Ciencia de datos, *analytics, machine learning*, inteligencia artificial, *data warehousing, business intelligence, big data*

Aplicación de técnicas de *machine learning* como soporte al control de calidad de interpretación de datos sísmicos

Por ***Juan Ignacio Gómez, Ignacio Rovira, Luis Alberto Vernengo y Juan Francisco Moirano***
(Pan American Energy)

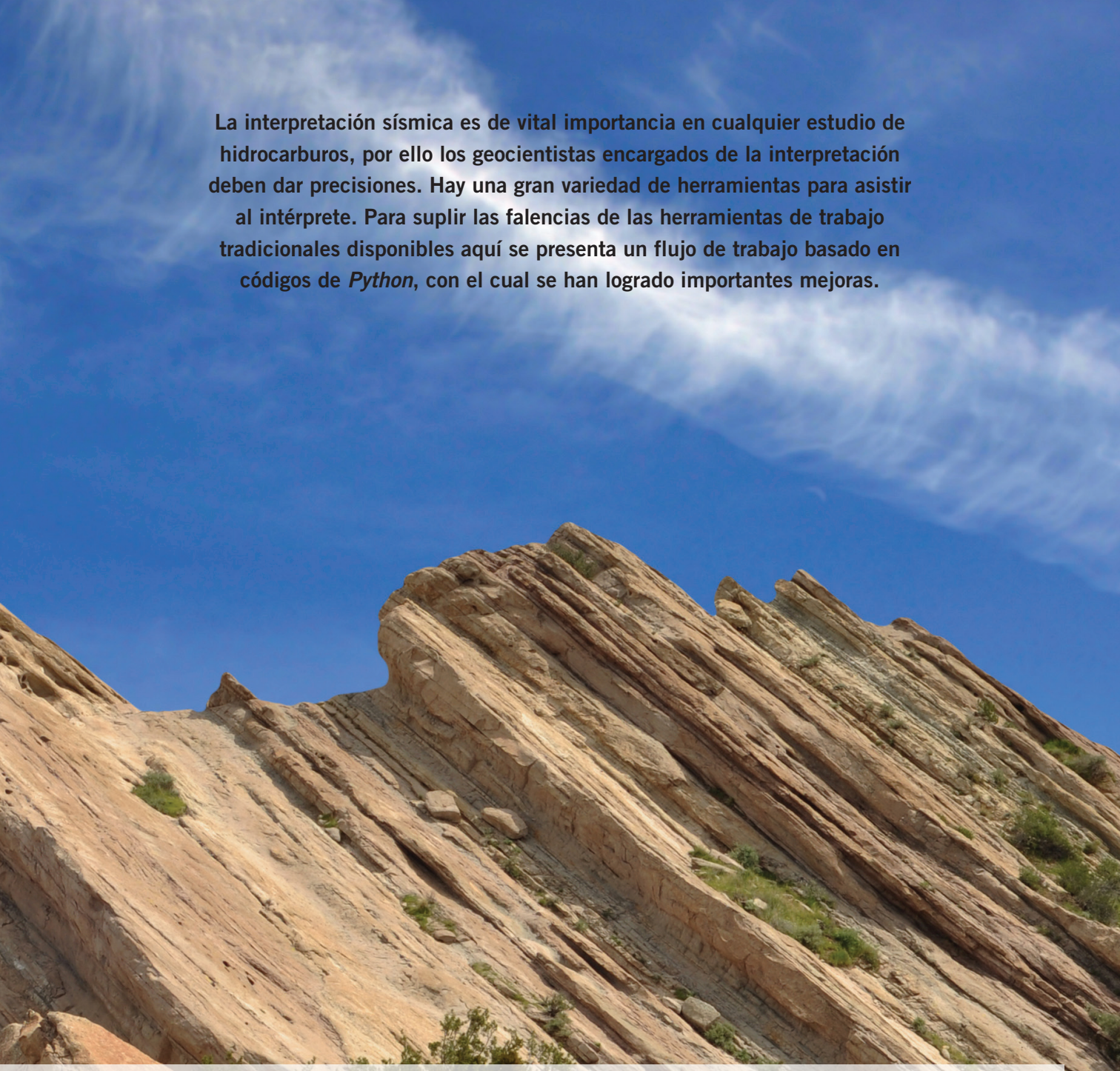
Debido a su vasta cobertura areal a bajo costo relativo, la adquisición de información sísmica otorga a los profesionales en geociencias una visión ideal de la corteza terrestre tanto para explorar como desarrollar yacimientos de hidrocarburos con una importante precisión. Por ello, el trabajo del geocientista, encargado de interpretar contextualmente estos datos, radica en identificar y mapear rasgos geológicos con interés económico, de manera rápida y eficiente. En este sentido, los horizontes y las fallas resultantes de la interpretación deben reflejar fielmente las morfologías y geometrías observadas en el dato sísmico con el fin de generar un modelo estático de la subsuperficie del área estudiada.

