vol. 43, n.° 1, enero-abril de 2021
ISSN impreso: 0120-0283 • ISSN en línea: 2145-8553

Inferencias del factor fotoeléctrico (PEF) en registros de pozo con *machine learning*

Mauro Felipe Pardo-Díaz[1,2*]; Carlos Alberto Vargas-Jimenez[1]

Forma de citar: Pardo-Díaz, M.F.; Vargas-Jimenez, C.A. (2021). Inferencias del factor fotoeléctrico (PEF) en registros de pozo con *machine learning. Boletín de Geología, 43*(1), 193-210. https://doi.org/10.18273/revbol.v43n1-2021010

Resumen

Los registros de pozo convencionales son importantes para la realización de análisis petrofísicos, amarres sísmicos y correlación estratigráfica. El presente estudio propone una metodología para realizar predicciones en estos registros haciendo uso de *machine learning* (ML), una herramienta altamente aplicada en múltiples disciplinas. El software de entrenamiento utilizado fue WEKA (*Waikato Environment for Knowledge Analysis*), en el que se generó un modelo para la predicción del registro de Absorción Fotoeléctrica (PDPE o PEF), a partir de tres atributos, los registros de Rayos Gamma (GRGC), Densidad (DEN) y Corrección de Densidad (DCOR). Esta metodología fue aplicada a registros de pozo de la Formación San Fernando, cuya unidad equivalente sería la Formación Mirador, en el sector sur-occidental de los Llanos Orientales de Colombia. Fueron usados los registros de trece pozos para hacer el entrenamiento del modelo y otros seis pozos fueron usados para evaluar el desempeño de este. Los resultados confirman la posibilidad de correlacionar registros que miden características diferentes en las rocas y evidencian que las inferencias en registros de pozo con ML requieren un filtrado minucioso para tomar la tendencia de los datos, y una optimización clara para prevenir el sobreentrenamiento en el modelo.

Palabras clave: Predicción de gráficas; Selección de atributos; Grilla de búsqueda; WEKA; Sobreentrenamiento.

Photoelectric factor (PEF) inferences in well logs with machine learning

Abstract

Conventional well logs are important for performing petrophysical analysis, seismic well ties and stratigraphic correlation. This study proposes a methodology to predict these types of logs using machine learning (ML), a tool highly applied in multiple disciplines. The training software used was WEKA (Waikato Environment for Knowledge Analysis), in which a model for the prediction of the Photoelectric Absorption log (PDPE) was generated, based on three attributes, the Gamma Ray log (GRGC), Density log (DEN) and Density Correction log (DCOR). This methodology was applied to well logs of San Fernando Formation, whose equivalent unit would be Mirador Formation, in the southwestern sector of the Llanos Basin, Colombia. Thirteen wells were used to train the model and six other wells were used to evaluate its performance. The results confirm the possibility of correlating logs that measure different characteristics in the rocks and show that inferences in well logs with ML require a detailed filtering to take the trend of the data, and a clear optimization to prevent overfitting in the model.

Keywords: Graph prediction; Attribute selection; Grid search; WEKA; Overfitting.

[1]Departamento de Geociencias, Universidad Nacional de Colombia, Bogotá, Colombia. (*) mfpardod@unal.edu.co; cavargasj@unal.edu.co
[2] Ecopetrol S.A., Bogotá, Colombia.

Introducción

Un problema presente al trabajar con registros de pozo es que algunos de ellos no están completos, ya sea porque no abarcan toda la profundidad requerida o porque carecen de registros eléctricos o "logs". Completar estos registros es importante para la realización de análisis petrofísicos, amarres sísmicos y correlación estratigráfica. Estudios anteriores, como el de Labani *et al*. (2010), han mostrado la posibilidad de estimar los parámetros de un registro de interés a partir de los registros convencionales de un pozo usando *machine learning* (ML). Los sistemas de ML permiten entrenar computadoras para realizar tareas de manera inteligente mediante el aprendizaje del entorno a través de repetidos ejemplos (El Naqa y Murphy, 2015).

Este estudio se realizó con registros de pozo correspondientes a la Formación San Fernando, en el sector sur-occidental de los Llanos Orientales. La unidad equivalente sería la Formación Mirador, conocida cómo Formación San Fernando según la nomenclatura tomada por Ecopetrol de edad Eoceno Tardío (Ballesteros y Torres, 2017). La ubicación del área de estudio no se especifica por restricción de uso de datos. Sin embargo, en la Figura 1 se puede observar la zona de estudio, con la distribución espacial de los pozos, horizontes y fallas asociadas a la unidad. Según Rojas *et al.* (2004), la Formación San Fernando corresponde a una secuencia masiva de arenitas cuarzosas de grano fino a grueso, moderadamente seleccionadas, localmente conglomerática, con secuencias grano decrecientes a la base e intercaladas con delgadas capas de arcilla, depositadas en un ambiente fluvial. Su espesor total varía entre 295 y 411 ft con un promedio de 340 ft de acuerdo a lo calculado gráficamente con ciertos pozos del estudio. Esta Formación está en contacto erosivo con la unidad infrayacente, la Formación Gacheta, que es una secuencia de edad Cenomaniano - Turoniano, litológicamente compuesta de arcillolitas negras, con alto contenido de materia orgánica, intercaladas con capas de arenitas y calizas, las cuales se depositaron en ambientes marinos de plataforma externa (Piedrahita, 2016). Como parte de una iniciativa académica, se pretende usar ML para inferir una curva faltante o de interés en el registro, a partir de otras curvas convencionales. Esto genera una nueva forma de abordar el estudio de registros de pozo, con implicaciones importantes, como cambiar la forma en que se obtienen ciertas curvas del registro, pues no se invertiría en obtener paquetes de registros con todas las curvas, sino sólo en aquellos que contengan las curvas necesarias para inferir el resto de curvas de un paquete estándar de registros. El objetivo de este estudio es inferir el registro de Absorción Fotoeléctrica (PDPE o PEF) a partir de las otras curvas básicas. Por lo tanto, para rocas reservorio, la metodología propuesta permite una buena estimación de litología basada en un registro PEF inferido, la cual se puede replicar en otras zonas de la cuenca de los Llanos.

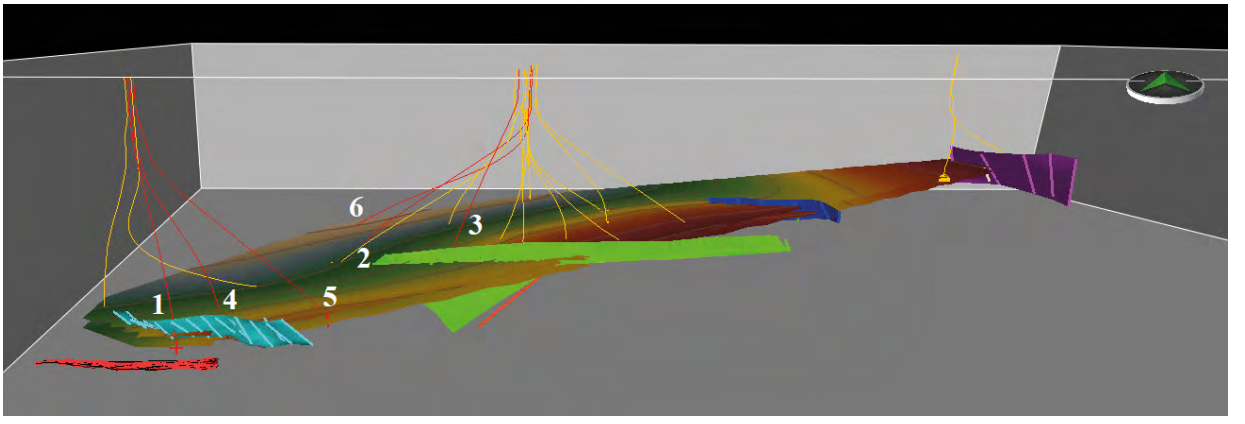

Figura 1. Bloque diagrama con el volumen de estudio y la distribución espacial de los pozos. Allí se muestran los pozos de entrenamiento en color amarillo y los pozos de prueba en rojo, los cuales están señalados del 1 al 6. Los horizontes representan los límites de la Formación San Fernando, cortados por las fallas que se muestran en diferentes tonos.

Metodología

Datos

A partir de veinticinco pozos disponibles, se seleccionó una muestra de veinte pozos para el estudio. Cada pozo tiene una cantidad diferente de registros (Tabla 1), no obstante, se utilizaron los pozos que contaban con seis registros básicos para procesar, en la Figura 2 es posible apreciar la tendencia típica de las curvas en la Formación San Fernando. Esta muestra se distribuyó en dos partes, catorce pozos para entrenamiento y seis pozos para evaluación (Tabla 2).

La profundidad que abarca cada registro es de alrededor de 500 ft para la mayoría de los pozos. Sin embargo, la Profundidad Vertical Verdadera (TVD, por sus siglas en inglés) del intervalo de interés va de los 8660 a los 8932 ft aproximadamente, abarcando un total de 272 ft.

Tabla 1. Registros disponibles para el estudio y la cantidad de pozos que los contienen.

