IDENTIFICACIÓN AUTOMÁTICA DE FACIES LITOLÓGICAS DE UNA SECUENCIA SEDIMENTARIA BASADO EN REGISTROS DE POZO

MAGÍSTER EN MATEMÁTICAS APLICADAS Y CIENCIAS DE LA COMPUTACIÓN

Tomas Andres Montealegre Pallares

Dirección:
Dr. John Jairo Villarejo Mayor

Universidad del Rosario
Escuela de Ingeniería, Ciencia y Tecnología
Maestría Matemáticas y Ciencias de la computación
2023

Agradecimientos

Quiero expresar mi profundo agradecimiento a mi madre, hermano y tía Esperanza por su invaluable apoyo a lo largo de mi tesis de maestría. Vuestra presencia constante, palabras de aliento y correcciones constructivas han sido fundamentales en mi camino hacia el logro de este objetivo.

A mi querida madre, mi mayor fuente de inspiración, te agradezco de corazón por estar siempre a mi lado, impulsándome a seguir adelante y celebrando mis triunfos. Tus palabras sabias y tu amor incondicional han sido un bálsamo en los momentos difíciles y me han dado la confianza para enfrentar cualquier desafío.

A mi hermano, aunque nuestras interacciones sean limitadas, siempre he sentido tu respaldo incondicional cuando te he necesitado. Tu disposición para ayudarme en momentos clave ha sido invaluable. Agradezco tu presencia silenciosa pero reconfortante.

Y a mi tía Esperanza, quiero expresar mi profundo agradecimiento por tu apoyo y preocupación constante por mi bienestar. Tus contribuciones han allanado el camino hacia mi éxito académico y me han brindado la tranquilidad necesaria para concentrarme en mis estudios.

Vuestra dedicación ha sido mi mayor impulso hacia el éxito. Sin ustedes, este logro no habría sido posible. Estoy profundamente agradecido por vuestra presencia constante en mi vida y por todo lo que han hecho por mí.

Resumen

La identificación precisa de la litología es esencial en la caracterización de yacimientos, ya que impacta significativamente la calidad de los yacimientos de petróleo y gas. La convencional interpretación manual de los datos de registro de pozo requiere un volumen masivo de datos y es subjetiva al depender de la experiencia del geofísico. En los últimos años se han desarrollado métodos automáticos basados en inteligencia artificial para identificar la litología mediante el análisis de los registros de pozos. No obstante, muchos de estos enfoques utilizan valores de una sola medición y tienen dificultades para distinguir las características de respuesta de las litologías, lo que lleva a predicciones inexactas. Este estudio tiene como objetivo desarrollar un modelo de aprendizaje automático efectivo para la clasificación de facies litológicas en pozos. Se propusieron modelos de redes neuronales como CNN1D y LSTM para aprovechar la naturaleza secuencial de los registros. Además, se exploraron modelos ramificados que combinan diferentes tipos de redes neuronales, incluyendo un mecanismo de autoatención. Comparando estos modelos con los enfoques tradicionales KNN y FC basada en una única medición se encontró que el CNN1D fue más efectivo en términos de métricas de evaluación, superando las limitaciones de los enfoques basados en datos puntuales. Además, un análisis de importancia de características reveló que todos los registros de pozo son relevantes en la clasificación, destacando GR, RDEP, RMED y DTC como los más influyentes. La importancia asignada a estos registros en el modelo propuesto coincidió con la atención dada por un petrofísico experto durante su identificación manual. Los resultados obtenidos con los modelos propuestos presentan alternativas eficientes y satisfactorias para su aplicación en el campo de la industria de gas y petróleo.

Registros de Pozo, Litología, Petrofísica, Caracterización de Yacimientos, Eliminación de Ruido, Importancia de las Características, Aprendizaje Automático

Abstract

The accurate identification of lithology is crucial in the characterization of reservoirs as it significantly impacts the quality of oil and gas fields. The conventional manual interpretation of well log data requires a massive volume of data and it is subjective, relying on the expertise of geophysicists. In recent years, automated methods have been developed to identify lithology by analyzing well log data based on artificial intelligence. Nevertheless, many of these approaches rely on single-measurement values and struggle to distinguish the response characteristics of different lithologies, leading to inaccurate predictions. This study aims on developing an effective machine learning model for the classification of lithological facies in wells. Neural network models were proposed, including CNN1D and LSTM, which leverage the sequential nature of the well log data. Furthermore, branched models combining different types of neural networks, including a self-attention mechanism, were explored. Comparing these models with traditional approaches such as KNN and a single-measurement-based FC, it was found that the CNN1D outperformed others in terms of evaluation metrics, surpassing the limitations of point-based approaches. Additionally, an analysis of feature importance revealed that all well log measurements were relevant in the classification process, with GR, RDEP, RMED, and DTC standing out as the most influential. The importance assigned to these measurements in the proposed model aligns with the attention given by expert petrophysicists during manual lithology identification. This convergence between automated approaches and human expertise reinforces confidence in the model's effectiveness. The achieved results with the proposed models present efficient and satisfactory alternatives for their application in the field of the oil and gas industry.

Well Logs, Lithology, Petrophysics, Reservoir Characterization, Noise Removal, Importance of Features, Machine Learning

Índice

Lista de tablas

Lista de figuras

1. JUSTIFICACIÓN Y DESCRIPCIÓN DEL PROBLEMA

El consumo de hidrocarburos ha aumentado continuamente en los últimos años y se espera que esta tendencia continué en el futuro cercano [20]. Sin embargo, cada vez es más difícil encontrar nuevos yacimientos convencionales, lo que obliga a las compañías a aventurarse a la exploración a mayores profundidades, en zonas con condiciones climáticas, geográficas y técnicas cada vez más desafiantes. Esto lleva a explorar nuevos tipos de yacimientos y optimizar los que actualmente se encuentran en producción, los cuales deben ser caracterizados para la implementación de procesos más eficientes.

La caracterización de yacimiento tiene como fin construir un modelo que incorpore todas las características que influyen en su capacidad para almacenar y producir hidrocarburos. Estos modelos son utilizados para simular el comportamiento de los fluidos dentro del yacimiento en diferentes escenarios para encontrar las estrategias de producción más eficientes para maximizar la extracción de recursos. La litología se encarga del estudio de las características de las rocas y los sedimentos que forman la corteza terrestre. Siendo la identificación de estas características uno de los pasos más importantes para la caracterización de un yacimiento, debido a que proporciona información valiosa sobre la composición y estructura de las rocas en el área, lo que puede ayudar a determinar la presencia y cantidad de recursos naturales en el área, y puede ser útil para identificar obstáculos y desafíos en la extracción de los recursos naturales, así como para predecir su comportamiento en el yacimiento durante la extracción.

La fuente primaria de información litológica en profundidad consiste en la perforación de la roca por varios cientos o miles de metros. Durante esta operación, se extraen cilindros de roca conocidos como núcleos o corazones que son analizados en superficie. Sin embargo, por razones técnicas y financieras, los núcleos no son usualmente extraídos y cuando se extraen es solo de un intervalo de interés. De la misma forma, durante el proceso de perforación se producen fragmentos rocosos que son transportados a la superficie en el fluido de perforación. Debido a su tamaño y a que durante el transporte

se mezclan ripios provenientes de diferentes profundidades, aportan poca información sobre la litología. El método más común para la identificación litológica consiste en descender una sonda equipada con instrumentos que miden diversas propiedades físicas de la roca y deducir la litología en función de los datos obtenidos. Estos registros dan un reconocimiento prácticamente continuo de las formaciones atravesadas por el pozo.

La interpretación de estos registros no es una tarea trivial dado que las características de respuesta de los registros de pozo para algunas litologías diferentes son similares e indistinguibles. Una práctica común en la interpretación de registros consiste en trazar los valores de más de un registro dentro de una misma gráfica y realizar una evaluación cuantitativa. Este es un proceso laborioso que implica el análisis de un gran volumen de datos y demanda un tiempo considerable por parte de un analista de registros experimentado, lo cual conduce a una alta incertidumbre en la interpretación y a posibles artefactos introducidos por el intérprete.

La automatización de esta tarea mediante análisis estadístico para reducir los tiempos de interpretación y obtener modelos más confiables ha sido un interés desde finales de la década de 1980 [9] [6]. En la actualidad, la identificación automática de facies litológicas se realiza principalmente por dos tipos de modelos basados en el aprendizaje de maquina: modelos de datos puntuales y modelos de sección de datos (Figura 1). El modelo de datos puntuales analiza los valores de cada uno de los registros para una única profundidad dada, mientras que el modelo de sección de datos analiza los valores de los registros para un rango específico de profundidades adyacentes.

Los procesos geológicos que dan origen a las rocas son continuos y las condiciones ambientales que permiten el depósito de una roca en particular varían lentamente a lo largo del tiempo, Debido a esto, una roca contiene un patrón de información de rocas depositadas previamente y de las que se depositarán en un futuro. Un modelo que usa un único dato para predecir sobre una serie puede ser menos preciso que un modelo que usa un conjunto de datos adyacentes. Esto se debe a que un único dato no puede proporcionar toda la información necesaria para hacer una predicción precisa, ya que

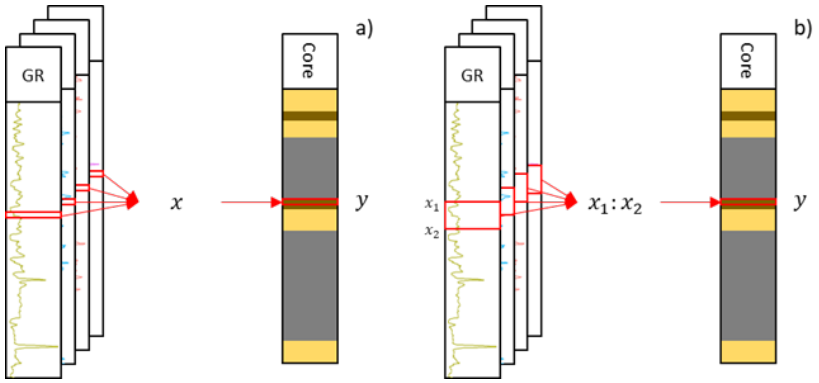

Figura 1: Comparación de tipos de modelos de aprendizaje de maquina: A) Modelos de datos puntuales, B) Modelo de sección de datos.

puede estar sesgado o no representar la tendencia general y patrones de la serie. Por esta razón y en conjunto a la similitud de la respuesta de los registros en diferentes litologías los modelos de datos puntuales pueden producir predicciones inexactas, donde, aunque la interpretación individual para cada una de las profundidades es válida, la secuencia de rocas como conjunto carece de sentido geológico. Debido a esto, los modelos de sección de datos son más apropiados que los modelos de datos puntuales para realizar una clasificación litológica.

Las señales obtenidas en profundidad de los registros de pozo son el insumo básico para la caracterización de rocas en profundidad. Sin embargo, aunque es sabido que estas señales contienen ruido y valores anómalos, en la mayoría de los casos de interpretación manual como automática se emplean los datos sin ningún tipo de condicionamiento. Son escasos los trabajos que han buscado caracterizar estos ruidos y valores anómalos, utilizando la transformada de Fourier o la transformada de Ondícula, para obtener datos geológicos más representativos [19].

De acuerdo con la revisión de literatura realizada por los autores, los modelos de datos puntuales son predominantes y solo existen unos escasos ejemplos de sección de datos [21] [34], Hasta la fecha, no se ha llevado a cabo ningún estudio que integre las metodologías de eliminación de ruido y detección de valores atípicos en un modelo de clasificación y predicción de litología. Además, no se ha considerado la importancia de cada registro en un conjunto específico de registros de pozo.

2. OBJETIVOS

2.1. Objetivo general

Proponer un modelo de identificación automática de diferentes facies litológicas de una secuencia sedimentaria siguiendo un contexto geológico, basado en un conjunto especifico de registros de pozo.

2.2. Objetivos específicos

1. Implementar técnicas de procesamiento de señales para la eliminación de ruido y detección de valores atípicos de registros de pozo.

2. Implementar modelos de identificación de facies litológicas basados en máquinas de aprendizaje, usando datos puntuales y secciones de datos.

3. Comparar los modelos de identificación de facies litológicas de acuerdo con el tipo de datos utilizados.

4. Determinar los registros de pozo más influyentes para identificar la litología de una roca en profundidad.

3. MARCO TEÓRICO

3.1. Litofacies

Una facie de roca o litofacie es un cuerpo de roca con características específicas. Puede ser una capa individual o un grupo de varias capas. Idealmente, debería ser una roca distintiva que se formó bajo ciertas condiciones de sedimentación, reflejando un proceso particular, conjunto de condiciones o ambiente de sedimentación [28]. Por lo tanto, una litofacie es un cuerpo de roca con unas características mineralógicas, granulométricas o deposicionales que permiten diferenciarlo de otros cuerpos rocosos adyacentes.

3.2. Pozo

Según [10] un pozo es un agujero que se perfora para descubrir o delimitar un depósito de petróleo y/o para producir petróleo o agua con fines de inyección, inyectar gas, agua u otro medio, o monitorear los parámetros del pozo. Hay varias categorías de pozos. Un pozo puede constar de uno o varios caminos de pozo y puede tener uno o varios puntos de terminación, pero para este estudio un pozo consta de un único camino y punto de terminación.

3.3. Registros de pozo

La mejor fuente de información de litofacies son las muestras de núcleos de roca de yacimiento de los pozos; sin embargo, los núcleos no se toman comúnmente debido a los costos [12]. Como alternativa a la extracción de núcleos, son necesarias otras técnicas para poder interpretar las litologías y características de un pozo. Actualmente la más utilizada son los registros de pozo, estos pueden ser descritos como un registro de las características de las formaciones rocosas atravesadas por un dispositivo de medición en el pozo [13].

Esta técnica consiste en descender por medio de un cable el pozo una sonda con diversos dispositivos de medición, algunas mediciones son pasivas mientras otras ejercen

algún cambio en el medio y miden la respuesta de este. Fue realizado por primera vez el 5 de septiembre de 1927, donde los hermanos Schlumberger hicieron una medición semicontinua de resistividad en un campo en Alsacia – Francia [17].

Algunas de las propiedades medidas dan un resultado un valor directo, pero en la mayoría de los casos es necesario realizar una interpretación de varios registros en conjunto para obtener un resultado, ya que, cada uno de los registros tiene sus limitaciones.

3.3.1. Caliper

El registro de Caliper (CALI) mide el diámetro en pulgadas (in) del pozo, el diámetro puede aumentar por el deslave o colapso de lutitas y rocas pobremente cementadas, o disminuir por la formación de un recubrimiento debido a la filtración de partículas de suspensión por una roca porosa, las respuestas típicas frente a diversas litologías son proporcionadas en Apéndice-Figura 10. Es utilizado para la corrección de medidas realizadas por otros registros afectados por el diámetro del pozo y como una aproximación a la litología identificando zonas porosas y permeables.

3.3.2. Potencial espontaneo

El registro de potencial espontaneo (SP) corresponde a la diferencia entre el potencial eléctrico de dos electrodos uno fijo en la superficie y uno móvil en pozo, en el rango de decenas a cientos de milivoltios (mV). Esta diferencia es atribuida a dos procesos que involucran el movimiento de iones principalmente Cl- y Na+ entre el fluido de perforación y el agua natural presente en las rocas. El fluido de perforación permea una roca porosa como una roca arenosa y dos soluciones de diferentes concentraciones entran en contacto, para balancear esta diferencia se produce una migración de iones desde las rocas hacia el pozo. Sin embargo, los iones no se mueven a la misma velocidad en un medio permeable. Los iones de Cl- migran más rápido que los iones de Na+, dando como resultado un flujo de carga negativa en dirección al pozo, generando una corriente eléctrica que fluye hacia las rocas esto es conocido como potencial de unión líquida [2], otra fuente corrientes eléctricas se presenta en la interfaz de una litología porosa y una lutita, actuando la lutita una membrana semipermeable que permite el

paso de iones Na+ pero no el de Cl-, produciendo así una migración de iones desde la capa permeable hacia el pozo por medio de la lutita, esto es conocido como potencial de membrana [2]. Sin embargo, este registro no puede ser medido en pozos rellenos de fluidos en base a aceite. El SP permite identificar presencia de hidrocarburos y capas permeables, las respuestas típicas son proporcionadas en Apéndice-Figura 11. Incluso una pequeña desviación en este registro indica que una capa tiene una permeabilidad razonable [16]. A partir de esta información es posible realizar una evaluación litología, correlacionar pozos y determinar la resistividad del agua que satura las rocas. Se ve afectado por presiones diferenciales entre los fluidos y la presencia de litologías con porosidades bajas.

3.3.3. Rayos Gamma

Los registros de rayos gamma (GR) miden la radioactividad natural de las rocas en grados API (gAPI). Principalmente son producidos por los isotopos ^{40}K, ^{232}Th y ^{238}U [31]. En una cuenca sedimentaria las rocas arenosas y carbonáticas contienen bajas concentraciones de material radioactivo y como resultado presentan lecturas bajas. Por otro lado, las lutitas presentan altas lecturas debido a su composición mineralógica. gracias a esto a partir GR se obtiene un índice de arcillosidad de la roca y una identificación inicial de la litología (Figura 2), que permite correlacionar la litología entre varios pozos. Sin embargo, la composición mineralógica de una roca arenosa puede variar o presentar mineralizaciones conteniendo altas concentraciones de material radioactivo, y algunas lutitas están compuestas por minerales no radioactivos, lo que puede causar interpretaciones erróneas del índice de arcillosidad cuando no es apoyado por otros registros. Diversos factores afectan la medición de GR como el diámetro del pozo, la posición del sensor con respecto al centro del pozo, el espesor de la capa, la densidad y composición del fluido de perforación, la cementación y revestimiento del pozo [31].

3.3.4. Resistivos

Los registros resistivos tales como resistividad la resistividad profunda (RDEP), resistividad intermedia (RMED), resistividad somera (RSHA), resistividad de la zona invadida (RXO), y micro resistividad (RMIC) miden la resistividad en ohmios por me-

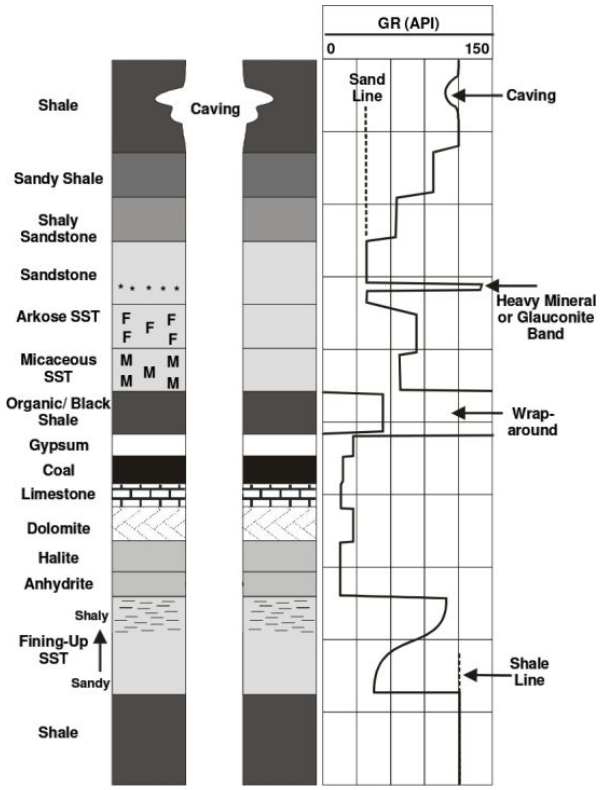

Figura 2: Efectos de diferentes litologías en el registo GR, tomado de [16].

tro (Ohm. m) usando uno o varios electrodos emisores envían una corriente eléctrica a las rocas adyacentes, estas corrientes se transmiten casi exclusivamente por los fluidos presentes a debido a que la mayoría de materiales rocosos son esencialmente aislantes mientras los fluidos atrapados en sus poros son conductores [29], y son recibidas por uno o varios electrodos receptores que miden la respuesta eléctrica, sus respuestas típicas son proporcionadas en Apéndice-Figura 12, variando la distancia entre los electrodos emisores es posible obtener información de diversas zonas del pozo, aumentando esta distancia se obtienen mediciones más alejadas del pozo a cambio de resolución vertical, de igual forma disminuyendo la distancia emisor – receptor se obtienen mediciones cercanas al pozo con una alta resolución vertical. Los registros resistivos tienen diversas aplicaciones. Mediante la comparación de RDEP, RMED y RSHA permite discernir entre zonas que contienen hidrocarburos y zonas que contiene agua, calcular la saturación de agua de la roca, correlaciones litológicas entre pozos, identificación de rocas productoras de hidrocarburos entre otros [16].

