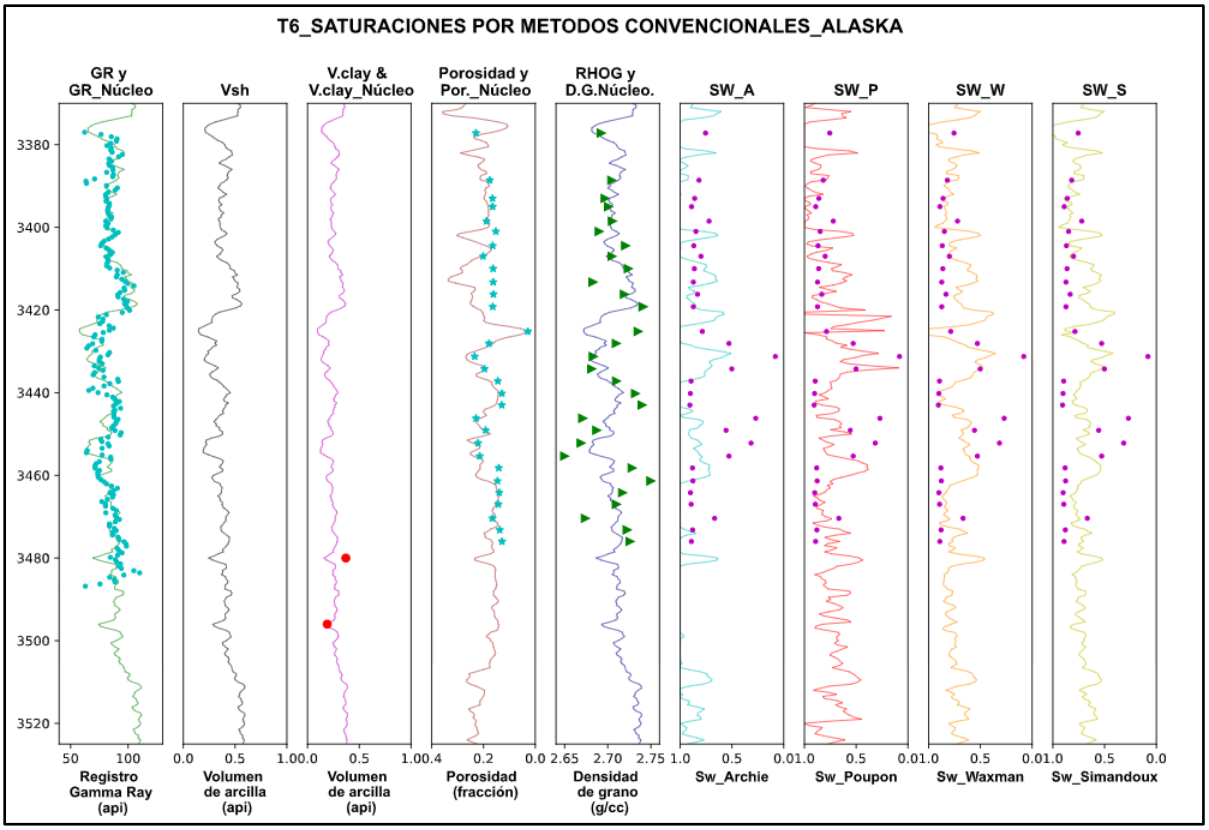
Tracks de los datos convencionales, de núcleos y de saturación- pozo T2.



Nota. La figura muestra nueve *tracks* de los valores calculados, de núcleo reportados y de saturación de agua con métodos convencionales para el pozo T2 en la zona de interés. De izquierda a derecha, se observan los cinco primeros *tracks* gráficados como se muestra en la Figura 33, los cuatro siguientes *tracks* muestran los datos continuos de: Sw por método Archie (línea cian), Sw por método Poupon (línea roja), Sw por método de Waxman-Smits (línea naranja) y Sw por método de Simandoux (línea amarilla). En los últimos 4 *tracks*, se gráfica la saturación de agua ajustada (SW_O) tomada de núcleos como datos discretos (puntos morados).

Figura 35.

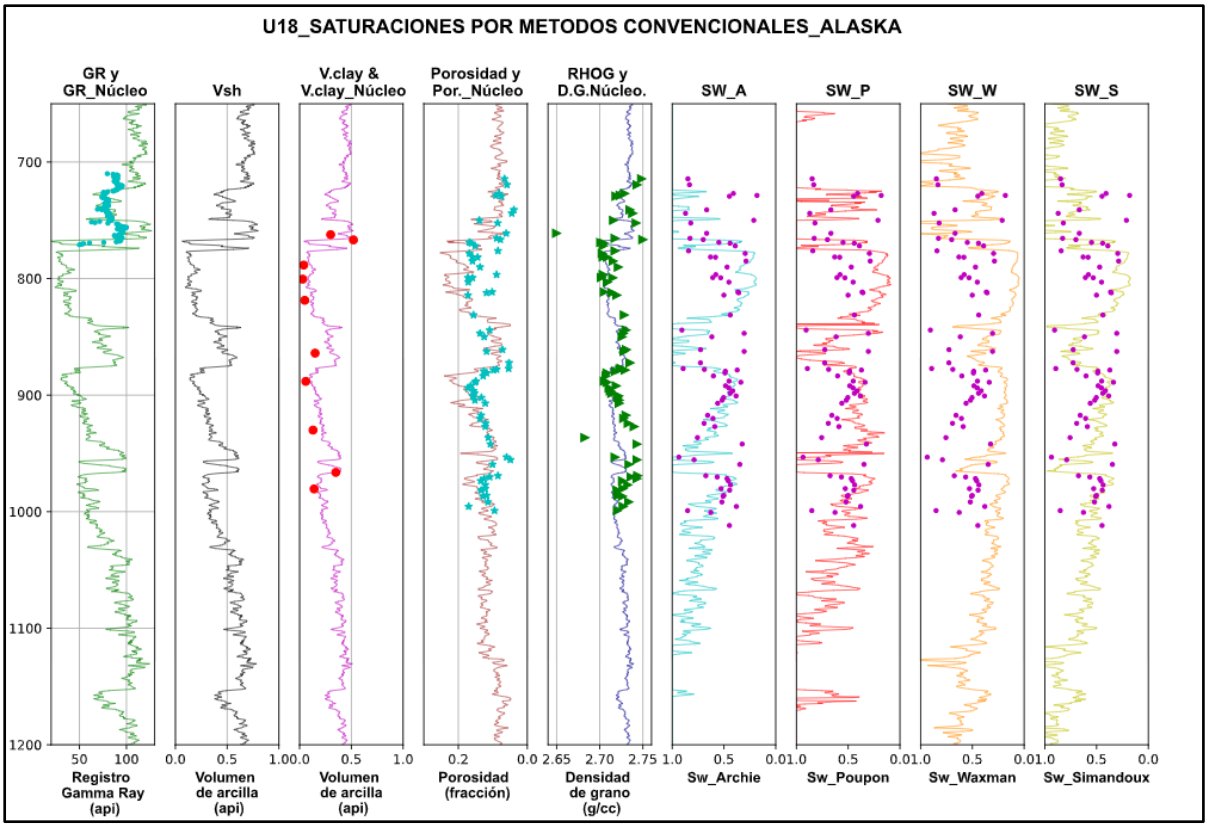
Tracks de los datos convencionales, de núcleos y de saturaciones- pozo T6.



Nota. La figura muestra nueve *tracks* de los valores calculados, de núcleo reportados y de saturación de agua calculada con métodos convencionales para el pozo T6 en la zona de interés. De izquierda a derecha, se observan los cinco primeros *tracks* gráficados de gamma ray, Vsh, Vclay , porosidad y densidad de grano; todos con su respectivo valor de núcleo. Los cuatro siguientes *tracks* muestran los datos continuos de: Sw por método Archie (línea cian), Sw por método Poupon (línea roja), Sw por método de Waxman-Smits (línea naranja) y Sw por método de Simandoux (línea amarilla). En los últimos 4 *tracks*, se gráfica la saturación de agua ajustada (SW_O) tomada de núcleos como datos discretos (puntos morados).

Figura 36.

Tracks de los datos convencionales, de núcleos y de saturaciones- pozo U18.



Nota. La figura muestra nueve *tracks* de los valores calculados, de núcleo reportados y de saturación de agua calculada con métodos convencionales para el pozo T6 en la zona de interés. De izquierda a derecha, se observan los cinco primeros *tracks* gráficados de gamma ray, Vsh, Vclay , porosidad y densidad de grano; todos con su respectivo valor de núcleo. Los cuatro siguientes *tracks* muestran los datos continuos de: Sw por método Archie (línea cian), Sw por método Poupon (línea roja), Sw por método de Waxman-Smits (línea naranja) y Sw por método de Simandoux (línea amarilla). En los últimos 4 *tracks*, se gráfica la saturación de agua ajustada (SW_O) tomada de núcleos como datos discretos (puntos morados).

3.2. Representación gráfica en Python del net Pay por métodos convencionales

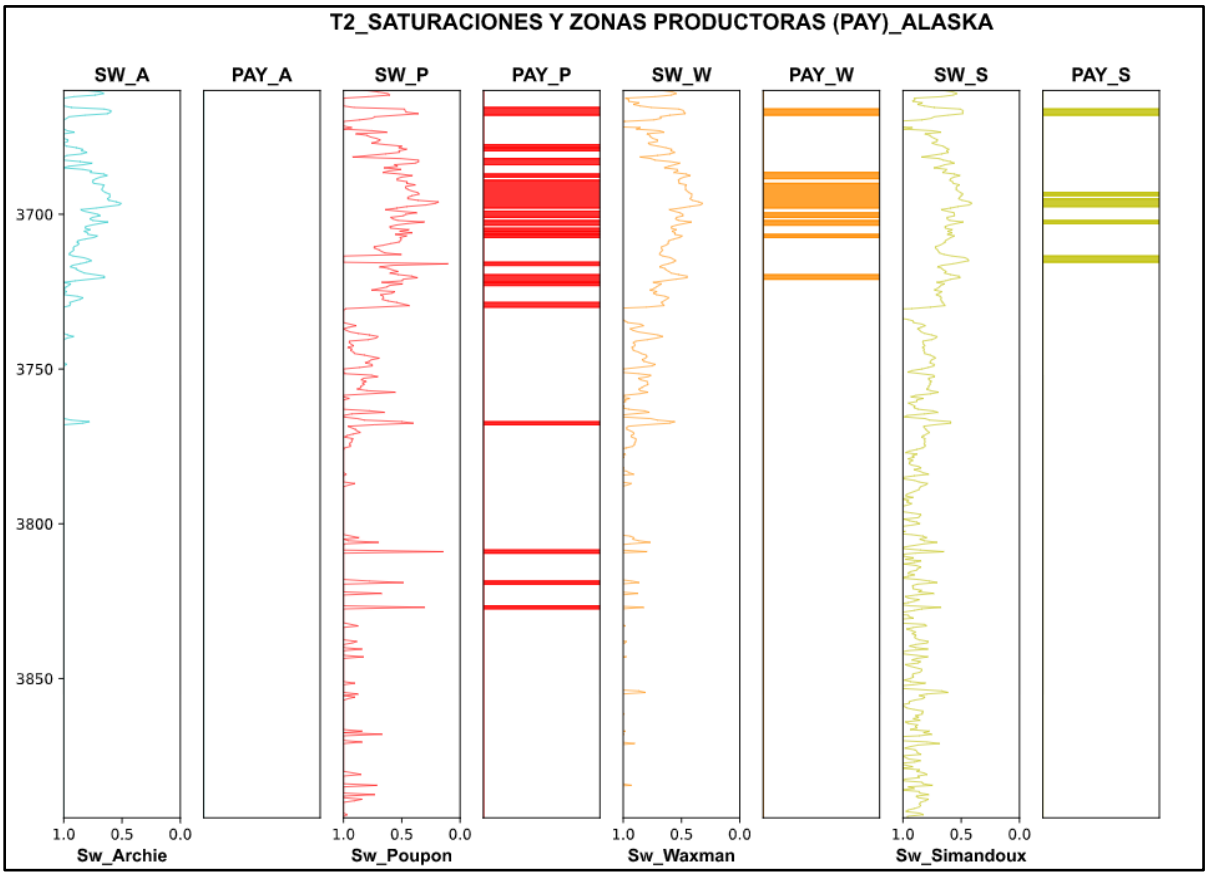
Se toman los valores de Sw de los métodos convencionales y se aplica el valor de *cutoff* del 50%. Como se muestra en las siguientes figuras, la saturación de Archie no muestra ninguna zona prospectiva para los pozos T2 y T6 mientras que para el pozo U18 logra identificar algunas zonas. En todos los pozos, el método de Puopon logra buenos resultados donde la arena es más limpia pero que hacia abajo logra perderse.

En la siguiente sección, se comparan estos resultados con las fotografías ultravioleta para determinar cuál método convencional es más representativo. Además, se usan estos

resultados para verificar la solidez del método de Machine Learning comparado con flujos de trabajo convencional.

Figura 37.