Para lograr este cometido, los intérpretes pueden encontrar soporte en diversas técnicas y algoritmos que abarcan desde el procesamiento y el cálculo de atributos de la señal sísmica y la interpretación asistida, hasta el análisis estadístico de datos para la confección de mapas.

Estos métodos tienen una amplia variedad de aplicaciones en distintas ramas de las geociencias, ciencias ambientales y agricultura, entre otras. En este sentido, la geoestadística se utiliza cuando, a partir de datos discretos, se desea conocer de forma continua la distribución espacial de una determinada variable. En el caso de las fallas y las estructuras geológicas interpretadas sobre la imagen de datos sísmicos es posible estimar su ubicación a partir de la interpretación discreta de algunos puntos en la zona de interés.

Respecto de las herramientas de trabajo disponibles, si bien algunas son muy avanzadas y complejas, todavía es posible detectar deficiencias en los procesos. Estas pueden dar lugar a resultados objetables e incluso a contratiempos en los flujos de trabajo, lo cual desvía el foco del proceso de interpretación, hacia la búsqueda de

La interpretación sísmica es de vital importancia en cualquier estudio de hidrocarburos, por ello los geocientistas encargados de la interpretación deben dar precisiones. Hay una gran variedad de herramientas para asistir al intérprete. Para suplir las falencias de las herramientas de trabajo tradicionales disponibles aquí se presenta un flujo de trabajo basado en códigos de *Python*, con el cual se han logrado importantes mejoras.



alternativas que permitan realizar la tarea propuesta.

Durante la década de 2020, el avance de nuevas tecnologías en el campo de la informática permitió un gran avance en la capacidad de cálculo de los procesadores, que resultó en un aumento del manejo digital de grandes volúmenes de información. Esto permitió que nuevas técnicas *data-driven* cobren una importancia vital para toda actividad científica y productiva. En este sentido, las técnicas de *Statistical Learning* y *Machine Learning se* han desempeñado muy bien en diversas ramas de la ciencia aplicada e industria (Goodfellow *et al*., 2016) y, en los últimos años, han cobrado una importante relevancia para las geociencias, especialmente, en la combinación de datos provenientes de distintas fuentes.

En particular, los algoritmos basados en árboles decisión como *Random Forest* (RF) (Breiman, 2001) *y Extremely Randomized Trees* (ERT) (Geurts *et al*., 2006), debido a su robustez predictiva y sencilla aplicación resultan muy prácticos para el manejo probabilístico de datos. Estos corresponden a una familia de técnicas de aprendizaje automatizado que puede describirse, en líneas generales, como una sucesión de instancias de decisión.

Adicionalmente, otros métodos como los *Support Vector Machine* (SVM) (Chang y Lin, 2011) poseen una robusta efectividad para realizar predicciones sobre un conjunto de datos tanto en problemas de clasificación como de regresión. El algoritmo *Support Vector Regressor* (SVR) equivalente para regresiones, busca ajustar un hiperplano de $P+1$ dimensiones, donde P es la cantidad de predictores por utilizar en el problema. Una constante $C > 0$ se encarga de controlar el ajuste del estimador y la mayor desviación tolerada.

Desde un punto de vista analítico, la interpretación sísmica de fallas es la determinación de tiempos *TWT*,

espacialmente referenciados, donde se entiende la existencia de una falla en un volumen de dato sísmico. De este modo es posible asumir que estos valores de *TWT* se componen del verdadero tiempo más una desviación:

$$Y=F(X)+e$$

En líneas generales, puede asumirse que *e* se distribuye normalmente con media cero y desviación estándar σ, y en su naturaleza se compone de ruido y desviaciones inducidas por otras fuentes. Modelando espacialmente F(x) mediante las técnicas de regresión mencionadas, el objetivo no es lograr en los modelos resultantes, sino encontrar el punto de equilibrio que permita una buena generalización de estos modelos. En este punto de equilibrio, al comparar las fallas interpretadas con los modelos, sería posible reconocer qué datos se desvían respecto de las tendencias generales de la interpretación, motivo del ruido gaussiano, y qué datos presentan desviaciones de otro tipo.