3.3.5. Acústicos

Los registros sónicos o acústicos como los de onda de cizalla (DTS) y de onda compresional (DTC) miden el tiempo de tránsito de una onda elástica. Para lograr esto, un transductor emisor que convierte la energía eléctrica en energía mecánica genera pulso muy corto y de gran amplitud. Este pulso viaja por la roca siendo dispersado y atenuado en el proceso hasta un transductor receptor realizando el proceso inverso obteniendo una señal eléctrica, el tiempo medido comienza desde la emisión del pulso hasta la recepción del primer arribo de la onda. Al propagarse por el medio los distintos tipos de onda viajan a diferentes velocidades y no toman los mismos caminos. Esto se debe a las diferentes direcciones de propagación de las ondas frente a las direcciones de desplazamiento. La dirección de propagación para las ondas compresionales es paralela a la dirección del desplazamiento de las partículas, esto permite su propagación en medios gaseosos, líquidos y sólidos. Por otro lado, para las ondas de cizalla su dirección de propagación es perpendicular a la dirección de desplazamiento de las partículas, por lo que su propagación está restringida a medios sólidos [2]. La información es registrada como el tiempo de viaje por pie recorrido (ms/ft). Sus respuestas típicas son proporcionadas en Apéndice-Figura 13. Son usados para la construcción de sismogramas sintéticos, calibración de información sísmica, determinación de la porosidad de una roca, correlaciones estratigráficas entre pozos, identificación de litologías, compactación de sedimentos, identificación de zonas de sobrepresión.

3.3.6. Factor fotoeléctrico

El factor fotoeléctrico (PEF) es un registro continuo del índice de sección transversal de absorción fotoeléctrica efectiva (Pe) de las rocas. El índice de absorción es fuertemente dependiente del número atómico promedio, Z, la complejidad atómica de los constituyentes de las rocas, lo que implica la composición y por inferencia, la litología, las respuestas más comunes son proporcionadas en Apéndice-Figura 14. El uso de este registro se encuentra severamente restringido por el hecho de que es ineficaz en pozos donde el fluido de perforación contiene barita, puesto que el índice de absorción fotoeléctrica es cerca de 150 veces el de la mayoría de los minerales más comunes presentes en las rocas [2]. Este registro es sensible solo al número atómico promedio de la

formación y es insensible a los cambios en la porosidad y la saturación de fluidos en la roca, eso lo hace una de las mejores aproximaciones al momento de determinar la litología de un pozo [16].

3.3.7. Densidad

Los registros de densidad como el registro de porosidad neutrón (NPHI) proveen un registro continuo de la respuesta de las rocas al bombardeo de neutrones rápidos en el rango de MeV, perdiendo energía colisionando con núcleos, esta pérdida de energía es máxima cuando colisiona con núcleos de la misma masa, la energía se reducirá al nivel de energía térmica (≈0.025eV) y el neutrón será absorbido por un núcleo emitiendo rayos gamma. los materiales modifican los neutrones rápidamente cuando contienen altas concentraciones de núcleos de hidrogeno, por tener la misma masa de los electrones, los cuales en el contexto geológico son aportados por agua (H_2O) atrapada en los poros, por lo tanto, el registro mide principalmente el contenido de agua de la roca, sus respuestas típicas son proporcionadas en Apéndice-Figura 15. Estos registros son utilizados para medir la porosidad de una roca, es un excelente discriminador entre gas y petróleo. Puede ser usado para identificar litólogas, evaporitas, minerales hidratados y rocas volcánicas[29].

3.4. Registros de pozo

Desde un punto de vista estadístico, los valores atípicos son puntos de datos que son significativamente diferentes de la tendencia general del conjunto de datos. Desde un punto de vista conceptual, una muestra se considera atípica cuando no representa el comportamiento del fenómeno/proceso representado por la mayoría de las muestras en un conjunto de datos. Los valores atípicos son indicativos de problemas en el procedimiento de recopilación/medición de datos o eventos inesperados en la operación/proceso que generó los datos [25]. Cuando un modelo es construido sobre datos que contienen valores atípicos su precisión y capacidad de generalizar disminuyen ya que el modelo identifica patrones no representativos de los datos presentes en los valores anómalos, por eso eliminar estos valores es una tarea primordial antes de empezar a entrenar un modelo.

3.4.1. Isolation Forest

Bosque de aislamiento o Isolation Forest (IF) es un algoritmo para la detección de datos anómalos construido sobre las bases de los árboles de decisión, identifica explícitamente las anomalías asumiendo que los datos anómalos tienden a encontrarse dispersos y se encuentran en zonas poco pobladas del espacio de características, frente a los valores típicos que siguiendo las tendencias generales se encuentran en zonas densamente pobladas. Para cumplir su tarea crea arboles aleatorios donde las particiones son creadas seleccionando aleatoriamente una característica y aleatoriamente selecciona un valor de corte entre el valor máximo y mínimo de la característica seleccionada, estas particiones son realizadas hasta que cada uno de los datos se encuentra aislado.

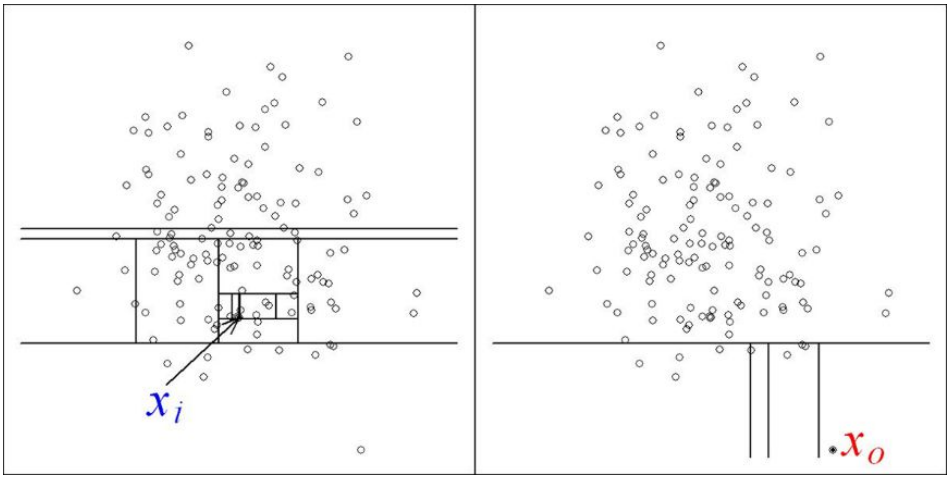

Figura 3: Aislamiento de dato típico vs atípico, tomado de [22].

En [22] se afirma que esta partición aleatoria produce caminos más cortos para las anomalías ya que la menor cantidad de instancias de anomalías da como resultado un número menor de particiones, y es más probable que los casos con valores alejados de la tendencia sean separadas en durante las primeras particiones (Figura 3). Por lo tanto, cuando un bosque de árboles aleatorios produce colectivamente longitudes de camino más cortas para algunos puntos en particular, es muy probable que se trate de anomalías.

IF es una técnica sólida y confiable para detectar valores anómalos en registro de pozo, ya que, exhibe un gran desempeño en detectar anomalías contextuales donde

hay zonas afectadas por malas condiciones de pozo, así como en zonas donde hay una mezcla de datos anómalos producidos por el ruido y por malas condiciones de pozo en presencia de un subgrupo diferente pero relevante que ocurre con poca frecuencia y no debe considerarse como atípico [25].

3.5. Eliminación de ruido

Cualquier curva de registro de pozo se puede considerar como la suma de una señal (propiedades reales de la roca), ruido aleatorio (incluidas las variaciones aleatorias en las tasas de conteo de las herramientas nucleares) y errores sistemáticos [32]. Eliminado el ruido se busca reconstruir una señal que refleje únicamente las propiedades reales.

3.5.1. Transformada de Fourier

La transformada de Fourier (FT) descrita en la ecuación 1, permite descomponer una señal en un conjunto de senos de diferentes frecuencias con una magnitud asociada, de esta forma nos permite conocer las frecuencias que constituyen una señal y su aporte individual, pero no proporciona donde en el tiempo existe una frecuencia, por esto, carece de capacidad para proporcionar información de frecuencia para una región de señal localizada en el tiempo.

Dado que cada sinusoide corresponde a una sola frecuencia, la representación de señales usando sinusoides nos lleva a la comprensión e interpretación de las señales en el dominio de la frecuencia. Una ventaja importante de la representación en el dominio de la frecuencia es que a menudo es fácil ver ciertas características de la señal cuando examinamos la señal en este dominio, en comparación con el dominio del tiempo [27].

$$F(f) = \int_{-\infty}^{+\infty} f(t) e^{-2\pi i f t} dt \tag{1}$$

La transformada de Fourier de la función $f(t)$ a la frecuencia f es dada por el numero complejo $F(f)$. al evaluar todos los valores de f produce la función de dominio de frecuencia. Mientras, la transformada de inversa de Fourier (IFT) descrita en la ecuación 2, permite reconstruir una señal $f(t)$ en el dominio del tiempo a partir de su

representación en el dominio de la frecuencia $F(f)$.

$$f(t) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} F(f) e^{-2\pi i f t} df \tag{2}$$

Para superar la pobre resolución en el tiempo de FT fue desarrollada la transformada de Fourier de tiempo corto (STFT), descrito en la ecuación 3. Nos proporciona una representación de tiempo y frecuencia de la señal. A partir de una ventana de ancho fijo que recorre toda la señal, donde se asume que una parte de la señal no estacionaria es estacionaria, se realiza una FT para cada ventana. Con esto no podemos conocer que frecuencias existen en un instante de tiempo, pero podemos conocer que bandas de frecuencia existen en que intervalos de tiempo.

$$F(\tau, f) = \int_{-\pi}^{+\pi} f(t) w(t - \tau) e^{-2\pi i f t} dt \tag{3}$$

Donde $F(\tau, f)$ es una función compleja que representa la fase y magnitud de la señal en el tiempo y la frecuencia, $f(t)$ señal en el tiempo a ser transformada, f es la frecuencia, $w(t - \tau)$ es la función de ventana, τ es el parámetro de traslado y nos proporciona información de localización en el tiempo.

La función de ventana es finita lo que tiene como consecuencia una disminución en la resolución de frecuencia. El tamaño de la ventana determine la resolución en frecuencia y tiempo. Por lo que, un cambio en la resolución temporal (Δt) produce un cambio en la resolución frecuencia (Δf) obedeciendo la desigualdad 4.

$$\Delta t \Delta f \geq \frac{1}{4\pi} \tag{4}$$

El cálculo del espectro de señales en ventanas de tiempo a lo largo de una señal da como resultado una gráfica tridimensional que representa las variaciones de energía del contenido frecuencial de una señal a lo largo del tiempo, conocido como espectrograma.

3.5.2. Transformada de ondícula

La transformada de la ondícula (WT) ecuación 5 es un método de análisis multi-resolución en el dominio de frecuencia localizado para señales variables en el tiempo,

que puede verse como una extensión de STFT. Analiza la señal en diferentes frecuencias a diferentes resoluciones temporales. Ya que, los componentes de baja frecuencia a menudo duran largos períodos de tiempo, por lo que una alta resolución en frecuencia es requerida. Por otro lado, los componentes de alta frecuencia a menudo aparecen en pequeños periodos de tiempo, necesitando una alta resolución en el tiempo. De esta forma WT analiza las bajas frecuencias con baja resolución temporal y alta resolución frecuencial, mientras, las altas frecuencias son analizadas con alta resolución temporal y baja resolución frecuencial. Garantizando una alta resolución en el dominio del tiempo y la frecuencia al mismo tiempo.

$$X(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} \Psi\left(\frac{t-\tau}{s}\right) x(t)dt \tag{5}$$

La función $\Psi(t)$ es conocida como ondícula, es una función que reemplaza a los senos y cosenos de la transformada de Fourier como la función base de la trasformación y actúa como ventana de análisis, el parámetro de escalamiento s equivalente a $1/f$ nos permite variar el ancho de la ondícula, así como su frecuencia central. Para valores grandes de s tenemos una ondícula expandida que analiza mejor los componentes de baja frecuencia de la señal. de igual forma, para valores pequeños de s tenemos una ondícula encogida que analiza mejor los componentes de alta frecuencia de la señal. τ es factor de traslado temporal que traslada la ondícula a lo largo de la señal.

3.6. Modelos de aprendizaje de maquina

3.6.1. K-Nearest Neighbors (KNN)

Intenta clasificar una muestra desconocida basándose en la clasificación conocida de sus vecinos. se pueden calcular todas las distancias entre la muestra desconocida y todas las muestras del conjunto de entrenamiento. La distancia con el valor más pequeño corresponde a la muestra del conjunto de entrenamiento más cercana a la muestra desconocida. Por lo tanto, la muestra desconocida puede clasificarse en función de la clasificación de sus K vecinos más cercanos, asignándole la clase que contenga a la mayoría de sus vecinos [8].

3.6.2. Artificial Neural Network (ANN)

Las redes neuronales artificiales se diseñaron para simular el sistema nervioso biológico, el adjetivo neural proviene de dos elementos importantes: la estructura interna de una unidad computacional básica y las interconexiones entre ellas. Una característica fundamental de estas redes es que las neuronas están completamente conectadas entre sí, lo que se conoce como *fully connected* (FC).

En una red neuronal, cada neurona está conectada con n canales de entrada, cada uno de ellos caracterizado por un peso sináptico. Para cada valor de entrada se multiplican por el peso sináptico correspondiente y se calcula su suma. A esta suma se le puede agregar un sesgo opcional. La suma resultante se filtra mediante una función de activación, de esta forma el valor de salida es producido [3].

Los valores de los pesos sinápticos y el sesgo son ajustados mediante el uso del algoritmo de retropropagación. Esta técnica permite que la red neuronal aprenda a partir de ejemplos y retroalimentación, ajustando gradualmente los pesos y mejorando su capacidad de realizar tareas específicas.

Estas neuronas se encuentran agrupadas en tres tipos de capas: capa de entrada que recibe los vectores de entrada y es responsable de pasarlos al resto de la red; capas ocultas, que son capas intermedias entre la capa de entrada y salida y procesan los datos aplicándoles funciones no lineales complejas. Estas capas son el componente clave que permite que una red neuronal aprenda tareas complejas, reconocer patrones y predecir datos; y una la capa de salida, que toma como entrada los datos procesados y produce los resultados finales.

3.6.3. Convolutional Neural Network (CNN)

Las redes neuronales convolucionales fueron propuestas por [15], son un tipo especializado de ANN para procesar datos que tiene una topología similar a una cuadrícula. Su nombre implica del uso de una operación matemática conocida como convolución,

Una operación de convolución usa un filtro de pesos del mismo número de dimensiones que la capa actual, pero con una extensión espacial más pequeña. El producto punto entre todos los pesos en el filtro y cualquier región del espacio del mismo tamaño del filtro en una capa define el valor del estado oculto en la siguiente capa, después de aplicar una función de activación. La operación entre el filtro y las regiones espaciales en una capa se realiza en todas las posiciones posibles para definir la siguiente capa, en la que las activaciones conservan sus relaciones espaciales de la capa anterior [1].

Las redes neuronales convolucionales se componen de tres tipos de capas principales. En primer lugar, están las capas convolucionales, donde se llevan a cabo las operaciones de convolución. Estas capas son fundamentales para extraer características y patrones relevantes de los datos. Luego, encontramos las capas de agrupación, cuya función es reducir la dimensionalidad mediante la selección del valor máximo o promediando los valores en una región. Estas capas ayudan a conservar la información esencial mientras reducen la cantidad de parámetros necesarios. Por último, se encuentran las capas completamente conectadas, donde todas las entradas de una capa están conectadas a cada unidad de activación de la siguiente capa. Estas capas finales permiten combinar y procesar la información extraída anteriormente para producir los resultados deseados. Su capacidad para capturar características locales y mantener las relaciones espaciales en los datos ha sido clave en su éxito en aplicaciones como el procesamiento de imágenes.

3.6.4. Long Short-Term Memory (LSTM)

Fueron propuestas en 1997 por Hochreiter et al. [18] para resolver el problema del desvanecimiento de gradiente presente en las redes neuronales recurrentes (RNN), el cual dificulta que un modelo pueda retener información relevante de eventos ocurridos en el pasado distante. Dentro de un LSTM la información puede ser almacenada, escrita y leída por medio de compuertas.

La compuerta del olvido juega un papel crucial al decidir qué información necesita atención y cuál puede ser ignorada. Esta compuerta determina el porcentaje de la memoria a largo plazo que debe mantenerse, utilizando tanto la información de la memoria

de corto plazo como la entrada actual. Por otro lado, la compuerta de entrada genera un valor potencial para la memoria a largo plazo, basándose en la memoria a corto plazo y la entrada actual. Simultáneamente, la compuerta de estado determina como debe ser actualizado el valor de la memoria a largo plazo y que porcentaje de valor potencial debe ser agregado. Además, la compuerta de salida actualiza el valor de la memoria a corto plazo utilizando la entrada actual, la memoria a corto plazo y la memoria a largo plazo. Por último, los valores de memoria a corto y largo plazo se envían a la siguiente iteración del proceso.

Gracias a su capacidad para aprender y retener dependencias a largo plazo, las LSTM son aplicables a una amplia gama de problemas de aprendizaje de secuencias. Estas redes neuronales han demostrado ser especialmente efectivas en tareas que involucran información contextual y dependencias temporales significativas.

3.6.5. Mecanismo de Autoatención (AT)

El mecanismo de autoatención, también conocido como self-attention, fue propuesto en el campo de la inteligencia artificial y el procesamiento del lenguaje natural por Vaswani et al. [33]. Este permite a los modelos de lenguaje capturar las relaciones contextuales entre las palabras en una secuencia. A diferencia de los enfoques tradicionales que utilizan contextos fijos o ventanas deslizantes para analizar las palabras, la autoatención se basa en la idea de que cada palabra puede interactuar con todas las demás palabras de la secuencia.

En la autoatención, cada palabra en la secuencia actúa simultáneamente como una consulta (Q), una clave (K) y un valor (V). La consulta se utiliza para obtener información relevante de las demás palabras, las claves ayudan a establecer las relaciones entre las palabras y los valores contienen la información asociada a cada palabra. Estas tres representaciones se obtienen a través de proyecciones lineales de la palabra de entrada.

Una vez que se han obtenido las consultas, claves y valores, se calcula la similitud entre las consultas (Q), y las claves (K), representada por QK^T, determina qué pala-

bras en la secuencia tienen una relación más fuerte en el contexto. Para evitar que estas similitudes crezcan desproporcionadamente al aumentar la dimensión de las claves, se normalizan mediante la raíz cuadrada de la dimensión de las claves d_k.

Posteriormente, las ponderaciones normalizadas, se multiplican elemento por elemento con los valores (V). Esta multiplicación pondera la contribución de cada valor en función de las ponderaciones obtenidas. Al combinar las ponderaciones y los valores mediante esta operación, se obtiene una representación contextualizada de cada palabra en la secuencia. Así, la función de atención, representada por la ecuación 6, permite capturar la información relevante proveniente de las demás palabras y enriquecer la representación de cada palabra en el contexto.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Además de la atención estándar, existe una variante conocida como atención multi-head, que ha demostrado ser especialmente efectiva en muchos modelos de lenguaje. La atención multi-head permite que el modelo aprenda diferentes representaciones de atención simultáneamente, utilizando múltiples proyecciones lineales. Esto proporciona al modelo una mayor capacidad para capturar relaciones y patrones complejos en los datos, mejorando así su desempeño en tareas de procesamiento del lenguaje natural.

3.7. Evaluación de los modelos

Para los algoritmos supervisados la evaluación de las tareas de clasificación normalmente se realiza dividiendo el conjunto de datos en un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El algoritmo se entrena en el primer conjunto, mientras que el conjunto de datos de prueba es usado para evaluar su rendimiento y permitir la comparación.

Los indicadores de rendimiento son calculados a partir de relaciones de los posibles resultados de clasificar un dato: Verdadero positivo (TP) la etiqueta es positiva y nuestro valor predicho también es positivo; Falso positivo (FP) la etiqueta es negativa pero

la predicción de nuestro modelo es positiva; Verdadero negativo (TN): la etiqueta es negativa y nuestro valor predicho también es negativo; Falso negativo (FN): la etiqueta es positiva pero la predicción de nuestro modelo es negativa.

Para los conjuntos de datos desequilibrados como es el caso las litofacies presentes en un pozo, algunas medidas no reflejan la verdadera precisión del modelo evaluado. Por lo tanto, es necesario incluir mediciones que no se vean afectadas por el desbalance de los datos y evaluar individualmente cada clase para describir el desempeño del modelo.

3.7.1. Matriz de confusión

También conocida como tabla de contingencia, es utilizada para ilustrar gráficamente el desempeño de un modelo predictivo, distingue entre predicciones TP, FP, TN y FN. Su principal desventaja es que requiere interpretación humana al evaluar un modelo.

3.7.2. Precisión

La precisión como lo muestra la ecuación 7 denota la proporción de casos predichos como positivos que son efectivamente son positivos reales [26], para nuestro caso la precisión responde a la pregunta de todas las instancias identificadas como lo la litofacie x, ¿cuántas son realmente x?