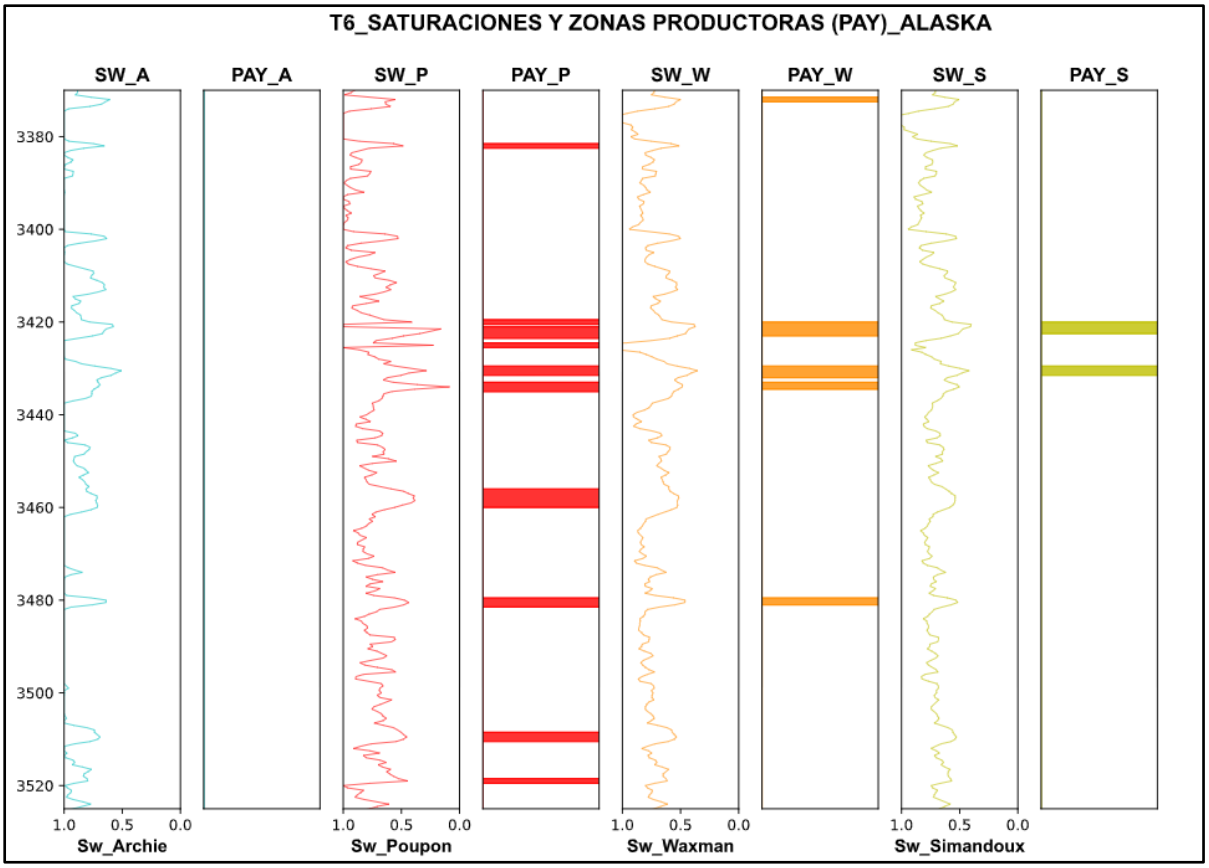
Tracks de Sw con métodos convencionales y zonas prospectivas (PAY)- T2.



Nota. La figura representa la saturación de agua y las zonas prospectivas por los cuatro modelos convencionales evaluados para el pozo T2. De izquierda a derecha, se encuentran graficados los valores calculados por método de Archie, que no muestra ninguna zona prospectiva, método de Poupon que indica mayores zonas prospectivas hacia el tope, método de Waxman-Smits muestra algunas zonas en el tope y hacia abajo ninguna. Por último, el método de Simandoux que muestra muy pocas zonas prospectivas.

Figura 38.

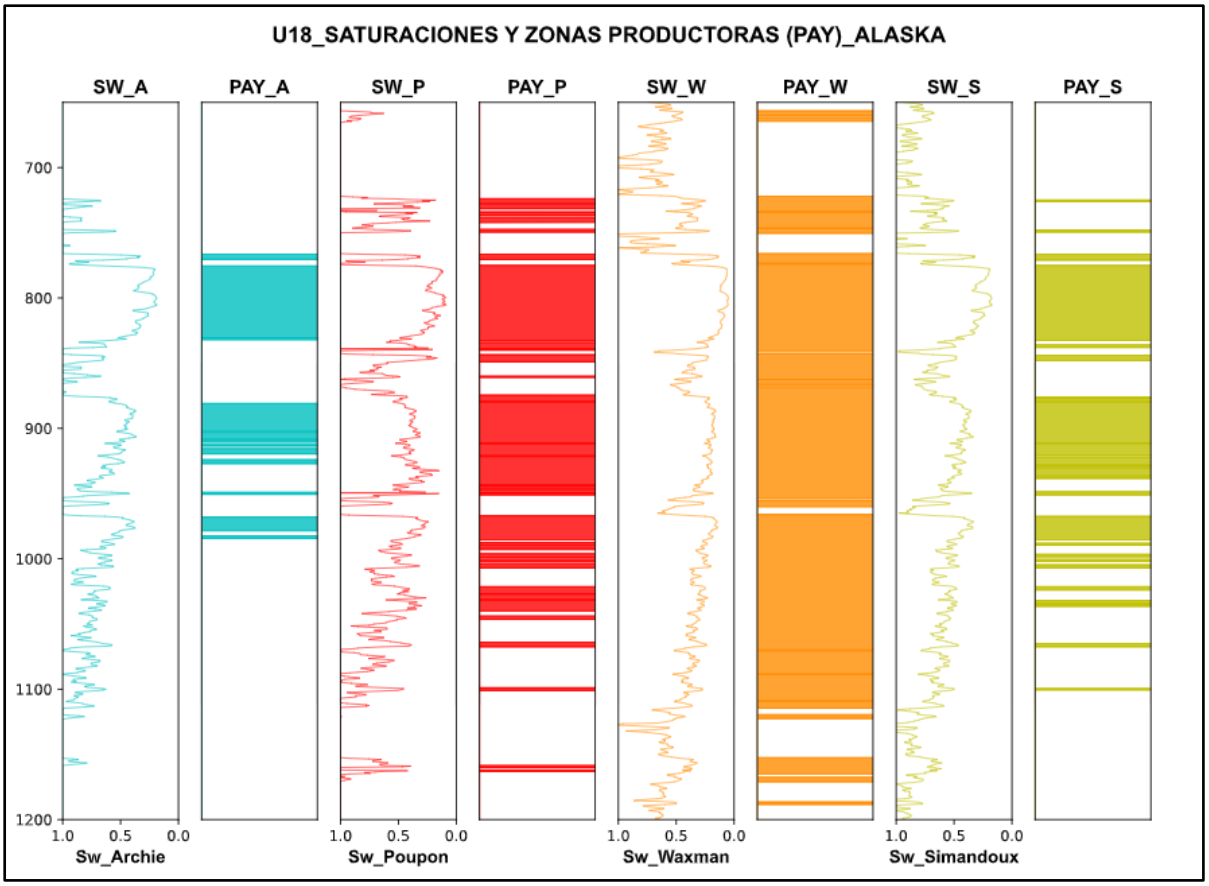
Tracks de Sw con métodos convencionales y zonas prospectivas (PAY)-T6.



Nota. La figura representa la saturación de agua y las zonas prospectivas por los cuatro modelos convencionales evaluados para el pozo T6. De izquierda a derecha, se encuentran gráficados los valores calculados por método de Archie, que no muestra ninguna zona prospectiva, método de Poupon que indica más zonas prospectivas que están distribuidas a lo largo del *track,* método de Waxman-Smits muestra muy pocas zonas prospectivas y por último, el método de Simandoux que muestra solo dos zonas prospectivas.

Figura 39.

Tracks de Sw con métodos convencionales y zonas prospectivas (PAY)-U18



Nota. La figura representa la saturación de agua y las zonas prospectivas por los cuatro modelos convencionales evaluados para el pozo U18. De izquierda a derecha, se encuentran gráficados los valores calculados por método de Archie, que muestra zonas prospectivas hacia el centro del *track*, método de Poupon muestra zonas prospectivas en gran parte de las arenas, método de Waxman-Smits muestra mayores zonas prospectivas hacia el tope y la base y por último, el método de Simandoux que muestra menos zonas prospectivas en comparación al método de Poupon.

3.3. Relación entre los registros de pozo, saturación calculada y procesamiento de imágenes

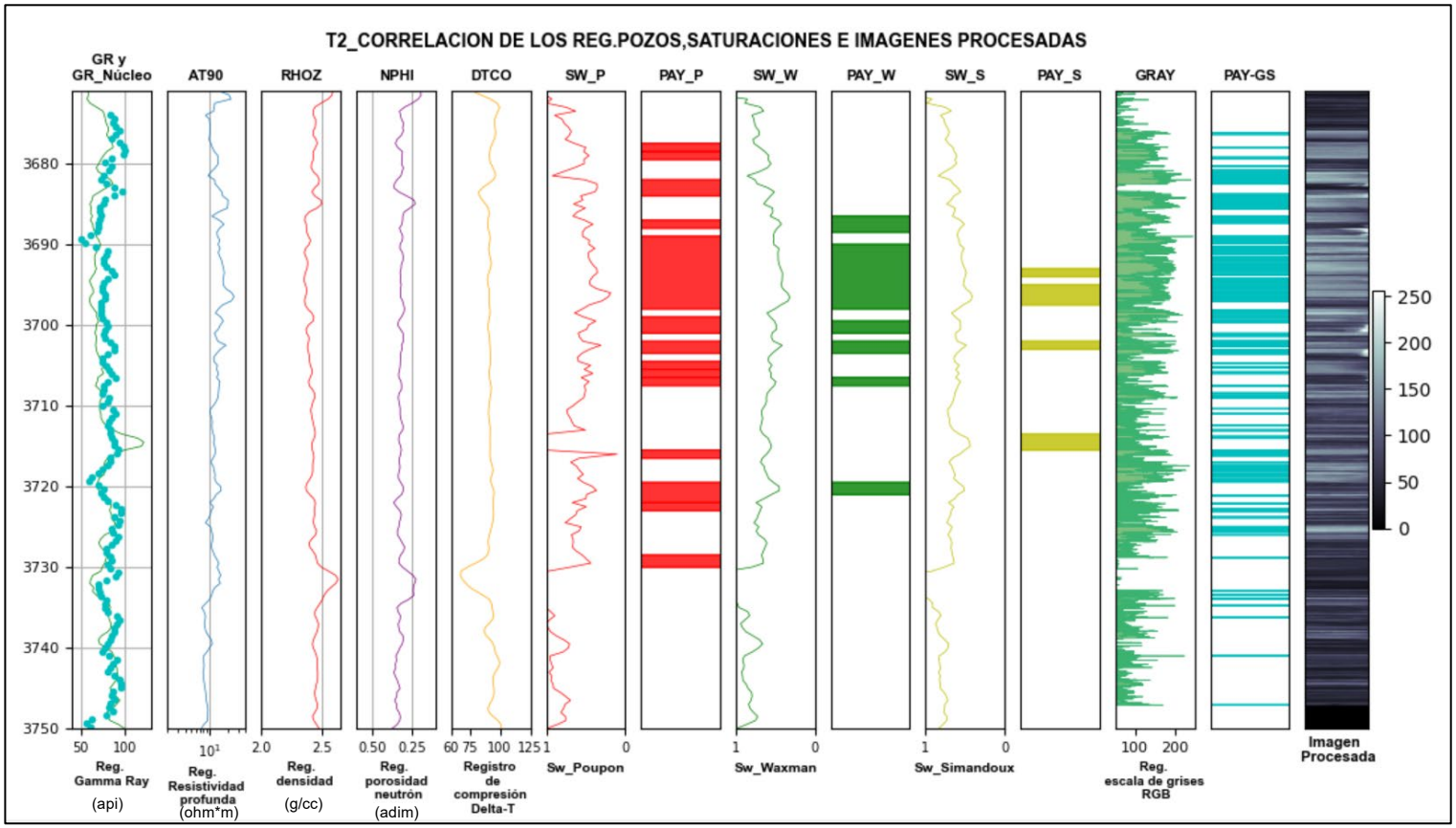
Cuando se termina el procesamiento de las imágenes, se hace una comparación con la saturación y las zonas prospectivas calculadas por los métodos convencionales en las profundidades del núcleo. Como se muestra en la Figura 40, se graficó el registro y el procesamiento de las imágenes con las zonas prospectivas (PAY-GS). Para obtener el track del Pay, se tuvo en cuenta el corte o "cutoff" de 170 (pozo T2) para determinar que todo valor por debajo del cutoff se convierte en una variable PAY-GS= 1 y de otra manera PAY-GS = 0.

Una vez obtenido el track del PAY-GS, se compara con las zonas prospectivas mostradas en cada uno de los tracks de PAY calculados por cada saturación. Se muestra que, aunque la saturación de Poupon muestra más zonas con aceite que la saturación de Simandoux, no se acerca a lo que fue interpretado con las imágenes procesadas de luz ultravioleta al existir zonas altamente laminadas. Esto explica que existan más hidrocarburos en cantidades explotables que no son percibidas por herramientas convencionales.

Todo el proceso anteriormente descrito, se realizó para los pozos T6 y U18 de la misma manera. El valor del corte (cutoff) varía muy poco para estos pozos (140 y 130).

Figura 40.

Correlación de registros, saturaciones y procesamiento de imágenes.



Nota. La figura representa la gráfica del procesamiento de imágenes junto a los Pay Flags de las saturaciones calculadas por métodos convencionales y los cinco registros básicos para el pozo T2. De izquierda a derecha, se encuentra en los cinco primeros *tracks* los registros de pozo (GR-GR_Núcleo (Figura 33) ,AT90,RHOZ,NPHI y DTCO) mostrados en la Figura 32. Luego se encuentran los *tracks* correspondientes a las zonas Pay calculadas por métodos convencionales mostradas en la Figura 37. Por último, se encuentran tres *tracks* del procesamiento de imágenes : el *track* del registro GS (verde), las zonas Pay mostradas por las fotos uv (PAY-GS) y el *track* de la imagen procesada en escala de grises (0 a 255).

3.4. Preparación para Machine Learning

La preparación de los datos es una de las fases del aprendizaje automatizado que requiere de tiempo. Existen diferentes desafíos que se encuentran al intentar construir un modelo óptimo. El conjunto de datos debe contener un amplio número de variables para que el entrenamiento y la prueba del modelo no presente dificultades [68].

3.4.1. Hacer "depth matching" de resolución de registros

Como se mencionó en el procesamiento de las imágenes, la resolución de una imagen depende de la cantidad de píxeles que esta contenga. En este caso, las imágenes cargadas a Python cuentan con una alta resolución, es decir, que contienen una gran cantidad de información capturada de los núcleos a frecuencias de profundidad más altas que los registros.

El depth matching es una práctica que consiste en desplazar las profundidades de varios conjuntos de datos a una profundidad diferente. Se realizó un reescalado del registro obtenido del procesamiento de las imágenes, a la misma escala de profundidad en la que se encontraron los registros eléctricos.