	Registro de Pozo	Siglas en inglés	Cantidad de pozos con el registro
1	Rayos gamma	GR	20
2	Densidad	DEN	20
3	PEF	PEF	20
4	Corrección de densidad	DCOR	20
5	Porosidad neutrón	NPOR	20
6	Potencial espontaneo	SP	20
7	Densidad-porosidad base	DPOR	15
8	Densidad-porosidad para arenitas	DPRS	15
9	Porosidad neutrón para arenitas	NPRS	15
10	Porosidad neutrón para dolomías	NPRD	13
11	Porosidad neutrón para calizas	NPRL	13
12	Densidad-porosidad para dolomías	DPRD	10
13	Densidad-porosidad para calizas	DPRL	10

Tabla 2. Resumen de los datos utilizados. Nótese que a partir de los seis registros básicos se definieron cuatro cómo atributos para entrenar el modelo predictivo.

Ítem	Cantidad
Pozos para entrenamiento	14
Pozos del set de entrenamiento (Al aplicar el filtro DCOR)	13
Registros básicos para seleccionar los pozos de la muestra	6
Registros definidos cómo atributos	4
Instancias del set de entrenamiento	6752
Pozos de evaluación	6

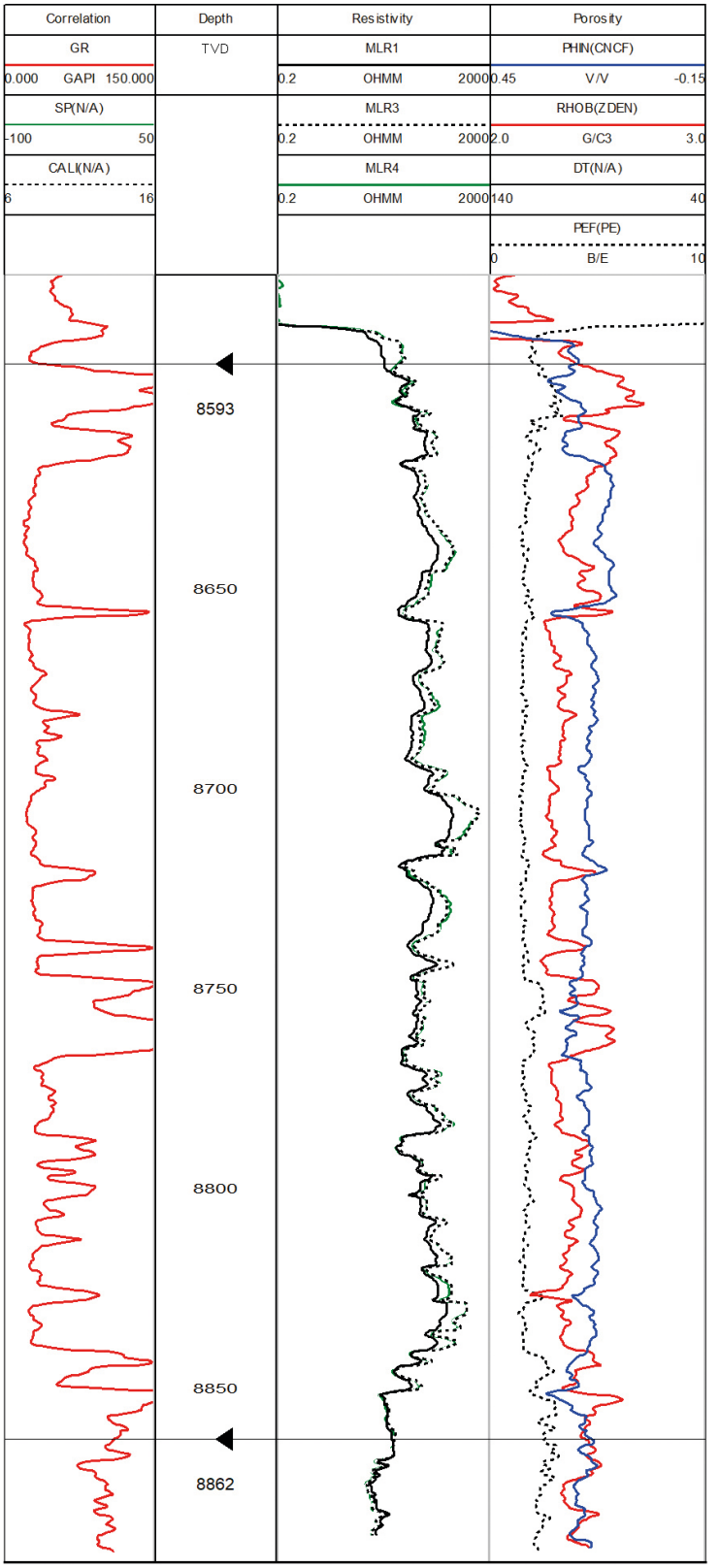

Figura 2. Registros típicos de la unidad San Fernando en el Pozo Y, con su tope y base marcados en la columna de profundidad. Por encima del tope se observa una porción anómala en los registros de porosidad (PHIN), densidad (RHOB) y PEF. Esta porción de los registros se omite más adelante para no ingresar valores inválidos en el set de entrenamiento.

Selección de pozos

La muestra está conformada por pozos que tenían las seis curvas básicas que se especifican a continuación:

Registro de densidad (DEN o RHOB): el registro de densidad mide la densidad del pozo y las rocas penetradas por la broca. La unidad de densidad es gramo por centímetro cúbico (Onajite, 2014). De acuerdo con Asquith y Krygowski (2004), este registro utiliza dos valores de densidad separados: la densidad aparente (ρ_b o RHOB) y la densidad de la matriz (ρ_{ma}). La densidad aparente es la de toda la formación (partes sólidas y fluidas) medida por la herramienta de registro. La densidad de la matriz es la densidad del armazón sólido de la roca.

Registro de corrección de densidad (DCOR o DRHO): esta curva indica la cantidad de corrección que se ha agregado a la curva de densidad durante el procesamiento, debido a los efectos del pozo (principalmente el espesor del lodo) y se usa como un indicador de control de calidad. Siempre que la curva de corrección exceda 0,20 g/cm^3 el valor de la densidad obtenida, debe considerarse sospechoso y posiblemente inválido (Asquith y Krygowski, 2004). Este registro se toma para dejar los valores más confiables de densidad.

Registro de las propiedades de absorción fotoeléctrica (PDPE o PEF): de acuerdo con el *Schlumberger Oilfield Glossary*, el registro PEF mide el factor de absorción fotoeléctrica (P_e) y hace parte de la medición de la densidad. P_e se define como $(Z/10)^{3,6}$ donde Z es el número atómico promedio de la formación. Teniendo en cuenta que los fluidos presentan números atómicos muy bajos, su influencia llega a ser despreciable, por lo tanto, el factor P_e es una medida de las propiedades de la matriz de la roca. La única excepción, de acuerdo con Glover (2000), es la salmuera altamente saturada que puede tener un valor significativo de P_e. Por otro lado, cuando se presentan picos aislados en el registro, estos pueden indicar la presencia de depósitos locales de minerales pesados.

El registro PEF es sensible a las diferencias en el número atómico promedio de una formación, pero no es sensible a los cambios en la porosidad y la saturación de fluidos en una litología. Esta combinación hace que este registro sea un muy buen indicador de litología.

Cómo se observa en la Figura 3, los valores de P_e para las litologías más comunes son menores a 6 barns/electrón. Para una arenita es de 1,8 barns/electrón, para una arcillolita es de 3,3 barns/electrón y para una roca caliza es de 5,08 barns/electrón. Sin embargo, la desventaja de este registro es que la herramienta de lito-densidad no puede usarse con lodo que contenga barita, ya que este mineral presenta un P_e de 267 barns/electrón (Glover, 2000).

Registro de rayos gamma (GRGC o GR): este registro mide la radiactividad natural en formaciones y puede usarse para identificar litologías y hacer correlaciones. También proporciona información para calcular el volumen de arcilla en una arenita o caliza (Asquith y Gibson, 1982). Las arenitas y calizas libres de arcilla tienen bajas concentraciones de material radiactivo y dan lecturas bajas de rayos gamma. A medida que aumenta el contenido de arcilla, la respuesta del registro aumenta debido a la concentración del material radiactivo. Sin embargo, una arenita limpia (es decir, con bajo contenido de minerales de arcilla) también podría producir una respuesta alta de rayos gamma si la roca contiene feldespatos de potasio, micas, glauconita o aguas ricas en uranio (Asquith y Krygowski, 2004). Las unidades de medida usadas pertenecen a la escala del Instituto Americano del Petróleo (API).

Porosidad neutrón (NPOR): los registros de porosidad neutrón miden la concentración de hidrógeno en una formación. En formaciones libres de arcilla donde la porosidad se llena con agua o aceite, el registro mide la porosidad saturada por el líquido (Asquith y Krygowski, 2004).

Potencial Espontáneo (SPDL): el registro SP es un registro del voltaje (o potencial) de corriente continua, que se desarrolla de forma natural (o espontánea) entre un electrodo móvil en el pozo y un electrodo fijo ubicado en la superficie (Doll, 1948 en Asquith y Krygowski, 2004).

Como menciona Glover (2000), el registro SP tiene cuatro usos principales:

- La detección de capas permeables.
- La determinación de la resistividad del agua de formación (Rw).
- La determinación del volumen de arcilla en las formaciones permeables.
- Correlación.

En la Figura 4 se muestra el desarrollo de la metodología, exponiendo cómo se escogieron los datos y cómo se trataron para el entrenamiento y las predicciones.

