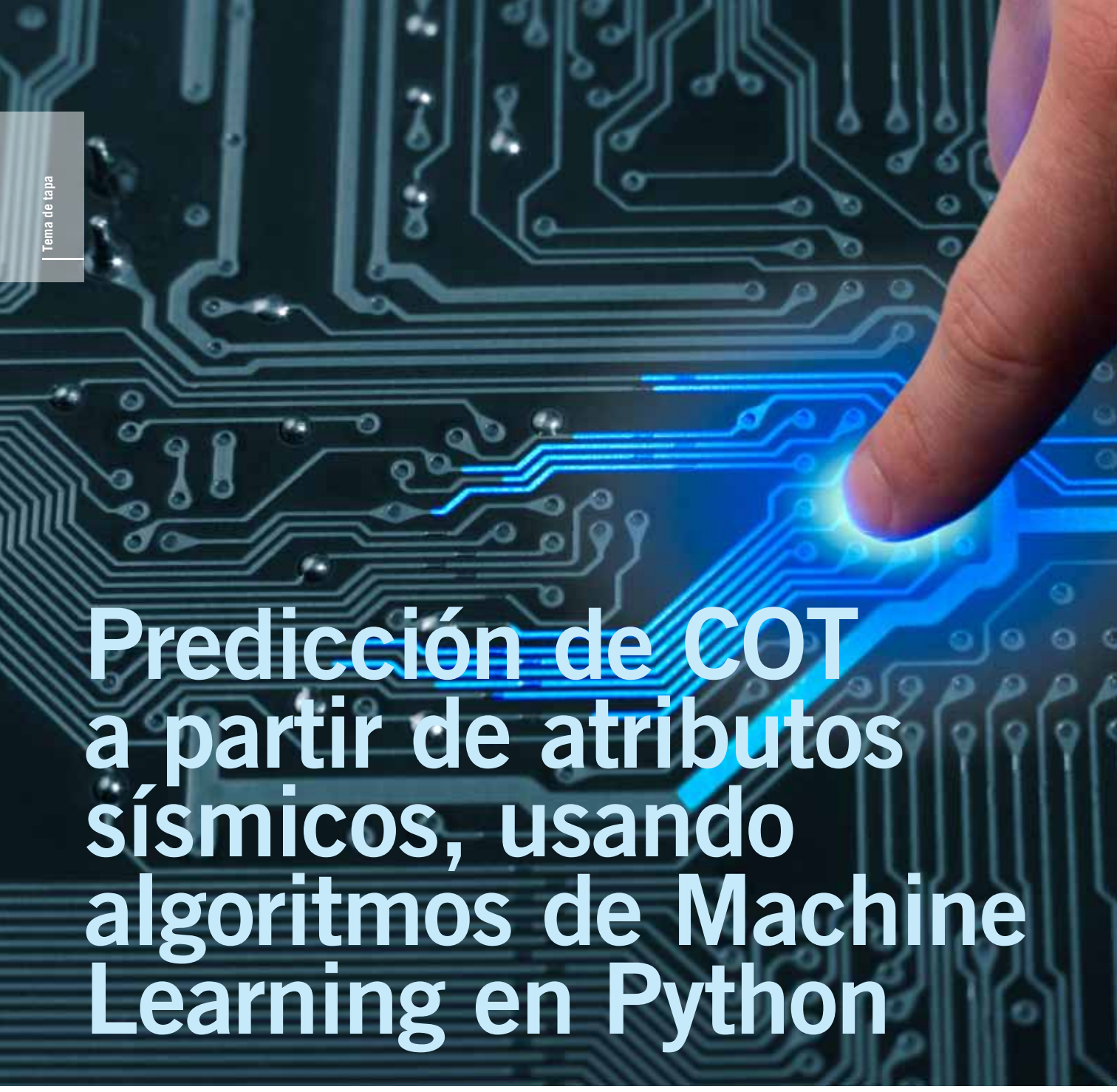
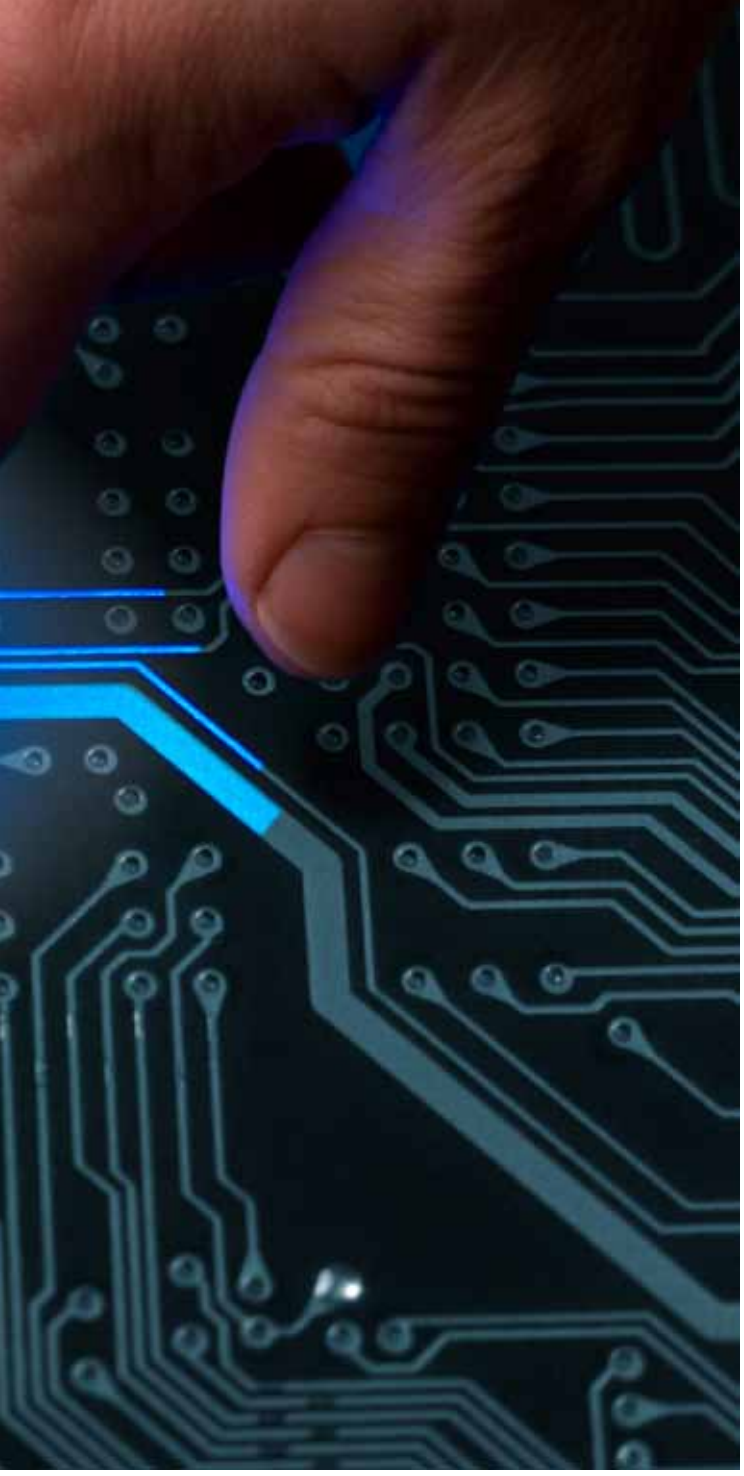