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

3.7.3. Exhaustividad

la exhaustividad o recall descrito en la ecuación 8 es la proporción de casos reales positivos que son correctamente pronosticados como positivos [26], en nuestro contexto responde a la pregunta de todas las instancias identificables de la litofacie x que existen, ¿cuántas identificamos?

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

3.7.4. F1-score

Es la media harmónica entre la exhaustividad y la precisión, ecuación 9, es elegida la media harmónica sobre la media aritmética ya que esta produce un puntaje alto solo si la precisión y la exhaustividad son altos.

$$\text{F1 score} = 2 \cdot \frac{(\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \tag{9}$$

3.7.5. Matthews's correlation coefficient (MCC)

Introducida por B.W. Matthews [24] y descrita mediante la ecuación 10, el coeficiente de correlación de Matthews es una medida que oscila entre 1 y -1. Un valor de 1 indica un clasificador perfecto, mientras que -1 indica un clasificador que comete errores en todas sus predicciones. Este coeficiente genera un puntaje alto únicamente si la predicción obtuvo buenos resultados en las cuatro categorías de la matriz de confusión. Además, se considera una medida equilibrada que puede utilizarse incluso cuando las clases tienen tamaños muy diferentes [5].

$$\text{MCC} = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

3.7.6. Validación cruzada

Es un método estadístico utilizado para estimar la habilidad de los modelos de aprendizaje automático. Donde inicialmente se divide todo el conjunto de datos en K partes y usar cada una de ellas secuencialmente como el conjunto de datos de prueba mientras se las demás se combinan para formar el conjunto datos de entrenamiento. Posteriormente, los indicadores de rendimiento se promedian para obtener la evaluación del modelo.

3.8. Bayesian Optimization (BO)

La optimización bayesiana es útil para optimizar funciones que tardan mucho tiempo en evaluarse, no tienen una expresión analítica y se desconoce su derivada, como es el caso de la función de costo de un modelo de aprendizaje automático, evitando

muestrear puntos innecesarios respondiendo a la pregunta según lo que sabemos hasta ahora, ¿qué punto deberíamos evaluar a continuación?

Comienza creando un sustituto para la función objetivo formada a partir de algunos puntos muestreados, y cuantifica la incertidumbre en ese sustituto usando una técnica de aprendizaje automático bayesiano, regresión de proceso gaussiana y luego usa una función de adquisición definida a partir de este sustituto para decidir dónde muestrear, actualizando la función sustituta [14]. Después de un cierto número de iteraciones, la función sustituta es una buena aproximación a la función objetivo, y podrá ser usada para encontrar el mínimo global.

3.9. Explicabilidad del modelo

La explicabilidad de un modelo de aprendizaje automático es fundamental para comprender y confiar en sus decisiones. Proporciona transparencia en el proceso que lleva a una predicción o resultado, permitiendo entender las relaciones causa-efecto entre las características y las decisiones del modelo. Además, la explicabilidad facilita la interpretación y comprensión de los procesos subyacentes, fomentando la generación de nuevos conocimientos y descubrimientos en el dominio de estudio. Asimismo, permite la identificación de características problemáticas o sesgadas y su influencia en las decisiones, lo que contribuye a la mejora continua del modelo.

3.9.1. Shapley Additive Explanations (SHAP)

SHAP es una metodología introducida en [23] desarrollada para explicar las predicciones de modelos de aprendizaje automático de manera individual y global, proporcionando una forma teóricamente sólida de atribuir la contribución de cada característica en la predicción de un modelo. El concepto clave en SHAP es el valor de Shapley, que se origina en la teoría de juegos y se adapta al contexto de la explicación de modelos de aprendizaje automático. Este cuantifica la contribución marginal de un jugador dentro de una coalición al generar un valor, teniendo en cuenta todas las posibles combinaciones de jugadores en el juego.

En el marco de SHAP, los jugadores son las características del modelo y la coalición se refiere a un conjunto de características. Un modelo explicativo, representado como $g(x')$, se define como una aproximación interpretable del modelo original $f(x)$. Aquí, x' es una versión simplificada de la entrada original x, obtenida a través de una función de mapeo $x = h_x(x')$. Esta simplificación proporciona una representación reducida de las características de entrada originales.

En el modelo explicativo $g(x')$, M es el número de características de la entrada simplificada. La salida nula o el término de sesgo, representado por Φ_0, corresponde a la predicción promedio del modelo cuando todas las características son nulas. Por otro lado, Φ_i es la atribución de la característica i, que se expresa mediante su valor de SHAP. Este valor de SHAP indica la contribución relativa de esa característica en la predicción del modelo.

$$g(x') = \Phi_0 + \sum_{i=1}^{M} \Phi_i + x'_i \tag{11}$$

En [23] mencionan que solo hay una solución al asignar los pesos Φ_i al modelo explicativo $g(x')$ cumpliendo las tres propiedades naturales precisión local (*local accuracy*), falta de características (*missingness*) y la consistencia (*consistency*).

1. Precisión local: Establece que el modelo explicativo $g(x')$ debe producir resultados que sean consistentes con los resultados del modelo original $f(x)$ para cada predicción que se está explicando. En otras palabras, las explicaciones deben reflejar la relación real entre las características de entrada y la salida del modelo para cada instancia en particular.

$$f(x) = g(x') = \Phi_0 + \sum_{i=1}^{M} \Phi_i + x'_i \tag{12}$$

2. Falta de características: Se puede definir como la condición en la que una característica ausente x'_i en el modelo original no tiene ningún impacto en el modelo explicativo $g(x')$. En otras palabras, si una característica no está presente en la instancia de entrada, su atribución de SHAP en el modelo explicativo $g(x')$ debe

ser cercana a cero, indicando que su ausencia no influye significativamente en la predicción del modelo.

$$x'_i = 0 \Rightarrow \Phi_i = 0 \tag{13}$$

3. Consistencia: Si al eliminar la característica i de la entrada, la diferencia en las predicciones del modelo f'_x es mayor o igual que la diferencia en las predicciones del modelo f_x, para todas las posibles combinaciones de características en la entrada, entonces la contribución de la característica i en el modelo f'_x es al menos tan grande como en el modelo f_x. Implicando que $\Phi_i(f', x) \geq \Phi_i(f, x)$.

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \tag{14}$$

La fórmula que calcula los valores de SHAP dada por la ecuación 15 utiliza comparaciones entre las predicciones realizadas por el modelo con y sin la presencia de una característica específica. Evalúa todas las combinaciones posibles de características para obtener las contribuciones individuales de cada una en una predicción. Esto permite cuantificar la contribución única de cada característica y cómo se combina con otras para influir en la predicción del modelo.

$$\Phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|\,!(M - |z'| - 1)!}{M!}[f_x(z') - f_x(z' \setminus i)] \tag{15}$$

Donde f es el modelo evaluado, x es la entrada del modelo, z' es un subconjunto de características de x, x' es una entrada local simplificada, $f_x(z')$ es la salida del modelo f cuando se utiliza la entrada x para el subconjunto de características z', y $f_x(z' \setminus i)$ es la salida del modelo f cuando se utiliza la entrada x para el subconjunto de características z' excluyendo la característica i. Además, M representa el número total de características del subconjunto z'.

El cálculo de los valores de SHAP puede ser costoso computacionalmente, pero existen algoritmos y enfoques eficientes para abordar este problema. Estos métodos aprovechan la estructura de las características y realizan cálculos simplificados.

4. ESTADO DEL ARTE

La interpretación litológica de forma automática a partir de los registros de pozo aún se encuentra en desarrollo y perfeccionamiento Uno de los primeros acercamientos realizados a finales de la década de 1980 usó el análisis de discriminante lineal donde las observaciones en una de un conjunto de categorías exhaustivas mutuamente excluyentes, basándose en las puntuaciones de una o más variables predictoras cuantitativas [9].

En los últimos años se han realizado diferentes investigaciones para automatizar la interpretación litológica utilizando el aprendizaje automático. En particular, el aprendizaje profundo ha demostrado que puede realizar tareas de clasificación a nivel humano, en lo que puede describirse como una revolución de la inteligencia artificial [7]. Recientemente se realizó un concurso de aprendizaje automático para la interpretación de litofacies por parte del SEG a partir de los registros de nueve pozos de los campos Hugoton y Panoma del sureste de Kansas y noroeste de Oklahoma – EEUU fue organizado un concurso de aprendizaje automático para la interpretación de litofacies por parte del SEG, en el cual participaron 40 equipos de ocho países durante cuatro meses, recibiendo 300 entradas, siendo que, los cinco primeros equipos utilizaron XGBoost [11]. Honório et al. [19] proponen un método basado en la transformada de la ondícula para eliminar el ruido de las señales de los registros de pozos. El objetivo principal de este método es obtener datos geológicos más representativos al analizar las señales de los registros de pozos. Al eliminar el ruido, se mejora la calidad de los registros y se obtienen mediciones más precisas y confiables. Además, esta eliminación de ruido resalta las características geológicas importantes y facilita su interpretación. El estudio también muestra una mejora significativa en la clasificación litológica utilizando el registro de rayos gamma y el algoritmo de k vecinos cercanos (KNN) en comparación con los datos originales. Adicionalmente, se han desarrollado diferentes aproximaciones para integrar los patrones de apilamiento de las rocas sedimentarias: Zhu et al. [34] propusieron un método basado en la descomposición de ondículas para construir una imagen multicapa para cada punto en profundidad, haciendo posible convertir el problema de registro de interpretación litológica en una tarea de reconocimiento de imagen supervisada que

puede ser resuelto usando una red CNN. Jaikla et al [21] convirtieron los registros en espectrogramas utilizando STFT para alimentar una red neuronal recurrente bidireccional (BRNN) logrando diferenciar facies reservorio y no reservorio, como lo son arenisca limpia y arenisca sucia, así como lutita y heterolíticos. Otras propuestas han utilizado información secuencial mediante RNN, como el estudio en [30] que utiliza pozos de la cuenca de Paraná – Brasil y usando una arquitectura BiLSTM, obteniendo mejores resultados frente a otros enfoques ampliamente utilizados como XGBoost, SVM o Random Forest.

El aprendizaje automático ha demostrado ser eficaz en la clasificación de litofacies en diversas cuencas a nivel mundial. Sin embargo, muchos estudios pasan por alto la consideración de los patrones de sedimentación en el proceso de clasificación. Es fundamental tener en cuenta que estos patrones desempeñan un papel crucial en la determinación de las litologías que son posibles en una secuencia sedimentaria continua. Al integrar los patrones de sedimentación en los modelos de aprendizaje automático, se puede mejorar la precisión de la clasificación de litologías y obtener una comprensión más completa de la heterogeneidad y la distribución espacial de las formaciones geológicas.

Al incorporar los patrones de sedimentación en los modelos de aprendizaje automático, se pueden capturar las relaciones y características intrínsecas de las litologías dentro del contexto sedimentario. Esto permite que el modelo aprenda a reconocer y diferenciar las litofacies en función de su posición y su relación con los patrones de deposición sedimentaria. Como resultado, se logra una clasificación más precisa y confiable de las litologías presentes en una secuencia estratigráfica.

5. METODOLOGÍA

El proyecto se centró en el análisis de datos almacenados en el repositorio [4], el cual alberga registros de pozos ubicados geográficamente en la plataforma continental noruega, específicamente en los bloques 7, 15, 16, 17, 25, 26, 29, 30, 31, 32, 33, 34, 35 y 36 del Mar de Barents y el Mar del Norte noruego. El objetivo principal del proyecto consiste en el modelamiento de datos mediante el aprendizaje a partir de registros proporcionados. Por lo tanto, la limpieza y reestructuración de los datos representan un papel fundamental en este proceso.

En este capítulo, se describe detallada la metodología utilizada en el análisis y modelado de los datos de registros de pozos en la plataforma continental noruega. El propósito de esta metodología es garantizar la calidad y coherencia de los datos, así como proporcionar una base sólida para el modelado y análisis subsiguientes.

5.0.1. Construcción de la base de datos

En total, se recopilaron 118 archivos en formato Log ASCII Standard (LAS), correspondientes a los 118 pozos analizados. Estos archivos contenían información proveniente de 24 registros diferentes, detallados en el Cuadro 1, así como la interpretación de las litofacies y la litología elaborada de forma manual utilizando geocientíficos expertos [4] Algunos de estos registros utilizan unidades de medida no tan comunes en el contexto general mientras que otros no reportan una unidad de medida, siendo denotados como 'ND', se describen brevemente estas unidades:

1. La resistividad eléctrica se mide en ohm.metro ($\Omega \cdot \mathrm{m}$), que indica la resistencia que presenta un material al paso de corriente eléctrica por unidad de longitud. Se utiliza para evaluar la resistividad en diferentes zonas del subsuelo.

2. La intensidad de la radiación gamma que emana de las formaciones geológicas se representa en la escala gAPI (gamma API). Cuanto mayor es el valor en gAPI, mayor es la radiactividad registrada.

3. El tiempo que tarda una onda sonora en recorrer una distancia específica en el

subsuelo se mide en microsegundos por pie (μs/ft). Esta unidad se utiliza para determinar la velocidad del sonido y las propiedades acústicas de las formaciones geológicas.

4. La porosidad de una formación geológica se describe mediante la relación de volumen de poros/volumen total (m^3/m^3). Esta relación indica el volumen de espacios vacíos (poros) en relación con el volumen total de la roca y se expresa sin unidades específicas debido a su carácter adimensional.

5. La eficiencia de interacción de los fotones con los electrones de las formaciones geológicas se representa en barn/electrón (b/e). El barn es una medida de área efectiva de interacción, y el electrón es una unidad de carga eléctrica. Cuanto mayor es el valor en b/e, mayor es la capacidad de la formación para absorber fotones.

Es importante destacar que, en conjuntos de datos de este tipo, no todos los pozos cuentan con información completa de todos los registros, lo cual es una situación común. Algunos registros pueden no proporcionar información para todos los puntos en profundidad.

5.0.2. Extracción y compilación de datos

La primera etapa de la metodología consistió en la extracción de datos de los archivos (.las) y su posterior compilación en un archivo .CSV. Para llevar a cabo esta tarea, se utilizó la librería Lasio en su versión 0.30 de Python. Esta librería permitió acceder y leer los datos de los archivos de manera eficiente. Durante este proceso, también se analizó de la consistencia en el espaciado de las observaciones, revelando un espaciado constante de 0.1524 m (0.5 ft), siendo este un estándar en la industria. Además, se verificó la consistencia de la unidad de medida para todos los registros.

5.0.3. Selección de registros relevantes

Se procedió a seleccionar los registros relevantes para el análisis de los datos de los pozos, teniendo en cuenta su relevancia en la caracterización de yacimientos. Se identificaron los siguientes registros clave: registro de rayos gamma (GR), medición de

Nombre	Descripción	Unidades
DEPTH_MD	Profundidad medida	m
CALI	Registro de caliper	in
DCAL	Registro de caliper diferencial	in
SP	Registro de potencial espontáneo	mV
RDEP	Medición de resistividad profunda	$\Omega \cdot m$
RSHA	Medición de resistividad superficial	$\Omega \cdot m$
RMED	Medición de resistividad intermedia	$\Omega \cdot m$
RXO	Medición de resistividad en zona infiltrada	$\Omega \cdot m$
RMIC	Medición de micro resistividad	$\Omega \cdot m$
GR	Registro de rayos gamma	gAPI
SGR	Registro de rayos gamma espectral	gAPI
DTS	Registro sónico de ondas de cizalla	μs/ft
DTC	Registro sónico de ondas de compresión	μs/ft
NPHI	Registro de porosidad de neutrones	m^3/m^3
RHOB	Registro de densidad aparente	g/cm^3
DRHO	Registro de corrección de densidad	g/cm^3
PEF	Registro de factor fotoeléctrico	b/e
BS	Tamaño del pozo	in
ROP	Tasa de penetración	m/h
ROPA	Tasa promedio de penetración	ND
MUDWEIGHT	Peso del lodo de perforación	ND
x_loc	Ubicación x de la muestra	ND
y_loc	Ubicación y de la muestra	ND
z_loc	Profundidad (TVDSS) de la muestra	ND

Cuadro 1: Descripción y unidades de los registros presentes en [4]

resistividad intermedia (RMED), medición de resistividad profunda (RDEP), registro sónico de ondas de compresión (DTC), registro de porosidad de neutrones (NPHI), registro de densidad aparente (RHOB) y registro de factor fotoeléctrico (PEF). Cabe destacar que se optó por utilizar los nombres de los registros en español para facilitar la comprensión del lector.

Todos los registros mencionados anteriormente son ampliamente utilizados en la industria petrolera y geológica debido a su relevancia para la caracterización de los yacimientos. Al centrarnos en estos registros, nos aseguramos de obtener la información necesaria para llevar a cabo el modelado y análisis subsiguientes. Además, se destacó que este conjunto de datos fue seleccionado porque proporcionaba la mayor cantidad de observaciones disponibles, lo cual es fundamental para un análisis exhaustivo y preciso.

5.0.4. Filtrado de observaciones

Se descartaron los pozos que carecían de información en los registros seleccionados. Durante el proceso de exploración de los datos, se observó que la falta de información generalmente se presentaba en secciones continuas en lugar de puntos aislados. Para mantener la integridad y coherencia de los datos utilizados en el análisis, se decidió eliminar las observaciones en profundidad que tenían información faltante en los registros seleccionados.

Los rangos de cada uno de los registros se determinaron visualmente y se basaron en estándares y consideraciones aceptadas en la industria petrolera Posteriormente, se procedió al filtrado de las observaciones que estaban fuera de los rangos predefinidos para cada registro, relacionados en la Cuadro 2. El objetivo fue eliminar observaciones con valores atípicos que pudieran afectar negativamente el análisis y modelado subsiguientes.

Registro	Min	Max
GR	0	150
RMED	0	100
RDEP	0	100
DTC	0	200
NPHI	0	1
RHOB	1.5	3
PEF	0	12

Cuadro 2: Rango de valores mínimos y máximos de los registros

El filtrado de observaciones fuera de rango resultó en un conjunto de datos coherente y consistente, asegurando la calidad y precisión de la información utilizada en el análisis. Al eliminar las observaciones atípicas y las secciones de datos con información faltante, evitamos sesgos y garantizamos una representación más precisa de los registros relevantes en los pozos.

5.0.5. Selección de pozos y categorías relevantes para el análisis

Con el fin de simplificar el problema y enfocar el análisis en las categorías relevantes, se llevó a cabo un proceso de unificación en el número de pozos y categorías considerados en el estudio.

En primer lugar, se limitó el número de pozos considerados en el estudio. Se realizó una evaluación exhaustiva de los pozos disponibles y se identificó que el bloque 35 contenía la mayor cantidad de pozos con todas las categorías presentes. Por lo tanto, se decidió centrar el análisis en este bloque específico, ya que proporcionaba una representación amplia y significativa de las categorías de interés.

Además, se realizó una evaluación de las categorías presentes en los datos. Se identificaron diferentes categorías como ' *Sandstone*' (Sa), ' *Shale*' (Sh), ' *Sandstone/Shale*' (Sa/Sh), ' *Limestone*' (Li) y ' *Marl*' (Ma). Sin embargo, se observó que algunas categorías como ' *Chalk*', ' *Tuff*', ' *Coal*', ' *Dolomite*', ' *Halite*', ' *Anhydrite*' y ' *Basement*' presentaban una representación insuficiente en los datos. Debido a esto, se decidió excluir estas categorías del análisis y simplificar el conjunto de datos concentrándose en las categorías más comunes y relevantes.

5.1. Descripción de los datos

Se procedió a graficar la distribución de las categorías presentes en los registros de pozo Figura 4. Este análisis puso de manifiesto un desbalance significativo entre las clases, donde "Sh" representó el 54.3 % de los datos, mientras que "Ma" y "Li" tienen una representación de 3.6 % y 3.0 % respectivamente. Dicho desbalance puede tener repercusiones en la construcción y evaluación de modelos de Machine Learning, particularmente en el desempeño de la clasificación precisa de las clases minoritarias, por lo tanto, es necesario considerar estrategias para abordar este desbalance y evaluar el rendimiento de manera equitativa en todas las categorías litológicas.