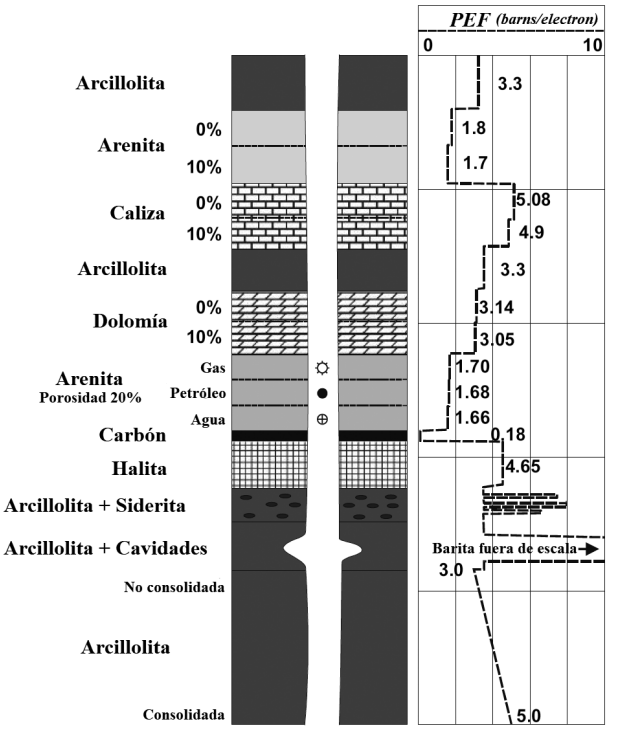

Figura 3. Respuestas del registro PEF para diferentes litologías (modificado de Glover, 2000). Se evidencia la diferencia entre los valores obtenidos para una arenita y una roca caliza. Los porcentajes se refieren a condiciones de porosidad.

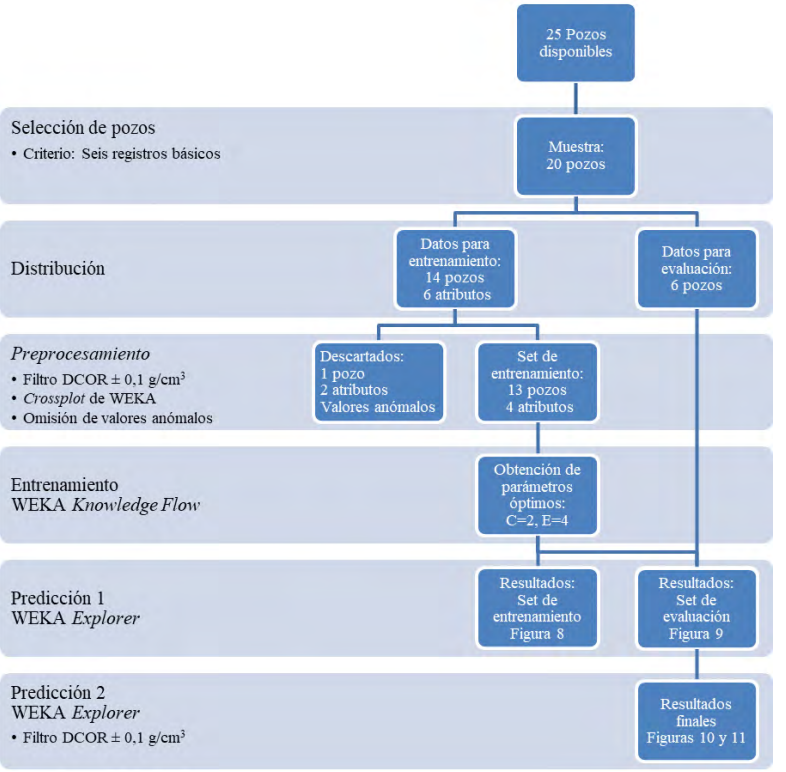

Figura 4. Esquema de la metodología desarrollada. Cada etiqueta menciona brevemente el proceso que dio paso a la jerarquía que contiene. Por ejemplo, el entrenamiento con WEKA *Knowledge Flow* dio paso a la obtención de los parámetros óptimos.

Preprocesamiento y selección de atributos

El conjunto de pozos para entrenamiento tuvo un procesamiento previo a la implementación de ML. En primera instancia, se realizó un filtrado al registro DCOR para descartar los valores errados de densidad, porque si la curva de corrección excede los 0,20 g/cm^3, la densidad en ese intervalo es considerada invalida según Asquith (1982) en Flores (2014). Entonces, para tener una base de datos de entrenamiento más certera, el filtro realizado consistió en reducir el rango de la curva al 20%, es decir, su rango normal es de -0,25 a 0,25 g/cm^3 y se reduce al de -0,05 a 0,05 g/cm^3 y así se descartan los valores de densidad anómalos o inválidos. Al hacer este filtro, el número de pozos para entrenamiento se redujo a trece. Luego, a partir de este conjunto de datos se eligieron los atributos válidos para entrenar el modelo.

El diagrama de *crossplot* presentado en la Figura 5 es una herramienta de visualización que ofrece el interfaz de WEKA *Explorer*, el cual permite revisar la tendencia de correlación entre los atributos del conjunto de datos. Verificando el *crossplot*, se definieron los atributos PDPE (PEF), DCOR, DEN y GRGC para entrenar el modelo. El registro de pozo con estos atributos se muestra en la Figura 6. Los registros NPOR y SP se omitieron debido a la baja correlación que presentaron con los otros atributos (Figura 5), y adicionalmente, en el caso del SP, se omitió por ser una variable dependiente de los fluidos.

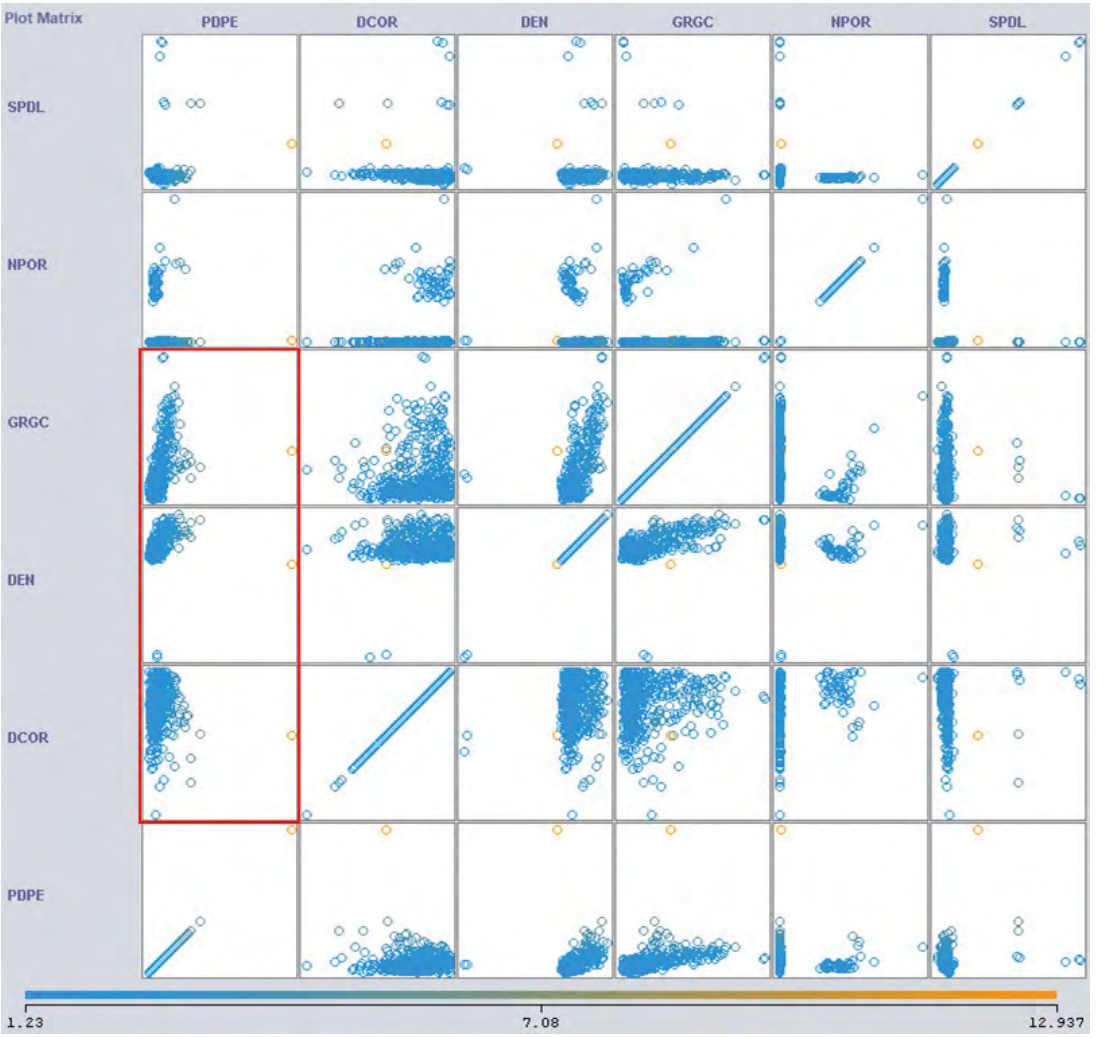

Figura 5. *Crossplot* para la visualización de la relación entre atributos en WEKA. El color de cada punto va de acuerdo al valor del registro PDPE (PEF), cuya escala numérica se muestra en la base de la figura. Los registros señalados en el recuadro rojo (GRGC, DEN y DCOR) fueron seleccionados como atributos por su tendencia en la correlación con el registro PEF. Mientras que los registros NPOR y SP fueron descartados, ya que no presentan una tendencia útil para ser incluidos en el entrenamiento.