Predicción de COT a partir de atributos sísmicos, usando algoritmos de Machine Learning en Python

Por ***Alejandro Bascur (Pampa Energía), Enzo Luna*** (Pampa Energía) y ***Hernán Merlino*** (IGPUBA – FIUBA).

El uso de machine learning sobre atributos sísmicos permite predecir el contenido orgánico total (COT) con mayor precisión en formaciones como Vaca Muerta. Esta metodología mejora la caracterización de reservorios no convencionales y amplía la cobertura areal de datos geoquímicos.

Este trabajo fue seleccionado en las 3º Jornadas de Revolución Digital para Petróleo y Gas.



El Contenido Orgánico Total (COT), esencial en la evaluación de reservorios no convencionales de shale, proporciona información sobre la cantidad de materia orgánica presente en la roca. El volumen de COT remanente se correlaciona con el volumen posible generado a una madurez dada, es decir, un COT elevado indica mayor capacidad de generación de hidrocarburos durante la maduración térmica, convirtiéndolo en uno de los indicadores para evaluar el potencial de producción en este tipo de reservorios.

El presente trabajo se enfoca en introducir una metodología innovadora, diferente a los enfoques convencionales, con el propósito de mejorar la estimación del COT en la Formación Vaca Muerta, empleando herramientas de ciencia de datos.