Todos los valores registrados, tanto para los registros básicos como para el registro obtenido del procesamiento de las imágenes, fueron puestos en una misma tabla como se muestra a continuación:

Figura 41.

Previsualización de la tabla de variables reescaladas en Python

	GR	RHOB	logRT	DTCO	NPHI	GRAY	Pay	DEPT	Well
5	57.4132	2.5472	1.358055	92.17439	0.2089	78.337261	0	3672.0	T2
6	60.4820	2.4898	1.091431	97.48821	0.2470	75.775341	0	3672.5	T2
7	68.2175	2.4371	1.081167	99.76886	0.2957	74.557063	0	3673.0	T2
8	68.2175	2.4371	1.081167	99.76886	0.2957	72.301991	0	3673.0	T2
9	68.2175	2.4371	1.081167	99.76886	0.2957	70.979323	0	3673.0	T2
...	...	...	...	...	...	...	...	...	...
100	64.4726	2.3719	1.370052	91.23937	0.3205	196.694397	1	3696.0	T2
101	64.4726	2.3719	1.370052	91.23937	0.3205	189.489088	1	3696.0	T2
102	61.0068	2.3591	1.408438	91.07550	0.3222	182.595187	1	3696.5	T2
103	60.0530	2.3532	1.366092	91.03039	0.3088	168.633173	0	3697.0	T2
104	60.0530	2.3532	1.366092	91.03039	0.3088	159.466614	0	3697.0	T2

Nota. La figura representa la previsualización de la tabla de datos generada en Python del reescalado del registro obtenido del procesamiento de las imágenes a las mismas profundidades de los registros eléctricos.

3.4.2. Filtrado (resolution match)

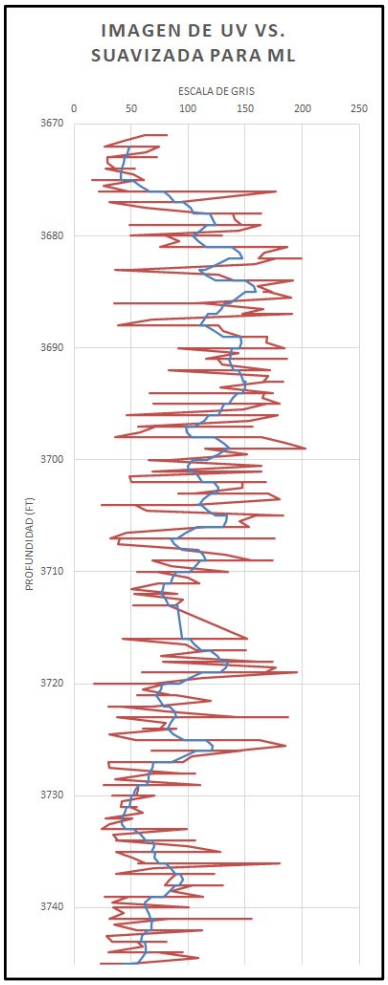
Los registros eléctricos poseen una baja resolución vertical comparados con una foto de un núcleo [70]. Usualmente las compañías de servicio proveen registros de alta resolución. Sin embargo, en este caso, los registros se encontraron con mediciones cada 0.5 ft ya que como se mencionó en este documento los datos se tomaron de una base de datos de acceso libre Esto implica que al relacionar los datos de registros con los de la imagen se debe hacer procesamiento.

El registro en escala de grises (GS), tiene una alta resolución debido a que contiene todos los cambios de color de la imagen procesada. Al poner este registro en un algoritmo de Machine Learning, es difícil correlacionar secciones de los registros eléctricos con el registro en escala de grises. El registro GS presenta mayor frecuencia de oscilaciones por unidad de profundidad que los registros, ya que estos promedian las propiedades de muchas capas de roca delgadas. En promedio, los registros de neutrón y densidad resuelven rocas de 1 pie de espesor y el registro de resistividad resuelve rocas de 2 a 3 pies dependiendo de la herramienta usada.

Para relacionar el registro GS y los registros básicos de pozo, se hace un preprocesamiento de la imagen de gris por medio de la función de suavizado (smoothing function) en Python. Esta función es una técnica que se usa para eliminar el ruido de un conjunto de datos con el promedio de los puntos y con otros puntos de la misma serie de datos. Este proceso tiene el efecto difuminado de los bordes en punta a un borde más redondeado. Generalmente al suavizado se le denomina “filtrado” por el hecho de suprimir señales de alta frecuencia y mejorar las de baja frecuencia [71]. Para este caso, se tomó un suavizado a 5 puntos por secciones para que facilite al algoritmo leer mejor el conjunto de datos.

Figura 42.

Previsualización - gráfica de imagen UV vs. imagen suavizada.



Nota. La figura representa la previsualización de los datos del registro en escala de grises sin preprocesamiento por profundidad (línea roja) y los valores obtenidos del preprocesamiento (suavizado) del registro de grises por profundidad (línea azul).

3.4.3. Equilibrio (Balancing)

Algunos conjuntos de datos tienen un desequilibrio de clases (forma de empaquetamiento de datos y sus valores), es decir, que generalmente existen mayor número de datos en un grupo o clase que en otras.

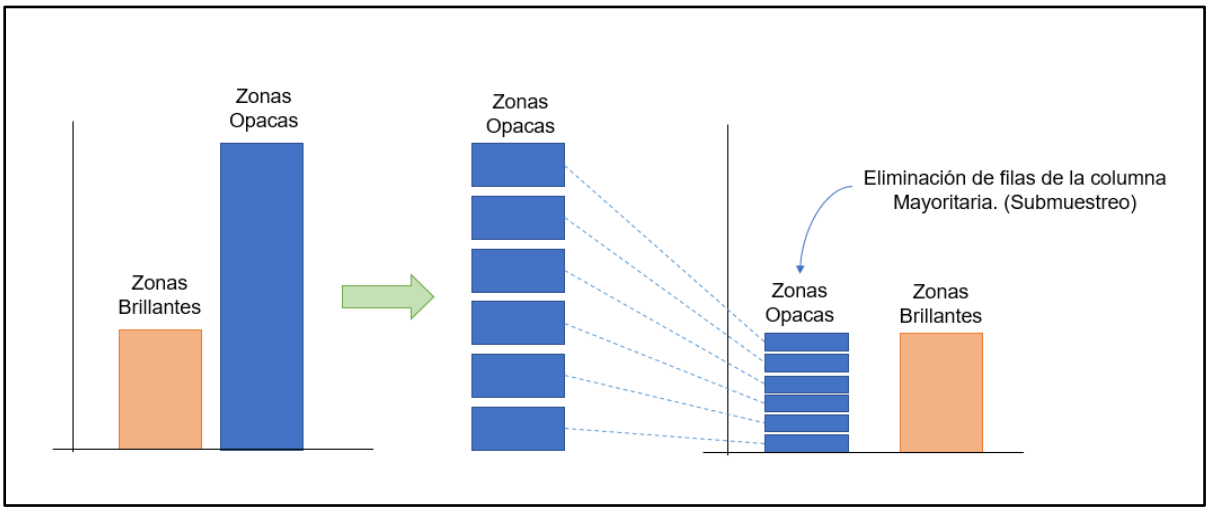
El desbalance de los datos es un problema muy común para trabajar en Machine Learning porque los algoritmos no logran manejarlos. Si no se genera el balance en los datos, es posible que un algoritmo tome todos los datos de una misma clase (mayoritaria) y ocasione una imprecisión durante la clasificación. El problema del desbalance en la regresión genera una

mayor dificultad que en los dominios de clasificación. Este problema en la regresión se debe a la complejidad que existe por el número de valores a tratar [75].

En las fotos UV hay muchas más zonas sin hidrocarburo que con hidrocarburo (mayores zonas de agua). Cuando se toma el conjunto de todas las fotos, se tiene una mayor proporción de datos sin Pay (zonas opacas) que zonas con Pay (brillantes). Al poner esta información en un algoritmo de Machine Learning, va a predecir las zonas opacas e ignorará las zonas brillantes (con fluorescencia). Para que pueda predecir las zonas de interés, es decir, las zonas de Pay es necesario lograr el equilibrio de los datos con un submuestreo de la clase mayoritaria.

Figura 43.

Esquema del balance de datos.

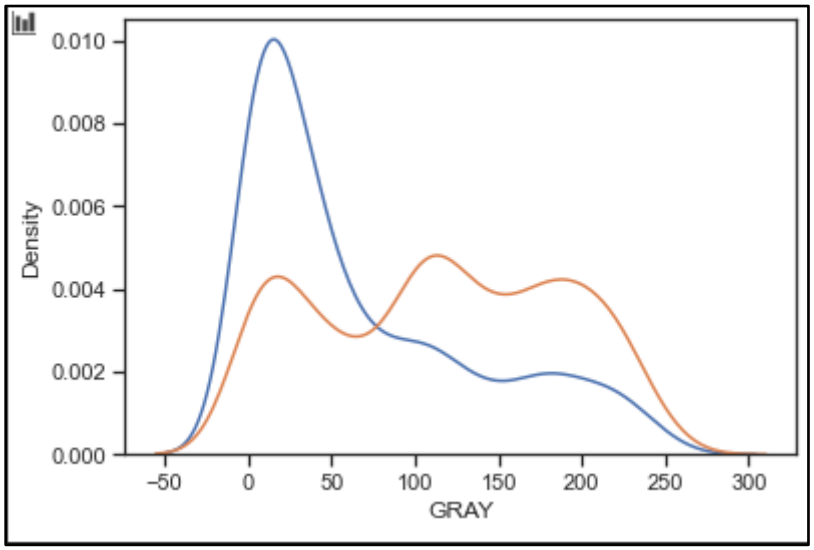


Nota. La figura representa el esquema del balance de datos entre las zonas opacas y las zonas brillantes (con fluorescencia) de las fotos UV. A la izquierda de la figura, se muestran dos barras: la barra naranja representa los datos contenidos de las zonas brillantes y la barra azul a los datos correspondientes a las zonas opacas de la imagen. En el centro de la figura, se muestra un seccionamiento del conjunto de datos o un remuestreo de las zonas opacas. A la derecha de la figura, se muestra el submuestreo de la clase mayoritaria para tener el balance de los datos entre las zonas opacas y las zonas brillantes.

Cuando el conjunto de datos se ingresa al algoritmo, inicialmente tiene una distribución asimétrica de los datos (más valores de zonas opacas que brillantes) y no permite hacer una predicción correcta de los datos totales. Para balancear los datos, el algoritmo internamente elimina secciones de los datos mayoritarios y los asigna a una clase minoritaria para que ambos conjuntos de datos queden al mismo nivel. El algoritmo va a poder hacer la predicción de una mejor forma y en el momento de hacer pruebas de error, den un valor similar.

Figura 44.

Curvas de desequilibrio y balance del conjunto de datos.



Nota. La figura representa las curvas de los datos desequilibrados y balanceados entre las zonas opacas y las zonas brillantes de la imagen UV. La curva azul, representa una distribución asimétrica de los datos donde los valores de las zonas opacas se encuentran en el pico más alto de la gráfica. La curva naranja, representa el conjunto de datos distribuidos de forma más homogénea para poder hacer una mejor predicción de todos los datos.

3.4.4. Escalado

La gran mayoría de los modelos de aprendizaje pueden trabajar de mejor manera cuando las variables o datos de entrada son transformados o escalados cuidadosamente antes del modelamiento.

La librería Scikit Learn, proporciona un módulo de preprocesamiento que permite hacer uso de funciones para el escalado de datos. La función StandardScaler permite transformar los datos de manera en que la distribución de estos tenga una media de 0 y una desviación estándar unitaria [74]. Como se muestra en la Figura 41, los valores de los 5 registros básicos son diferentes en cuanto a su magnitud. Al hacer el entrenamiento, el algoritmo va a darle mayor importancia a los valores de mayor magnitud que a los de menor magnitud.

Los algoritmos solo reconocen números más no qué significado tiene cada valor para cada variable. En este caso, tenemos algunos valores de GR por encima de 100, valores de NPHI de 0.2 a 0.5 aprox, valores de RHOB de 2 en adelante , etc. El algoritmo va a suponer que los valores de mayor importancia son los de GR y trabajará en función de estos ignorando los demás. El escalado permite que se tenga una igualdad de magnitud en los valores y no de una mayor importancia a una sola columna de valores.

En Machine Learning, el escalado es uno de los procesos más críticos en el preprocesamiento antes de construir un modelo. Este proceso toma todas las variables independientes y las ajusta a una escala específica para que todos los valores se cambien a un rango de distribución generalmente entre 0-1. Esto se hace con el fin de que el cambio sea proporcional al resultado y evitar que el algoritmo no funcione por la diferencia en los features (variables independientes).

Para nuestro conjunto de datos, se hizo el escalado a 2 fases. La primera fase se hizo por pozo y la segunda para todo el set de entrenamiento como se muestra a continuación:

3.4.4.a.Fase 1: escalado del registro en escala de grises por pozo. Para esta fase, se hizo un escalamiento del registro GS por pozo a la escala de los registros básicos usados. Algunos de los tipos de escalado son: escalador mínimo-máximo y escalador estándar.

El escalador mínimo-máximo transforma las características haciendo un escalado dentro de un rango determinado que puede estar entre [0,1] o [-1,1]. El escalador toma cada característica de forma individual (valor - valor mínimo) y lo prueba en el conjunto de entrenamiento verificando que esté dentro del rango (mínimo-máximo) [76]. En el caso del escalador estándar se asume que los datos están distribuidos normalmente en cada columna de valores y los escala de tal forma que tenga una media de 0 y una distribución estándar de 1.

En la sección 3.4.3 Balancing, una vez se tienen suavizados los valores se hace el escalado para poder comparar las imágenes de un pozo a otro. Este escalado se hace con la función scaler.fit, que transforma los datos contenidos en las variables de la Figura 45 en un tamaño o escala entre el rango de [-1,1] como se muestra en la Figura 46.

Figura 45.