Después de analizar la distribución de las clases en los registros de pozo, se examinaron las relaciones entre las variables mediante el análisis de la matriz de correlación, como se muestra en la Figura 5. Esta matriz revela las correlaciones lineales entre las

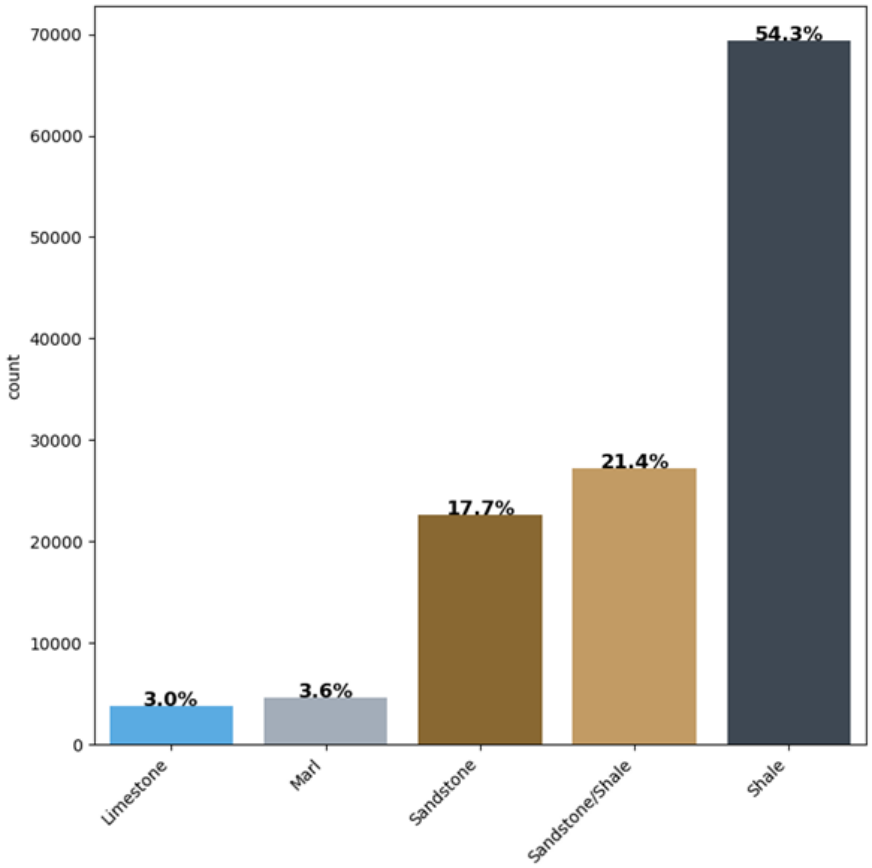

Figura 4: Distribución de litologías.

diferentes variables, donde los valores cercanos a 1 indican una correlación alta positiva, mientras que los valores cercanos a -1 indican una correlación alta negativa.

Durante el análisis de la matriz de correlación, se hicieron varios descubrimientos significativos. En primer lugar, se encontró una correlación positiva entre NPHI y DTC, que sugiere que a medida que aumenta la porosidad efectiva, también aumenta la velocidad de la onda acústica compresiva. Esto podría indicar que las formaciones con mayor porosidad efectiva exhiben una velocidad de propagación de la onda compresiva más alta. Además, se observó una correlación negativa entre RHOB y DTC, lo que indica que a medida que aumenta la densidad aparente, la velocidad de la onda acústica compresiva tiende a disminuir. También se encontró una correlación negativa entre RHOB y NPHI, lo que sugiere que a medida que aumenta la densidad, la porosidad efectiva tiende a disminuir. Otro descubrimiento relevante fue la correlación negativa entre DTC y los registros resistivos (RDEP y RMED), que indica que cuando la velocidad de la onda

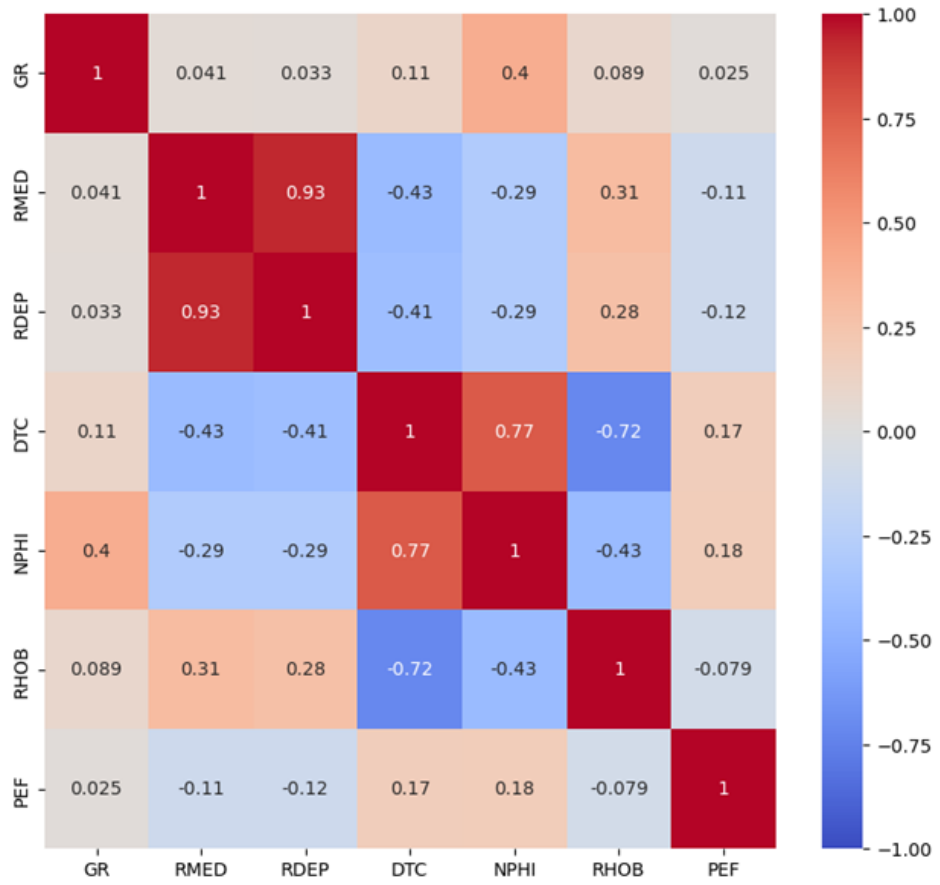

Figura 5: Matriz de correlación entre los registros.

acústica compresiva disminuye, la resistividad también tiende a disminuir. Además, se identificó una correlación positiva entre GR y NPHI, lo que sugiere que la porosidad efectiva tiende a aumentar con el incremento de la radiación natural. Por último, se observó que el factor fotoeléctrico (PEF) no presentó una correlación significativa con los demás registros.

Después de analizar las correlaciones entre las variables, se llevó a cabo la generación de un pair plot Figura 6 para obtener una visualización más detallada de las relaciones y dispersiones de los datos. Este análisis reveló un solapamiento, en las distribuciones de las categorías de todos los pares de variables, indicando una superposición significativa en las características entre clases. Estos solapamientos podrían dificultar la distinción de las clases basándose únicamente en las características individuales.

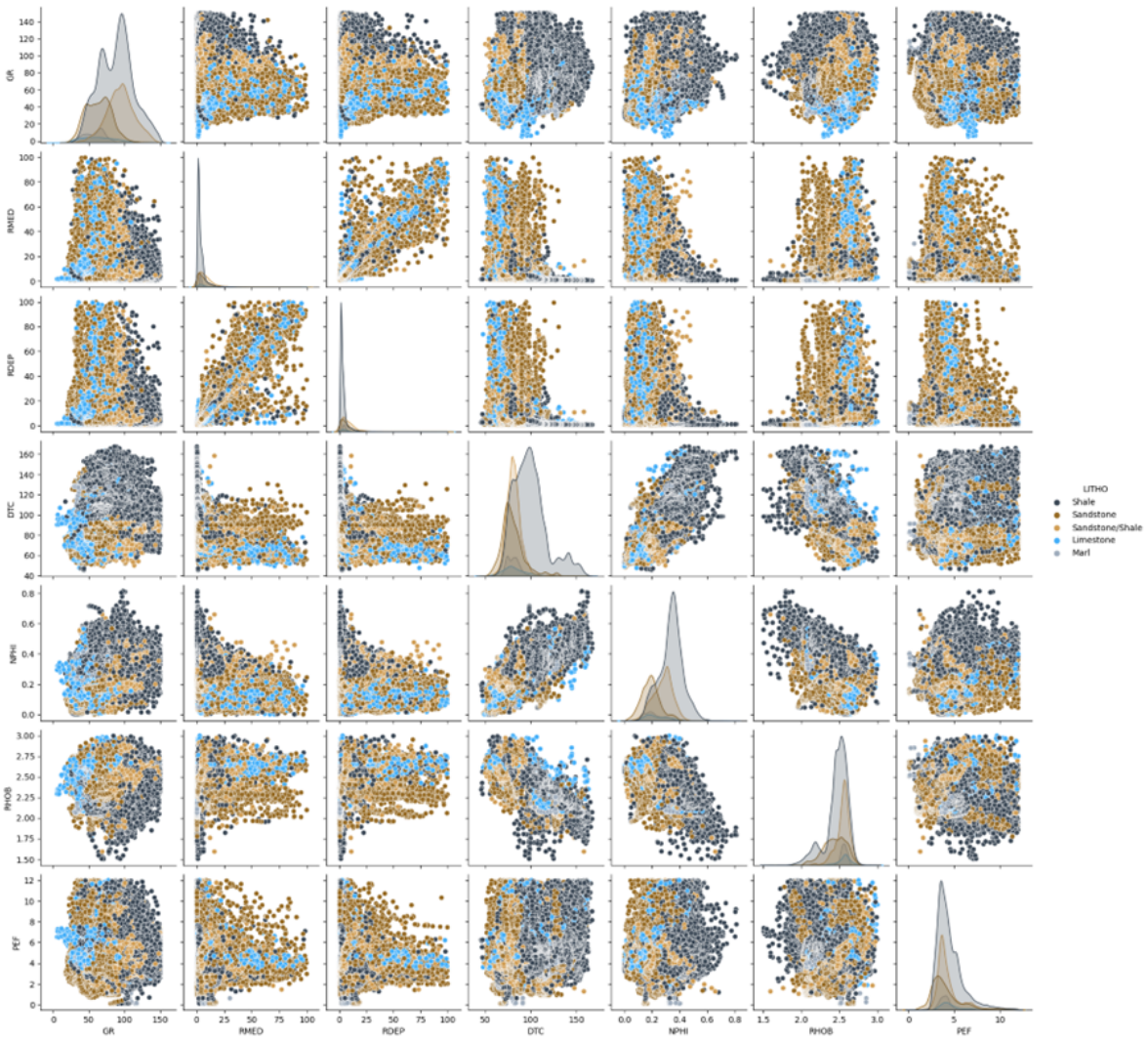

Figura 6: Pair plot entre las variables.

5.2. Preprocesamiento

5.2.1. División de los datos

Con el objetivo de evaluar y validar los modelos desarrollados en este estudio se implementó una estrategia de validación cruzada. La validación cruzada es una técnica ampliamente utilizada en el aprendizaje automático que permite estimar el rendimiento y la generalización de un modelo utilizando los datos disponibles.

Para llevar a cabo la validación cruzada, la base de datos se dividió en un número k de conjuntos que se definen sistemáticamente como entrenamiento y validación. En este estudio se definió un valor de k igual a 5, lo que significa que la base de datos se dividió

en cinco partes o pliegues. De esta forma, el proceso de validación cruzada determina cinco conjuntos de entrenamiento y validación diferentes.

El proceso de división se realizó de la siguiente manera: cada pozo fue dividido en k partes, garantizando que cada parte tuviera un número similar de observaciones. Luego, de manera aleatoria, se seleccionó una sección de cada pozo para formar el conjunto de validación, mientras que las secciones restantes se utilizaron como conjunto de entrenamiento. Este proceso se repitió k veces, asegurando que cada parte de los pozos se asignara al conjunto de validación exactamente una vez. Finalmente, los resultados parciales de cada iteración son promediados para obtener el resultado general.

5.2.2. Transformación de los datos

Se aplicó una normalización logarítmica dada por la ecuación 16 a los registros resistivos (RDEP y RMED). Estos registros presentan un comportamiento logarítmico en su distribución de valores, por lo que la normalización logarítmica permitió transformarlos en una escala más adecuada para su procesamiento.

$$x' = \log_{10} x \tag{16}$$

Posteriormente, se realizó una estandarización de los datos con el objetivo de que los datos tuvieran una media de 0 y una desviación estándar de 1 (ecuación 17). Esta técnica es comúnmente utilizada en el aprendizaje automático para normalizar los datos y garantizar que todas las variables tengan un impacto similar en el modelo. La estandarización permitió comparar y combinar los diferentes registros de manera más efectiva durante el proceso de modelado.

$$x' = \frac{x - \mu}{\sigma} \tag{17}$$

Finalmente, se aplicó una normalización min-max simétrica descrita en (ecuación 18), para ajustar los datos en un rango específico. En este caso, se utilizó un rango de -1 a 1. La normalización min-max simétrica permitió escalar los datos de manera proporcional dentro del rango establecido, lo que facilitó la comparación y la interpretación

de los resultados.

$$x' = \left(\frac{x - \text{mín}(x)}{\text{máx}(x) - \text{mín}(x)} \right) \cdot 2 - 1 \tag{18}$$

Es importante destacar que el preprocesamiento de los datos se llevó a cabo de manera secuencial y los datos de entrenamiento fueron preprocesados primero. Los parámetros utilizados para la estandarización y la normalización se calcularon a partir de los datos de entrenamiento y se aplicaron posteriormente a los datos de validación. Esto garantizó que el procesamiento de los datos fuera coherente y consistente entre el conjunto de entrenamiento y el conjunto de validación.

5.2.3. Purgado de datos anómalos

En este procedimiento se llevó a cabo la identificación y eliminación de valores atípicos y observaciones anómalas. Para ello, se utilizaron técnicas de detección de anomalías, como el algoritmo de Isolation Forest, que permite identificar patrones anómalos en los datos. Se determinó que el 5 % de los datos eran anómalos y se procedió a eliminarlos de los conjuntos de entrenamiento y validación correspondientes. Esto garantizó que los modelos se entrenaran y evaluaran utilizando datos confiables y representativos.

Es importante destacar que el purgado de datos anómalos se realizó exclusivamente en el conjunto de entrenamiento, con el objetivo de asegurar que los datos de validación reflejaran de manera más precisa el comportamiento general de los registros

5.2.4. Identificación de secciones continuas

Con el objetivo de garantizar la continuidad necesaria para los modelos LSTM (Long Short-Term Memory) y CNN (Convolutional Neural Network), se llevó a cabo la identificación de secciones continuas de información a nivel de pozo. Durante este proceso, se examinaron los registros de cada pozo y se buscó identificar aquellas secciones que carecieran de interrupciones significativas, teniendo en cuenta las diferentes longitudes y características de los datos disponibles para cada pozo.

Se estableció que una sección se consideraba continua si superaba un umbral mínimo

de 650 observaciones, lo que aproximadamente equivale a unos 100 metros de datos. Aquellas secciones que no alcanzaron este umbral fueron descartadas, ya que no proporcionaban la cantidad suficiente de información para llevar a cabo un análisis preciso.

5.2.5. Reestructuración de los datos

La matriz de características X a lo largo de profundidad puede ser expresada por la (ecuación 19), donde n es el número de registros y d es el número de observaciones continuas en profundidad medidas a una distancia constante.

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_i^1 & x_i^2 & \cdots & x_i^n \\ \vdots & \vdots & \ddots & \vdots \\ x_d^1 & x_d^2 & \cdots & x_d^n \end{bmatrix}_{(d \times n)} \tag{19}$$

Las redes completamente conectadas (FC) reciben como entrada un arreglo bidimensional. El primer eje corresponde al tamaño del lote, es decir, la cantidad de observaciones que se procesan simultáneamente. El segundo eje indica la dimensionalidad de cada observación, es decir, el número de características o variables en cada profundidad. De esta forma la observación a la profundidad i es un tensor de tamaño $1 \times n$ (ecuación 20).

$$x_i = [x_i^1, x_i^2, \cdots, x_i^n]_{(1 \times n)} \tag{20}$$

A diferencia de las FC. Tanto las redes LSTM como las redes CNN1D requieren entradas en forma de arreglos tridimensionales para su adecuado procesamiento de la información. Estas dimensiones representan características específicas de los datos.

En una red LSTM. El primer eje corresponde al tamaño del lote, es decir, la cantidad de secuencias que se procesan simultáneamente. El segundo eje se refiere a la longitud de cada secuencia, es decir, el número de pasos en profundidad en la secuencia. Por último, el tercer eje indica la dimensionalidad de cada paso en profundidad de la secuencia, es

decir, el número de características o variables en cada profundidad. De esta forma la observación a la profundidad i es un tensor de tamaño $(1 + (w \times 2)) \times n$ (ecuación 21), donde w es el número de observaciones consideradas antes y después del punto analizado, en este trabajo se seleccionó un w de 10, obteniendo una longitud de secuencia de 3.2 m

$$x_i = \begin{bmatrix} x^1_{(i-w)} & x^2_{(i-w)} & \cdots & x^n_{(i-w)} \\ \vdots & \vdots & \vdots & \vdots \\ x^1_i & x^2_i & \cdots & x^n_i \\ \vdots & \vdots & \vdots & \vdots \\ x^1_{(i+w)} & x^2_{(i+w)} & \cdots & x^n_{(i+w)} \end{bmatrix}_{((1+(w\times 2))\times n)} \tag{21}$$

En el contexto de los registros de pozo, una red LSTM analiza la información de todos los registros en un punto de profundidad, para una secuencia. Esto implica que se considera la secuencia de registros a lo largo de la profundidad para capturar las relaciones temporales y extraer patrones relevantes.

Por otro lado, en una red CNN1D El primer eje sigue representando el tamaño del lote. El segundo eje se refiere al número de canales de información, es decir, el número de dimensiones de características. Y el tercer eje representa la longitud de la secuencia o el número de puntos en la secuencia. De esta forma la observación a la profundidad i es un tensor de tamaño $n \times (1 + (w \times 2))$ (ecuación 22).

$$x_i = \begin{bmatrix} x^{(i-w)}_1 & \cdots & x^i_1 & \cdots & x^{(i+w)}_1 \\ x^{(i-w)}_2 & \cdots & x^i_2 & \cdots & x^{(i+w)}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x^{(i-w)}_n & \cdots & x^i_n & \cdots & x^{(i+w)}_n \end{bmatrix}_{(n\times(1+(w\times 2)))} \tag{22}$$

En este contexto, las redes CNN1D analizan la información de todos los puntos de profundidad para un único registro, para todos los registros. Esto permite capturar relaciones espaciales entre los diferentes puntos de profundidad y extraer características significativas en la dimensión vertical.

5.3. Optimización y Entrenamiento Modelos

5.3.1. Hardware y Software utilizado

Los modelos se desarrollaron utilizando Python 3.10.11 junto con el framework PyTorch 2.0.1. El software se implementó en un sistema con las siguientes especificaciones de hardware: un procesador AMD Ryzen 9 3900X 12-Core funcionando a 3.80 GHz, una GPU Nvidia GeForce 2060 Super con 8 GB de memoria de video y 16 GB de memoria RAM. Se aprovechó la capacidad de cálculo paralelo de la GPU mediante el uso de CUDA para acelerar los tiempos de entrenamiento.

5.3.2. Función de perdida y matriz de penalización

Los datos presentaron un desbalance significativo en la distribución de las clases de nuestros datos. La clase mayoritaria representa el 54.3 % de las muestras, mientras que la clase minoritaria solo constituye el 3 %. Esta disparidad puede llevar a que el modelo tenga dificultades para aprender patrones y realice predicciones sesgadas hacia la clase dominante.

Para abordar este desafío, se ha propuesto un enfoque basado en dos estrategias complementarias. En primer lugar, se ha introducido un factor de ponderación w dado por la ecuación 23 en la función de pérdida.

$$w_i = \frac{N}{C \times F_i} \tag{23}$$

Donde w_i es el factor de ponderación correspondiente a la clase i, N es el numero total de observaciones en el conjunto de datos, C es el número de clases y F_i Es el número de observaciones en el conjunto de datos que pertenecen a la clase i.

Este factor ponderado tiene en cuenta el peso relativo de cada clase en la tarea de clasificación. Así, se asigna un mayor peso a la clase minoritaria y un peso menor a la clase mayoritaria, con el objetivo de reducir el sesgo hacia la clase dominante y mejorar la capacidad del modelo para aprender de manera equilibrada.

Además, se ha empleado una matriz de penalización P (ecuación 24) para incorporar información adicional sobre las relaciones entre las diferentes clases. Esta matriz de penalización permite asignar penalidades diferentes cuando se comete un error específico en la clasificación. Por ejemplo, se puede penalizar más clasificar la categoría A como B en lugar de clasificarla como C, según la dificultad percibida por un experto para distinguir entre dos categorías específicas. De esta manera, se busca proporcionar al modelo una señal adicional que refuerce la capacidad de discriminación entre clases desafiantes.

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} \end{bmatrix}_{(i \times j)} \tag{24}$$

La matriz de penalización P es de tamaño $i \times j$ donde i representa las etiquetas reales y j representa las predicciones del modelo. En esta matriz de penalización, cada elemento $p_{i,j}$ representa la penalización asociada cuando la etiqueta real es i y la predicción del modelo es j. Es importante destacar que cuando $i = j$, el valor de $p_{i,j} = 1$.