A pesar de que la Impedancia P se ha establecido como el atributo principal para la estimación del COT en shale, la inclusión de otros productos sísmicos y la aplicación de esta nueva metodología han permitido reducir los errores por debajo de los niveles alcanzados con métodos de ajuste lineales.

Para este fin, se emplearon datos sísmicos pre-stack y post-stack, así como productos derivados de inversiones sísmicas, en conjunción con mediciones de COT obtenidas directamente de pozos (mediante muestras de cutting y coronas).

La integración de estos conjuntos de datos sísmicos, combinada con técnicas de modelado y análisis de datos utilizando algoritmos de machine learning, ha permitido obtener una representación más completa y detallada de las propiedades de la roca madre y su capacidad de generación de hidrocarburos. Este enfoque ha evidenciado el potencial y la utilidad de la ciencia de datos en la evaluación de reservorios de shale, culminando en una mayor precisión y ajuste, lo que representó un avance significativo en la caracterización, comprensión y precisión de los datos.

Introducción

La formación Vaca Muerta se destaca como una de las formaciones geológicas más relevantes para la explotación de hidrocarburos no convencionales, tanto el shale oil como el shale gas. En este marco, la geoquímica y la geomecánica juegan un papel fundamental. La integración de estudios geoquímicos y geomecánicos proporciona una comprensión completa de la formación Vaca Muerta.

La integración de datos y modelos de ambas disciplinas optimiza la explotación de los recursos no convencionales presentes en esta formación geológica. Dentro de la amplia gama de estudios ofrecidos por la geomecánica y la geoquímica, se optó en enfocarse en el análisis de la materia orgánica, en particular, el carbono orgánico total (COT).

La medición y comprensión de estos valores se realiza de diversas maneras, como la extracción de coronas y el cutting. Sin embargo, estas mediciones suelen ser puntuales y están limitadas a las ubicaciones de los pozos. Aunque proporcionan un excelente nivel de detalle en la dimensión vertical, su alcance en la dimensión horizontal es limitado.

Aquí es donde la sísmica adquiere relevancia, ya que ofrece una distribución horizontal amplia. No obstante, su resolución vertical suele ser más baja como contrapartida. La adquisición y el análisis de datos sísmicos son vitales para comprender la estructura y composición de los yacimientos. Estos datos ofrecen una perspectiva invaluable para la exploración y producción, pero su análisis e interpretación pueden ser desafiantes debido a la complejidad de las formaciones geológicas y la gran cantidad de información recopilada (varios atributos sísmicos).

En este punto entra en juego el Machine Learning (ML). Los algoritmos avanzados de ML, como las redes neuronales y los modelos de aprendizaje profundo, se presentan como herramientas poderosas para analizar

tanto datos de pozos como datos sísmicos, y realizar predicciones sobre las propiedades de los reservorios.

Objetivo del proyecto

El objetivo de este estudio es evaluar las herramientas y metodologías de la ciencia de datos aplicadas a datos sísmicos, con el propósito de mejorar la predicción de las propiedades del reservorio. La propiedad seleccionada es el carbono orgánico total (COT), y el objetivo específico es predecir a partir de datos obtenidos en pozos, utilizando la sísmica como medio para la propagación areal.

Desarrollo

La elección del COT como propiedad de la roca a predecir se justifica por su baja complejidad estructural y el amplio conocimiento disponible sobre esta característica en la Cuenca Neuquina. Además, la disponibilidad de un modelo previo nos permitió utilizarlo como punto de referencia para comparar los resultados obtenidos en este trabajo.

La zona seleccionada corresponde al área El Mangrullo (255 Km2 aprox.) debido a la calidad destacada de la información sísmica disponible.

El flujo de trabajo desarrollado, detallado en la Figura 1, representa esencialmente un proceso estándar para la construcción de un modelo de Machine Learning (ML), que utilizamos como referencia para nuestro estudio.