En cada registro de los pozos de la muestra se eliminaron los extremos superiores e inferiores, esto con el fin de remover los valores anómalos o inválidos. Estos valores se pueden observar en la Figura 2, en este ejemplo, los registros de porosidad, densidad y PEF son anómalos en el extremo superior del pozo, por encima del intervalo de interés. De esta forma, se obtuvo el set de datos de entrenamiento, el cual está conformado por trece pozos con cuatro atributos (registros seleccionados). Este set cuenta con 6752 instancias, cada una consta de un dato de GR, DCOR, DEN y PEF, obtenidos a una profundidad específica.

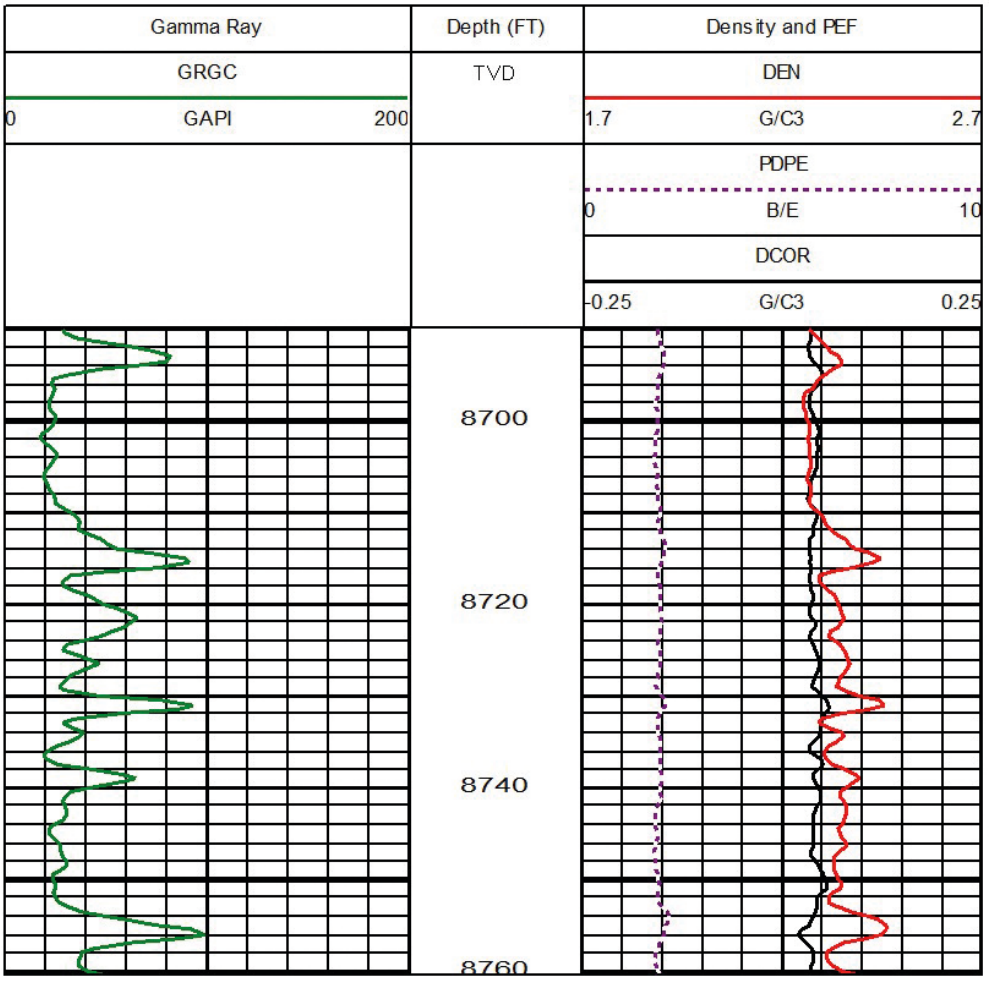

Figura 6. Registros seleccionados para el entrenamiento, los cuales hacen parte del Pozo X del set de entrenamiento.

Machine learning con WEKA

El software de entrenamiento utilizado fue WEKA (*Waikato Environment for Knowledge Analysis*). Es un software libre desarrollado por la Universidad de Waikato de Nueva Zelanda e implementado en Java. El proyecto WEKA tiene como objetivo proporcionar una colección completa de algoritmos de aprendizaje automático y preprocesamiento de datos. Permite probar y comparar rápidamente diferentes tipos de aprendizaje automático en nuevos conjuntos de datos (Hall *et al.,* 2009).

Se utilizó el interfaz de *Knowledge Flow* de WEKA (Figura 7) para realizar el entrenamiento mediante validaciones automatizadas. De este proceso, se obtuvieron el coeficiente de correlación, el error medio absoluto y el error medio cuadrático. Cada uno de estos valores se ubicó en una grilla de búsqueda (Tabla 3, 4 y 5), que resume los resultados de múltiples validaciones.

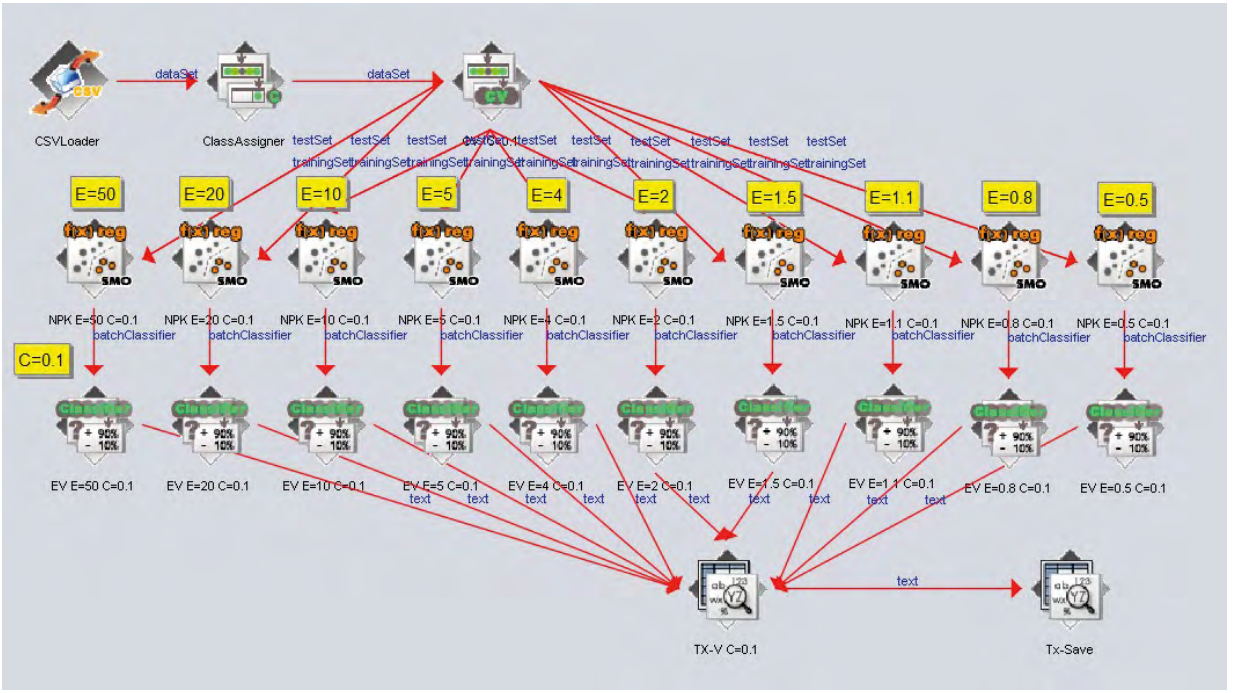

Figura 7. Interfaz de WEKA *Knowledge Flow*. Este es el flujo de trabajo del software, donde se muestra el procedimiento interno de cada validación. Esto lo realiza el programa por cada valor de complejidad (C), en esta figura se muestran las validaciones de cada exponente con C = 0,1. De aquí se obtienen los resultados consignados en las Tablas 3, 4 y 5.

Tabla 3. Grilla de búsqueda del coeficiente de correlación más indicado. Sirve para hallar los parámetros óptimos necesarios para realizar el entrenamiento, los cuales son Complejidad (C) igual a 2 y Exponente (E) igual a 4. Los valores más altos de correlación que están en el extremo inferior derecho no se tienen en cuenta, ya que el hecho de estar en las posiciones de mayor exponente y complejidad significa sobreentrenamiento.

Coeficiente de correlación		Exponente									
		0,5	0,8	1,1	1,5	2	4	5	10	20	50
Factor de complejidad	0,1	0,551	0,5621	0,5799	0,6084	0,6336	0,6645	0,6674	0,6677	0,666	0,6683
	0,5	0,4421	0,4901	0,5884	0,6428	0,6587	0,6695	0,6695	0,668	0,6673	0,6719
	0,8	0,3088	0,4251	0,596	0,6501	0,6605	0,6698	0,6697	0,6669	0,6676	0,673
	1	0,2218	0,3847	0,6	0,6527	0,6606	0,6698	0,6699	0,6666	0,6677	0,6739
	2	-0,026	0,2058	0,6196	0,6524	0,6615	0,6702	0,6696	0,666	0,6675	0,6763
	4	-0,137	0,0495	0,6373	0,6485	0,6618	0,67	0,6681	0,6671	0,6672	0,6768
	5	-0,151	0,0198	0,6414	0,6424	0,662	0,6694	0,6678	0,6675	0,6672	0,6766
	10	-0,189	-0,027	0,6441	0,5913	0,6619	0,6684	0,6661	0,6678	0,6674	0,6776
	20	-0,204	-0,044	0,6293	0,4335	0,662	0,6666	0,6655	0,6677	0,6679	0,679
	50	-0,214	-0,06	0,4975	0,1707	0,6619	0,6657	0,6657	0,6671	0,6681	0,6769

Tabla 4. Grilla de búsqueda del error medio absoluto, con el cual, es posible medir directamente cuanto se desvía la predicción. El error encontrado con los parámetros seleccionados, quiere decir que los valores predichos se estimaron con 0,22 barns/electrón de error absoluto en el set de entrenamiento (Ochoa *et al.,* 2018). Lo cual puede ser una aproximación, ya que esta tabla no contiene una tendencia central, a diferencia de las otras dos grillas que si la presentan.