En consecuencia, se ha modificado la función de pérdida de entropía cruzada para incluir tanto el factor de ponderación que aborda el desbalance de clases como el factor de penalización que refleja la dificultad relativa de diferenciación entre clases (ecuación 25). Estas adaptaciones buscan mejorar la capacidad del modelo para realizar predicciones equilibradas y precisas en presencia de desbalance de clases y desafíos de discriminación.

$$l = -\log\left(\frac{\exp(x_i)}{\sum_{c=1}^{C}\exp(x_c)}\right) \times w_i \times P_{i,j} \tag{25}$$

5.3.3. Arquitectura de los modelos

Se proporcionará una descripción detallada de los cinco modelos distintos utilizados en este estudio. Cada modelo ha sido diseñado con el objetivo de evaluar y comparar diferentes enfoques de aprendizaje automático para abordar el problema.

- Modelo KNN (K Nearest Neighbors): Se destaca como un enfoque clásico y notablemente simple en comparación con las redes neuronales. Se ha incluido intencionalmente como punto de comparación frente a las redes neuronales con el propósito de evaluar si un enfoque más sencillo es suficiente para abordar el problema en cuestión. Su incorporación como referencia permite contrastar su rendimiento con las arquitecturas más complejas de las redes neuronales, permitiéndonos determinar si el costo adicional de utilizar una red neuronal se justifica en términos de una mejora significativa en la capacidad para capturar patrones complejos en los datos.

- Modelo FC (Fully Connected): Se caracteriza por su simplicidad al contar únicamente con una capa fully connected (totalmente conectada) seguida de una activación ReLU y una capa de dropout. Esta arquitectura lineal nos permite examinar el rendimiento de un modelo básico sin la incorporación de operaciones de convolución o estructuras recurrentes. Aunque puede parecer limitado en su capacidad para capturar patrones complejos, su enfoque directo y lineal puede proporcionar información valiosa sobre la idoneidad de modelos más simples para la tarea en cuestión.

- Modelo CNN1D (Convolutional Neural Network 1D): Se basa en una única capa convolucional en una dimensión, seguida de una activación ReLU, una capa de maxpooling en una dimensión y una capa de dropout. Esta arquitectura ha demostrado ser eficaz en la extracción de características locales y la detección de patrones en datos unidimensionales, como series temporales o señales. Al aprovechar filtros convolucionales y operaciones de pooling, el modelo es capaz de capturar relaciones espaciales y detectar características relevantes en diferentes regiones de los datos de entrada.

- Modelo LSTM (Long Short-Term Memory): Esta arquitectura se basa en una capa LSTM seguida de una capa de dropout. Las LSTMs son diseñadas para capturar y procesar secuencias de datos a lo largo del tiempo. Son capaces de modelar dependencias a largo plazo y retener información contextual relevante, lo que las hace especialmente adecuadas para el modelado de series temporales y

secuencias con dependencias a largo plazo.

- Modelo CNN1D-LSTM (Convolutional Neural Network 1D - Long Short-Term Memory): Para explorar la combinación de características de diferentes enfoques, se ha desarrollado el modelo CNN1D-LSTM. Este modelo consta de dos ramas principales: una rama LSTM y una rama CNN1D. Estas ramas operan en paralelo, extrayendo características en diferentes niveles de abstracción. La fusión de las salidas de ambas ramas se realiza mediante una capa totalmente conectada seguida de una activación ReLU. Esta aproximación permite aprovechar la capacidad de la LSTM para modelar dependencias a largo plazo y, al mismo tiempo, capturar patrones locales y características relevantes mediante la CNN1D.

- Modelo CNN1D-LSTM-AT (Convolutional Neural Network 1D - Long Short-Term Memory - Attention): Es una extensión de la arquitectura anterior que incorpora un bloque de autoatención en la rama LSTM antes de fusionar las ramas LSTM y CNN1D. Esta adición permite que el modelo enfoque su atención en partes específicas y relevantes de la secuencia de datos, mejorando su capacidad para capturar relaciones y patrones significativos. El bloque de autoatención proporciona adaptabilidad, flexibilidad y una mayor interpretabilidad al asignar pesos diferenciados a diferentes segmentos de la secuencia.

La elección de estas arquitecturas ofrece un enfoque exhaustivo para explorar diferentes niveles de complejidad y combinaciones en la modelización de los datos. Al analizar y comparar el rendimiento de estos modelos, se busca identificar la arquitectura más adecuada y efectiva en la tarea de identificar litología.

5.3.4. Optimización de hiperparametros

En esta sección se describe el proceso de optimización de los modelos de aprendizaje profundo utilizando la biblioteca Optuna 3.1.0. Esta implementa técnicas de optimización bayesiana para buscar de manera eficiente los hiperparámetros óptimos de un modelo. El objetivo principal fue encontrar los conjuntos de hiperparámetros que minimicen el valor de una función objetivo (ecuación 26). Esta función objetivo integra

la minimización de la perdida y la maximización del MCC en un conjunto de datos de validación.

$$f(x) = \text{loss} \times (1 - \text{mcc}) \tag{26}$$

Esta aproximación holística permite alcanzar un equilibrio óptimo entre la precisión de la predicción y la capacidad de clasificación del modelo, lo que resulta en una mayor efectividad general.

Un paso crucial en el proceso de optimización fue definir un espacio de búsqueda adecuado para los hiperparámetros de cada modelo. Se realizó un análisis previo para identificar los hiperparámetros relevantes, incluyendo la tasa de aprendizaje, el número de capas ocultas, el tamaño de la red neuronal, la regularización y otros hiperparámetros específicos de cada arquitectura.

Basándonos en este análisis, se definió un amplio espacio de búsqueda que cubría una amplia gama de configuraciones. Se prestó especial atención a la selección de rangos y distribuciones apropiadas para evitar explorar configuraciones poco realistas o beneficiosas. Para cada hiperparámetro, se establecieron distribuciones continuas o discretas, eligiendo entre distribuciones uniformes o logarítmicas según la naturaleza específica del hiperparámetro en cuestión. Esta estrategia permitió una exploración eficiente del espacio de hiperparámetros y el descubrimiento de combinaciones prometedoras. Las distribuciones utilizadas se encuentran resumidas en Apéndice-Cuadro 6.

Este enfoque cuidadosamente diseñado para definir el espacio de búsqueda de hiperparámetros permitió una exploración sistemática y exhaustiva de las opciones disponibles, maximizando así las posibilidades de encontrar combinaciones óptimas que impulsen el rendimiento del modelo.

En esta etapa del experimento, para cada una de las arquitecturas se entrenaron 100 modelos utilizando diferentes combinaciones de hiperparámetros a lo largo de 50 épocas. Cada modelo se ajustó utilizando el conjunto de entrenamiento y se evaluó su

rendimiento en términos de la función objetivo (ecuación 26).

Posteriormente, se realizó una inspección visual de las métricas de rendimiento obtenidas en el conjunto de entrenamiento para evitar selecciones basadas en aumentos locales de la variable a optimizar.

El modelo que presentó el valor más bajo y consistente fue seleccionado como el mejor modelo. Se utilizaron los hiperparámetros correspondientes a este modelo seleccionado para las evaluaciones posteriores y las predicciones en el conjunto de prueba. Los hiperparámetros que han sido ajustados de manera óptima para cada una de las arquitecturas se encuentran registrados en Apéndice-Cuadro 7.

5.3.5. Validación cruzada de los modelos con hiperparámetros optimizados

En esta sección, se llevaron a cabo pruebas exhaustivas de validación cruzada para evaluar el rendimiento de los modelos con hiperparámetros optimizados. Cada uno de los cinco modelos fue entrenado durante 100 épocas en cada uno de los cinco pliegues de datos. Durante el proceso de entrenamiento, se registró y almacenó el mejor modelo obtenido en términos del MCC en los datos de evaluación.

Una vez completado el entrenamiento en todos los pliegues, se realizó un promedio del MCC obtenido en los datos de evaluación para tener una medida representativa del desempeño general de cada modelo. Este enfoque de validación cruzada garantiza una evaluación rigurosa y confiable de los modelos, al tener en cuenta la variabilidad de los datos de entrenamiento y evaluación en diferentes pliegues.

El uso de esta estrategia permite obtener una visión más completa y robusta del rendimiento de los modelos, al considerar su capacidad para generalizar y adaptarse a diferentes conjuntos de datos.

6. RESULTADOS

6.1. Comparación de modelos puntuales y secciones de datos

En esta sección, presentamos y analizamos los resultados obtenidos al comparar el desempeño de diferentes modelos en la tarea de clasificación de facies litológicas. Los modelos evaluados incluyen aquellos basados en datos puntuales, como KNN (K Nearest Neighbors) y FC (Fully Connected), así como modelos que consideran secciones de datos, como CNN1D (Convolutional Neural Network 1D), LSTM (Long Short-Term Memory), CNN1D -LSTM (Convolutional Neural Network 1D - Long Short-Term Memory) y CNN1D -LSTM-AT (Convolutional Neural Network 1D - Long Short-Term Memory - Attention).

Para cada uno de los modelos se crearon cinco instancias y se entrenaron durante 100 épocas cada una. Cada instancia fue entrenada utilizando un pliegue diferente de datos mediante la metodología de validación cruzada. De esta manera, se obtuvo una evaluación más robusta y representativa del rendimiento de los modelos. El modelo que mostró el mejor MCC fue almacenado para su posterior análisis y comparación.

Los resultados se evaluaron utilizando métricas comúnmente utilizadas en problemas de clasificación, incluyendo la precisión, el F1-score macro y el MCC promedio de todas las categorías. Estas métricas proporcionan una medida del rendimiento global del modelo, su capacidad para capturar patrones complejos y su capacidad para realizar predicciones precisas en cada categoría de facies litológicas.

Los resultados obtenidos para cada modelo y cada métrica se analizaron y se compararon entre sí. Se buscó identificar aquellos modelos que lograron un mejor desempeño en términos de clasificación precisa de las facies litológicas. Además, se evaluó si los modelos basados en secciones de datos mostraron una mejora significativa en comparación con los modelos basados en datos puntuales.

A continuación, presentaremos los resultados obtenidos para cada uno de los modelos evaluados en Cuadro 3, Cuadro 4 y Cuadro 5.

Fold	KNN	FC	CNN1D	LSTM	CNN1D LSTM	CNN1D LSTM AT
1	0.776	0.745	0.756	0.740	0.778	0.770
2	0.733	0.719	0.748	0.725	0.729	0.721
3	0.679	0.650	0.684	0.653	0.703	0.647
4	0.736	0.671	0.717	0.711	0.742	0.734
5	0.698	0.610	0.650	0.670	0.683	0.666
Promedio	0.724	0.679	0.711	0.700	0.727	0.708

Cuadro 3: Presicion (%) de todos los modelos en cada pliegue

Modelo	KNN	FC	CNN1D	LSTM	CNN1D LSTM	CNN1D LSTM AT
1	0.539	0.607	0.650	0.594	0.653	0.657
2	0.554	0.592	0.654	0.624	0.647	0.638
3	0.495	0.515	0.565	0.528	0.598	0.553
4	0.506	0.545	0.599	0.569	0.598	0.602
5	0.531	0.477	0.522	0.517	0.546	0.525
Promedio	0.525	0.547	0.598	0.566	0.608	0.595

Cuadro 4: Macro f1-score de todos los modelos en cada pliegue

Modelo	KNN	FC	CNN1D	LSTM	CNN1D LSTM	CNN1D LSTM AT
1	0.459	0.522	0.573	0.517	0.589	0.588
2	0.478	0.519	0.590	0.550	0.579	0.570
3	0.411	0.442	0.494	0.458	0.526	0.479
4	0.430	0.479	0.536	0.498	0.536	0.541
5	0.463	0.424	0.481	0.473	0.495	0.472
Promedio	0.448	0.477	0.535	0.499	0.545	0.530

Cuadro 5: Matthews Correlation Coefficient de todos los modelos en cada pliegue

La comparación de métricas entre los modelos basados en datos puntuales (KNN y FC) y los modelos de secciones de datos (CNN1D, LSTM, CNN1D-LSTM, CNN1D-LSTM-AT) reveló diferencias significativas en su desempeño.

En términos de precisión, los modelos de secciones de datos superaron a los modelos puntuales. Específicamente, los modelos CNN1D-LSTM y CNN1D-LSTM-AT lograron la precisión más alta, seguidos por el CNN1D. Estos resultados indicaron que la capacidad de capturar características locales y patrones espaciales en las secciones de datos mejora la precisión de la clasificación en comparación con los modelos basados en información de un único punto.

En cuanto al F1-score macro, nuevamente los modelos de secciones de datos mostraron un mejor desempeño. Los modelos CNN1D-LSTM y CNN1D obtuvieron los valores más altos de F1-score, demostrando una mejor capacidad para clasificar correctamente todas las categorías de facies litológicas.

El MCC también favorece a los modelos de secciones de datos. Los modelos CNN1D-LSTM y CNN1D-LSTM-AT obtuvieron los valores más altos de MCC, seguidos por el CNN1D.

En general, al comparar los modelos basados en datos puntuales con los modelos de secciones de datos, se observó un claro patrón de desempeño superior en los modelos de secciones de datos. Estos modelos, como el CNN1D-LSTM, mostraron mejores resultados en términos de precisión, F1-score y MCC en la clasificación de facies litológicas. Estos modelos aprovechan la capacidad de las redes neuronales convolucionales y las LSTM para capturar patrones locales y relaciones espaciales en los datos, lo que mejora significativamente el rendimiento de la clasificación en comparación con los modelos más simples basados en un único dato. Estos hallazgos respaldan la idea de que la incorporación de arquitecturas más complejas en los modelos puede mejorar la capacidad de capturar patrones complejos y, en consecuencia, aumentar la precisión en la clasificación de facies litológicas.

En adición al análisis de las métricas de rendimiento, se realizó el cálculo de las matrices de confusión para cada pliegue en todos los modelos evaluados, proporcionando una visión detallada de las clasificaciones realizadas por los modelos y revelando el grado de acierto y error en la clasificación de cada categoría. Con el fin de obtener una evaluación más precisa y generalizada del rendimiento de los modelos, se aplicó el promedio de las matrices de confusión entre los diferentes pliegues.

El enfoque de promediar las matrices de confusión mediante la validación cruzada ofrece beneficios sustanciales al proporcionar una evaluación más amplia y sólida del rendimiento de los modelos. Al suavizar las variaciones entre los diferentes pliegues, se

obtuvo una medida más estable y representativa del desempeño promedio. Esta estrategia no solo permite estimar el rendimiento esperado del modelo en datos desconocidos, sino que también reduce el sesgo asociado a un pliegue específico. Además, facilita la identificación tanto de las fortalezas como las limitaciones de cada enfoque al destacar aquellas áreas en las que cada modelo demuestra un rendimiento sobresaliente, así como los casos en los que enfrenta desafíos significativos en la clasificación de ciertas categorías, aportando una comprensión más clara de la capacidad del modelo para manejar diferentes clases.

En la Figura 7 se muestran los resultados promedio de las matrices de confusión correspondientes a cada uno de los modelos evaluados. Estas matrices de confusión representan visualmente las clasificaciones realizadas por los modelos, permitiendo un análisis detallado de su desempeño en la clasificación de cada categoría.

En el análisis de los resultados obtenidos para el modelo KNN, se observó una alta precisión en la clasificación de la categoría 'Sh', con un porcentaje de aciertos superior al 80 %. Además, se encontró una precisión aceptable en la clasificación de las categorías 'Sa' y 'Sa/Sh', con porcentajes de aciertos superiores al 60 %. Sin embargo, se identificaron algunas dificultades en la clasificación de la categoría 'Ma', donde se observó un porcentaje de acierto de tan solo 43 %. En esta categoría se observó una confusión con la categoría 'Sh'. Por otro lado, el modelo mostró un desempeño muy pobre al clasificar la categoría 'Li', donde se encontró un porcentaje de acierto cercano al 10 %. En este caso, se evidenció una alta confusión entre esta categoría y la clase 'Sh'. Estos patrones revelan una inclinación hacia la predicción de la categoría sobrerrepresentada 'Sh'.

Con relación al modelo FC se observó un rendimiento equilibrado en la mayoría de las categorías, con tasas de acierto que oscilaron entre el 60 % y 80 %. No obstante, se identificaron desafíos particulares en la clasificación de la categoría 'Li', donde se obtuvieron tasas de error más altas en comparación con las demás categorías. Al analizar las predicciones para las observaciones de esta categoría, se encontró una distribución similar respecto de las demás, comparable a una predicción al azar, lo que sugiere una

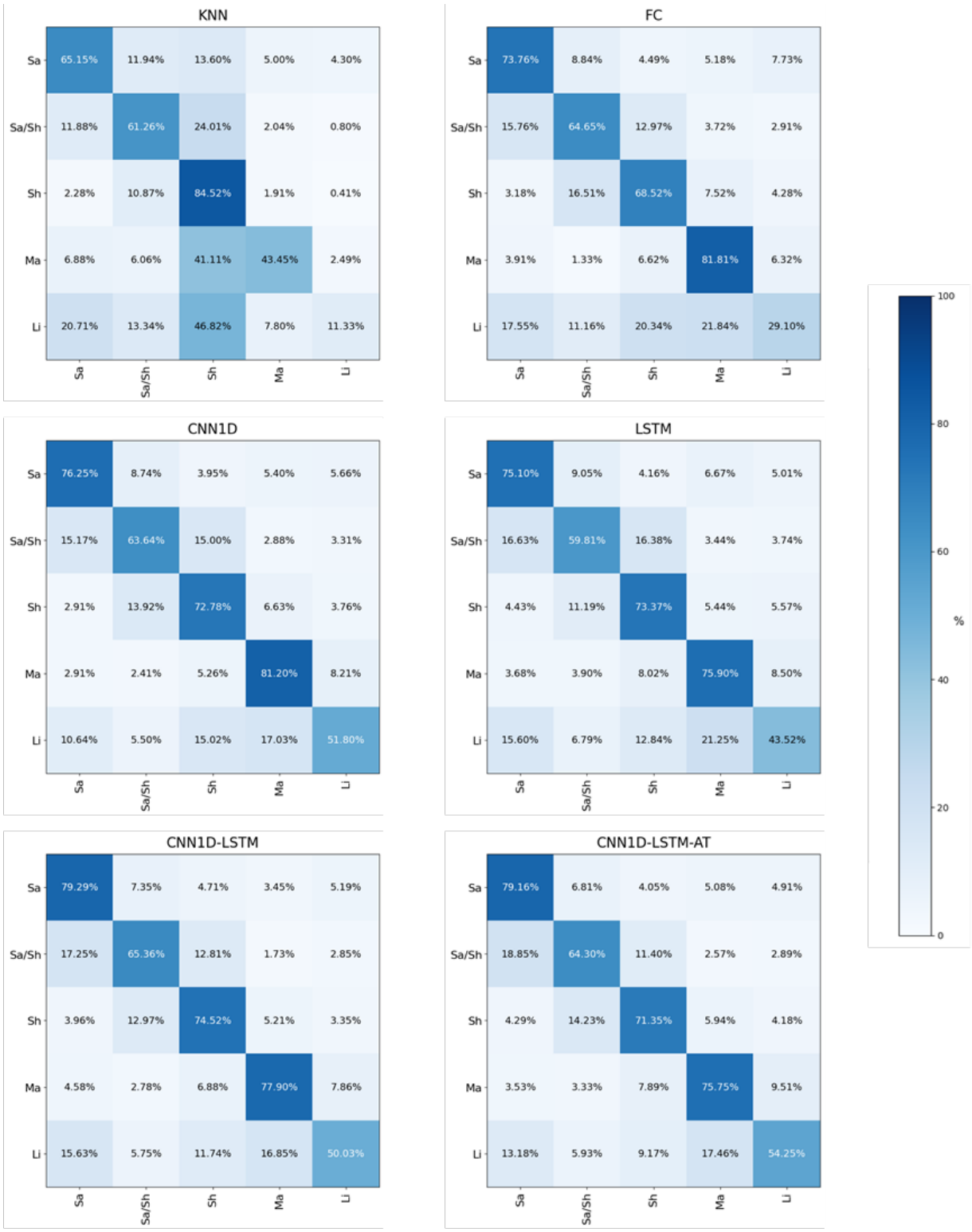

Figura 7: Matrices de correlación promediadas para cada pliegue en los modelos evaluados, para las categorías: Sandstone (Sa), Sandstone/Shale (Sa/Sh), Shale (Sh), Marl (Ma) y Limestone (Li).

posible dificultad del modelo en identificar características diferenciables.

Respecto a los modelos CNN1D, LSTM, CNN1D-LSTM y CNN1D-LSTM-AT, se obtuvieron resultados muy similares. Se observó un alto grado de precisión para las categorías 'Ma' y 'Sa', con porcentajes de aciertos superiores al 75 % en todos los casos. Por otro lado, se obtuvieron desempeños ligeramente inferiores para la categoría 'Sh'. La categoría 'Sa/Sh' presentó valores cercanos al 60 % de precisión, mientras que la categoría 'Li' tuvo precisiones superiores al 50 % en todos los modelos, excepto en el modelo LSTM, lo que indica cierta capacidad de diferenciación para esta categoría, aunque con algunas limitaciones en el modelo LSTM.