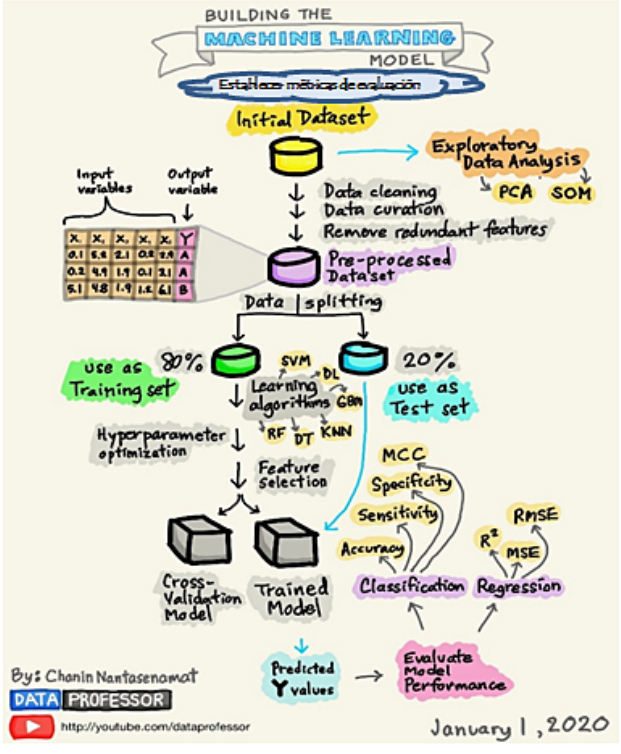


Figura 1. ***Cartoon Infographic on building the machine learning model (Draw by Chanin Nantasemanat).***

Selección de Datos (Initial Dataset)

En el área de estudio se encontraron 7 pozos con mediciones de COT, de los cuales 6 se obtuvieron mediante cutting y 1 a través de coronas. Además, se recopilaron 35 productos sísmicos (Pre-stack, Post-stack y de inversión) para utilizar como datos de entrada en el modelo. Cada uno de estos productos fue seleccionado por su relación con el parámetro que se busca predecir.

Métricas de Evaluación

Seleccionar métricas de evaluación antes de construir el modelo es crucial para diseñarlo y optimizarlo eficazmente, asegurando que cumpla con los objetivos del proyecto y las necesidades del negocio. Esto proporciona claridad en los objetivos, facilita la comparación entre modelos, se alinea con los requisitos del negocio y previene sorpresas desagradables en el futuro. Las métricas que nosotros seleccionamos son:

Error absoluto medio (MAE)

Es una medida del tamaño medio de los errores en una colección de predicciones, sin tener en cuenta su dirección. Se mide como la diferencia absoluta promedio entre los valores predichos vs los valores reales y se utiliza para evaluar la efectividad de un modelo de regresión.

$$MAE = (1/n)\ \Sigma(i=1\ an)\ |y\_i - \ddot{y}\_i|$$

n = número de observaciones en el conjunto de datos.
y_i =valor verdadero.
ÿ_i =valor previsto.

R^2 (R cuadrado)

Muestra qué tan bien un modelo de regresión (variable independiente) predice el resultado de los datos observados (variable dependiente). R^2 también se conoce comúnmente como coeficiente de determinación. Es un modelo de bondad de ajuste para análisis de regresión lineal.

$$R^2 = 1 - RSS/TSS$$

RSS = suma de cuadrados de residuos.
TSS = suma total de cuadrados.

Adecuación de datos

El preprocesamiento de datos es un paso fundamental antes de aplicar un modelo de machine learning, ya que mejora la calidad, consistencia y relevancia de los datos, reduce el riesgo de sobreajuste y contribuye a un mejor rendimiento general del modelo. Algunos de los pasos más relevantes que realizamos fueron:

- Calidad de los datos: Identificación de valores atípicos (outliers) (Figura 2).
- Consistencia de los datos: Estandarización de escalas y normalización de distribuciones.