Previsualización de código - suavizado de valores del registro GS

```
scaler = MinMaxScaler ()
window = 11
T2.GRAY = T2.GRAY.rolling(window , center=True,win_type='gaussian').mean(std=3)
T6.GRAY = T6.GRAY.rolling(window , center=True,win_type='gaussian').mean(std=3)
U18.GRAY= U18.GRAY.rolling(window , center=True,win_type='gaussian').mean(std=3)
```

Nota. La figura representa la sección de código realizada para el suavizado de los valores del registro GS. La variable scaler, guarda los valores mínimos y máximos del escalado. La variable window, es el número de observaciones utilizadas para el cálculo. Las variables T2.GRAY, T6.GRAY y U18.GRAY, almacenan los valores nuevos con una desviación estándar de 3 y en una ventana tipo gaussiana para realizar el promedio de la curva (Figura 44)[84].

Figura 46.

Previsualización de código - escalado en rango determinado por pozo.

```
T2['GRAY'] = 255*scaler.fit_transform(T2['GRAY']. values.reshape(-1,1))
T6['GRAY'] = 255*scaler.fit_transform(T6['GRAY'].values.reshape(-1,1))
U18['GRAY'] = 255*scaler.fit_transform(U18['GRAY'].values.reshape(-1,1))
```

Nota. La figura representa la previsualización de la sección de código del escalado por pozo. La función scaler.fit calcula la media y la desviación estándar que se utilizaran para el escalado. la función values.reshape, reescala los valores a un rango de [-1,1] que son almacenadas en las variables T2.[‘GRAY’], T6.[‘GRAY’] y U18.[‘GRAY’][84].

3.4.4.b.Fase 2: escalado del set de entrenamiento. En esta fase, el escalado se realiza como en la sección anterior, pero para todo el set de entrenamiento haciendo uso del escalador estándar.

Figura 47.

Previsualización de código - escalado del conjunto de datos de entrenamiento**.**

```
scaler = StandardScaler()
print("Test Dataset Shape  :::::: ", test.shape)
print("Train Dataset Shape :::::: ", train.shape)

train.to_excel('./ML_Results/train.xlsx','Train')
test.to_excel('./ML_Results/test.xlsx','Test')

option = 'Gray' #or 'Pay'
if option == 'Gray':
   target_col = 5
else:
   target_col = 6

X = train.iloc[:, [0,1,2,3,4]]
y = train.iloc[:, [target_col]]   #5 is GRAY, #6 is Pay

X_test = test.iloc[:, [0,1,2,3,4]]
y_test = test.iloc[:, [target_col]]

# Scaling
#Find scaling parameters based on training data only
scaler = StandardScaler()
scaler.fit(X)
print(scaler.mean_)
X = scaler.transform(X)
X_test = scaler.transform(X_test)
```

Nota. La figura representa la previsualización de la sección de código de la segunda fase de escalado basado solo en el set de entrenamiento. La función print, imprime el tamaño del conjunto de datos de entrenamiento y prueba. Esos valores son llevados a un Excel con las variables train.to_excel y test.to_excel respectivamente. La variable option, guarda los valores de las imágenes en escala de grises de la columna 5 (target_col) y/o almacena los valores de la columna 6 (Pay). Las variables X y Y, ubican y almacenan los valores tanto de prueba como de entrenamiento. La variable X ubica las columnas de los registros mientras que la variable Y ubica las columnas objetivo de GRAY y PAY [84].

3.5. Análisis de los gráficos

En esta sección, se grafican diferentes variables para entender la relación entre las mismas, su distribución y detección de posibles valores anómalos que deban corregirse antes de comenzar la fase de modelamiento. Se tomaron las variables de GR y Escala de grises (procesamiento de fotos UV) para ser representadas en dos diferentes histogramas. Adicionalmente, se hace un diagrama de parejas que muestra la relación que existe entre cada una de las diferentes variables.

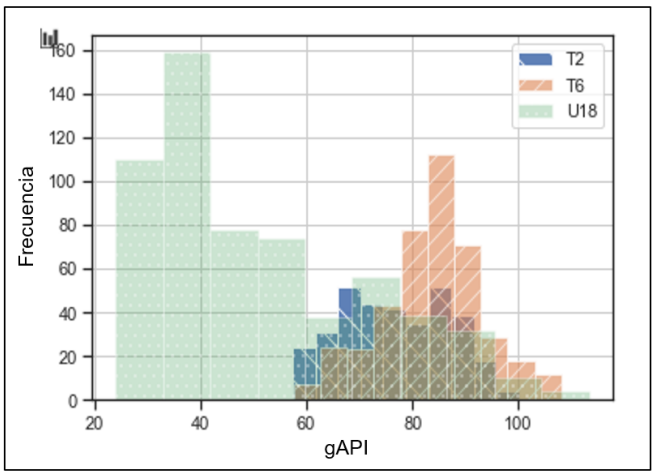
3.5.1. Histogramas

Los histogramas permiten mostrar la tendencia, dispersión y forma de distribución de los datos recopilados para los 3 pozos (T2, T6, U18). Para estos histogramas se toman los valores del registro de gamma ray y de la escala de grises como se muestra en las Figuras 48 y 49.

3.5.1.a.Registro Gamma Ray (GR). En la Figura 48 se muestra el histograma con la relación que existe entre los tres pozos y el registro de GR (GR_EDTC). En el eje x, se tienen los valores de los grados api (gAPI) en un rango de 0 a 150. En el eje y, está la frecuencia absoluta de la muestra, es decir, el número de veces que se repite un resultado en el conjunto de los valores de gAPI.

Figura 48.

Histograma - Gamma Ray en función de la frecuencia absoluta.



Nota. La figura representa el histograma de la relación entre el registro gamma ray y la frecuencia absoluta de los rayos emitidos para los pozos T2 (azul oscuro), T6 (naranja) y U18 (verde).

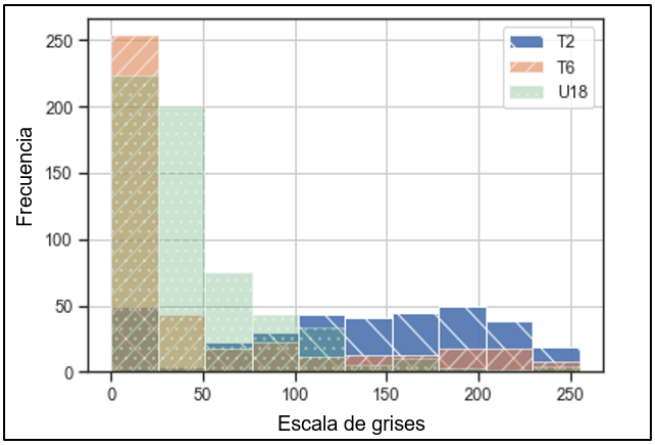
El histograma, muestra que el pozo U18 tiene muchos más valores que emiten una menor cantidad de rayos gamma en comparación al pozo T6. Esto sucede porque en la formación donde se perforó el pozo U18, hay mayor contenido de grano fino (lutitas). Aunque los tres pozos se encuentran en la misma cuenca, el pozo U18 fue perforado más lejos de los pozos T2 y T6 y por ende, fue sometido a procesos geológicos diferentes lo que hace que contenga menos contenido arcilla y esté más compacto.

3.5.1.b.Registro de Escala de Grises. En base a la intensidad del color de la imagen, es posible clasificar el conjunto de datos en una escala que va entre 0 a 255 en escala de grises [16] como se explica en la sección 2.6.4.

En el histograma presentado en la Figura 49, se relaciona la escala de grises en el eje "x" y en el eje "y" la frecuencia absoluta de la muestra por cada pozo. Se muestra que la mayoría de los datos del T6 y U18, se encuentran en un rango de 0 a 75 (tonos oscuros) que indica ausencia de hidrocarburo. Algunos de los datos del pozo T6 avanzan en gran proporción hacia las tonalidades brillantes en un rango de 100 a 255 (presencia de hidrocarburos) como pasa con el pozo T2.

Figura 49.

Histograma- escala de grises en función de la frecuencia absoluta.



Nota. La figura representa el histograma de la variación de la escala de grises en función de la frecuencia absoluta de los datos para los pozos T2 (azul oscuro), T6 (naranja) y U18 (verde) Elaboración propia.

3.5.2. Diagrama de parejas

La función *Pairplot* o "de parejas", es una de las funciones utilizadas para la exploración de datos de las librerías Seaborn. Esta función permite visualizar un conjunto de gráficos en un

solo gráfico haciendo más fácil el poder correlacionar diferentes variables en un análisis completo. Este tipo de conjuntos de gráficos, son útiles para determinar el comportamiento de las variables antes de hacer cualquier proceso con *Data Science* [69].

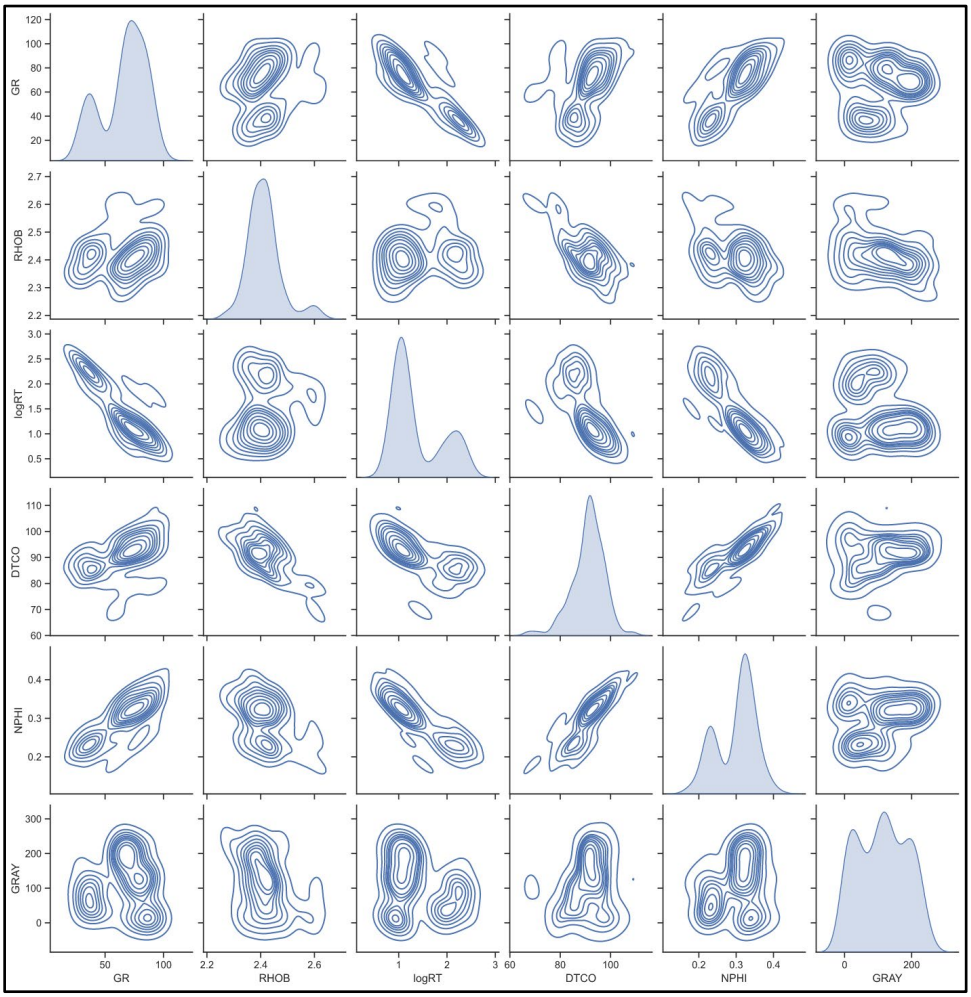
La figura 50 nos permite ver rápidamente relaciones importantes entre las variables independientes (los registros) y la variable objetivo (GRAY). Los gráficos de Gray contra los registros muestran dos zonas (*clústeres*) de distribución de los datos. La razón de estos dos agrupamientos es que el pozo U18, que se encuentra a 96,3 km de los pozos T2 y T6, penetra arenas que estuvieron sometidas a procesos geológicos diferentes, lo cual se evidencia en un menor contenido de arcilla en comparación con los otros dos pozos. Estas diferencias en propiedades petrofísicas causan lecturas diferentes de los registros.

La similitud entre la respuesta de los registros de los pozos T2 y T6 y su diferencia respecto al pozo U18 es ideal para el problema que se intenta resolver. Si tomamos tres pozos muy cercanos, no se cubre un rango geográfico suficientemente grande para que el modelo que se va a entrenar tenga aplicación regional. En este caso, el uso del pozo U18 permite capturar en rango más amplio de rocas que podrían estar presentes en la formación Nanushuk regionalmente.

Además, se logró verificar relaciones fuertes entre los registros de pozo y la variable GRAY derivada de las imágenes de fluorescencia. En las regiones brillantes de las imágenes (altos valores de GRAY) se reflejan altos valores de resistividad, y bajos valores de GR, hecho consistente con lo que se espera de una arena con poca arcilla y saturada de hidrocarburo en la formación Nanushuk.

Figura 50.

Diagrama de parejas.



Nota. La figura representa el diagrama de parejas de variables. La diagonal muestra los histogramas de cada variable y cada gráfica no diagonal muestra un *crossplot* entre variables con líneas de contorno para indicar la concentración de los datos. El componente principal de la gráfica es la última fila, que muestra la relación entre la variable Gray (objetivo) y todos los registros de pozo (*features*).