El objetivo de este trabajo fue encontrar el mejor algoritmo que pueda modelar espacialmente las fallas estudiadas a través de la interpretación sísmica, utilizando técnicas de regresión de aprendizaje automatizado para emplear estos modelos en la realización del control de calidad y densificación de los datos interpretados.

Datos y metodología

Para realizar este trabajo se analizó un set de fallas interpretadas sobre un volumen sísmico correspondiente a yacimientos productivos de la Cuenca del Golfo San Jorge. El set de datos se compuso de 10 fallas normales, 6 de ellas interpretadas regularmente cada 2, 5 y 10 líneas, mientras que las 4 restantes fueron interpretadas en intervalos no regulares.

En primer lugar, se realizó un control de calidad visual cualitativo de los datos interpretados a partir de vistas en planta, en sección y tridimensionales. A continuación, se procedió al análisis numérico o cuantitativo, foco principal de esta contribución.

La estructura de los datos de entrada de los algoritmos de regresión utilizados es un conjunto del tipo $D_f=\{(x_1, y_1),\ldots,(x_N,y_N)\}$ con *N* número de muestras o puntos interpretados para la falla *f*. Cada muestra es un vector de dimensión 1x*P*, donde *P* es el número de predictores por considerar para la regresión, *e* es un valor escalar de tiempo (o profundidad) interpretado para la falla. Para este trabajo, los vectores descriptos se componen como x_i=(Longitud, Latitud), donde ambas variables responden a una representación de coordenadas planas. Mientras que es el tiempo sísmico *two-way-time (TWT)* al cual se interpreta la ocurrencia de la falla *f* para el punto x_i.

Dado que los algoritmos implementados corresponden a la categoría de aprendizaje supervisado, el conjunto de datos se dividió en dos subconjuntos, uno de entrenamiento, D_{Train}, y otro de evaluación o testeo, D_{Test}. El primero de estos responde a los datos utilizados para el entrenamiento o el ajuste de un modelo. Antes de realizar la división se mezclaron los datos, luego, de manera aleatoria, se separaron en los subconjuntos mencionados.

Cabe resaltar el hecho de que la precisión y la exactitud del modelo entrenado se encuentra estrechamente relacionada con la calidad y cantidad de datos utilizados para su entrenamiento. En consecuencia, esta etapa del proceso debe realizarse con la mayor cantidad de muestras posible. Al mismo tiempo, es importante que la evaluación del modelo sea realizada con una cantidad de muestras suficientes como para ser representativas de la distribución general de la variable que se intenta predecir.

Teniendo en cuenta estas consideraciones, la división del conjunto de datos total se realizó respetando una relación porcentual de 80/20 para la cantidad de muestras. Es decir, para cada modelo construido, el 80% de los datos se utilizaron para el entrenamiento y el 20% restante, para la evaluación.

Como métrica de evaluación para los modelos entrenados se utilizó el Error Cuadrático Medio (por sus siglas en inglés *MSE*):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

En donde es el valor predicho por el estimador, o ensamble de ellos, para la muestra perteneciente a D_{Test}.

A partir del conjunto de datos D_{Train} se entrenaron cuatro modelos diferentes, cada uno de ellos basado en algoritmo regresivo distinto. Estos fueron *Single Decision Tree*, *Random Forest, Extremely Randomized Trees* y *Support Vector Machines*. A continuación, haciendo uso de D_{Test}, se evaluaron los resultados obtenidos.

En una primera etapa de pruebas y ajuste de hiperparámetros, se repitió este proceso a razón de 50 corridas por falla con el objetivo de reconocer las tendencias generales en las respuestas de los modelos entrenados. Como consecuencia del carácter aleatorio del muestreo de los conjuntos de entrenamiento y evaluación, es esperable que cada modelo prediga diferentes tiempos *TWT* para un mismo par coordenado , dando lugar a diferentes errores de evaluación.

Dada la naturaleza del problema por resolver, la cantidad de muestras con las que se cuenta es baja. Por esta razón, debido al funcionamiento interno de los algoritmos utilizados, resulta probable caer en un caso de *overfitting* (*Low Bias-High Variance*). Con el fin de reducir estos efectos en los modelos entrenados, se utilizaron métodos de ensamble como *Bostrapping Aggregating* (*Bagging*). Una discusión con mayor profundidad sobre este tema excede los propósitos de este trabajo. En este sentido, y si se desean mayores referencias, se recomienda consultar la propuesta de Belkin *et al*. (2019).