Un patrón interesante que se observó en todos los modelos es la tendencia a confundir las categorías 'Sa' y 'Sh' con la categoría 'Sa/Sh'. Del mismo modo, la categoría 'Sa/Sh' es confundida con las categorías 'Sa' y 'Sh'. Esto se debe a que estas categorías se encuentran en un espectro donde 'Sa' y 'Sh' representan los extremos, mientras que 'Sa/Sh' se sitúa en el medio, lo que dificulta establecer una línea clara de separación. Además, se observó que la categoría 'Li' es comúnmente confundida con la categoría 'Ma', debido a las mismas razones anteriormente mencionadas. También se encontró que la categoría 'Li' es comúnmente confundida con la categoría 'Sa'. Estos resultados reflejan la dificultad inherente a la similitud de los registros de pozo utilizados en su caracterización, los cuales pueden presentar solapamientos en sus valores. Esta superposición dificulta una clasificación precisa y una distinción clara entre ambas categorías.

6.2. Selección del modelo

Tras un análisis exhaustivo de las métricas de evaluación recopiladas en las Cuadro 3, Cuadro 4 y Cuadro 5, junto con la revisión de las matrices de confusión presentadas en la Figura 7, se puede concluir que el modelo CNN1D ofrece el mejor rendimiento general en términos de precisión, capacidad de clasificación y correlación de las predicciones con los valores reales. A diferencia de los modelos FC y KNN, este modelo no presenta dificultades en la identificación de categorías. Además, aunque ofrece un rendimiento similar a los modelos compuestos como CNN1D-LSTM y CNN1D-LSTM-AT, el modelo

CNN1D es significativamente más sencillo en su estructura. Cabe destacar que se ha observado un desempeño ligeramente superior del modelo CNN1D en comparación con el modelo LSTM.

6.3. Eliminación de ruido

En esta sección, se exploró el impacto del suavizado en el rendimiento del modelo CNN1D mediante la técnica de eliminación de ruido utilizando la descomposición de la ondícula. La descomposición se realizó en el número máximo de niveles permitidos para la longitud de la señal de datos, utilizando la ondícula 'db4'. Este proceso se aplicó a todos los registros del conjunto de datos.

Con el objetivo de evaluar la efectividad del suavizado, se realizaron 12 experimentos comparativos en los que se estableció un punto de referencia utilizando los datos originales sin suavizar (experimento '00'). En cada experimento, se establecieron en cero los coeficientes correspondientes a los niveles seleccionados, y se entrenó una nueva instancia del modelo CNN1D utilizando validación cruzada.

El primer experimento consistió en establecer en cero los coeficientes desde el nivel 1 hasta el último nivel de descomposición. A medida que avanzamos en los experimentos, se incorporaron niveles de descomposición más altos. En el experimento 12, se establecieron en cero los coeficientes desde el nivel 12 hasta el último nivel de descomposición.

Posteriormente, se evaluaron las predicciones de los modelos ya entrenados utilizando los datos de validación en todos los pliegues. Para medir el desempeño de cada modelo, se utilizaron tres métricas: accuracy, F1-score macro y MCC promedio por categoría.

Al analizar los resultados obtenidos con estas métricas, podremos determinar cómo el suavizado afecta la capacidad del modelo CNN1D para realizar predicciones precisas y discernir si existe un equilibrio óptimo entre la reducción de ruido y la preservación de la información relevante en nuestros datos.

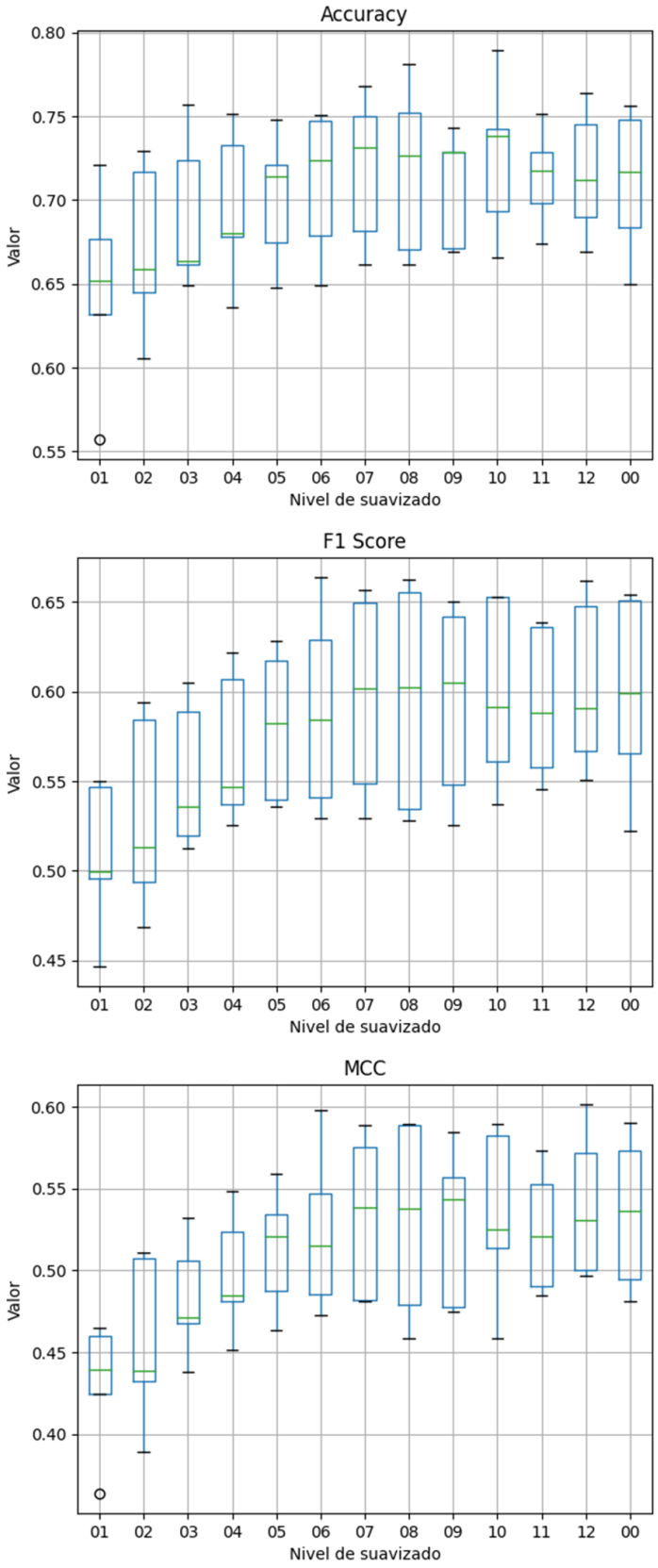

Figura 8: Desempeño del modelo CNN1D a distintos niveles de suavizado.

Los resultados de los experimentos resumidos en la Figura 8 se pueden consultar en las tablas completas en el Apéndice-Cuadro 8. Al analizar detenidamente estos resultados, se observó que los experimentos del 01 al 06 mostraron valores promedio más bajos en las tres métricas de evaluación en comparación con el modelo sin suavizado. Esto sugiere que, en general, el suavizado tuvo un impacto negativo en el rendimiento del modelo. Por el contrario, los experimentos del 07 al 12 mostraron valores promedio de MCC, F1-score y accuracy que son comparables o incluso ligeramente mejores que los del modelo sin suavizado. Estos hallazgos revelan que, en estos experimentos particulares, el suavizado tuvo un impacto neutro o incluso ligeramente benéfico en el rendimiento del modelo, resultando en un desempeño similar al del modelo sin suavizado.

6.4. Explicabilidad del modelo

Para analizar la explicabilidad de nuestro modelo de clasificación multiclase, empleamos el método SHAP (SHapley Additive exPlanations). Para este análisis, seleccionamos un subconjunto de datos balanceado, compuesto exclusivamente por observaciones en las que el modelo realizó predicciones correctas.

Dado que se trabaja con datos secuenciales, se requiere un enfoque especial para calcular los valores SHAP de los registros. Para cada observación en el subconjunto balanceado, se calcularon los valores SHAP correspondientes a cada uno de los registros en cada punto de la secuencia. Luego, se obtuvo el promedio de los valores absolutos de los valores SHAP por registro para tener un único valor que represente la importancia de cada variable para cada observación.

Además, se realizó un promedio adicional por registro entre todas las observaciones del subconjunto balanceado. De esta manera, se obtuvo un único valor representativo de la importancia de cada registro para el modelo CNN1D, para cada una de las categorías y en total. Este enfoque nos permite tener una visión general de la importancia relativa de los registros en la tarea de clasificación. Este proceso se realizó para cada uno de los

pliegues y se promediaron los resultados para obtener una mejor representatividad, los resultados se encuentran resumidos en la Figura 9

Durante el análisis de importancia de variables en el modelo de clasificación multiclase, se evidenciaron diferencias significativas en la relevancia de cada una de ellas. Entre los registros examinados, 'GR' sobresalió como el más importante. La presencia de radiación gamma emitida por las formaciones rocosas se consideró como un indicio relevante de material radiactivo y tuvo una influencia significativa en las predicciones del modelo.

Asimismo, los registros resistivos 'RMED/RDEP' demostraron una importancia relevante. Estas variables representan la resistividad de las capas intermedias y más profundas de la formación. Los resultados sugieren que la resistividad en diferentes zonas de la formación juega un papel crucial en la capacidad del modelo para realizar predicciones precisas.

El registro 'DTC' también se destacó como importante en el modelo. Una mayor velocidad de propagación indicaría la presencia de formaciones compactas o rocas densas, lo cual resulta valioso para la identificación de ciertos tipos de formaciones geológicas.

Por otro lado, las variables 'NPHI', 'RHOB' y 'PEF' mostraron una importancia relativamente menor en el modelo. Si bien la porosidad 'NPHI' es un factor relevante en la clasificación, no existe una relación directa y única entre la porosidad y un tipo particular de formación geológica, ya que puede variar dentro de una misma formación debido a diversos factores como la compactación, cementación, fracturación y la historia geológica específica.

En cuanto a 'RHOB', aunque proporciona información sobre la masa y compacidad de la roca, su correlación con tipos específicos de formaciones geológicas no es directa. Su contribución a las predicciones correctas resultó relativamente menor en comparación con otras variables.

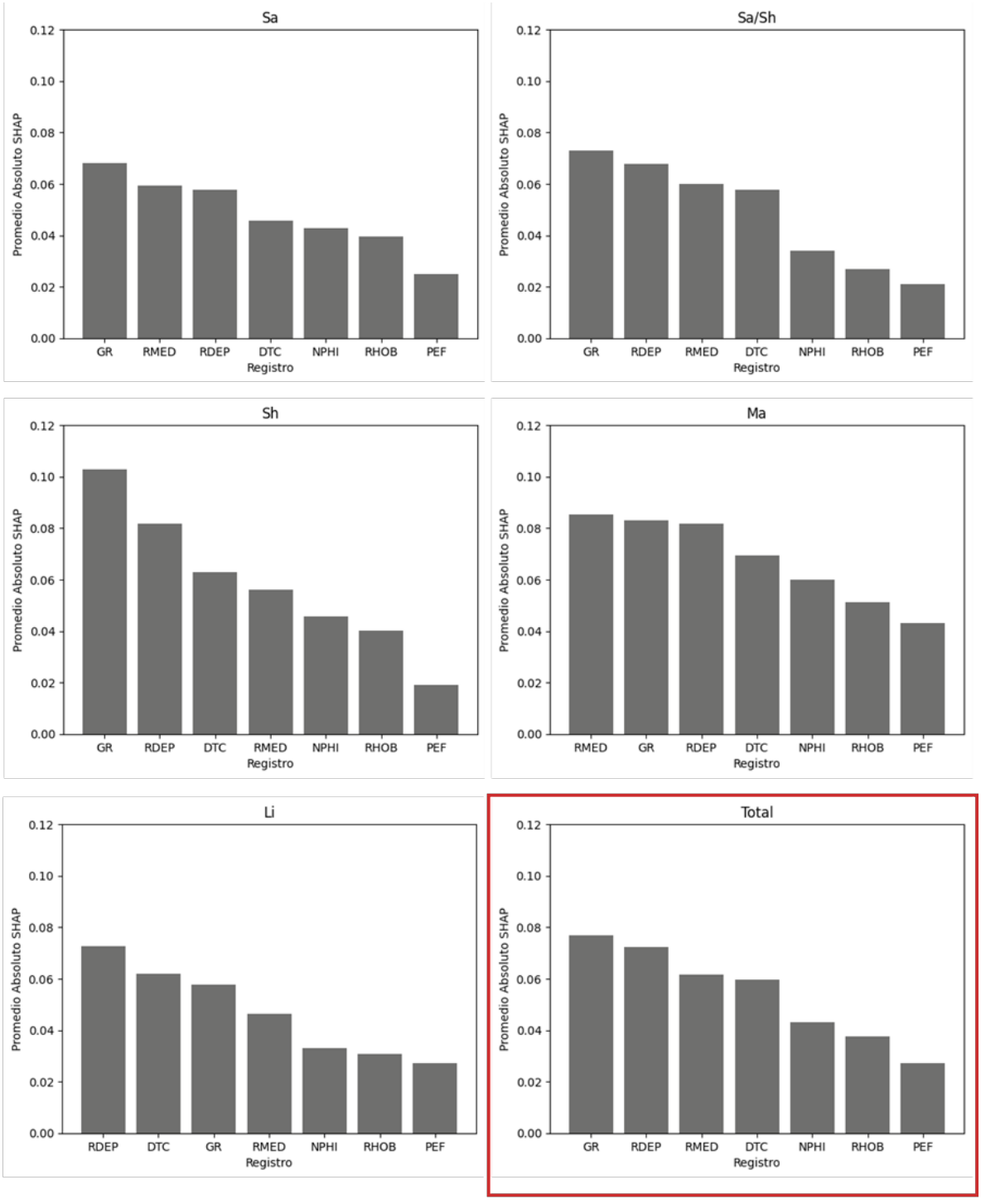

Figura 9: Valor SHAP promedio para el modelo CNN1D para la importancia de los registros por categoría y total.

Finalmente, 'PEF' se destacó como el registro de menor importancia. Aunque puede proporcionar información sobre la composición y las propiedades de la formación, se considera un indicador complementario en el análisis de las propiedades de la roca. Su influencia en la clasificación precisa resultó relativamente menor en comparación con otras variables.

Después de analizar la importancia de cada variable en el modelo de clasificación en su conjunto, se llevó a cabo un análisis adicional para evaluar la importancia de las variables dentro de cada categoría específica. Esta aproximación permitió entender cómo cada variable contribuye a la clasificación precisa de cada categoría en particular.

El análisis de los valores SHAP revela que todas las variables desempeñan un papel relevante en la clasificación de las diferentes categorías. Sin embargo, la importancia relativa de cada variable varía dependiendo de la categoría considerada. Este descubrimiento sugiere que cada categoría puede estar asociada con patrones y características específicas, lo que resulta en la importancia diferencial de las variables en su clasificación.

7. DISCUSION

Durante la comparación de los modelos puntuales KNN y FC con los modelos de secciones CNN1D, LSTM, CNN1D-LSTM y CNN1D-LSTM-AT, se pudo observar que, en términos de las métricas de desempeño utilizadas en este estudio (precisión, F1-score y MCC), ambos tipos de modelos presentaron desempeños muy similares. Sin embargo, al analizar las matrices de confusión se evidenció que el modelo KNN mostró un sobreajuste en la predicción de la categoría sobrerrepresentada 'Sh', mientras que el modelo FC tuvo un buen desempeño en general, a excepción de la categoría de menor representación, 'Li', donde las clasificaciones fueron igualmente distribuidas entre todas las categorías. Los modelos de datos puntuales mostraron limitaciones al capturar los patrones necesarios para la clasificación precisa de todas las categorías. Estos modelos no lograron abarcar la complejidad y las características distintivas de cada categoría de forma adecuada, lo que condujo a un desempeño subóptimo en la clasificación de todas las clases.

Con relación a los modelos de secciones de datos, tanto los modelos simples (CNN1D y LSTM) como los modelos compuestos (CNN1D-LSTM y CNN1D-LSTM-AT) presentaron un desempeño muy similar. Esto sugiere que los patrones presentes en los datos fueron capturados de manera efectiva por los modelos más simples, y no se encontró una mejora significativa al aumentar la complejidad del modelo o al utilizar mecanismos de atención adicionales. Por lo tanto, se decidió seleccionar el modelo CNN1D debido a su desempeño ligeramente superior en comparación con el modelo LSTM.

Al analizar la matriz de confusión del modelo propuesto se observó que éste no presentaba sobreajustes en las categorías sobrerrepresentadas. La categoría 'Ma' fue la mejor clasificada, a pesar de ser una de las menos representadas. Le siguieron en desempeño las categorías 'Sa' y 'Sh', mientras que la categoría 'Li' mostró un desempeño aceptable. Al analizar la matriz de confusión, se identificó que muchos de los errores en la clasificación se debían a la dificultad de establecer límites claros dentro de las categorías, dado que estas pertenecían a un espectro. Por ejemplo, en el espectro 'Sa - Sa/Sh - Sh', se observó que la categoría 'Sa/Sh' era la que generaba más confusión con

las categorías 'Sa' y 'Sh', mientras que ésta última también era clasificada incorrectamente como 'Sa' o 'Sh'. Asimismo, se identificó una tendencia de confusión entre las categorías 'Li' y 'Ma', donde ambas se clasificaban erróneamente una como la otra.

Es importante destacar que, a pesar del desbalance inicial entre las categorías 'Sh' y 'Li' en una proporción de 18:1 en los datos originales, se aplicaron estrategias como la modificación de la función de pérdida de entropía cruzada para tener en cuenta este desbalance y la utilización de una pérdida diferencial. Estas medidas permitieron abordar de manera efectiva el desbalance y mejorar el desempeño general de la clasificación.

Se realizó un proceso de eliminación de ruido mediante la transformada de la ondícula, descartando los coeficientes de detalle en diferentes niveles. El objetivo era encontrar un equilibrio óptimo entre la eliminación del ruido y la preservación de los patrones relevantes en los datos. Sin embargo, los resultados revelaron que, en el mejor de los casos, la eliminación del ruido no tuvo ningún efecto en la mejora de la clasificación. En el peor de los casos, incluso tuvo un efecto contraproducente en la precisión de la clasificación. Estos hallazgos sugieren que, a pesar del ruido presente en los datos, el modelo es capaz de identificar y capturar los patrones necesarios para la clasificación. Además, se evidenció que esta técnica de eliminación de ruido es inefectiva en modelos tan complejos como el propuesto, y puede tener un impacto significativo únicamente en modelos más simples, como el KNN.

Se realizó un análisis de importancia de características utilizando SHAP, el cual reveló que todas las variables son relevantes en la clasificación, pero algunos registros son más importantes que otros. Específicamente, los registros más destacados en términos de importancia son GR, RDEP, RMED y DTC. Por otro lado, los registros NPHI, RHOB y PEF contribuyen de manera menos significativa en la clasificación de las muestras.

Además, se llevó a cabo un análisis para determinar las variables más importantes en la identificación de muestras de categorías específicas. En este sentido, al igual que

un petrofísico experto que se centra en el registro GR para identificar categorías de rocas siliciclásticas como 'Sa', 'Sa/Sh' y 'Sh', el modelo también muestra una atención especial hacia este registro. Del mismo modo, al momento de clasificar rocas calcáreas como 'Li' y 'Ma', el modelo da mayor importancia al registro DTC. Este paralelismo entre el enfoque del modelo y el de un experto petrofísico resulta interesante y puede proporcionar una mayor confianza en los resultados del modelo.

Una posible limitación de este estudio radica en la falta de representatividad de la muestra recolectada, la cual se limitó a una región geográfica específica. Esto implica que los resultados obtenidos pueden no ser generalizables a otras áreas con distintas características geológicas. Además, es importante mencionar que la selección de los rangos de los registros y el umbral mínimo de longitud de datos continuos también podría haber introducido sesgos en el análisis.

Es relevante destacar que este estudio se centró en las categorías más comúnmente importantes para la caracterización de yacimientos, como 'Sandstone', 'Shale', 'Sandstone/Shale', 'Limestone' y 'Marl'. Sin embargo, no se incluyeron categorías como 'Chalk', 'Tuff', 'Coal', 'Dolomite', 'Halite', 'Anhydrite' y 'Basement', las cuales pueden ser críticas en la caracterización de yacimientos en diferentes contextos geológicos. Por lo tanto, es necesario tener en cuenta esta limitación y considerar la relevancia de estas categorías adicionales en futuros estudios para una caracterización más completa y precisa de los yacimientos.

Los hallazgos de este estudio tienen implicaciones prácticas significativas para la industria petrolera, optimizando la selección de parámetros en futuros estudios y exploraciones. Esta optimización permite asignar recursos de manera más eficiente, maximizando la eficiencia y reduciendo los costos asociados. Además, estos resultados contribuyen al avance del conocimiento en petrofísica al revelar similitudes entre el enfoque del modelo propuesto y el enfoque tradicional de los expertos en la identificación de formaciones. Esto fortalece la validez y confiabilidad de los métodos de inteligencia artificial aplicados, impactando las decisiones estratégicas y técnicas en la industria

petrolera al mejorar las evaluaciones de yacimientos y la planificación de perforaciones.