Error medio absoluto		Exponente									
		0,5	0,8	1,1	1,5	2	4	5	10	20	50
Factor de complejidad	0,1	0,2456	0,2432	0,2407	0,2374	0,2341	0,228	0,2263	0,2236	0,2221	0,2197
	0,5	0,2528	0,245	0,2372	0,232	0,2298	0,2251	0,2244	0,2225	0,2215	0,2185
	0,8	0,2625	0,2482	0,2359	0,231	0,2295	0,2246	0,2241	0,2225	0,2213	0,218
	1	0,2723	0,2505	0,2354	0,2306	0,2295	0,2245	0,2239	0,2225	0,2212	0,2177
	2	0,355	0,2692	0,2333	0,2308	0,2292	0,2241	0,2235	0,2224	0,2212	0,2171
	4	0,5986	0,3505	0,2315	0,2318	0,2291	0,2237	0,2234	0,2221	0,221	0,2167
	5	0,722	0,4049	0,2311	0,2324	0,229	0,2237	0,2233	0,222	0,2209	0,2166
	10	1,3616	0,7173	0,2312	0,2352	0,229	0,2235	0,2232	0,2217	0,2204	0,2161
	20	2,6617	1,3773	0,2326	0,2471	0,229	0,2235	0,2231	0,2216	0,22	0,2157
	50	6,5835	3,3743	0,2415	0,286	0,229	0,2232	0,2228	0,2217	0,2196	0,2157

Tabla 5. Grilla de búsqueda del error medio cuadrático. Se evidencia la correspondencia con la grilla del coeficiente de correlación (Tabla 3), ya que el valor más aceptable de error se encuentra con los mismos parámetros de exponente y complejidad seleccionados.

Error medio cuadrático		Exponente									
		0,5	0,8	1,1	1,5	2	4	5	10	20	50
Factor de complejidad	0,1	0,3526	0,349	0,3437	0,3343	0,3249	0,3121	0,3105	0,3103	0,3112	0,3099
	0,5	0,3799	0,3677	0,3388	0,3203	0,314	0,3094	0,3095	0,3105	0,3111	0,3084
	0,8	0,4224	0,3896	0,3362	0,3173	0,3131	0,3093	0,3095	0,311	0,311	0,3081
	1	0,4645	0,4066	0,3349	0,3162	0,313	0,3093	0,3094	0,3112	0,3111	0,3079
	2	0,8101	0,5391	0,3282	0,3161	0,3126	0,3092	0,3096	0,3118	0,3111	0,307
	4	1,6507	1,1061	0,3219	0,3177	0,3124	0,3094	0,3104	0,3116	0,3112	0,3069
	5	2,0626	1,4287	0,3203	0,3201	0,3123	0,3097	0,3107	0,3115	0,3111	0,3071
	10	4,1527	2,8898	0,3193	0,3444	0,3123	0,3103	0,3116	0,3113	0,3109	0,3066
	20	8,257	5,8395	0,3256	0,4477	0,3123	0,3114	0,3122	0,3114	0,3107	0,3061
	50	20,5087	14,532	0,3973	1,02	0,3123	0,3121	0,3123	0,3115	0,3106	0,3071

La grilla de búsqueda sirve para hallar el valor óptimo de los dos parámetros con los que se entrena el modelo. El primero es el factor de complejidad, de acuerdo con Brownlee (2016), este parámetro controla la flexibilidad del proceso que traza la línea que se ajusta a los datos, y el segundo es el exponente del Kernel Polinómico Normalizado, el cual define qué tan ondulada es la línea de ajuste, mientras mayor es el exponente, más ondulada es esta línea.

Si la complejidad es mayor a la adecuada, se puede producir sobreentrenamiento (*overfitting*), que es cuando el modelo se ajusta a las particularidades de los datos de entrenamiento en vez de encontrar una tendencia predictiva general (Dietterich, 1995). Los efectos del sobreentrenamiento se evidencian por una disminución en la exactitud de la clasificación (Drazin y Montag, 2012). La delimitación de los datos en cuanto a los rangos y atributos más adecuados, es de gran importancia para evitar el sobreajuste. Esto se evidenció en este estudio ya que las primeras validaciones presentaron sobreentrenamiento. Por esta razón, fue necesario evaluar los diferentes registros, cambiar los filtros DCOR y omitir algunos atributos y

valores anómalos, para obtener finalmente las grillas de búsqueda presentadas (Tabla 3, 4 y 5). Todos estos procedimientos permitieron descubrir, aclarar y definir la metodología para realizar las predicciones (Figura 4).

Cabe mencionar el estudio realizado por Labani *et al.* (2010), aplicado en formaciones que consisten principalmente de caliza, anhidrita y dolomía, en el que los autores integran sistemas inteligentes usando el concepto de *committee machine*, así combinan los resultados de diferentes algoritmos para la estimación de los parámetros de un registro a partir de otros registros convencionales, esta metodología guarda cierta similitud con la propuesta en este artículo. Así pues, se entiende que el procedimiento desarrollado con WEKA se asemeja al del estudio mencionado, el cual tiene un modelo preciso y efectivo.

Ahora bien, se cree que es apropiado usar el método aquí descrito (Figura 4) para una litología siliciclástica, y que funciona en un sistema fluvial, porque el método de *machine learning* permite extraer la tendencia de los datos aceptando su particularidad. Un punto importante dentro de este método, es el preprocesamiento de los datos, mediante el cual se pueden disponer para facilitar el entrenamiento.

Por otra parte, continuando con la metodología, sobre la base de los parámetros más óptimos obtenidos (C = 2 y E = 4) se procede a clasificar instancias utilizando el set de entrenamiento, a través de la herramienta WEKA *Explorer classify*, esto equivale a predecir el registro PEF del mismo set de datos con el que se entrenó el modelo. La utilidad de este procedimiento se aclara en los resultados. En el interfaz de WEKA se selecciona el atributo PDPE (PEF) para clasificar instancias y así generar la predicción. La función clasificadora utilizada es *SMOreg* con un *Normalized Polykernel*. Esta técnica también se utiliza para el resto de las predicciones, y se configuró así teniendo en cuenta la asesoría del autor que realizó la estimación de un epicentro utilizando el mismo programa (Ochoa *et al.,* 2018). Los resultados obtenidos para el set de entrenamiento se muestran en la Figura 8.

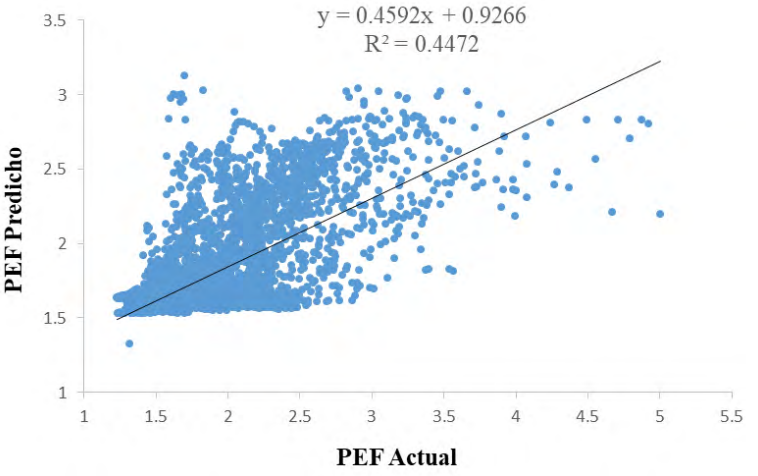

Figura 8. Valores de PEF actuales frente a los predichos del set de entrenamiento.

Usando los mismos parámetros (C = 2 y E = 4), se llevó a cabo la predicción con cada uno de los seis pozos del set de evaluación. La Figura 9 muestra los valores de PEF actuales frente a los predichos para cada pozo de prueba. Sin embargo, fue necesario realizar una segunda predicción para eliminar valores atípicos extremos presentes en algunas de las nubes de puntos (Figura 9A, 9B, 9D). Para la segunda predicción se aplicó un filtro previo para cada pozo de prueba, este consistió en definir la corrección de densidad (DCOR) en el rango de -0,1 a 0,1 g/cm^3. Esto con el fin de omitir valores de densidad inválidos que podrían haber provocado los valores atípicos. Los resultados de este segundo proceso de predicción se muestran en la Figura 10.

De esta forma, fue desarrollada la metodología plasmada en la Figura 4, llevando a cabo cada procedimiento y buscando las mejoras en los resultados.