Una vez determinados los parámetros más relevantes de entrenamiento, para independizar este análisis de la varianza inducida por la división aleatoria de los conjuntos de datos mencionados, se realizaron pruebas de 100 y 1000 corridas por falla, para cada modelo. Luego, se promediaron los errores de cada modelo sobre las corridas realizadas, para obtener el valor esperado del error de prueba.

Por último, fue posible realizar una densificación de las fallas, interpolando los datos validados a partir de los modelos mencionados. Una dificultad adicional radicó en el hecho de que las interpretaciones no se realizaron

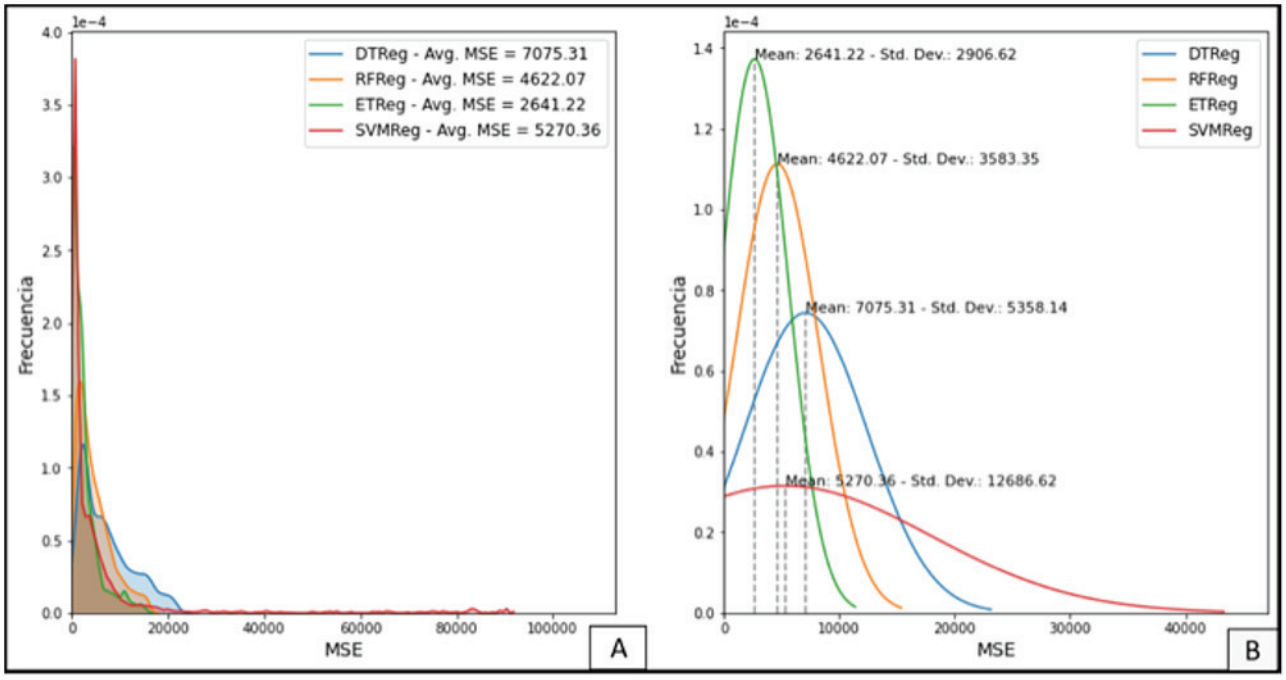


Figura 1. **Las curvas azules corresponden al regresor Single Decision Trees; las amarillas, a los *Random Forest*; las verdes, a los Extremely Randomized Trees y las rojas, a los Support Vector Machines. A. Distribución del MSE para cada uno de los regresores utilizados a lo largo de 1000 corridas. B. Representación de los resultados obtenidos como distribuciones normales.**

sobre una distribución regular de puntos en el espacio. A pesar de que una de las coordenadas suele mostrar un incremento regular, a medida que se visualiza línea a línea un volumen sísmico, la otra coordenada no respondía a un muestreo regular, sino que quedaba supeditada a la interpretación del geocientista.