La implementación de modelos como el presentado en este estudio brinda un valioso apoyo a los petrofísicos al acelerar el proceso de análisis y proporcionar resultados consistentes y precisos de manera eficiente. Utilizando técnicas de inteligencia artificial y aprendizaje automático, estos modelos procesan grandes volúmenes de datos, capturando patrones sutiles y complejos que podrían pasarse por alto en un análisis manual. Esto mejora la calidad y confiabilidad de las interpretaciones al proporcionar información detallada y completa.

Además, los modelos superan los sesgos y subjetividades inherentes a la interpretación humana al ofrecer una base objetiva y consistente para la toma de decisiones. Esto aumenta la confiabilidad y mejora la reproducibilidad y comparabilidad de los análisis entre diferentes expertos.

En última instancia, la aplicación de estos hallazgos puede traducirse en una mayor precisión en la identificación y caracterización de formaciones geológicas, lo que resulta en una mejora en la estimación de reservas, la reducción de riesgos operativos y la optimización de los procesos de extracción de hidrocarburos. Esto no solo beneficia a las empresas de la industria petrolera en términos de rentabilidad y competitividad, sino que también tiene un impacto económico y ambiental más amplio al garantizar una gestión más eficiente y sostenible de los recursos energéticos.

Para las futuras investigaciones, se plantean diversas direcciones interesantes. En primer lugar, sería enriquecedor realizar clasificaciones litológicas utilizando conjuntos de datos provenientes de distintas ubicaciones geográficas. Este enfoque permitiría generalizar las predicciones y evitar el sesgo asociado a una cuenca específica. Además, se sugiere explorar la aplicabilidad de otros modelos dentro del ámbito de las redes neuronales, así como también considerar la combinación de modelos mediante técnicas como XGBoost u otras opciones prometedoras.

Un área de investigación intrigante es la utilización de técnicas de transferencia de aprendizaje, aprovechando los conocimientos adquiridos en dominios relacionados. Esta estrategia podría fortalecer la capacidad de los modelos para clasificar litologías con mayor precisión y eficiencia.

Asimismo, se recomienda explorar enfoques para abordar el desbalance de los datos, como el uso de redes generativas adversariales (GAN). Estas redes podrían contribuir a equilibrar la distribución de las clases en el conjunto de datos y mejorar la capacidad de los modelos para manejar dicha variabilidad.

Además, sería sumamente interesante analizar la importancia de los registros en múltiples cuencas y diversas litologías. Esta amplia exploración permitiría comprender de manera más precisa y completa la relevancia real de cada característica en la tarea de clasificación litológica. Al estudiar múltiples contextos geológicos, se podrían identificar patrones consistentes y determinar qué características son realmente fundamentales en la identificación de las litologías. Este enfoque enriquecería nuestro entendimiento y proporcionaría una base más sólida para la interpretación y aplicación de los modelos en diferentes entornos geológicos.

Por último, se plantea la posibilidad de investigar la función de pérdida modificada que se propone en este estudio. Esta función puede ser especialmente útil al tratar con desafíos relacionados con el desbalance de clases y abordar categorías difíciles de clasificar. Adaptar la función de pérdida según las características de estas categorías problemáticas puede mejorar la precisión y la capacidad de los modelos para distinguir entre litologías complejas o similares.

8. CONCLUSIONES

En este estudio, se propuso un modelo de identificación automática de facies litológicas basado en un conjunto específico de registros de pozo y se plantearon varios objetivos específicos para lograrlo. Se implementaron técnicas de procesamiento de señales para eliminar el ruido y detectar valores atípicos en los registros de pozo, y se utilizaron modelos de aprendizaje automático basados en datos puntuales y secciones de datos para identificar las facies litológicas.

Al comparar diferentes modelos, se encontró que los modelos puntuales y los modelos de secciones de datos presentaban desempeños similares en términos de métricas de desempeño, pero los modelos puntuales mostraban limitaciones al capturar la complejidad y las características distintivas de cada categoría. Por otro lado, los modelos de secciones de datos más simples fueron tan efectivos como los modelos más complejos, lo que sugiere que no había una mejora significativa al aumentar la complejidad del modelo. En consecuencia, se seleccionó el modelo CNN1D como el más adecuado para este estudio.

El análisis de la matriz de confusión reveló que el modelo propuesto no presentaba sobreajuste en las categorías sobrerrepresentadas y mostraba un desempeño aceptable en general. Sin embargo, se identificaron desafíos en la clasificación de categorías que pertenecían a un espectro y presentaban límites difusos. A pesar de estas dificultades, el modelo logró capturar los patrones necesarios para la clasificación de las facies litológicas.

Se implementaron estrategias para abordar el desbalance de clases, como la modificación de la función de pérdida y el uso de una pérdida diferencial, y se encontró que estas medidas fueron efectivas para mejorar el desempeño general de la clasificación. Además, se realizó un proceso de eliminación de ruido, pero se encontró que no tuvo un impacto significativo en la mejora de la clasificación.

El análisis de importancia de características mostró que todos los registros de pozo

eran relevantes en la clasificación, pero algunos sobresalieron ante. Los registros GR, RDEP, RMED y DTC se destacaron como los más influyentes en la identificación de las litologías. Además, se observó una coincidencia entre la importancia de ciertos registros en el modelo y la atención que un petrofísico experto presta a esos mismos registros en la identificación de las litologías.

Es importante destacar que este estudio se centró en categorías litológicas específicas y se reconoce la importancia de incluir otras categorías en futuros estudios para una caracterización más completa de los yacimientos.

Los hallazgos de este estudio tienen implicaciones prácticas significativas para la industria petrolera, al proporcionar un apoyo valioso a los petrofísicos en el análisis de facies litológicas. Los modelos propuestos aceleran el proceso de análisis, ofrecen resultados consistentes y precisos, y superan las limitaciones asociadas con la interpretación humana.

Se proponen diversas direcciones para futuras investigaciones con el objetivo de mejorar este estudio. Estas incluyen el uso de conjuntos de datos de diferentes ubicaciones geográficas, explorar modelos adicionales a las redes neuronales, aplicar técnicas de transferencia de aprendizaje, balancear los datos mediante redes generativas adversarias (GAN) y estudiar la función de pérdida modificada propuesta en este estudio.

9. REFERENCIAS

[1] Charu C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Google-Books-ID: achqDwAAQBAJ. Springer, 25 de ago. de 2018. 512 págs. ISBN: 978-3-319-94463-0.

[2] Richard M. Bateman. *Openhole Log Analysis and Formation Evaluation*. Google-Books-ID: GomDNAEACAAJ. Society of Petroleum Engineers, 2012. 653 págs. ISBN: 978-1-61399-156-5.

[3] Giuseppe Bonaccorso. *Machine Learning Algorithms - Second Edition: Popular Algorithms for Data Science and Machine Learning, 2nd Edition*. Google-Books-ID: 4xjAuQEACAAJ. Packt Publishing, 30 de ago. de 2018. 522 págs. ISBN: 978-1-78934-799-9.

[4] Peter Bormann et al. "FORCE 2020 Well well log and lithofacies dataset for machine learning competition [Data set]". En: (18 de dic. de 2020). DOI: 10.5281/zenodo.4351156.

[5] Sabri Boughorbel, Fethi Jarray y Mohammed El-Anbari. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". En: *PLoS ONE* 12.6 (2 de jun. de 2017), e0177678. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0177678. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5456046/ (visitado 17-11-2022).

[6] J. M. Busch, W. G. Fortney y L. N. Berry. "Determination of Lithology From Well Logs by Statistical Analysis". En: *SPE Formation Evaluation* 2.4 (1 de dic. de 1987), págs. 412-418. ISSN: 0885-923X. DOI: 10.2118/14301-PA. URL: https://doi.org/10.2118/14301-PA (visitado 06-11-2022).

[7] Francois Chollet, Tomasz Kalinowski y J. J. Allaire. *Deep Learning with R, Second Edition*. Google-Books-ID: 4GF6EAAAQBAJ. Simon y Schuster, 26 de jul. de 2022. 566 págs. ISBN: 978-1-63343-984-9.

[8] T. Cover y P. Hart. "Nearest neighbor pattern classification". En: *IEEE Transactions on Information Theory* 13.1 (ene. de 1967). Conference Name: IEEE Transac-

tions on Information Theory, págs. 21-27. ISSN: 1557-9654. DOI: 10.1109/TIT.1967.1053964.

[9] Pierre Delfiner, Olivier Peyret y Oberto Serra. "Automatic Determination of Lithology From Well Logs". En: *SPE Formation Evaluation* 2.3 (1 de sep. de 1987), págs. 303-310. ISSN: 0885-923X. DOI: 10.2118/13290-PA. URL: https://doi.org/10.2118/13290-PA (visitado 06-11-2022).

[10] The Norwegian Petroleum Directorate. *Well definitions*. Well definitions. URL: https://www.npd.no/en/facts/wells/well-definitions/ (visitado 15-11-2022).

[11] *Distributed collaborative prediction: Results of the machine learning contest — The Leading Edge*. URL: https://library.seg.org/doi/10.1190/tle36030267.1 (visitado 15-11-2022).

[12] Martin K. Dubois, Geoffrey C. Bohling y Swapan Chakrabarti. "Comparison of four approaches to a rock facies classification problem". En: *Computers & Geosciences* 33.5 (1 de mayo de 2007), págs. 599-617. ISSN: 0098-3004. DOI: 10.1016/j.cageo.2006.08.011. URL: https://www.sciencedirect.com/science/article/pii/S0098300406001956 (visitado 11-11-2022).

[13] Darwin V. Ellis y Julian M. Singer. *Well Logging for Earth Scientists*. Dordrecht, 19 de oct. de 2010. 728 págs. ISBN: 978-90-481-6947-4.

[14] Peter I. Frazier. *A Tutorial on Bayesian Optimization*. 8 de jul. de 2018. DOI: 10.48550/arXiv.1807.02811. arXiv: 1807.02811[cs,math,stat]. URL: http://arxiv.org/abs/1807.02811 (visitado 18-11-2022).

[15] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". En: *Biological Cybernetics* 36.4 (1 de abr. de 1980), págs. 193-202. ISSN: 1432-0770. DOI: 10.1007/BF00344251. URL: https://doi.org/10.1007/BF00344251 (visitado 15-11-2022).

[16] Paul Glover. *Petrophysics MSc Course Notes*. University of Aberdeen, 2001.

[17] Hingle. "The use of logs in exploration problems". En: *SEG 29th Annual Meeting*. (1959).

[18] Sepp Hochreiter y Jürgen Schmidhuber. "Long Short-term Memory". En: *Neural computation* 9 (1 de dic. de 1997), págs. 1735-80. DOI: 10.1162/neco.1997.9.8.1735.

[19] Bruno Honório et al. "Well Log Denoising and Geological Enhancement Based on Discrete Wavelet Transform and Hybrid Thresholding". En: *Energy Exploration & Exploitation* 30 (5 de jun. de 2012), págs. 417-434. DOI: 10.1260/0144-5987.30.3.417.

[20] IEA. *Oil 2021*. 2021. URL: https://iea.blob.core.windows.net/assets/1fa45234-bac5-4d89-a532-768960f99d07/Oil_2021-PDF.pdf.

[21] Chayawan Jaikla et al. "FaciesNet: Machine Learning Applications for Facies Classification in Well Logs". En: 20 de nov. de 2019.

[22] Fei Tony Liu, Kai Ming Ting y Zhi-Hua Zhou. "Isolation Forest". En: *2008 Eighth IEEE International Conference on Data Mining*. 2008 Eighth IEEE International Conference on Data Mining. ISSN: 2374-8486. Dic. de 2008, págs. 413-422. DOI: 10.1109/ICDM.2008.17.

[23] Scott M. Lundberg y Su-In Lee. "A unified approach to interpreting model predictions". En: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 4 de dic. de 2017, págs. 4768-4777. ISBN: 978-1-5108-6096-4. (Visitado 11-06-2023).

[24] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". En: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2 (20 de oct. de 1975), págs. 442-451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9. URL: https://www.sciencedirect.com/science/article/pii/0005279575901099 (visitado 15-11-2022).

[25] Siddharth Misra, Hao Li y Jiabo He. *Machine Learning for Subsurface Characterization*. Google-Books-ID: WdO1DwAAQBAJ. Gulf Professional Publishing, 12 de oct. de 2019. 442 págs. ISBN: 978-0-12-817737-2.

[26] David Powers. "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation". En: *Mach. Learn. Technol.* 2 (1 de ene. de 2008).

[27] Sadasivan Puthusserypady. *Applied Signal Processing*. Accepted: 2021-04-20T08:09:59Z. Now Publishers, 2021. ISBN: 978-1-68083-979-1 978-1-68083-978-4. DOI: 10.1561/9781680839791. URL: https://library.oapen.org/handle/20.500.12657/47870 (visitado 15-11-2022).

[28] Harold G. Reading. *Sedimentary Environments: Processes, Facies and Stratigraphy*. Google-Books-ID: CBdgt_qxnrYC. Wiley, 9 de dic. de 1996. 2 págs. ISBN: 978-0-632-03627-1.

[29] M. H. Rider. *The Geological Interpretation of Well Logs*. Google-Books-ID: mt2UAAAACAAJ. Whittles Publishing, 1996. 280 págs. ISBN: 978-1-870325-36-3.

[30] Daniel Theisges dos Santos, Mauro Roisenberg y Marivaldo dos Santos Nascimento. "Deep Recurrent Neural Networks Approach to Sedimentary Facies Classification Using Well Logs". En: *IEEE Geoscience and Remote Sensing Letters* 19 (2022). Conference Name: IEEE Geoscience and Remote Sensing Letters, págs. 1-5. ISSN: 1558-0571. DOI: 10.1109/LGRS.2021.3053383.

[31] Oberto Serra y Lorenzo Serra. *Well Logging: Data Acquisition and Applications*. Google-Books-ID: iAacQAAACAAJ. Serralog, 2004. 674 págs. ISBN: 978-2-9515612-5-0.

[32] D.E. Shier. "Well log normalization: Methods and guidelines". En: *Petrophysics* 45 (1 de mayo de 2004), págs. 268-280.

[33] Ashish Vaswani et al. *Attention Is All You Need*. 5 de dic. de 2017. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762[cs]. URL: http://arxiv.org/abs/1706.03762 (visitado 08-06-2023).

[34] Liping Zhu et al. "Intelligent Logging Lithological Interpretation With Convolution Neural Networks". En: *Petrophysics* 59 (1 de dic. de 2018), págs. 799-810. DOI: 10.30632/PJV59N6-2018a5.

10. APÉNDICE

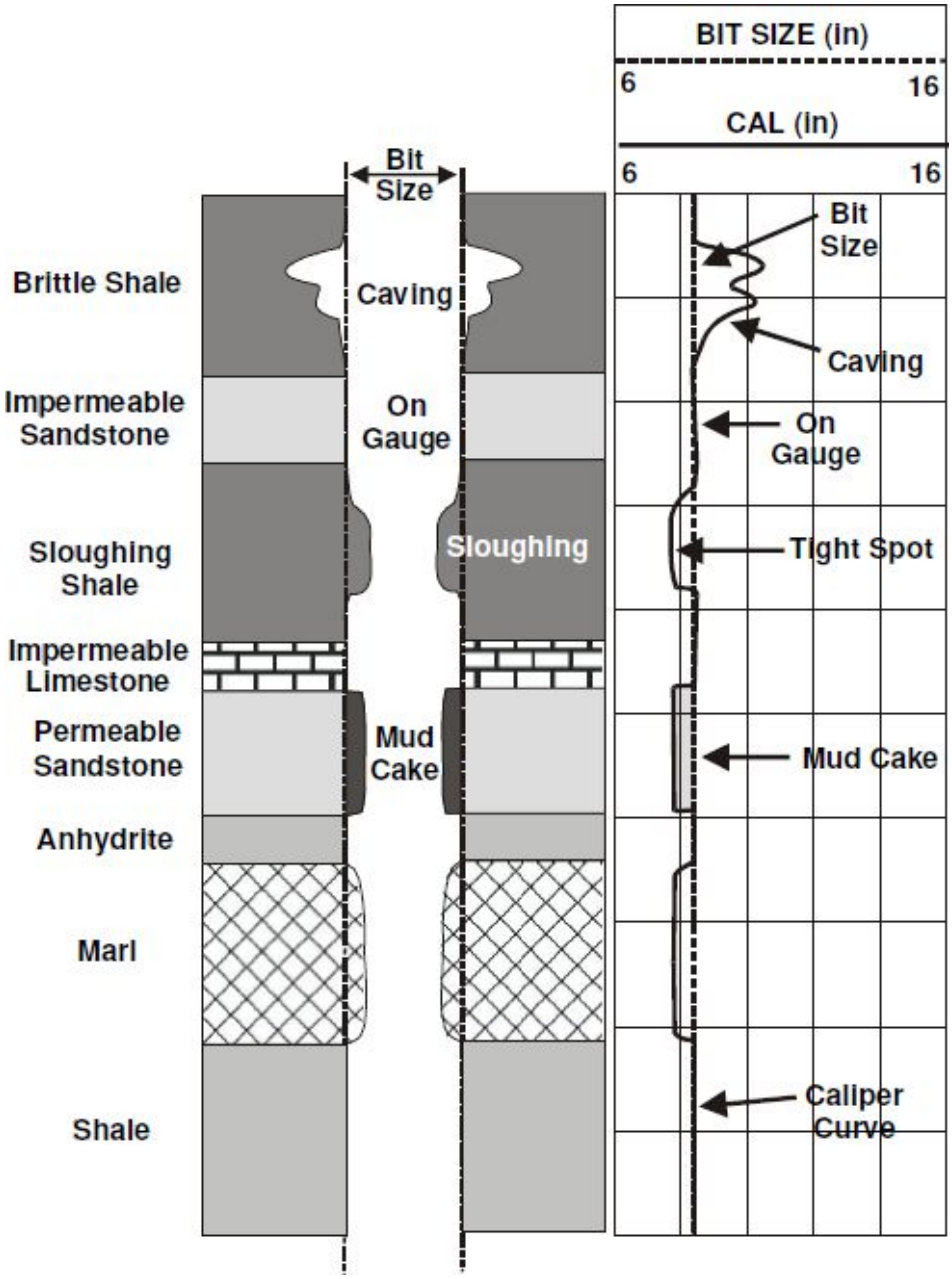

Figura 10: Respuestas típicas del registro CALI a diferentes litologías, tomado de [16].

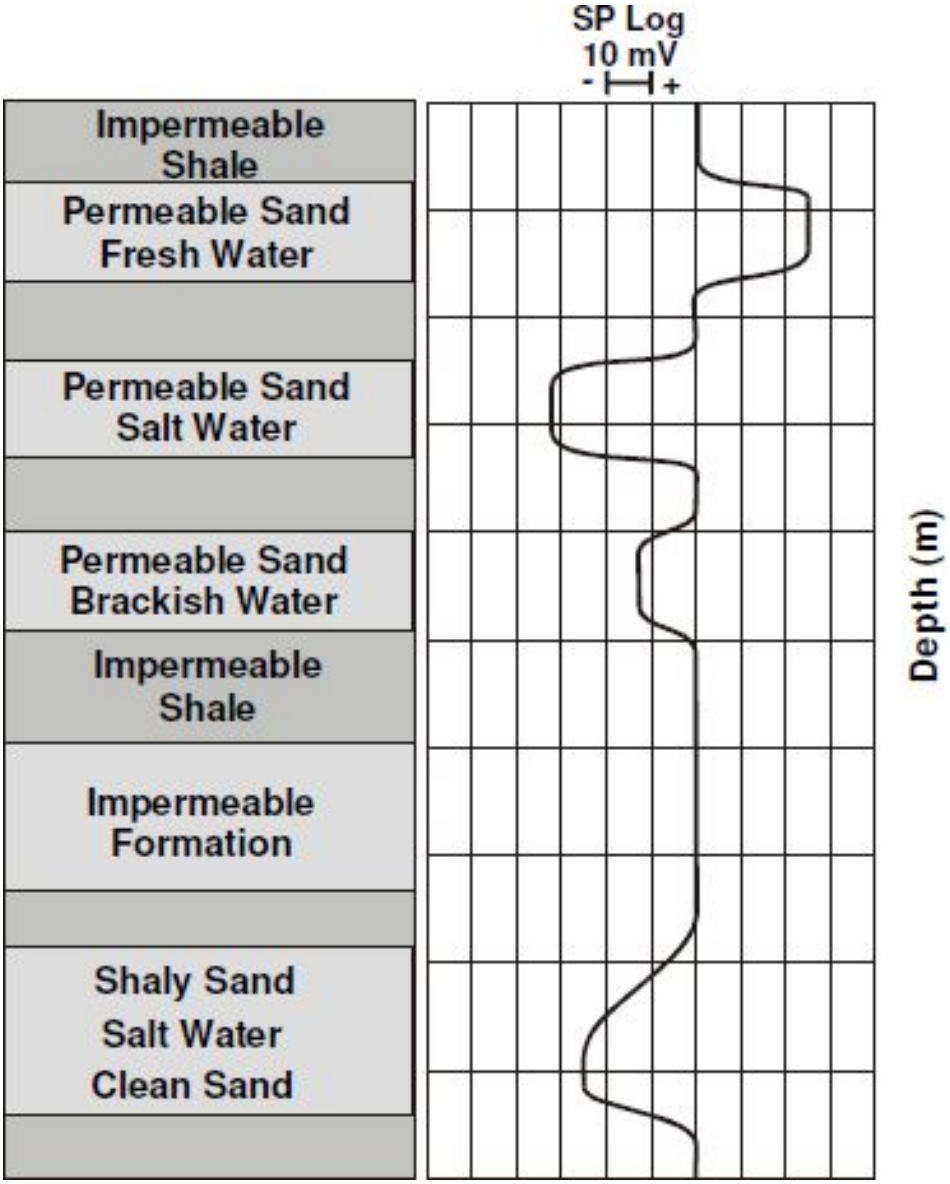

Figura 11: Respuestas típicas del registro SP, tomado de [16].