3.5.3. Predicción de cada modelo

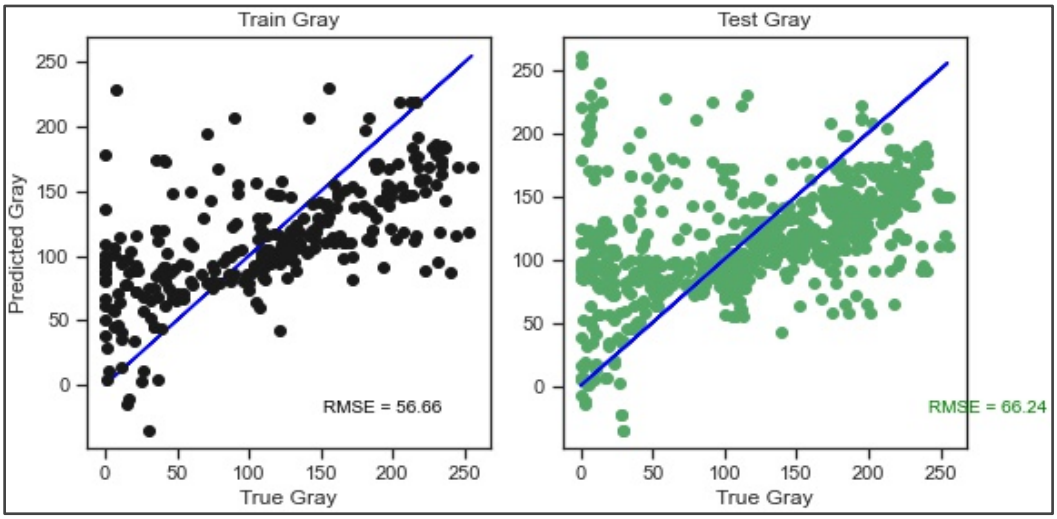
En esta sección, se evaluaron todos los modelos relacionados en la sección 2.8 mediante el usó la técnica de Grid Search descrita en la sección 2.9. Como se mencionó antes, no solo se minimiza la función de costo (Cost Function) sino que también se selecciona la arquitectura óptima para cada tipo de algoritmo. Las siguientes subsecciones muestran los resultados en entrenamiento y prueba para todos los algoritmos descritos en este documento. El objetivo es probar cómo se comporta cada mdoelo con pocas variables independientes y número escaso de muestras.

En las siguientes figuras, se muestran los resultados de entrenamiento (gráfica izquierda, puntos negros) y prueba (gráfica derecha, puntos verdes). La línea azul representa una línea de pendiente 1 para la cual el error equivalente es de 0. En cada gráfica, se muestra el valor del error de la media cuadrática (RMSE) para comparar los rendimientos de cada modelo y hacer la selección del mejor modelo.

3.5.3.a.Lasso. Este algoritmo hace que los pesos de algunas características disminuyan a cero en un punto determinado y eliminar de forma efectiva aquellas características que causan mayor variación y problemas de sobreajuste.

Figura 51.

Diagrama de dispersión - modelo Lasso.



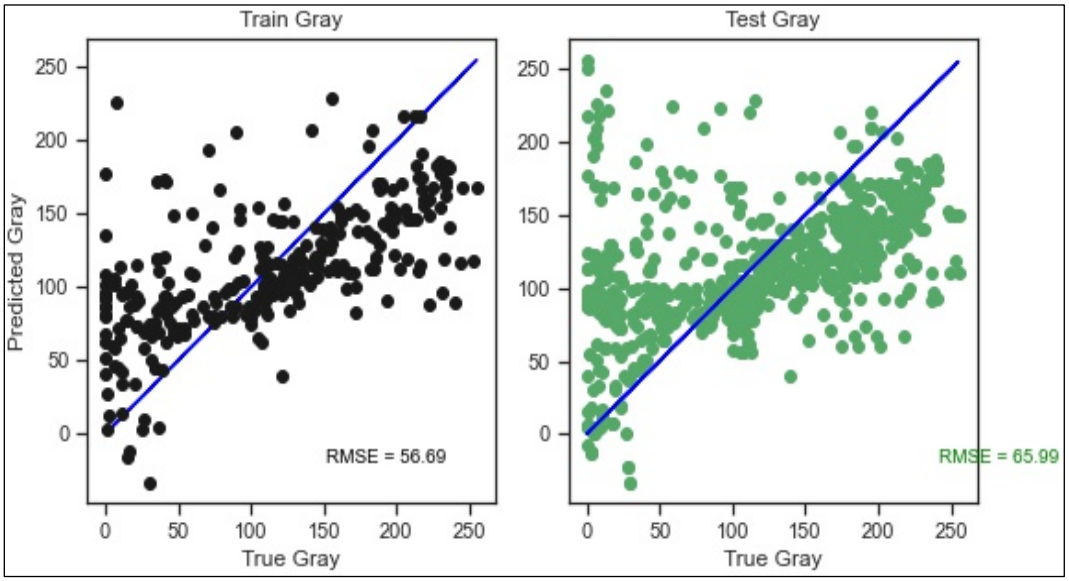
Nota. La figura representa el diagrama de dispersión de los datos del registro en escala de grises (GRAY) en función de la predicción del modelo Lasso. En el lado izquierdo de la figura, se representa la predicción del modelo Lasso entre el entrenamiento y el valor True GRAY con un porcentaje de error (RMSE) del 56.66. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un porcentaje de error (RMSE) del 66.24. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

En la figura anterior, el *subplot* de la derecha muestra el comportamiento del entrenamiento del modelo con una mayor dispersion de los datos y dificil agrupamiento de los valores en rangos de GRAY (entre 0-50). En la prueba del modelo (*subplot derecho*), muchos de los puntos se logran agrupar pero en un rango de 0 a 100 siguen presentando una mayor dispersion. En ambos casos hace que se genere un porcentaje de error (RMSE) mayor en prueba que en entrenamiento.

3.5.3.b.ElasticNet. En la Figura 52, se muestra el comportamiento de los datos para entrenamiento y prueba del algoritmo en funcion de las escala de grises obtenida del registro GS. En la figura, se observa que en la prueba las zonas opacas (0-50) se generan valores bastante diferentes al verdadero. Los valores de RMSE muestran la distancia de error en la que se encuentra el conjunto de valores tanto para entrenamiento como para prueba con un valor de 56.69 y de 65.99 , respectivamente.

Figura 52.

Diagrama de dispersión -modelo ElasticNet.



Nota. La figura representa el diagrama de dispersión de los datos del registro en escala de grises (GRAY) en función de la predicción del modelo ElasticNet. En el lado izquierdo de la figura, se representa la predicción del modelo en el entrenamiento y el valor True GRAY con un porcentaje de error (RMSE) del 56.69. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un porcentaje de error (RMSE) del 65.99. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

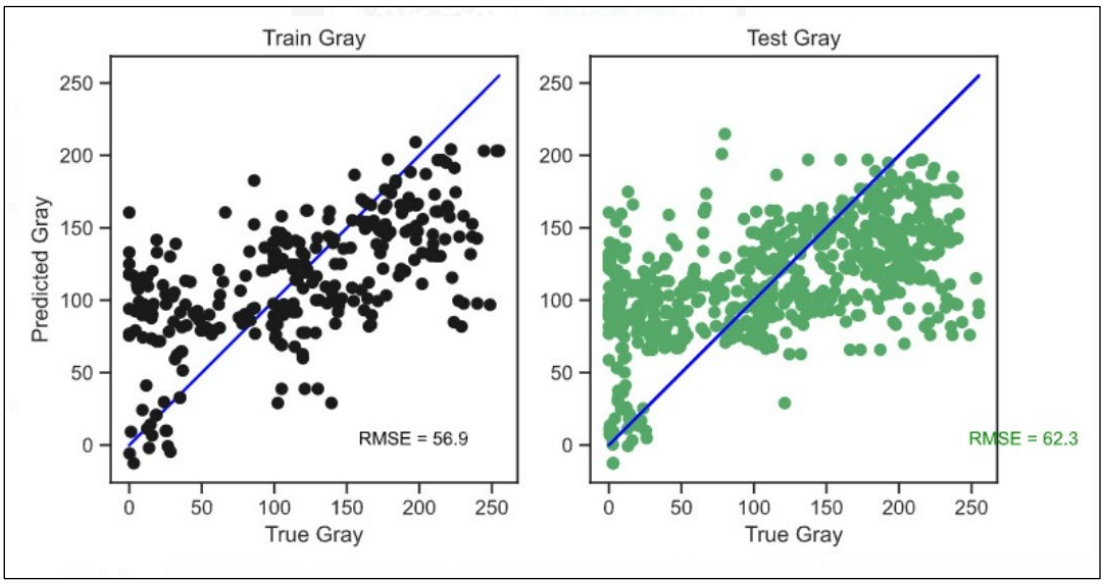
3.5.3.c.Ridge Regression. Este algoritmo es eficiente para conjuntos de datos lineales, ya que reduce el tamaño total de los valores de peso durante la optimización y reduce el sobreajuste. Cuando el algoritmo trabaja con valores donde una línea no abarca la mayoría de los datos, el modelo no se ajusta efectivamente.

La Figura 53 muestra la distribucion de los datos que hay en las gráficas de *train* y *test* de dispersion de los puntos.En el entrenamiento (puntos negros), los datos se dispersaron mas mientras que en la prueba (puntos verdes) los datos se agruparon dando un valor mucho menor respecto al predecido con el modelo.

Se muestra ademas, que el modelo logra predecir mejor en el conjunto de datos de prueba los valores verdaderos respecto a los de entrenamiento. El error de media cuadratica, indica que en el set de prueba los datos a pesar de estar mas agrupados no logran un buen ajuste sino que sub-ajusta los datos sin generar una buena salida de prediccion.

Figura 53.

Diagrama de dispersión -modelo Ridge Regression.

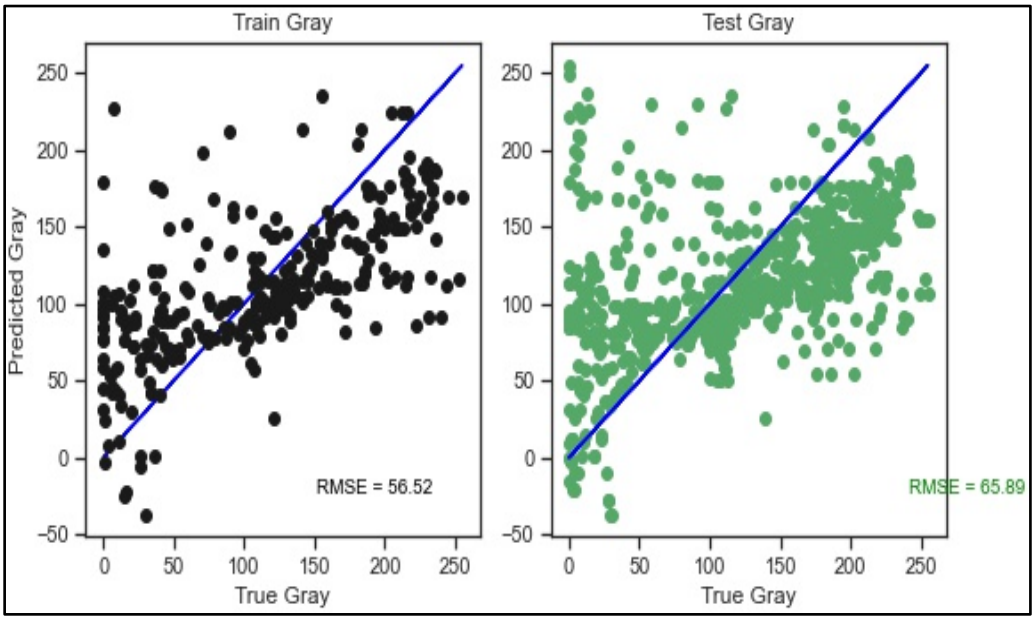


Nota. La figura representa el diagrama de dispersión entre los datos del registro en escala de grises (GRAY) en función de la predicción del modelo Ridge Regression. En el lado izquierdo de la figura, se representa la predicción del modelo en el entrenamiento y el valor True GRAY con un RMSE del 56.9. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un RMSE del 62.3. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

3.5.3.d.Support Vector Regression (SVR). Al analizar la Figura 54, se muestra que en el rango de 0-50 de GRAY se obtiene varios datos con error, y es un error persistente con mayor frecuencia en el conjunto de test, por lo tanto se evidencia que el modelo no logra una buena generalización de los datos que están en ese rango, algo similar ocurre para el rango 150-250. Así que el modelo demuestra que hubo un subajuste de los datos.

Figura 54.

Diagrama de dispersión - modelo Support Vector Regression (SVR).

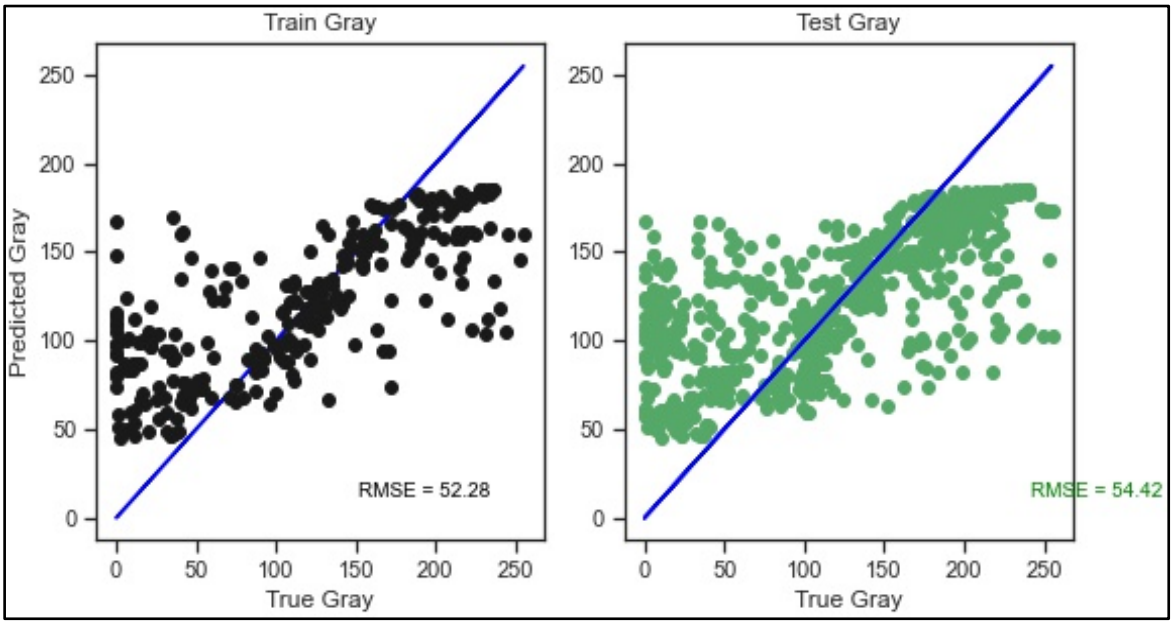


Nota. La figura representa el diagrama de dispersión entre los datos del registro en escala de grises (GRAY) en función de la predicción por el modelo Support Vector Regression (SVR). En el lado izquierdo de la figura, se representa la predicción del modelo en el entrenamiento y el valor True GRAY con un porcentaje de error (RMSE) del 56.52. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un porcentaje de error (RMSE) del 65.89. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

3.5.3.e.Gradient Boosting Regression. Al observar la Figura 55, existe una diferencia de error reducida entre train y test, donde se ve una tendencia y forma de los gráficos similar. Aunque son parecidos ambos subplot, estos indican que la mayoría de los datos están teniendo una variedad en la intensidad de color mayor de 52 lo cual es una evidencia considerable a la hora de seleccionar el mejor modelo.