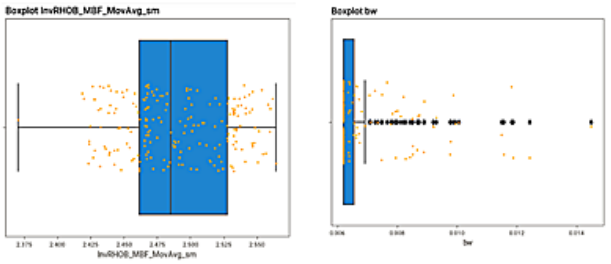


Figura 2. ***Outliers.***

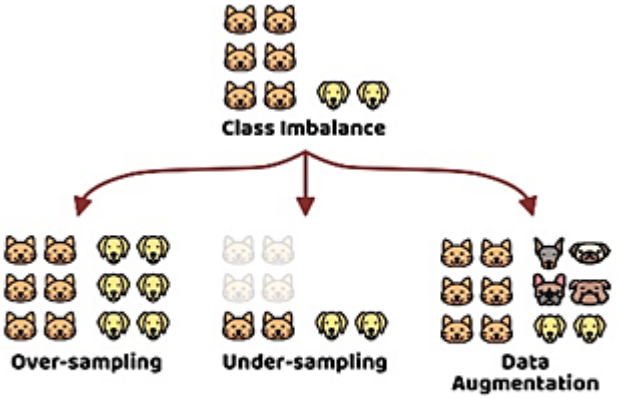


Figura 3. ***Oversampling.***

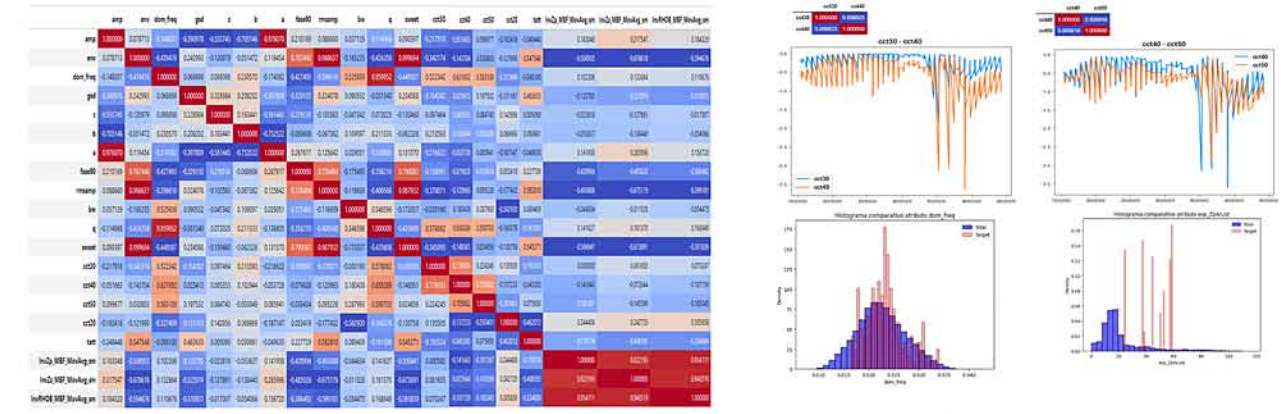


Figura 4. ***Matriz de Correlación y reducción de Dimensionalidad.***

- Balance de Clases: Utilizamos el método ADASYN (Adaptive Synthetic Sampling) para generar muestras sintéticas de forma adaptativa (Figura 3).
- Reducción de Dimensionalidad: Empleamos el análisis de componentes principales (PCA) para la selección de características (Figura 4).
- Prevención de Sobreajuste: Aplicamos técnicas de validación cruzada (k-fold), dividiendo el conjunto de datos en k subconjuntos ("folds"). El modelo se entrena k veces, cada vez utilizando k-1 folds como datos de entrenamiento y el fold restante como datos de prueba. El rendimiento del modelo se calcula promediando los resultados de las k iteraciones.

Modelado

Para iniciar el modelado, optamos por utilizar bibliotecas de aprendizaje automático (AutoML). Herramientas como AutoKeras, TPOT y PyCaret están diseñadas para simplificar la construcción, ajuste y evaluación de modelos de machine learning. A continuación, se destacan algunas de las diferencias clave entre ellas:

	AutoKeras	**TPOT (Tree-based Pipeline Optimization Tool)**	**Pycarets**
Enfoque	Centrada en la automatización del diseño y la selección de arquitecturas de modelos de deep learning.	Utiliza algoritmos genéticos para explorar automáticamente una amplia gama de posibles pipelines de machine learning y encontrar la mejor combinación de preprocesamiento de datos y algoritmos de modelado.	Diseñada para simplificar el flujo de trabajo de aprendizaje automático desde la preparación de datos hasta la implementación del modelo.
Funcionalidad	interfaz de alto nivel para la creación de modelos de deep learning con pocas líneas de código. Utiliza técnicas como búsqueda bayesiana y la búsqueda aleatoria para la hiperparametrizacion y armado del modelo.	Búsquedas exhaustivas y automatizadas de pipelines de machine learning, incluyendo preprocesamiento de datos, selección de características y ajuste de hiperparámetros.	Ofrece amplia gama de funciones de aprendizaje automático, como la preparación de datos, la selección y la evaluación de modelos y la interpretación de resultados. Permite la personalización.
Flexibilidad	Orientado a deep learning , gran flexibilidad para ajustar y personalizar modelos de redes neuronales convolucionales, recurrentes y otros tipos de arquitecturas.	Proporciona una variedad de opciones de configuración para controlar el proceso de optimización, incluyendo el número de generaciones, la población inicial y los operadores genéticos.	Interfaz de usuario fácil de usar y una API programática para adaptarse a diferentes niveles de experiencia en aprendizaje automático.