Resultados y discusión

Se realizó una evaluación de 1000 corridas para cada algoritmo de regresión sobre cada falla. Para cada una se evaluó el error cuadrático medio *(MSE)*. Finalmente, se analizó la frecuencia de ocurrencia de cada uno de estos errores para cada uno de los tipos de regresores utilizados (Figura 1.A).

A partir de los resultados obtenidos, se calcularon el primer y segundo momento estadístico de cada una de estas distribuciones. Con estos valores, se representaron las distribuciones obtenidas como distribuciones normales (Figura 1.B).

En la figura 1 se puede observar que, según los momentos estadísticos mencionados, los árboles de decisión simples pueden ser considerados como la forma más sencilla de modelar las fallas interpretadas. Esto, sin tener aplicación directa sobre el control de calidad de los datos, puede ser tomado como punto de partida para enriquecer los modelos.

Para el caso conjunto de las técnicas de *Random Forest* y *Extremely Randomized Trees*, se observó una marcada disminución en la media y desviación estándar del *MSE*. Esto queda reflejado con claridad en la reproducción de los resultados a través de distribuciones normales, donde las curvas para este tipo de modelos son más angostas y cercanas al origen, respectivamente. Esto podría deberse a la aplicación de la técnica de *bagging* por parte de estos algoritmos. La misma actúa como regularizador de los modelos entrenados previo al ensamble, logrando así una mejor generalización del modelo final.

Además, pudo corroborarse visualmente el acuerdo entre las predicciones realizadas por estos modelos y el dato interpretado (Figura 2).

En cuanto al rendimiento de los modelos de *Support Vector Machines*, al analizar individualmente cada una de las fallas modeladas, se observó una dependencia muy marcada respecto de la cantidad de muestras disponibles y su distribución espacial. Para los mejores modelos construidos, en términos del *MSE*, es posible destacar la correspondencia con las fallas que presentaban una mayor cantidad de puntos interpretados y una distribución espacial aproximadamente homogénea. En contraposición, las fallas cuya interpretación constaba de unos pocos puntos, los modelos de *SVM* resultantes mostraron una gran dispersión y poca fidelidad a los rasgos que se intentaban replicar.

El marcado contraste entre los resultados obtenidos para los últimos modelos explica el gran ancho de la campana correspondiente en su representación bajo una distribución normal.

Conclusiones

En base a la teoría de elección de modelos estadísticos propuesta por Hastie *et al.* (2009), los algoritmos implementados han sido exitosos tanto desde el punto de vista numérico como visual. Paralelamente, puede verse una muy buena generalización de las superficies de falla generadas a partir de los modelos regresivos entrenados.

A pesar de que la regresión espacial es un proceso capaz de devolver resultados en principio no confirmados,