Figura 55.

Diagrama de dispersión -modelo Gradient Boosting Regression.



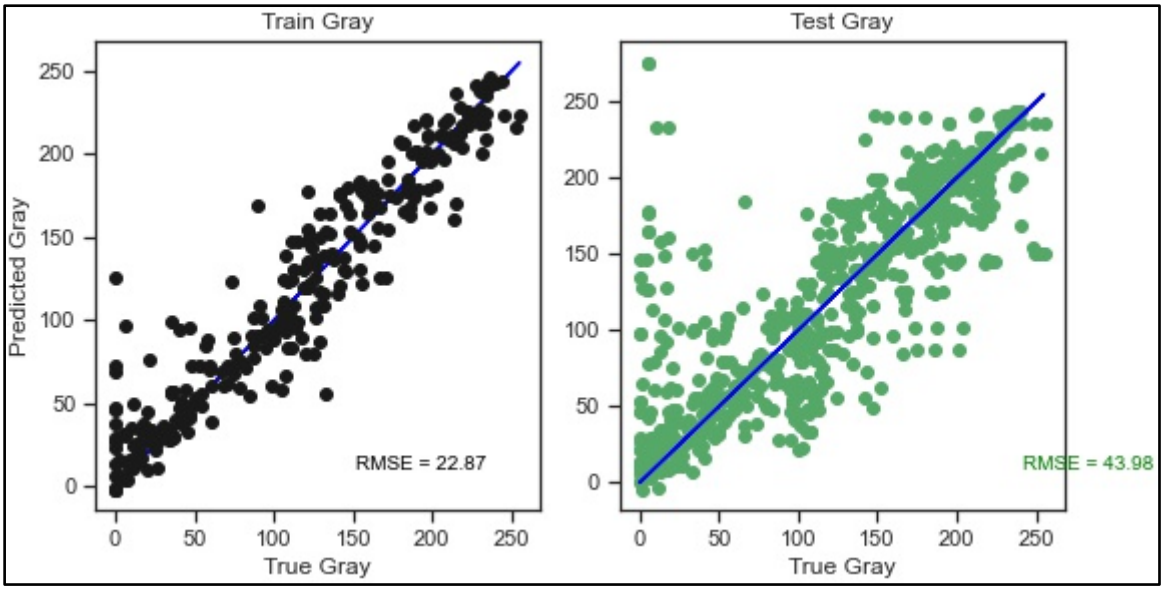
Nota. La figura representa el diagrama de dispersión entre los datos del registro en escala de grises (GRAY) en función de la predicción por el modelo Gradient Boosting Regression. En el lado izquierdo de la figura, se representa la predicción del modelo en el entrenamiento y el valor True GRAY con un porcentaje de error (RMSE) del 52.28. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un porcentaje de error (RMSE) del 54.42. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

3.5.3.f.Multilayer Perceptron (MLP) - Neural Network. Estos sistemas aprenden y se forman asimismo en lugar de ser programados de forma explícita, de tal manera que se puede reconocer patrones, clasificar datos y pronosticar eventos futuros, consiguiendo así aprender de sus propios datos. Al mismo tiempo sobresale en áreas donde la obtención de soluciones es difícil de expresar con la programación convencional [61]. Aunque el funcionamiento del modelo Neural Network es efectivo o ideal, este requiere de una cantidad significativa de parámetros para obtener un resultado óptimo.

Analizando la Figura 56 se observa que el entrenamiento de este modelo tiene un comportamiento ideal, ya que la distancia de error es de solo 22.87 de Gray. Comparándolo con el subplot de prueba, es notable ver que el modelo se estaba sobre ajustando, debido a que los datos varían de forma significativa especialmente entre el rango de 0-50. Sin embargo, la variación en la intensidad de color es de solo 43.98 logrando que este modelo sea considerado como uno de los mejores.

Figura 56.

Diagrama de dispersión - modelo Multilayer Perceptron (MLP) y Neural Network.



Nota. La figura representa el diagrama de dispersión entre los datos del registro en escala de grises (GRAY) en función de la predicción por la combinación de los modelos Multilayer Perceptron (MLP) - Neural Network. En el lado izquierdo de la figura, se representa la predicción del modelo en el entrenamiento y el valor True GRAY con un porcentaje de error (RMSE) del 22.87. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un porcentaje de error (RMSE) del 43.98. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

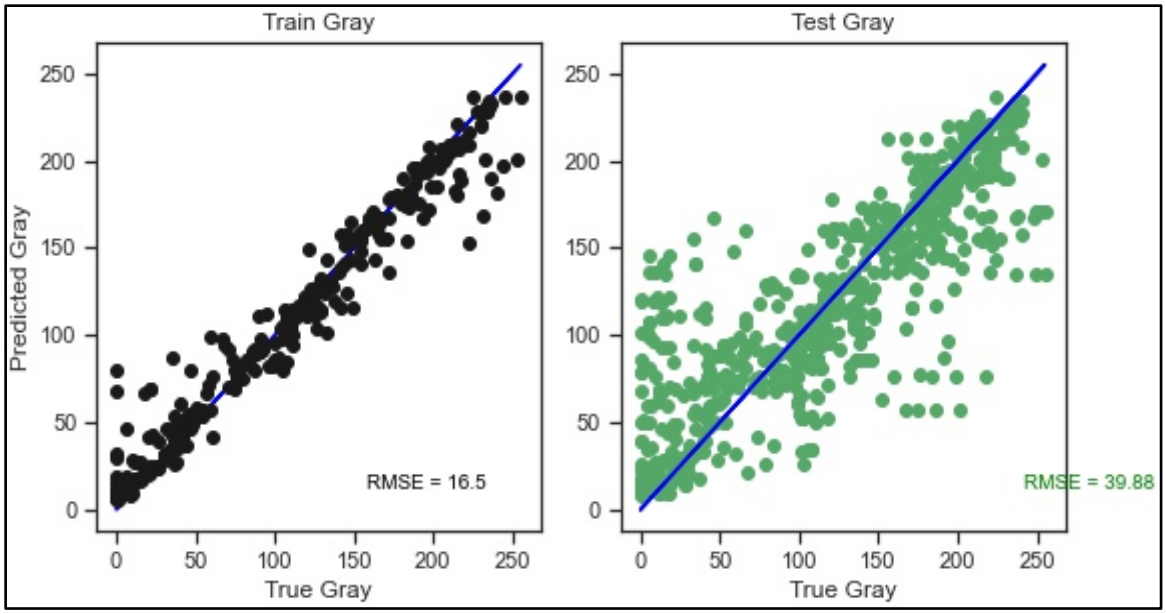
3.5.3.g.Random Forest. Es un algoritmo de aprendizaje de tipo supervisado y que puede ser implementado tanto para regresión como para clasificación. Este tipo de algoritmo trabaja de forma aleatoria con los datos creando árboles de decisión y uniéndose de forma secuencial para obtener como resultado un bosque. Este modelo se desempeña de manera eficiente en la estimación de las variables relevantes y mantiene una precisión en un escenario con datos insuficientes. Sin embargo, el desarrollo ideal del modelo dependerá de la cantidad de árboles de decisión que requiera. De esta forma, se va a predecir de manera correcta un conjunto de datos o de lo contrario el árbol de decisión tenderá a sobre ajustarse [55].

Esta sección muestra el mejor modelo alcanzado de todos los probados, él modelo de Random Forest. La Figura 57 muestra en la parte izquierda un bajo margen de error ya que el modelo sigue acertadamente la tendencia de la línea azul (RMSE = 16.5), lo que significa que, en el entrenamiento, el modelo se adaptó correctamente. En el set de prueba es evidente que el error aumenta; el modelo se sobre ajusta a los datos de entrenamiento y al mostrarle nuevos

datos la generalización no resulta tan acertado. Sin embargo, el algoritmo de Random Forest logra obtener el error más bajo de todos los algoritmos, RMSE de 39.88 en prueba.

Figura 57.

Diagrama de dispersión - modelo Random Forest.



Nota. La figura representa el diagrama de dispersión entre los datos del registro en escala de grises (GRAY) en función de la predicción por el modelo Random Forest. En el lado izquierdo de la figura, se representa la predicción del modelo en el entrenamiento y el valor True GRAY con un porcentaje de error (RMSE) del 16.5. En el lado derecho de la figura, se representa el conjunto de datos de prueba con un porcentaje de error (RMSE) del 39.88. La línea azul, representa la distancia promedio de error del conjunto de datos predichos respecto a los valores de fluorescencia de las imágenes.

Los resultados obtenidos están basados en los parámetros seleccionados para cada modelo buscando optimizar la arquitectura, es decir el nivel de diseño de la estructura, funcionamiento e interacción entre las partes de la programación. Donde el modelo seleccionado se manejaron opciones para el diseño de la arquitectura en donde se realizó la sintonización de los parámetros utilizando Grid search. Para este caso puntual, se decidió variar el número de estimadores (*n_estimators*) el cual indica la cantidad de árboles que tiene el modelo Random Forest, figura 58. El número de árboles que tenga el modelo depende del desempeño de las predicciones.

Figura 58.

Previsualización de código - Grid search para Random Forest .

```
tuned_parameters = [{'n_estimators':np.array ([10,50,100,500, 1000])}]
gsr= GridSearchCV(rgr, tuned_parameters, scoring=score_RMSE, refit=True, verbose=True)
```

Nota. La figura muestra la previsualización de la sección de código del ajuste de los parametros de Random Forest. La variable tuned_parameters, guarda el parámetro que genera el número de arboles (*n_estimators*) y los valores a probar (10, 50, 100, 500,1000). La variable gsr, almacena todos los parametros a evaluar por gridsearch que son: *rgr* (contiene el modelo a evaluar), *scoring* (estrategia para evaluar el rendimiento del modelo) , *refit* (reentrena el modelo utilizando los mejores parámetros hasta el momento que se evalua) y *verbose* (muestra la información que evalua del entrenamiento y el modelo).

Es importante observar que en todos los modelos probados el error en prueba siempre es mayor que en entrenamiento. Los datos del set de entrenamiento son los que el modelo está "viendo", por lo cual está dando la respuesta al modelo para que ajuste los parámetros hasta reproducirla. En cambio, la prueba o *test* es una prueba a ciegas del modelo resultante, por lo tanto el modelo no ha visto esos datos y por ende el error es más alto.

3.5.4. Resultados obtenidos de las métricas de regresión

En la siguiente tabla se muestra un resumen de los resultados de las métrica de regresión (distancia de error) para cada uno de los algoritmos evaluados anteriormente.

Tabla 11.

Métrica de los métodos de Machine Learning.

	Random Forest	Lasso	ElasticNet	Ridge	SVR	Gradient Bosting	MLP
RMSE	**40**	66	66	66	66	54	44
MSE	**1590**	4387	4355	4355	4342	2962	1934

Nota. Esta tabla muestra la comparación entre las métricas de regresión del conjunto de datos de prueba para los siete modelos de machine learning. La primera columna, tiene los dos tipos de errores calculados: error del cuadrado medio de la raíz (RMSE) y error cuadrático medio (MSE) . El valor más bajo calculado, se encuentra en la segunda columna que corresponde al modelo Random Forest con un error RMSE de 40 y un MSE de 1590 en comparación a los demás modelos.

Los modelos lineales mencionados en la sección 2.8.1, no son viables de implementar ya que los datos que se tienen son altamente no lineales y no se pueden predecir con modelos

de regresión multilíneal (Ridge Regressor, ElasticNet, Lasso). Esto sucede porque la resolución de los registros es bastante compleja como para poder solucionarlo con modelos básicos.

Existen efectos no lineales, que se deben capturar con otros modelos más robustos como Random Forest o Neural Network. El modelo Neural Network, requiere de una significativa cantidad de parámetros para lograr hacer una buena predicción pero que al tener una cantidad limitada de datos no le hace posible ejecutarse de manera óptima; mientras que el modelo Random Forest, cumple con un mejor balance de los datos disponibles en comparación con todos los demás modelos anteriormente descritos.