Los resultados obtenidos al comparar los modelos mediante las métricas de evaluación se muestran en la siguiente tabla:

	MAE	R^2	Modelo Seleccionado
AutoKeras	0.59	0.62	Light Gradient Boosting Machine
TPOT	0.26	0.93	Random Forest Regression
Pycarets	0.98	0.72	Redes neuronales convolucionales

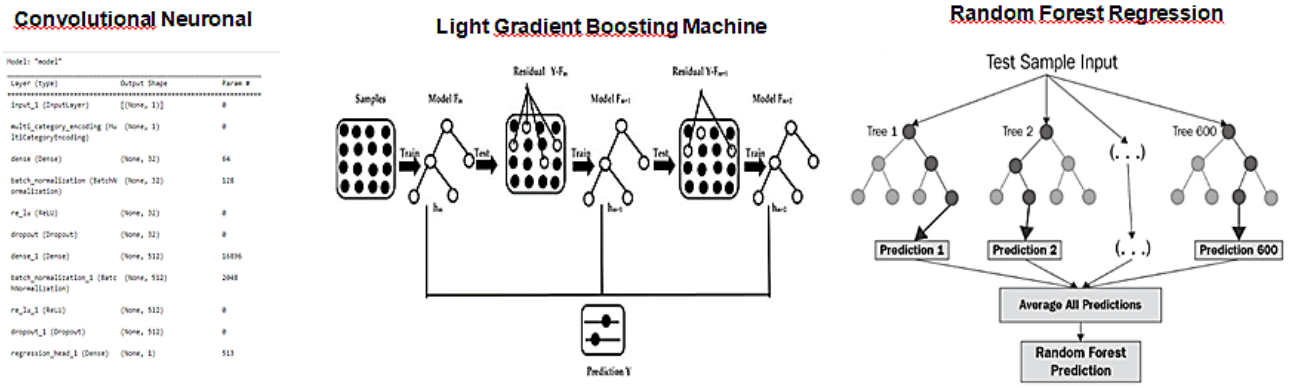


Figura 5. ***Modelos Candidatos***

La Figura 5 muestra el funcionamiento de los distintos modelos seleccionados por cada biblioteca como mejor candidato.

Comparación de modelos

Al comparar los tres modelos, observamos que tanto el modelo obtenido con PyCaret como el obtenido con TPOT muestran buenos resultados. Sin embargo, optamos por el modelo de TPOT debido a la gran correlación observada entre los valores medidos y los valores predichos, con los puntos muy cerca de la recta de identidad (Figura 6).

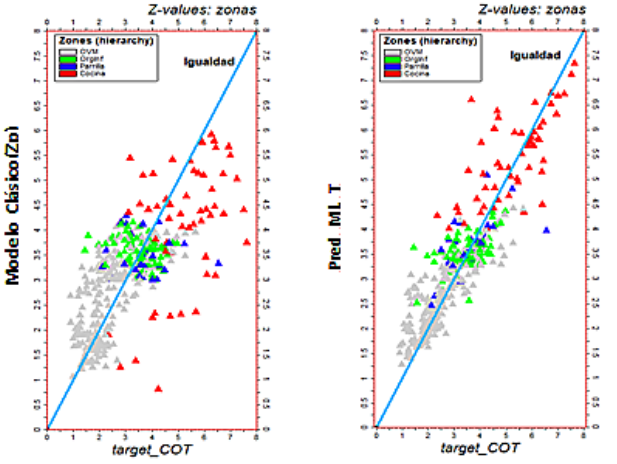


Figura 7. ***Comparación de modelo de referencia (clásico) vs. modelo de predicción.***