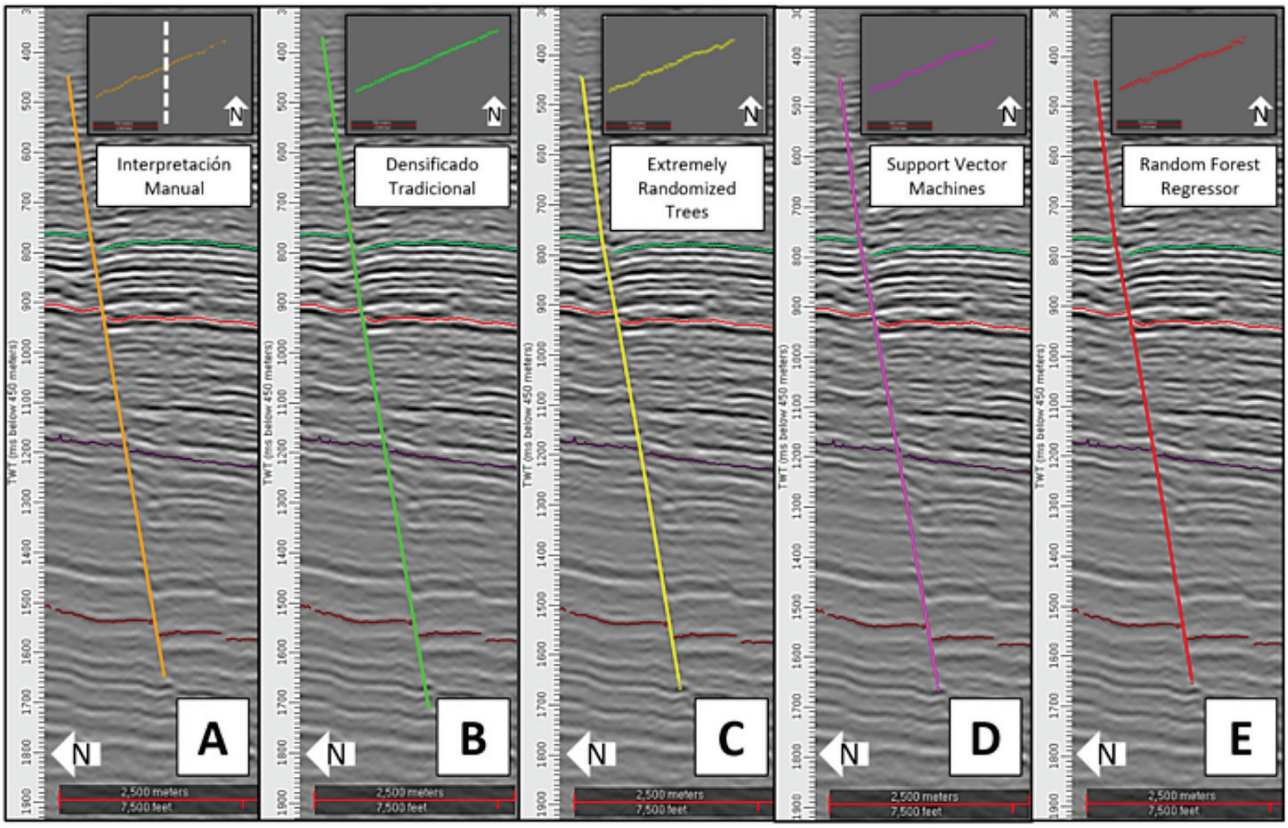


Figura 2. **Se observa una sección sísmica repetida con la falla interpretada y los respectivos modelos entrenados, a través de distintos algoritmos regresivos, para esa falla. En la parte superior de cada una se observan mapas con su distribución areal. A. La falla interpretada. B, C, D y E. Distintos modelos regresivos entrenados.**

los modelos construidos para las distintas fallas estudiadas en este trabajo respetan las tendencias de las interpretaciones originales. Esto también se observó en zonas donde la interpretación debía ser desviada respecto de las tendencias generales.

En este sentido, se recomienda una examinación y un repaso visual de los datos generados, para verificar rápidamente la coherencia y la consistencia del modelo de reproducción del comportamiento de las fallas elegido, particularmente en los sectores donde la falla no ha sido interpretada. Esto resulta fundamental, independientemente de la capacidad de predicción de los algoritmos implementados.

A partir de la comparación entre el dato interpretado y el modelo que mejor generalice las tendencias de la falla estudiada, resulta sencillo reconocer los puntos interpretados cuya desviación no sea despreciable. Una vez localizados los puntos anómalos, es posible proceder a una revisión más detallada del motivo.

Además, también fue posible comparar dichos resultados ante los obtenidos a partir de los métodos geoestadísticos de interpolación tradicionales.

Sobre la base de estos resultados, se concluye que los modelos son comparables a los primeros y pueden ser utilizados en su lugar.

Para finalizar, remarcamos el valor agregado de la metodología propuesta como parte del flujo de trabajo de interpretación, la herramienta resulta de utilidad al momento de interpolar datos irregularmente distribuidos en el espacio. Esto permite incrementar la densidad de interpretación de las fallas estudiadas dentro de las áreas de interés y brinda mayor robustez a los modelos estáticos construidos.

Agradecimientos

Agradecemos a la vicepresidencia de Desarrollo de Reservas de Pan American Energy Group por autorizar el uso de los datos mencionados en este trabajo.

Referencias bibliográficas

Belkin, M., Hsu, D., Ma, S. y Mandal, S. (2019). Reconciling modern machine-learning practice and the classical biasvariance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15.849-15.854.

Breiman, L. (2001). *Random forests*. *Machine Learning*, *45*(1), 5-32.

Burrough, P. A. y McDonnell, R. A. (1998). *Principles of geographical information systems: Spatial information systems and geostatistics*. Oxford University Press.

Chang, C. C. y Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM.

Transactions on intelligent systems and technology (TIST), *2*(3), 1-27.

Cortes, C. y Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Geurts, P., Ernst, D. y Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*,*63*(1), 3-42.

Goodfellow, I., Bengio, Y. y Courville, A. (2016). *Deep learning*. MIT press.

Hastie, T., Tibshirani, R. y Friedman, J. (2009). The elements of statistical learning. *Springer*. 757.