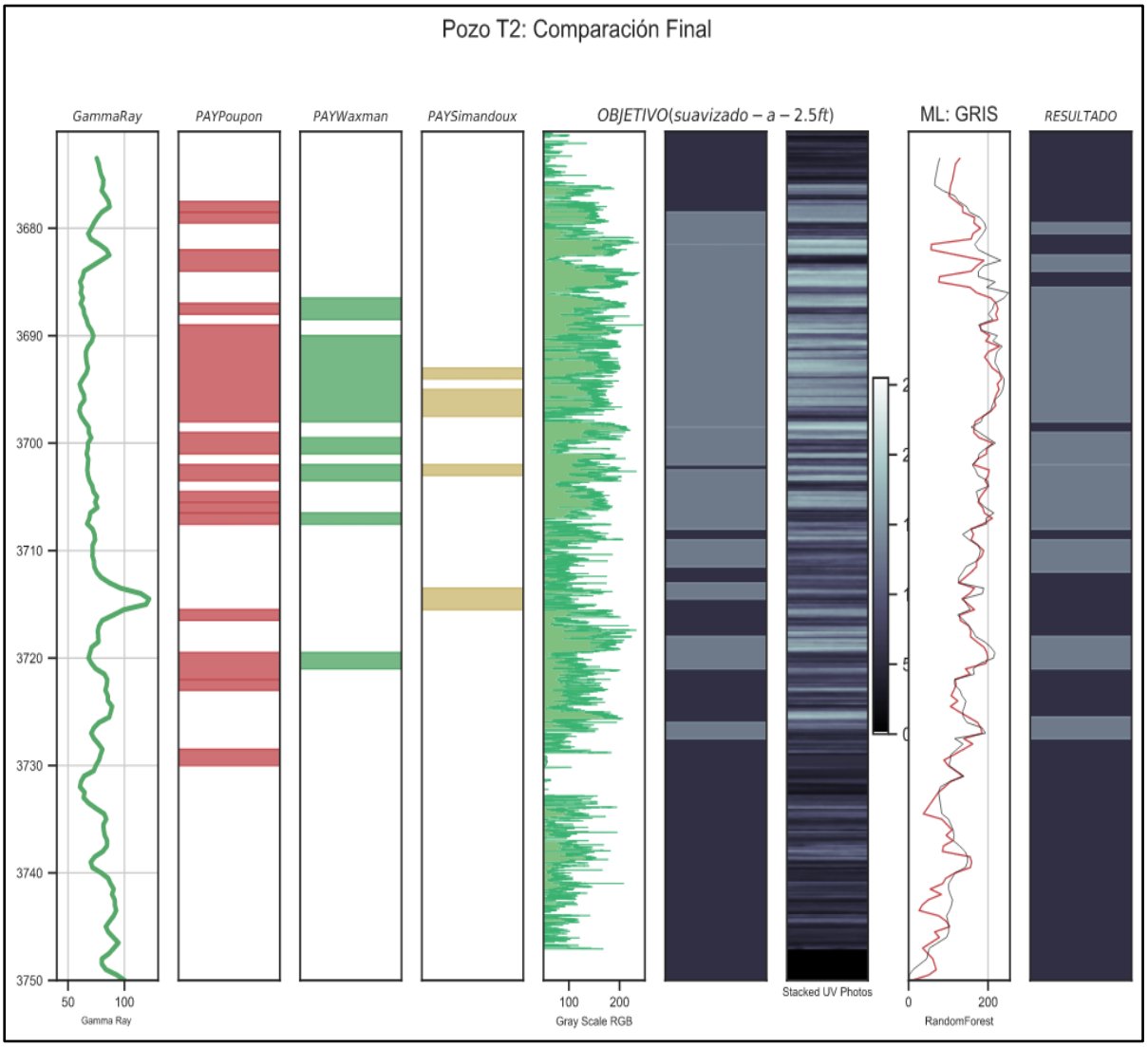
3.5.5. Comparación del mejor modelo de ML con modelos convencionales

Para esta sección, se tomó el pozo T2 (Pozo de prueba) ya que los pozos T6 y U18 se confirmaron similares geológicamente para poder aplicar el modelo a nivel regional. Las zonas que contienen hidrocarburos (*Pay Flags*) fueron calculadas a través de los métodos convencionales como Archie, Poupon, Simandoux y Waxman Smits. Estas zonas generadas por los métodos convencionales (Figuras 20-22) son comparadas con los resultados obtenidos con el mejor modelo de predicción Random Forest.

El porcentaje de los datos utilizados, para el conjunto de prueba fue del 70% tras haber recopilado todos los datos de los 3 pozos en un solo conjunto. Esto permite al modelo Random Forest dar los mejores resultados en comparación con los demás modelos implementados durante el proceso. El porcentaje de error de media cuadrática (RMSE) para este modelo, es un valor más bajo lo que indica un mejor ajuste de los datos. El RMSE es una muy buena medida de la precisión con la que el modelo predice la respuesta y es el criterio de ajuste más importante para el modelo.

Figura 59.

Resultados.



Nota. La figura representa las gráficas obtenidas del resultado final de la comparación del modelo Random Forest utilizando el 70% del total del set de datos para prueba vs. métodos convencionales. De izquierda a derecha, el primer *track* es el registro gamma ray en función de la profundidad, del segundo al cuarto *track* están los Pay calculados por los métodos convencionales (Poupon, Waxman-Smits y Simandoux). Para los siguientes tres *tracks*, está el registro en escala de grises, el objetivo suavizado a 2.5ft y el procesamiento imágenes en escala de grises de 0 a 255. Los dos últimos *tracks*, representan los resultados de la predicción en machine learning. El penúltimo *track*, contiene la predicción realizada con los registros y con Random Forest entrenado (línea roja), la línea negra representa el suavizado del registro en grises a 2.5 ft. Para el último *track,* se muestra el resultado de las zonas Pay y no Pay luego de haber sido probado y validado en el pozo T2.

3.5.5.a.Resultados del error de media cuadrática del Pay Neto calculado. Para esta sección, se muestran los resultados de RMSE según el Pay neto calculado por los métodos convencionales y el modelo de machine learning. El procesamiento de imágenes (True_UV) , tiene un % de error de 0 porque es el valor real al que se puede aproximarse los métodos o el modelo como se muestra en la siguiente tabla:

Tabla 12.

Error de media cuadrática.

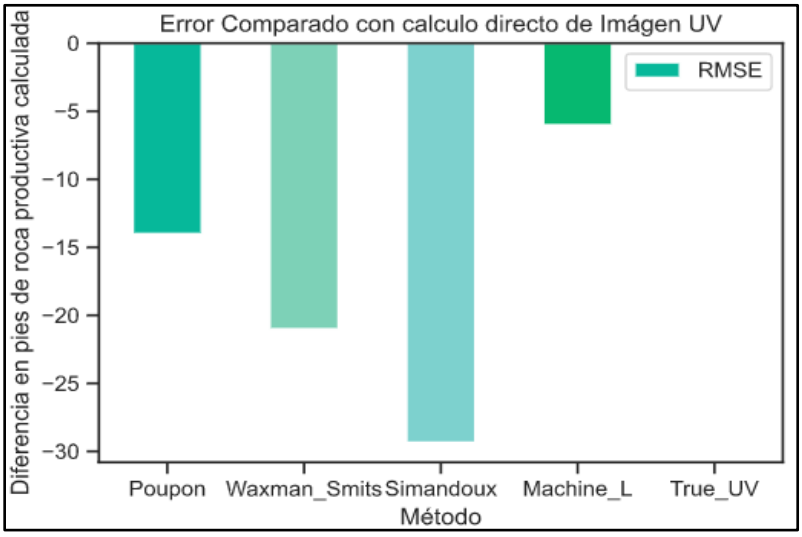
	Poupon	Waxman-Smits	Simandoux	Machine_L	True_UV
FT	19.67	12.67	4.33	27.67	33.67
RMSE	- 14	- 21	- 29.34	- 6	0

Nota. Esta tabla muestra los valores del error de media cuadrática (RMSE) y el total de Pay por los 3 métodos convencionales, ML y el procesamiento de las imágenes UV. El valor verdadero de fluorescencia (*True_UV*) es el valor de referencia de comparación para los modelos de predicción. En la tabla, el valor más cercano al verdadero fue el de Machine_L con un valor de -6 seguido por el método de Poupon con un valor de -14.

En la tabla anterior, se observa los resultados obtenidos para los métodos convencionales y el modelo de machine learning de mejor ajuste. Los métodos convencionales y el modelo de ML subestiman la cantidad de arena saturada comparado con True_UV (valor verdadero de fluorescencia de las imágenes -GS suavizado). El modelo (Random Forest) de ML, refleja que es la mejor solución al momento de trabajar formaciones de arena intercalada con arcilla por el valor tan aproximado al valor verdadero.

Figura 60.

Error comparado de los Modelos de ML y métodos convencionales con el cálculo de imagen UV.



Nota. La figura representa el error de los modelos de machine learning y de los métodos convencionales en función de los Pay totales calculados comparados con los cálculos de las imágenes UV. De izquierda a derecha de la figura, se observa que el método de Poupon como es el que más zonas Pay cálculo durante el proceso y que mantiene el menor error producido en comparación con los otros dos métodos convencionales. La cuarta barra, corresponde al modelo de ML que tiene el menor error y que se acerca a 0.

4. CONCLUSIONES

Una de las fases más importantes de este proyecto fue el condicionamiento, procesamiento, y apilamiento de las imágenes UV. En este proceso se logró obtener un registro en escala de grises (0 (negro) - 255 (blanco)) para identificar zonas Pay (reales) por pozo siendo comparado con los métodos convencionales y con el modelo de Machine Learning seleccionado.

El procesamiento completo de las imágenes UV, se realiza una sola vez sin importar el número de fotos que hayan disponobles por pozo. Todas las imágenes de corazones son procesadas de la misma forma a través del algoritmo que se creó y esta disponible al publico [84].

El escalado de los datos registró oscilaciones de la imagen UV en un rango de 20 a 210 en la escala de grises mientras que el filtrado (suavizado) disminuyó estas oscilaciones a un rango de 40 a 160 para poder ser leida por los modelos de ML. Visualmente, la distribución de datos de la imagen UV registra un valor de 0.010 y 0.003 de las zonas opacas y brillantes respectivamente. Luego del balance, los valores disminuyen a 0.004 y 0.005 obtiendose una distribución multimodal (homogenea) para una mejor predicción de los modelos de ML de las zonas con hidrocarburos.

Se demuestra que los métodos convencionales identifican menor cantidad de zonas con hidrocarburos (“Pay”) debido a la supresión real de los registros de resistividad que impactan en los cálculos de la saturación de agua. El método de Archie no identifica ninguna zona Pay (0 ft) mientras que los métodos de Poupon, Waxman- Smits y Simandoux identifican más zonas en arenas con precencia de arcilla con valores de 19.67,12.67 y 4.33 ft, respectivamente. Esto demuestra que Archie no es una opción viable al momento de evaluar formaciones complejas de estudio.

El modelo selecciónado de Machine learning fue Random Forest logrando subestimar el Pay en tan solo 6 pies (27.6 pies calculado vs 33.6 pies real) o 18% de error, mientras que el mejor modelo petrofísico convencional (Poupon) logró subestimar el Pay en 14 pies -es decir un error 24% más alto en Pay respecto al modelo de Machine Learning.

Se demostró que el uso de Machine Learning logra predecir la existencia de hidrocarburos en formaciones complejas de estudio únicamente con el uso de los registros de pozo básicos (GR, RHOZ, NPHI, AT90, DTCO) y fotos de corazones en luz UV. Las compañías petroleras no tendrán que invertir en numerosas y costosas pruebas de laboratorio para la interpretación petrofísica convencional.

Para el conjunto de prueba con un tamaño de datos del 70% sobre el total, se mostró que no necesariamente el modelo de ML más complejo es el mejor. El modelo de Random Forest registró un error de media cuadratica (RMSE) de 40 mostrando un mejor ajuste en comparación con el modelo de Neural Network con un RMSE de 44. Tambien se muestra que el modelo SVR a pesar de ser un modelo robusto registró un valor de RMSE (66) igual al de los modelos lineales.

BIBLIOGRAFÍA

[1] P.L. Decker, "Nanushuk Formation Discoveries: World-class exploration potential in a newly proven stratigraphic play, Alaska North Slope" AAPG ACE, pp.9, mayo 2017.

[2] D. W. Houseknecht, K J. Whidden, C. D. Connors, R. O. Lease, C. J. Schenk, T. J. Mercier, W. A. Rouse, P. J. Botterell, R. A. Smith, M. M. Sanders, W. H. Craddock, C. A. DeVera, C. P. Garrity, M. L. Buursink, C. O. Karacan, S. J. Heller, T. E. Moore, J. A. Dumoulin, M. E. Tennyson, K. L. French, C. A. Woodall, R. M. Drake II, K. R. Marra, T. M. Finn, S. A. Kinney, and C. M. Shorten, "Assessment of undiscovered oil and gas resources in the central North Slope of Alaska, 2020";U.S. Geological Survey,Reston VA, Report, Rep. 2020-3001, 2020.

[3] A. C. Huffman, jr. "Introduction to the Geology of the Nanushuk Group and Related Rocks, North Slope, Alaska", United States Geological Survey bulletin: v. 129 p, pp. 1-6, 1641.

[4] S. Cray, S. Jacobsen, J. Rasmus, R. Spaeth, "Effect of Resistive Invasion on Resistivity Logs" SPE, SPE 71708, pp. 1-3, octubre 2001.

[5] D. F Castellanos; "técnicas para determinar la distribución de la saturación de aceite remanente durante el periodo de producción primaria de un yacimiento", Tesis, universidad Industrial de Santander UIS, Bucaramanga, Col, 2008.

[6] C. Y. Sánchez; ""Evolución de los registros de resistividad y su aplicación en la estimación de la saturación de fluidos (agua e hidrocarburos)". tesis, UNAM; Ciudad de México, MX; 2012.

[7] M. A. Andersen, B. Duncan, R. McLin, "Core Truth in Formation Evaluation", Oilfield Review, vol.25, no. 2, pp. 19-20, 2013.

[8] C.V. Zertuche, " Cálculo de saturación de agua en yacimientos de hidrocarburos por medio de una herramienta dieléctrica de última generación", Tesis, ESIA, México, Ciudad de México, 2017.

[9] E. R. Crain (01, Jan, 2015). laminated reservoirs. [En línea]. Disponible en: https://www.spec2000.net/17-speclam.htm

[10] G.S Barrero, "Construcción del modelo de saturación de agua en un yacimiento de crudo pesado en la formación Mirador con agua de formación dulce", trabajo fin de máster, Departamento de Procesos y Energía, U. Nacional, Medellín, Colombia,2016.

[11] C.G Mull, "Cretaceous Tectonics, Depositional Cycles, and the Nanushuk Group, Brooks Range and Arctic Slope, Alaska", U.S Geological Survey Bulletin, v.129p, pp.7-9,1641.

[12] M. A. Martínez. "Desarrollo y aplicaciones de los registros acústicos". Tesis, UNAM, MX D.F. México, 2012.

[13] Br.A. Urdaneta, “Estimación de la curva de saturación de agua a partir de registros de pozos”, Tesis de grado, Universidad de Los Andes Mérida, Venezuela,2009.

[14] Mkinga, O.J., Skogen, E. & Kleppe, J. “Petrophysical interpretation in shaly sand formation of a gas field in Tanzania”. J Petrol Explor Prod Technol, v10, pp.1207-1213 (2020). https://doi.org/10.1007/s13202-019-00819-x

[15] M. Madrid. (2017, jul 10). Saturación de Fluidos en Yacimiento [en línea]. Disponible en:https://www.portaldelpetroleo.com/2017/07/saturacion-de-fluidos-en-yacimiento_10.html

[16] C. Krall, “Colores html y css. rgb decimal o porcentual. códigos de colores hexadecimales”, $apr^2.com$, nº19,2019.