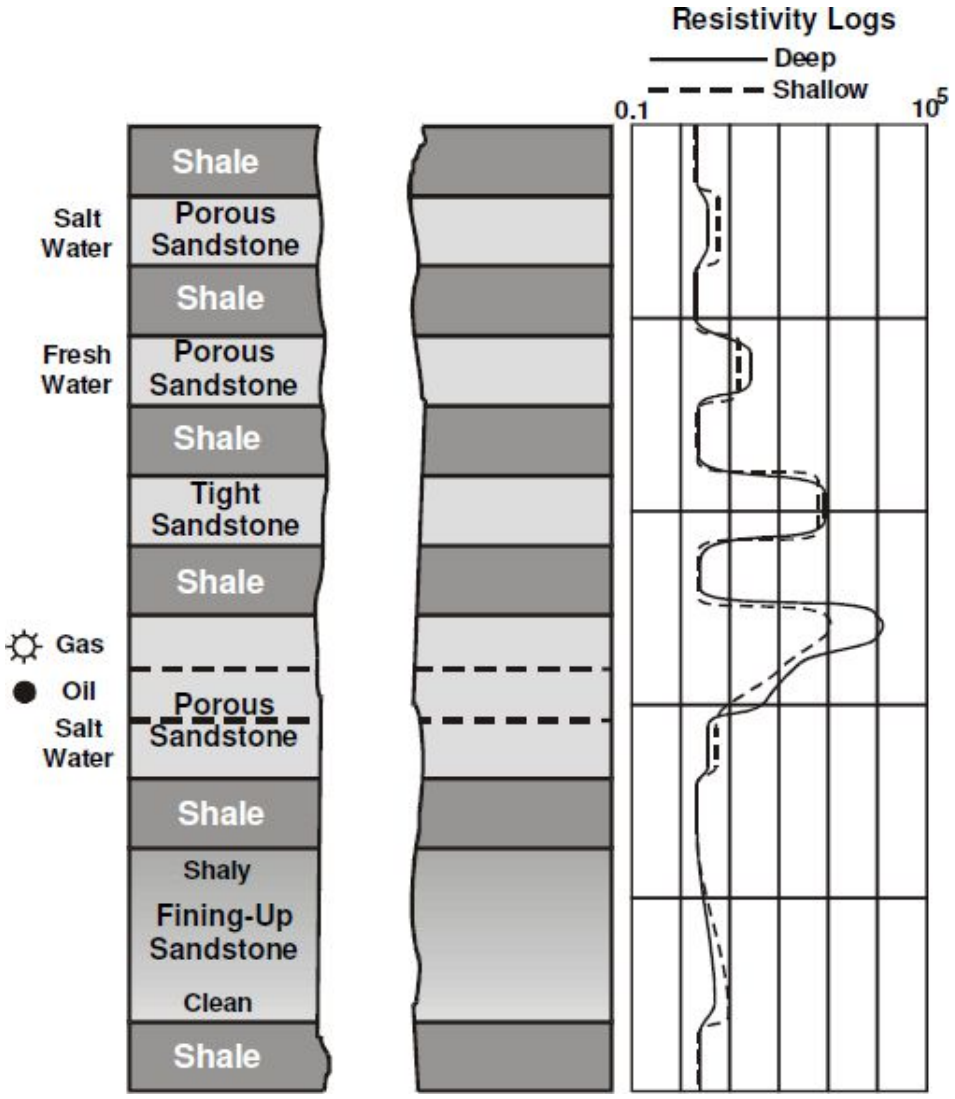

Figura 12: Respuestas típicas de los registros resistivos, tomado de [16].

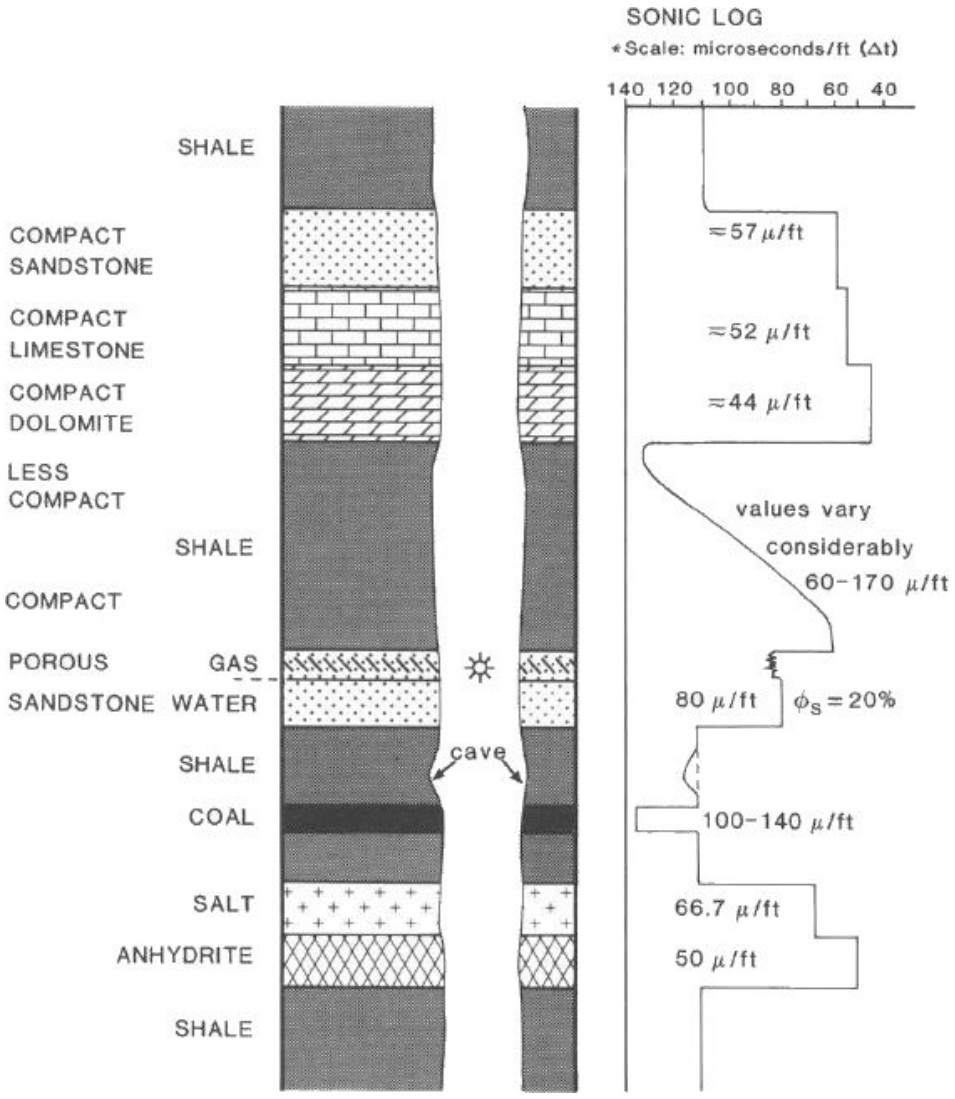

Figura 13: Respuestas típicas de los registros acústicos, tomado de [2].

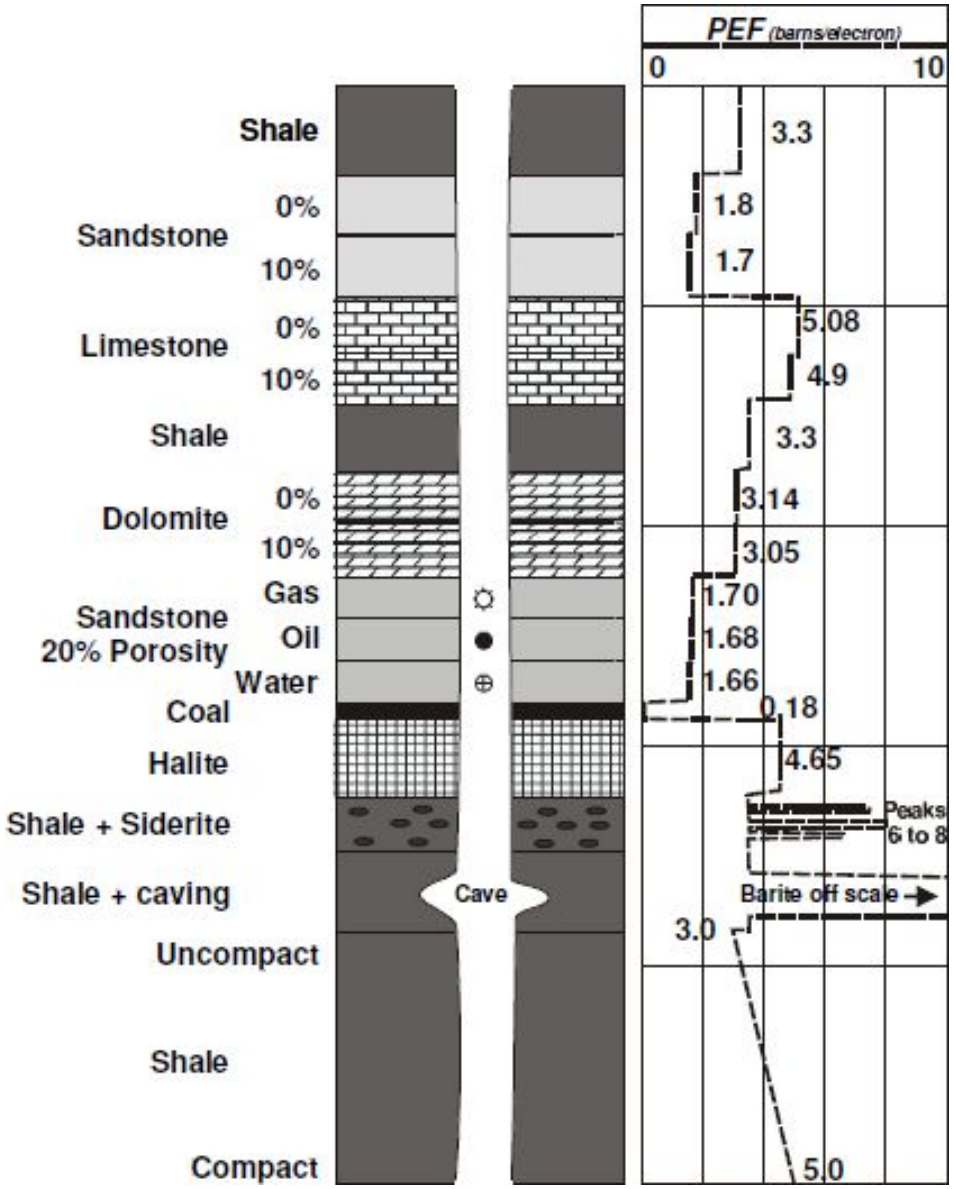

Figura 14: Mediciones del registro de PEF para litologías comunes, tomado de [16]

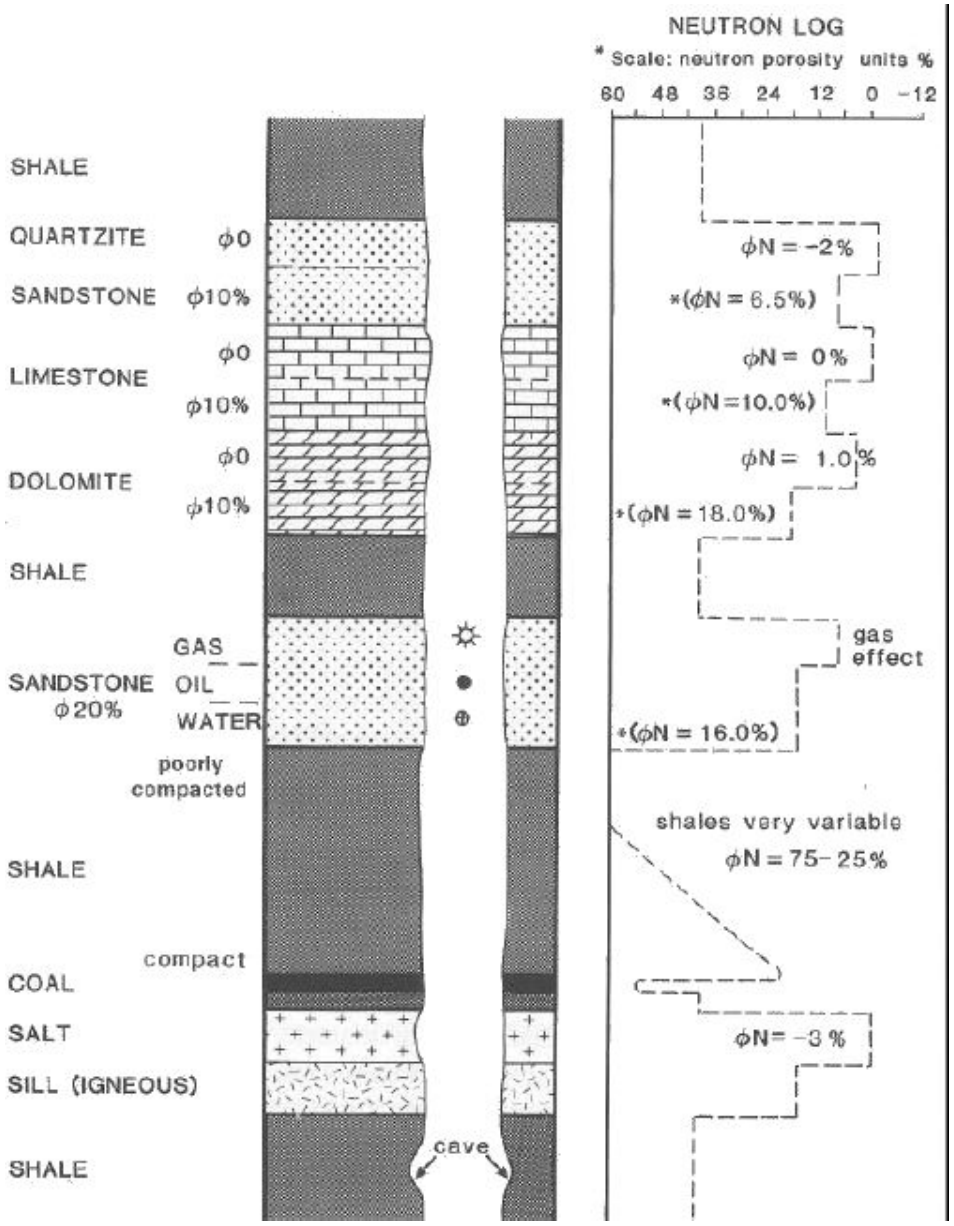

Figura 15: Respuestas típicas del registro NPHI en litologías comunes, tomado de [16]

	FC		CNN1D		LSTM		CNN1D LSTM		CNN1D LSTM AT	
Hiperparametro	min	max	min	max	min	max	min	max	min	max
seq len			8	64	8	64	8	64	8	64
batch size	8	128	8	128	8	128	8	128	8	128
learning rate	0.00001	0.001	0.00001	0.001	0.00001	0.001	0.00001	0.001	0.00001	0.001
cnn channel size			8	128			8	128	8	128
cnn kernel size			3	5			3	5	3	5
cnn padding size			1	1			1	1	1	1
cnn stride size			1	2			1	2	1	2
cnn dropout			0.0	0.5			0.0	0.5	0.0	0.5
lstm hidden size					8	128	8	128	8	128
lstm layer size					1	1	1	1	1	1
lstm bidirectional					0	1	0	1	0	1
lstm dropout					0.0	0.5	0.0	0.5	0.0	0.5
lstm attention heads									2	4
fc hidden size	8	128					8	128	8	128
fc dropout	0.0	0.5					0.0	0.5	0.0	0.5

Cuadro 6: Tabla con distribuciones usadas en la exploración para cada uno de los parámetros en la optimización de los modelos

Hiperparametro	FC	CNN1D	LSTM	CNN1D LSTM	CNN1D LSTM AT
seq len		21	21	21	21
batch size	128	128	128	128	128
learning rate	0.00075	0.00047	0.00079	0.00006	0.00014
cnn channel size		110		47	101
cnn kernel size		3		5	5
cnn padding size		1		1	1
cnn stride size		1		1	1
cnn dropout		0.19		0.48	0.41
lstm hidden size			33	10	28
lstm layer size			1	1	1
lstm bidirectional			1	1	1
lstm dropout			0.43	0.34	0.46
lstm attention heads					4
fc hidden size	98				42
fc dropout	0.32				0.36

Cuadro 7: Tabla con distribuciones usadas en la exploración para cada uno de los parámetros en la optimización de los modelos

	Accuracy												
Fold	1	2	3	4	5	6	7	8	9	10	11	12	0
0	0.7209	0.7294	0.7573	0.7517	0.7482	0.7505	0.7682	0.7815	0.7289	0.7895	0.7511	0.7638	0.7560
1	0.6769	0.7172	0.7239	0.7328	0.7207	0.7474	0.7498	0.7521	0.7434	0.7424	0.7287	0.7453	0.7480
2	0.5573	0.6057	0.6491	0.6357	0.6476	0.6792	0.6819	0.6615	0.6712	0.6655	0.6740	0.6901	0.6840
3	0.6321	0.6587	0.6613	0.6778	0.7138	0.7237	0.7310	0.7262	0.7286	0.7385	0.7177	0.7123	0.7170
4	0.6517	0.6447	0.6639	0.6804	0.6750	0.6489	0.6617	0.6703	0.6691	0.6934	0.6980	0.6691	0.6500
Promedio	0.6478	0.6711	0.6911	0.6957	0.7011	0.7099	0.7185	0.7183	0.7082	0.7259	0.7139	0.7161	0.7110
	F1 Score Macro												
Fold	1	2	3	4	5	6	7	8	9	10	11	12	0
0	0.5469	0.5941	0.5887	0.6069	0.6169	0.6286	0.6564	0.6625	0.6419	0.6527	0.6361	0.6476	0.6505
1	0.5501	0.5843	0.6051	0.6216	0.6280	0.6637	0.6492	0.6553	0.6500	0.6530	0.6385	0.6619	0.6541
2	0.4465	0.4682	0.5122	0.5253	0.5360	0.5408	0.5488	0.5278	0.5481	0.5368	0.5575	0.5668	0.5653
3	0.4954	0.5130	0.5360	0.5466	0.5820	0.5841	0.6017	0.6025	0.6047	0.5915	0.5883	0.5904	0.5990
4	0.4993	0.4934	0.5192	0.5369	0.5395	0.5292	0.5290	0.5343	0.5254	0.5612	0.5455	0.5503	0.5221
Promedio	0.5077	0.5306	0.5523	0.5675	0.5805	0.5893	0.5970	0.5965	0.5940	0.5990	0.5932	0.6034	0.5982
	MCC promedio												
Fold	1	2	3	4	5	6	7	8	9	10	11	12	0
0	0.4601	0.5074	0.5062	0.5233	0.5342	0.5472	0.5750	0.5889	0.5570	0.5822	0.5526	0.5717	0.5734
1	0.4652	0.5111	0.5324	0.5483	0.5589	0.5977	0.5888	0.5894	0.5846	0.5891	0.5734	0.6014	0.5902
2	0.3638	0.3892	0.4380	0.4516	0.4634	0.4725	0.4808	0.4587	0.4746	0.4583	0.4846	0.5002	0.4945
3	0.4245	0.4388	0.4714	0.4809	0.5210	0.5150	0.5387	0.5381	0.5436	0.5250	0.5207	0.5303	0.5364
4	0.4393	0.4327	0.4674	0.4849	0.4877	0.4854	0.4818	0.4788	0.4779	0.5136	0.4902	0.4969	0.4812
Promedio	0.4306	0.4558	0.4831	0.4978	0.5131	0.5236	0.5330	0.5308	0.5275	0.5336	0.5243	0.5401	0.5351

Cuadro 8: Resultados de las métricas acurracy, f1-score macro y mcc para los experimentos de suavizamiento