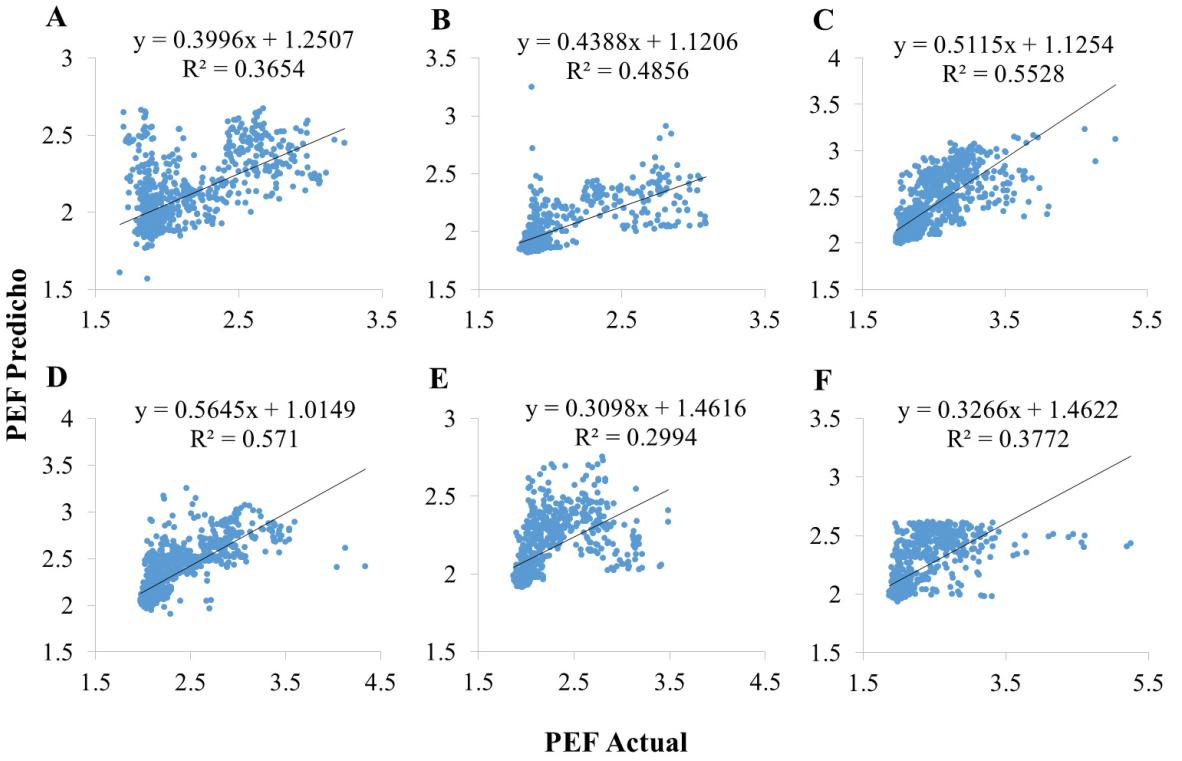

Figura 9. Valores de PEF actuales contra los predichos para los pozos de evaluación. **A.** Pozo 1. **B.** Pozo 2. **C.** Pozo 3. **D.** Pozo 4. **E.** Pozo 5. **F.** Pozo 6.

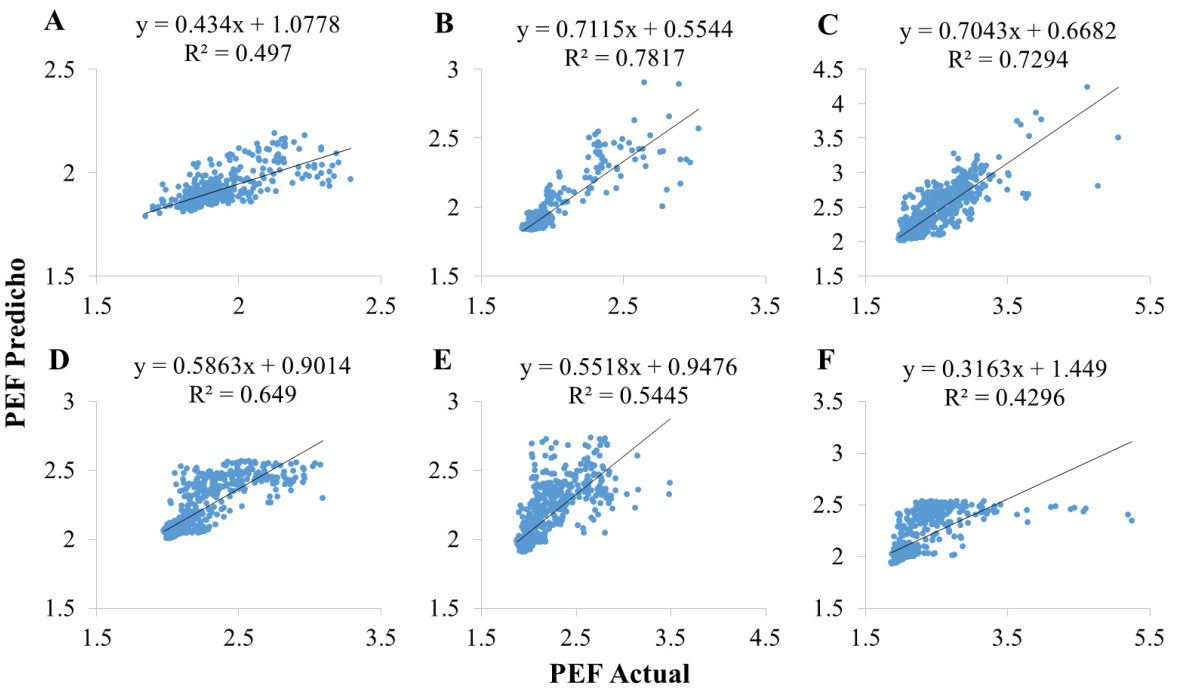

Figura 10. Valores de PEF predichos a partir de los seis pozos de evaluación, procesados previamente con el filtro de corrección de densidad. **A.** Pozo 1. **B.** Pozo 2. **C.** Pozo 3. **D.** Pozo 4. **E.** Pozo 5. **F.** Pozo 6.

Resultados

Las grillas de búsqueda (Tablas 3, 4 y 5) muestran el resumen de los valores obtenidos de coeficiente de correlación, error medio absoluto y error medio cuadrático. El mayor coeficiente de correlación de esta grilla que se puede aceptar es el de 0,6702, ya que los otros valores mayores exhiben una tendencia de sobreentrenamiento, debido a su incremento hacia los factores de complejidad y exponentes más altos (Tabla 3). El coeficiente aceptado fue alcanzado con los siguientes parámetros, un factor de complejidad igual a 2 (C = 2) y un exponente de 4 (E = 4), los cuales están resaltados en la Tabla 3. En dicha tabla se tomó el mayor valor de correlación en el centro de la grilla, porque, aunque hay valores más altos en el extremo inferior derecho, estos son los que muestran sobreentrenamiento y por lo tanto no son aceptables. Estás validaciones se usaron para obtener los parámetros más óptimos de complejidad y exponente para realizar las predicciones.

Los resultados del set de entrenamiento (Figura 8) permitieron confirmar que es posible correlacionar registros que miden diferentes características en el subsuelo, y fueron útiles para verificar que no hubiese sobreentrenamiento, en cuyo caso el valor resultante de R^2 sería muy cercano a 1,0 en esta figura y la predicción tendría una alta exactitud.

La Tabla 6 reúne los resultados de las predicciones obtenidas con la herramienta clasificadora de WEKA *Explorer*, en ella se muestran los valores de coeficiente de correlación, error medio absoluto y error medio cuadrático para cada etapa. En el caso de las predicciones 1 y 2, se presenta el promedio obtenido con los seis pozos de prueba.

Tabla 6. Resumen de los resultados de las predicciones realizadas con WEKA *Explorer*.

Clasificación	Coeficiente de Correlación	Error Medio Absoluto	Error Medio Cuadrático
Set de entrenamiento	0,669	0,219	0,315
Predicción 1	0,660	0,166	0,283
Predicción 2	0,774	0,095	0,172

Nota. Los valores de las predicciones 1 y 2 resultan del promedio de los pozos del set de evaluación. Los filtros aplicados en cada etapa se enuncian a continuación.

Set de entrenamiento: Con filtro DCOR ± 0,05 g/cm^3.

Predicción 1: Sin filtro.

Predicción 2: Con filtro DCOR ± 0,1 g/cm^3.

De acuerdo con la Tabla 6, el filtro realizado para la última etapa produjo una mejora en la correlación y el error, no obstante, se debe tener en cuenta que a mayor restricción en el filtro DCOR, menor es la cantidad de instancias que se pueden predecir.

Con los valores obtenidos en la segunda predicción (Figura 10), se graficaron los registros PEF para cada pozo de prueba en la Figura 11, allí se observa que la predicción es útil, ya que guarda la tendencia general del registro actual, sin embargo, presenta desfases en las partes profundas del registro donde aumenta pronunciadamente el PEF actual.

Al mirar el intervalo de interés aproximado (primeros 272 ft de cada registro en la Figura 11), se verifica que la curva predicha corresponde en buena medida a la actual. Después de dicho intervalo aproximado, comienza la Formación Gacheta, que presenta diferencias litológicas respecto a nuestra unidad de estudio suprayacente, esto se evidencia en las partes B, C, D, E y F de la Figura 11, donde hay una mayor discrepancia entre el PEF predicho y el actual, el cual alcanza valores de 5,0 en algunos casos (Figura 11C, 11F).

La Formación San Fernando muestra patrones de rayos gamma cilíndricos (Figura 2), los cuales, según Rojas *et al.* (2004), son típicos de las arenitas de esta unidad y pueden interpretarse como arenitas fluviales de tipo anastomosado con intervalos de arcillolitas de llanura de inundación. Por otra parte, la Formación Gacheta está compuesta por arcillolitas ricas en materia orgánica con algunas intercalaciones de arenitas y calizas, las cuales fueron depositadas en ambientes marinos de plataforma externa (Piedrahita, 2016). Este cambio de facie sedimentaria explica los desfases en los registros de la Figura 11, ya que el PEF incrementa con la presencia de carbonatos como la calcita y dolomita (Figura 3).