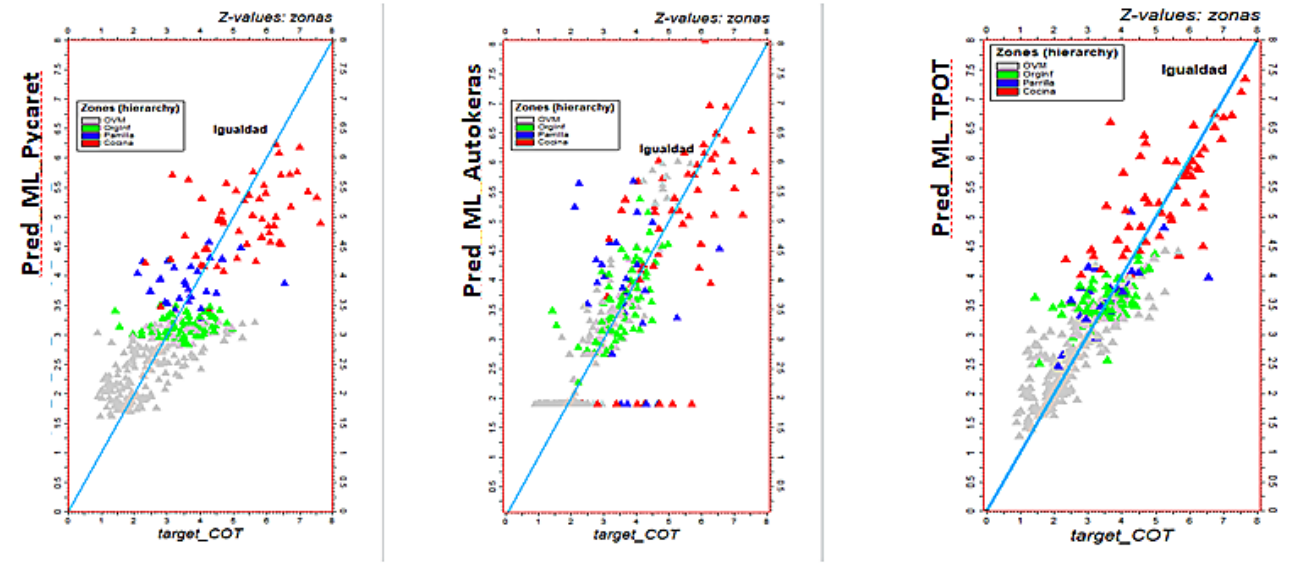


Figura 6. ***Comparación de los 3 modelo de predicción.***

Resumen

En resumen, al comparar el modelo de referencia con el modelo generado utilizando múltiples datos sísmicos, observamos que este último se ajusta considerablemente mejor, mostrando una dispersión de datos más estrecha (Figura 7). Además, el modelo muestra una buena separación de las distintas zonas.