[17] H.M. Quinche, “Los registros eléctricos y sus aplicaciones en minería y obras civiles”, Log-Tech, Perú, Marzo 2019.

[18] Gong, B., Keele, D., Toumelin, E., & Clinch, S. “Estimating Net Sand from Borehole Images in Laminated Deepwater Reservoirs with a Neural Network”. PETROPHYSICS vol.60, No. 5, pp. 596–604, October 2019. DOI: PJV60N5-2019a4

[19] T. Bradner. (22, mayo, 2019). Not all Slope wells are gushers [En línea]. Disponible en: https://www.anchoragepress.com/news/not-all-slope-wells-are-gushers/article_eaab4a9a-7c9f-11e9-bf39-37ab681ec828.html

[20] Difracción de rayos X (XRD)", *Sgs.co*, 2021. [En línea]. Disponible en: https://www.sgs.co/es-es/mining/metallurgy-and-process-design/high-definition-mineralogy/x-ray-diffraction-xrd.

[21] P. Decker, "Brookian Topset Stratigraphic Play: Petroleum Systems Elements", Alaska Geological Society Meeting, December 2018. Disponible en : https://dog.dnr.alaska.gov/Documents/ResourceEvaluation/20181213_BrookianTopsetStratPlay_PetrolSysElements_AGS.PDF

[22]. P. Decker, “Nanushuk Formation Discoveries: World-class exploration potential in a newly proven stratigraphic play, Alaska North Slope”, Alaska Geological Society Meeting, Mayo 2018.

[23] J. M. Coleman, D. B. Prior, 1982. "Deltaic Environments of Deposition", Sandstone Depositional Environments, Peter A. Scholle, Darwin Spearing.

[24] C.V. Zertuche, “ Cálculo de saturación de agua en yacimientos de hidrocarburos por medio de una herramienta dieléctrica de última generación”, Tesis, ESIA, México, Ciudad de México, 2017.

[25] S. limited, "Deltaico", *Schlumberger Oilfield Glossary*, 2021. [En línea]. Disponible en: https://glossary.oilfield.slb.com/es/terms/d/deltaic.

[26] Giosan, L., and Goodbred, S.L., 2007. Deltaic Environments. In Elias, S.A. (ed.), Encyclopedia of Quaternary Science. Elsevier, p. 704-716.

[27] Simon P. Neill, M. Reza Hashemi, in Fundamentals of Ocean Renewable Energy, 2018

[28] BitDegree. (25. Nov 2019). "Splitting Datasets with the Sklearn train_test_split Function". [En línea]. Disponible: https://www.bitdegree.org/learn/train-test-split#what-sklearn-and-model-selection-are

[29] X. Ying, «An Overview of Overfitting and its Solutions, » Research Gate, pp. 1-7, 2019.

[30] Kenneth J. Bird, Cornelius M. Molenaar (1992). "The North Slope Foreland Basin, Alaska: Chapter 13".

[31] "Hidrocarburos | Textos Científicos", *Textoscientificos.com*, 2005. [En línea]. Disponible en: https://www.textoscientificos.com/quimica/hidrocarburos.

[32] Caballero, C.. Parasecuencias (y Secuencias Deposicionales) [Diapositivas de PowerPoint]. Recuperado de http://usuarios.geofisica.unam.mx/cecilia/CT-SeEs/63Paraseq.pdf

[33] Steffens J, Landulfo E, Courrol LC, Guardani R. Application of fluorescence to the study of crude petroleum. J Fluoresc. 2011 May;21(3):859-64. doi: 10.1007/s10895-009-0586-4. Epub 2010 Jan 29. PMID: 20111988.

[34] K. Inverarity, "lasio", *PyPI- Read/write well data from Log ASCII Standard (LAS) files*, 2021. [En línea]. Disponible: https://pypi.org/project/lasio/.

[35] J.E. Torres, L.D.Cupil.(05. Febrero 16).Clasificación de tipos de registro.[En línea]. Disponible: https://www.academia.edu/22096369/Clasificaci%C3%B3n_de_tipos_de_registro_Petrof%C3%ADsica_y_registros_de_pozos?auto=download

[36]"Python for Loop", *Programiz.com*, 2021. [En línea]. Disponible: https://www.programiz.com/python-programming/for-loop.

[37] P. Decker. (13, Dic 2018),"Brookian Topset Stratigraphic Play: Petroleum Systems Elements",[En línea]. Disponible: https://dog.dnr.alaska.gov/Documents/ResourceEvaluation/20181213_BrookianTopsetStratPlay_PetrolSysElements_AGS.PDF

[38] IBERDROLA, "Descubre los principales beneficios del 'Machine Learning",[En línea]. Disponible: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

[39] Crain, P. Eng. Crain's Petrophysical Handbook [online]. 2015. WATER RESISTIVITY FROM CATALOG OR DST RECOVERY.. Available from: https://www.spec2000.net/05-5rwtemperature.htm

[40] M. Usman and A. Haris, "Reservoir Characterization Sandstone Reservoir Based on Wireline Log", IOP Conference Series: Materials Science and Engineering, vol. 546, p. 2, 2019. Available: https://iopscience.iop.org/article/10.1088/1757-899X/546/7/072011.

[41] Ho, Tin Kam (1998). The Random Subspace Method for Constructing Decision Forests

[42] O. Rotimi, B. Ako and Z. Wang, "Reservoir Characterization and Modeling of Lateral Heterogeneity Using Multivariate Analysis", *Energy Exploration & Exploitation*, vol. 32, no. 3, p. 530, 2014. Disponible en: https://journals.sagepub.com/doi/abs/10.1260/0144-5987.32.3.527.

[43] J. Amat(Nov 20),"Regularización Ridge, Lasso y Elastic Net con Python",[En línea]. Disponible: https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html

[44] J.Avendaño, "Análisis de modelos petrofísicos para formaciones clásticas", (tesis), ESIA,México,2015.

[45] C. Forero, J. Riveros, "VALIDACIÓN DEL VOLUMEN TÉCNICO PETROLÍFERO DE LA SUB-UNIDAD PRODUCTORA C7-X DEL CAMPO ASTRO EN LA CUENCA DE LOS LLANOS ORIENTALES", tesis de grado, Facultad de ingenierías, F.U.A, Bogotá, Col, 2018.

[46] ScikitLearn. "Linear Models", [En línea]. Disponible: https://scikit-learn.org/stable/modules/linear_model.html#elastic-net

[47] ScikitLearn. "sklearn.linear_model.Lasso", [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

[48] Scikit-learn. "RandomForest Regressor ", [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[49] Scikit-learn. , "sklearn.linear_model.ElasticNet", [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

[50] Scikit-learn., "Metrics and scoring: quantifying the quality of predictions", [En línea]. Disponible: 3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.24.2 documentation

[51]"Los Minerales Fluorescentes", *Foro de Minerales*, 2015. [En línea]. Disponible: https://www.forodeminerales.com/2015/05/los-minerales-fluorescentes.html.

[52] T. Oliphant, "What is NumPy? — NumPy v1.20 Manual", *Numpy.org*, 2021. [En línea]. Disponible en : https://numpy.org/doc/stable/user/whatisnumpy.html.

[53] “glob — Unix style pathname pattern expansion — Python 3.9.5 documentation", *Docs.python.org*, 2021. [En línea]. Disponible: https://docs.python.org/3/library/glob.html.

[54] J. Hunter, D. Dale, E. Firing, M. Droettboom and e. al., "Matplotlib: Python plotting — Matplotlib 3.4.2 documentation", *Matplotlib.org*, 2012. [En línea]. Disponible: https://matplotlib.org/

[55] RandomForest. (08, mayo, 2013). “Definición Random Forest”. [En línea]. http://randomforest2013.blogspot.com/2013/05/randomforest-definicion-random-forests.html

[56] Scikit-learn., “Support Vector Machines“, [En línea]. Disponible: https://scikit-learn.org/stable/modules/svm.html

[57] Scikit-learn. , “sklearn.ensemble.GradientBoostingRegressor“, [En línea]. Disponible:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[58] Scikit-learn. , “sklearn.neural_network.MLPRegressor“, [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

[59] D. Mesquita , “Python AI: How to Build a Neural Network & Make Predictions “, [En línea]. Disponible: https://realpython.com/python-ai-neural-network/

[60] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

[61] Scikit-learn., “Neural network models (supervised) “, [En línea]. Disponible: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[62] "os — Interfaces misceláneas del sistema operativo — documentación de Python - 3.10.0b3", Docs.python.org, 2021. [En línea]. Disponible: https://docs.python.org/es/3.10/library/os.html.

[63] J. Castillo, "RGB qué es esto y para qué se utiliza en informática", Profesional Review, 2019. [En línea]. Disponible: https://www.profesionalreview.com/2019/01/20/rgb-que-es/.

[64] AFPD, R., 2019. *¿Cuáles son los tipos de algoritmos del machine learning?*. [En línea] APD España. Disponible en: https://www.apd.es/algoritmos-del-machine-learning

[65] N. de Juárez, "¿Qué es la conductividad?", HACH COMPANY, México, circuito científico, 22, 2017.

[66] I. Education, "What is Machine Learning?", Ibm.com, 2020. [En línea]. Disponible: https://www.ibm.com/cloud/learn/machine-learning.

[67] F. Pedregosa, G. Varoquaux and e. al., "Choosing the right estimator — scikit-learn 0.24.2 documentation", Scikit-learn.org, 2011. [En línea]. Disponible: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

[68] J. Heras, "Las 7 Fases del Proceso de Machine Learning - IArtificial.net", IArtificial.net, 2020. [En línea]. Disponible: https://www.iartificial.net/fases-del-proceso-de-machine-learning/.

[69] seaborn "Seaborn. Pairplot", [En línea]. Disponible:https://seaborn.pydata.org/generated/seaborn.pairplot.html

[70] S. limited, "resolución vertical", *Schlumberger Oilfield Glossary*, 2021. [En línea]. Disponible en: https://glossary.oilfield.slb.com/es/terms/v/vertical_resolution

[71] M. Brett, "An introduction to smoothing — Tutorials on imaging, computing and mathematics", Matthew-brett.github.io, 2016. [En línea]. Disponible en: https://matthew-brett.github.io/teaching/smoothing_intro.html.

[72] Khan Academy, "Introducción a la estadística: media, mediana y moda", [En línea]. Disponible en: https://es.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/mean-and-median/v/statistics-intro-mean-median-and-mode

[73] Greelane, (03 Sep, 2018) "¿Cuáles son el máximo y el mínimo de un conjunto de datos?",[En línea]. Disponible en: https://www.greelane.com/es/ciencia-tecnolog%c3%ada-matem%c3%a1ticas/mates/what-are-the-maximum-and-minimum-3126236/

[74] F. Pedregosa, G. Varoquaux and e. al., "sklearn.preprocessing.StandardScaler", Scikit-learn.org, 2011. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[75] A. Santacruz, "Por qué es importante trabajar con datos balanceados para clasificación", *Amsantac.co*, 2016. [En línea]. Disponible en:http://amsantac.co/blog/es/2016/09/20/balanced-image-classification-r-es.html.

[76] *MinMaxScaler | Interactive Chaos*. (2021). Home page | Interactive Chaos. [En línea]. Disponible en: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/minmaxscaler

[77] MapMaker Interactive. [En línea]. Disponible en: https://mapmaker.nationalgeographic.org/eQzLQJf2nMH7AMoMDrFRnx//?edit=hxaqMVdFYb2soXutWTZQtt

[78] MapMaker Interactive. [En línea]. Disponible en: https://mapmaker.nationalgeographic.org/ccQJpfu9H6bWim3l85ZN5e//?edit=fZwtreyD4khUC3356wGeA2

[79] "LAS Format", Usgs.gov, 2021. [En línea]. Disponible en: https://www.usgs.gov/core-science-systems/national-geological-and-geophysical-data-preservation-program/las-format.

[80] Crain, P. Eng. Crain's Petrophysical Handbook 2015. WATER SATURATION FROM WAXMAN-SMITS METHOD. [En línea]. Disponible en: https://www.spec2000.net/14-swws.htm

[81] Crain, P. Eng. Crain's Petrophysical Handbook 2015, WATER SATURATION FROM SIMANDOUX METHOD. [En línea]. Disponible en: https://www.spec2000.net/14-sws.htm

[82] Scikit-learn, "sklearn.model_selection.GridSearchCV", [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[83] D. Ellis (USA), J. Singer (UK), *Well Logging for Earth Scientists.* 2nd edition. y Elsevier NY, 2008.

[84] GitHub, edwino26, Repositorio de GitHub - Código Machine Learning, [En línea]. Disponible en: https://github.com/edwino26/CoreImages/blob/master/ML_EO.py

[85] Roman, V. (2019, 12 de junio). *Aprendizaje No Supervisado en Machine Learning: Agrupación*. Médium. [En línea] Disponible en : https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc

[86] Manual Estructuración del Trabajo de Grado. Fundación Universidad de América, 2021. [PDF].

RECOMENDACIONES

En base a los resultados y conclusiones expuestas, presentamos las siguientes recomendaciones:

Usar la herramienta de procesamiento de fotos de núcleo que fue creada y se hizo disponible al público en Github para que futuros trabajos en esta área tengan una base. (GitHub, edwino26, Repositorio de GitHub - Código Machine Learning, [En línea]. Disponible en: https://github.com/edwino26/CoreImages/blob/master/ML_EO.py) .

Recopilar la mayor cantidad de información del area de estudio para obtener una base de datos más amplia y generar resultados más aproximados al implementar un modelo de predicción.

Establecer diferentes subconjuntos de división de entrenamiento y prueba de los datos en base a los criterios de selección y la información recopilada para obtener una mayor ampliación de resultados para una óptima elección.

Implementar la metodología descrita a lo largo del trabajo para lograr optimizar tiempos e inversiones de las compañías en caracterización de rocas complejas en diferentes campos de estudio.

Crear una herramienta digital (software, aplicativo, etc) que facilite el uso del modelo que creamos de Machine learing (disponible en Github) a cualquier persona sin conocimientos previos en programación y poderlo implementar en cualquier parte del mundo.