La información de cada pozo del set de entrenamiento corresponde prácticamente al intervalo de interés, debido a la omisión de los extremos superiores e inferiores de

cada registro, lo que permite entrenar específicamente la unidad que se está trabajando, es decir, la Formación San Fernando. Por consiguiente, el modelo no genera la predicción esperada para los valores altos de PEF que están fuera de dicho intervalo y que además hacen parte de otra unidad litoestratigráfica (Figura 11B, 11C, 11D, 11E, 11F). A pesar de ello, para el set de evaluación no se omitieron los extremos de los registros, en virtud de que son datos que se usan para probar si el entrenamiento funciona y por este motivo se evitó su alteración, salvo por el filtro DCOR que se introdujo con el propósito de retirar los valores atípicos para la predicción final.

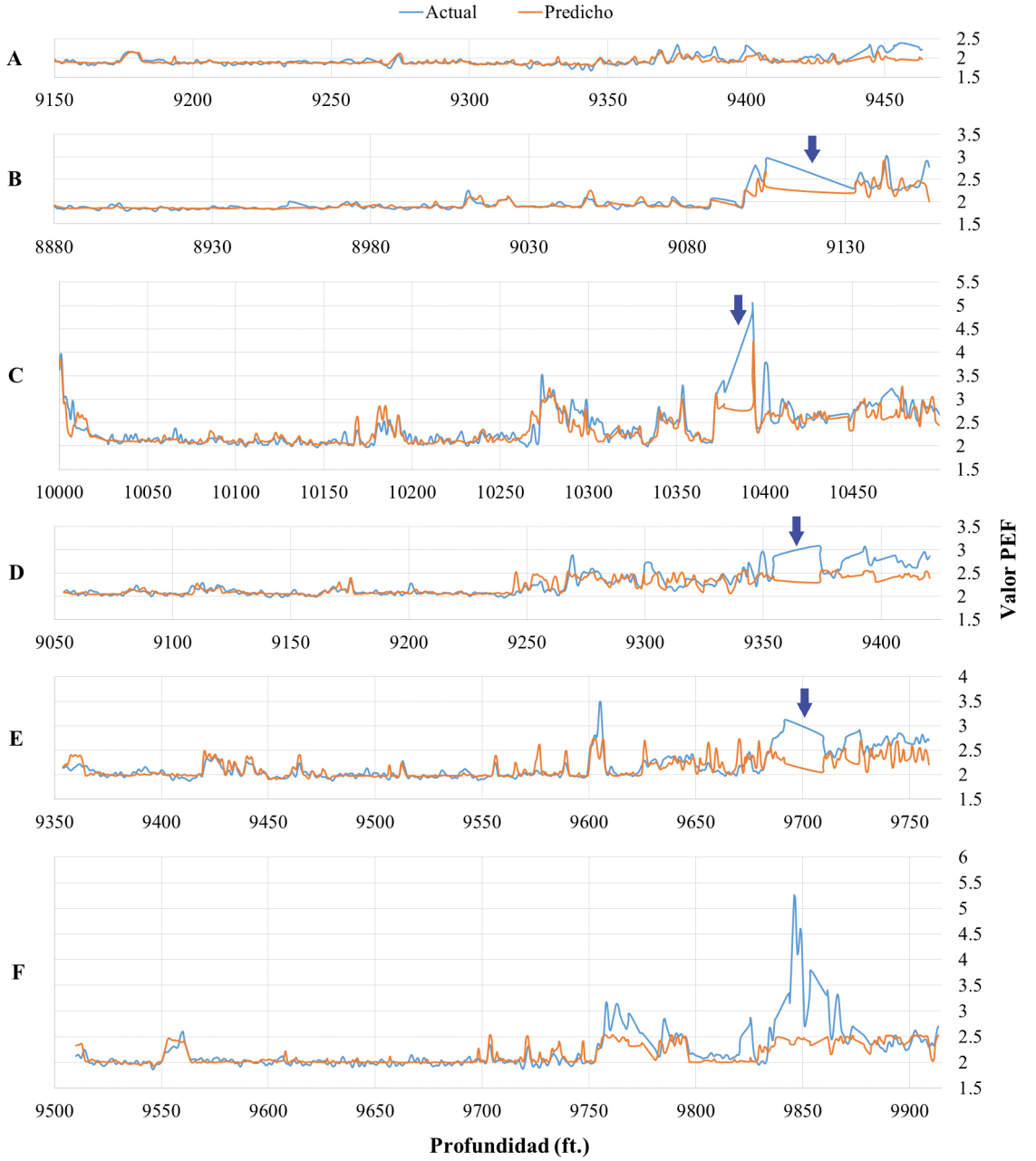

Figura 11. Registro PEF actual (azul) y predicho (naranja) para los pozos de prueba, cuya profundidad es en *Measured Depth* (MD). El intervalo de interés que comprende la Formación San Fernando, corresponde aproximadamente a los primeros 272 ft de cada registro de pozo. Luego de este intervalo, comienza la Formación Gacheta. Por ejemplo, en el Pozo 3 (C) la base de la Formación San Fernando está a una profundidad aproximada de 10272 ft. Las flechas señalan las porciones con pendientes rectas que están relacionadas con picos altos en el registro DCOR. **A.** Pozo 1. **B.** Pozo 2. **C.** Pozo 3. **D.** Pozo 4. **E.** Pozo 5. **F.** Pozo 6.

No obstante, la metodología no excluye ningún tipo de litología, de modo que debe ser aplicable para cualquier tipo de unidad sin importar si es homogénea o heterogénea. En este sentido, el modelo tiene una buena respuesta a las zonas heterogéneas de la Formación de estudio, por ejemplo, en la parte C de la Figura 11 en la porción central del registro, el PEF varía de 2,0 a 3,0 barns/electrón y la predicción logra ajustarse a los datos actuales. Este es uno de los segmentos donde hay mayor variación dentro del intervalo de interés, por esta razón, la metodología puede acomodarse a unidades con este tipo de heterogeneidad.

En otro orden de ideas, al revisar las respuestas de los demás registros básicos en los pozos de prueba, se notó que el registro DCOR presenta picos altos a las profundidades donde están los desfases del PEF, y es por esta causa que la mayoría de estas disparidades exhiben líneas rectas en pendiente (Figura 11B, 11C, 11D, 11E), dado que las porciones con picos (de 2,0 y hasta 4,0 g/cm^3) en la curva DCOR fueron omitidas gracias al filtro (el cual restringió los datos del set de evaluación hasta un máximo valor de 0,1 g/cm^3), en consecuencia el registro PEF actual conecta los puntos entre los cuales hubo omisión de datos y esto es lo que produce las pendientes rectas en los desfases.

Debido a estas omisiones, es posible que la validación para la predicción 2 en WEKA *Explorer classify,* no haya logrado tomar la tendencia correcta de los datos en dichos segmentos, y por este motivo queda subestimada la predicción. Si bien, las discrepancias fuera del intervalo de interés son notorias, al evaluar el error medio absoluto, se puede decir que no es significativo puesto que su valor es de 0,095 barns/electrón, esto es comparable con la influencia despreciable que tienen los fluidos en el valor del factor P_e.

De acuerdo a estos resultados y al registro PEF inferido, se evidencia que esta predicción es funcional en cuanto a la formación estudiada.

Discusión

La diferencia en los segmentos con desfases en la Figura 11 también podría deberse a que el set de entrenamiento está compuesto principalmente por valores menores de PEF, como se evidencia en la Figura 5. Por otro lado, en los pozos de prueba no se omitieron los extremos de los registros como se hizo con los pozos del set de entrenamiento, lo cual, respalda la idea que los desfases corresponden al tope de la formación que subyace a la unidad de estudio (Figura 11).

Estudios como el de Labani *et al.* (2010) confirman que es factible obtener mejores valores de correlación cuando el set de datos de entrenamiento está compuesto solo por un par de pozos, pero esto puede provocar el sobreentrenamiento del modelo, o la generación de uno aplicable solo para registros particulares con características muy similares. Si se disminuye el número de pozos se puede esperar una mejor correlación sin caer en el sobreajuste, siempre y cuando, se rectifique esto en las grillas de búsqueda y en las gráficas de predicción. Haciendo validaciones consecutivas se podría llegar al número ideal de pozos para un escenario cómo el de este estudio. El hecho de utilizar más pozos sencillamente hace que el modelo esté mejor alimentado, y que pueda admitir un mayor grado de variación en los datos ingresados para predecir.

Los resultados de la primera predicción del set de prueba (Figura 9), muestran que la predicción desviada de los valores actuales ocasiona una reducción considerable en el valor de regresión, lo cual mejoró al procesar los pozos de prueba con el filtro DCOR y predecir nuevamente. Este filtro (±0,1 g/cm^3) se amplió respecto al realizado para el entrenamiento (±0,05 g/cm^3), es decir que para entrenar se tuvo una mayor restricción en los atributos, y esto permitió que se tuvieran datos más certeros para alimentar el modelo. En la segunda predicción se obtuvieron regresiones mayores (Figura 10), que dieron paso a la representación gráfica de los registros (Figura 11).