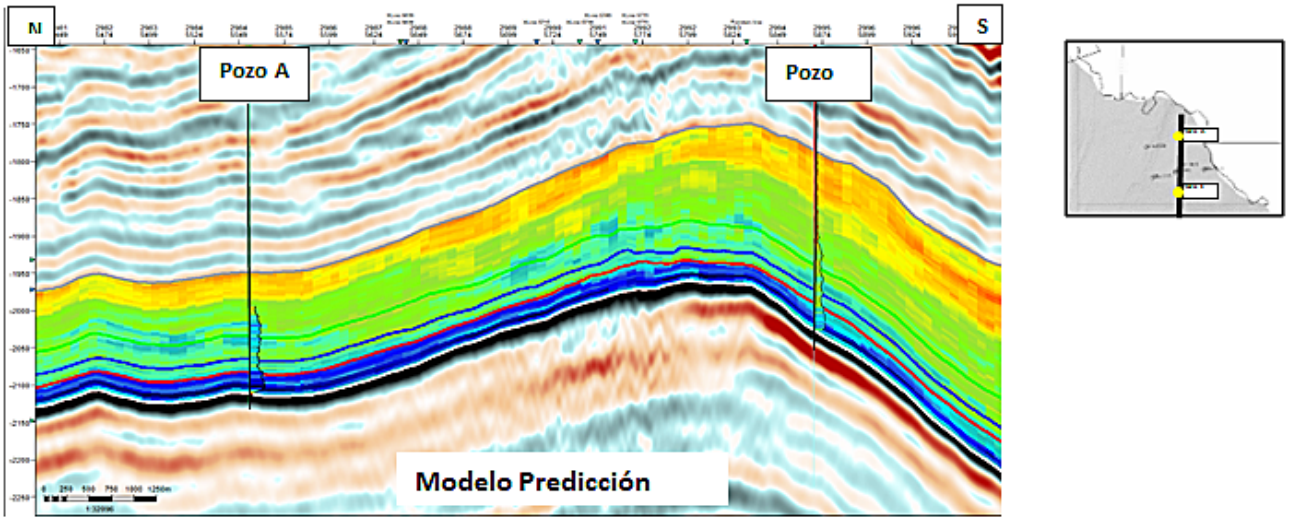


Figura 8. ***Corte comparando valor de COT en pozo vs. Predicción.***

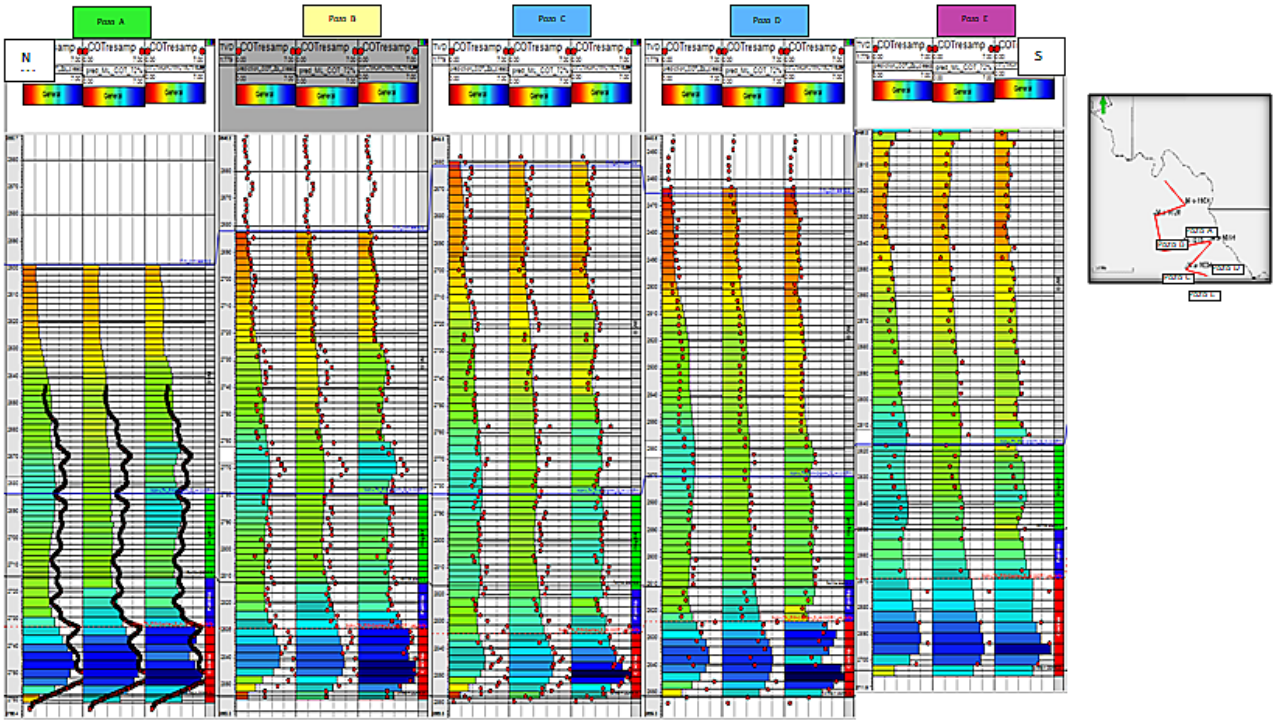


Figura 9. ***Comparación de puntos medidos vs. Valores predichos.***

Este resultado refleja una evaluación altamente satisfactoria de las herramientas y técnicas de ciencia de datos para la estimación del TOC a partir de atributos sísmicos y mediciones de pozo. En la Figura 8 se puede observar la superposición del modelo de predicción obtenido con los datos medidos en pozo, donde se destaca que, a pesar de su baja resolución vertical, el modelo identifica correctamente los rasgos

En la Figura 9 se presenta un corte estratigráfico que compara los resultados de tres modelos por pozo: el modelo de referencia (primero a la izquierda), el modelo obtenido con PyCaret (al centro) y el modelo obtenido con TPOT (a la derecha). Los puntos medidos en el pozo están superpuestos en color rojo. Es notable cómo el modelo generado con TPOT muestra un mejor ajuste a los datos de entrada en comparación con los otros dos modelos.

Otra ventaja importante es que nos proporciona un valioso conocimiento práctico (know-how) que nos permitirá aplicar estas herramientas y técnicas a otros activos y abordar diversas problemáticas con mayor confianza. Como próximo objetivo es aplicar esta metodología a la problemática de la sismoestratigrafia.