Respecto a la litología, es una variable implícita en los registros, y podría considerarse una ventaja el hecho que corresponda a arenitas cuarzosas predominantes con delgadas intercalaciones de arcillolitas. Sin embargo, si se estudiaran unidades más heterogéneas o litologías diferentes, habría que examinar cuáles ajustes serían necesarios en el procesamiento y las predicciones, ya que esta fue una de las partes más importantes de este estudio. Esta afirmación sobre la aplicabilidad de la metodología a cualquier tipo de unidad, debería verificarse mediante la realización de pruebas en distintas formaciones. A pesar de esto, el estudio muestra cierta semejanza con el procedimiento realizado por Labani *et al.* (2010) para rocas calizas. Lo cual muestra que dicha metodología puede ser ajustable a las variaciones geológicas, considerando el procesamiento previo de los datos cómo un paso importante.

Por otro lado, no se conocen aplicaciones de esta metodología en la misma cuenca, aunque se han aplicado métodos cómo el de redes neuronales artificiales en otros lugares para generar registros de pozo sintéticos, por ejemplo, el estudio de Parapuram *et al.* (2018) se desarrolló con datos de 112 pozos de la Formación Bakken en Dakota del Norte, y según los autores sus modelos funcionan y pueden predecir las propiedades geomecánicas de los futuros pozos de la Formación. Otro caso, es el de Rolon *et al.* (2009) que utiliza registros de rocas siliciclásticas del Devónico Superior del sur de Pensilvania. Así mismo, Zhang *et al.* (2018) trabajaron con la Formación Eagle Ford Shale en los EE.UU. y su método de *machine learning* proporciona una forma precisa y rentable para la generación de registros sintéticos. Por último, Khandelwal y Singh (2010) se dedicaron a predecir los valores de porosidad neutrón y densidad empleando los parámetros de entrada de los registros de rayos gamma, resistividad y sónico. Esto sin mencionar las investigaciones realizadas para automatizar la interpretación estratigráfica (Igbokwe, 2011; Wu y Nyland, 1987) y el procesamiento de registros de pozo (Wu *et al.,* 2018).

Conclusiones

En este ejercicio se lograron unir diferentes disciplinas para obtener una innovación en la predicción de información geológica, y dar un aporte al conocimiento que se tiene sobre *machine learning* en el campo de la geología del petróleo. Las inferencias en registros de pozo con ML requieren un filtrado minucioso para tomar la tendencia de los datos, y una optimización clara para prevenir el sobreentrenamiento del modelo. La metodología expuesta, muestra que se pueden obtener cada vez mejores modelos, y es útil para aplicar a conjuntos de datos donde se necesite realizar selección de atributos y predicción de gráficas.

El enfoque dado al procesamiento de los datos fue el de predecir el registro PEF, que a su vez es un buen indicador de la litología. La entidad que suministró los datos fue un canal que permitió desarrollar esta iniciativa académica. Básicamente, para el campo estudiado esto fue un ejercicio práctico, ya que se conoce la litología de la unidad, pero esta es una metodología que está disponible para ser aplicada en la industria, y podría servir en otros campos petrolíferos que requieren del completamiento de información.

Para su implementación sería necesario realizar pruebas en pozos carentes del registro PEF, y compararlos con aquellos que estén completos en la misma unidad. Esto implicaría obtener un nuevo acceso a la información requerida, para posteriormente tratar los datos, entrenar un modelo y confirmar su viabilidad. Puesto que esta aplicación no se ha realizado en la entidad, aún no se conoce el alcance que podría tener en la producción, pero cabe aclarar que de acuerdo a los resultados y al desarrollo de estudios paralelos, se visualiza la inserción de esta técnica en la industria del petróleo, la cual se vería beneficiada si se logra completar información crucial y se evita incurrir en gastos de perforación. Dicha información puede ser la que normalmente se requiere para los análisis petrofísicos, amarres sísmicos, correlación estratigráfica y la toma de decisiones para explotación y desarrollo. Adicionalmente, esta técnica se puede proyectar para descubrir nuevos usos, como predecir el potencial de nuevos prospectos y la tasa de producción de un campo en desarrollo.

Agradecimientos

Esta investigación se realizó en Ecopetrol S.A. en la Vicepresidencia de Desarrollo, con el apoyo de la Universidad Nacional de Colombia sede Bogotá. Agradecemos a cada uno de los que colaboraron para que este trabajo se llevara a cabo. Al profesor Luis Ochoa, al ingeniero Pedro Solórzano, al geólogo Simón Hincapié, al geólogo Javier Rojas y al geofísico Franklyn Sáez. A los evaluadores Lucia Torrado y William Agudelo.

Referencias

Asquith, G.B.; Gibson, C.R. (1982). *Basic well log analysis for geologists*. 2nd edition. AAPG.

Asquith, G.; Krygowski, D. (2004). *Basic well log analysis.* AAPG. Methods in Exploration Series, No 16.

Ballesteros, C.A.; Torres, C.A. (2017). Análisis conceptual del impacto de procesos térmicos de recobro mejorado, en completamientos convencionales y multiobjetivo con crudo extra-pesado en la Formación San Fernando, en la cuenca de los llanos Orientales. Tesis, Fundación Universidad de América, Bogotá, Colombia.

Brownlee, J. (2016). *Machine Learning Mastery with Weka: Analyze Data, Develop Models, and Work Through Projects*. Ebook Edition.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326-327. https://doi.org/10.1145/212094.212114

Drazin, S.; Montag, M. (2012). Decision tree analysis using weka. Machine Learning-Project II. University of Miami. Internal report.

El Naqa, I.; Murphy, M.J. (2015). What is Machine Learning? In: I. El Naqa; R. Li; M.J. Murphy (eds). *Machine Learning in Radiation Oncology* (pp. 3-11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Flores, R. (2014). Coalbed Gas Production. In: *Coal and Coalbed Gas* (pp. 369-436). Elsevier. https://doi.org/10.1016/B978-0-12-396972-9.00007-0

Glover, P. (2000). Petrophysics MSc Course Notes Wireline Logging Dr Paul Glover Page 59. Consultado el 4 de Noviembre del 2020. https://www.coursehero.com/file/p46psk5/Petrophysics-MSc-Course-Notes-Wireline Logging-Dr-Paul-Glover-Page-59-Data-is/

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18. https://doi.org/10.1145/1656274.1656278

Igbokwe, O.A. (2011). Stratigraphic interpretation of Well-Log data of the Athabasca Oil Sands of Alberta Canada through Pattern recognition and Artificial Intelligence. Master's thesis. Westfälische Wilhelms-Universität Münster, Münster, Alemania.

Khandelwal, M.; Singh, T.N. (2010). Artificial Neural Networks as a valuable tool for well log interpretation. *Petroleum Science and Technology*, *28*(14), 1381-1393. https://doi.org/10.1080/10916460903030482

Labani, M.M.; Kadkhodaie-Ilkhchi, A.; Salahshoor, K. (2010). Estimation of NMR log parameters from conventional well log data using a committee machine with intelligent systems: A case study from the Iranian part of the South Pars gas field, Persian Gulf Basin. *Journal of Petroleum Science and Engineering*, *72*(1-2), 175-185. https://doi.org/10.1016/j.petrol.2010.03.015

Ochoa, L.H.; Niño, L.F.; Vargas, C.A.; (2018). Fast estimation of earthquake epicenter distance using a single seismological station with machine learning techniques. *DYNA, 85*(204), 161-168. https://doi.org/10.15446/dyna.v85n204.68408

Onajite, E. (2014). Understanding seismic interpretation methodology. In: *Seismic Data Analysis Techniques in Hydrocarbon Exploration* (pp. 177-211). Elsevier.

Parapuram, G.; Mokhtari, M.; Ben Hmida, J. (2018). An artificially intelligent technique to generate synthetic geomechanical well logs for the Bakken Formation. *Energies*, *11*(3). https://doi.org/10.3390/en11030680

Piedrahita, J. (2016). Caracterización petrofísica de un área en el bloque CPO 16 en la cuenca de los Llanos Orientales – Colombia. Tesis, Universidad EAFIT, Medellín, Colombia.

Schlumberger Oilfield Glossary. Consultado el 26 de Diciembre del 2019. https://www.glossary.oilfield.slb.com/en/Terms/p/pef.aspx

Rojas, N.; Cardona, P.; Arango, S.; Florez, A. (2004). Redescubriendo un gigante: yacimiento San Fernando campo Chichimene - Cuenca Llanos orientales. *III Convención Técnica ACGGP*. Bogotá, Colombia.

Rolon, L.; Mohaghegh, S.D.; Ameri, S.; Gaskari, R.; McDaniel, B. (2009). Using artificial neural networks to generate synthetic well logs. *Journal of Natural Gas Science and Engineering*, *1*(4-5), 118-133. https://doi.org/10.1016/j.jngse.2009.08.003

Wu, X.; Nyland, E. (1987). Automated stratigraphic interpretation of well-log data. *Geophysics*, *52*(12), 1665-1676. https://doi.org/10.1190/1.1442283

Wu, P.Y.; Jain, V.; Kulkarni, M.S.; Abubakar, A. (2018). Machine learning-based method for automated well-log processing and interpretation. *SEG Technical Program Expanded Abstracts 2018,* 2041-2045. https://doi.org/10.1190/segam2018-2996973.1

Zhang, D.; Chen, Y.; Meng, J. (2018). Synthetic well logs generation via Recurrent Neural Networks. *Petroleum Exploration and Development*, *45*(4), 629-639. https://doi.org/10.1016/S1876-3804(18)30068-5

Fecha de recibido: 08 de abril de 2020
Fecha de aprobado: 07 de diciembre de 2020