MONITOREO ALTERNATIVO DE PRODUCCIÓN DE GAS CON TÉCNICAS DE *MACHINE LEARNING*

Axel Canatelli[1], Godefroy Dupuis[1], Lorenzo Parigini[1]

1: TotalEnergies. axel.canatelli@totalenergies.com, godefroy.dupuis@totalenergies.com, lorenzo.parigini@totalenergies.com

Palabras clave: Monitoring, *Machine Learning*, Aguada Pichana Este, Vaca Muerta

ABSTRACT

Fluid flow quantification is crucial for an accurate determination of individual well productivity. In blocks where production is mostly based on dry/wet gas, this information is obtained by dedicated flow quantity indicators (FQI) measurements. As an alternative to FQI measurements to estimate fluid flow, a common mathematical expression is the ***choke*** formula (CF). However, this formula is not always usable due to the assumption of a monophasic fluid flow (only gas). As a matter of fact, from own experience, for shale gas wells ***choke*** formula seems to perform typically very well in decline stages of production but poorly in ramp up stages where water production is not negligible.

As part of monitoring procedures during the productive time of a well, many other variables are measured with an hourly frequency, such as: ***choke***, pressures after and before ***choke***, and temperatures after and before ***choke***. This available information is part of all wells' production history. The objective of the present work was to make use of other wells' production history data for gas flow estimation in horizontal wells in order to avoid redundant FQI measurements and achieve cost reduction. For that, a *Machine Learning* (ML) based model was proposed, which was able to capture decline trends from historical production data. ML model's performance was compared to CF estimations as a proxy for acceptable accuracy levels according to the Oil & Gas industry standards.

The results obtained with ML model proved to estimate individual gas flow in a wide range of wells with an acceptable performance that, in 82% of total cases, was more accurate than ***choke*** formula based on R2 scores. In positive prediction cases, accumulated gas estimation with ML model showed a lower error than CF estimation, with errors being below the 5% tolerance margin. These results suggest an alternative way to exploit monitoring data to estimate gas flow production in gas fields.

Achieved results demonstrate the high prediction potential of AI algorithms when combined with proper use of data. Further efforts may lead to a more robust model for a larger number of wells within the same fluid window.

INTRODUCCIÓN

Los proyectos de explotación de yacimientos no convencionales (NOC) implican la constante perforación, estimulación hidráulica y finalmente puesta en producción del pozo en cuestión. Dado que naturalmente estos reservorios se depletan con mayor rapidez, el

nivel de actividad asociado a la puesta en producción de nuevos pozos debe ser tal que mantenga los niveles de producción establecidos en el proyecto de desarrollo. Al poner un pozo en producción, se instalan en superficie diversos sensores con la finalidad de monitorear con determinada frecuencia un cierto número de variables que indican el estado de operación del mismo. De esta manera, es factible poder realizar un seguimiento del desempeño del pozo y con ello detectar desviaciones de la productividad con respecto a lo estipulado en el plan de desarrollo. Tanto la frecuencia de medición como el número y tipo de variables a medir dependen del caso en cuestión, pero estas no suelen variar dentro de un mismo proyecto.

En bloques donde la producción se basa mayoritariamente en gas seco/húmedo, la medición de caudales de producción se logra mediante *Flow Quantity Indicators* (FQI). Particularmente en Aguada Pichana, como parte del monitoreo de los pozos, las siguientes variables se miden constantemente y con una frecuencia horaria:

- Presión de cabeza de pozo antes del *choke* (WHP)
- Presión de cabeza de pozo después del *choke* (WHPAO)
- Temperatura antes del *choke* (WHT)
- Temperatura después del *choke* (WHTAO)

Las mediciones de FQI implican un costo adicional como parte del CAPEX/OPEX que podría reducirse si se aplicara otra metodología para la determinación del flujo de gas. En este sentido, la única herramienta matemática aplicable hasta el momento es la fórmula del *choke*, que no siempre se puede utilizar debido a la suposición de un fluido monofásico (solo gas). En base a propia experiencia, la fórmula del *choke* suele funcionar muy bien en las etapas de declinación para la mayoría de los casos, pero no así en las etapas de apertura del pozo donde la producción de agua no es despreciable.

Las herramientas de aprendizaje automático o *Machine Learning* (ML), como una subclase de algoritmos de inteligencia artificial (IA), han demostrado ser útiles en muchas áreas para las que se dispone de datos relacionados con el historial de un proceso (Hegde *et al.*, 2015). Un mayor conocimiento sobre las aplicaciones de estas técnicas al historial de producción de los pozos podría eventualmente simplificar la implementación de elementos de medición, contribuyendo a la optimización del desarrollo y la reducción de costos (Liao *et al.*, 2020).

El objetivo del presente trabajo es hacer uso de datos de producción de pozos para la predicción de caudales de gas mediante la implementación de algoritmos de IA para con ello evitar mediciones redundantes con FQI y lograr una reducción de costos.

METODOLOGÍA

ADQUISICIÓN DE DATOS

Los datos de monitoreo de pozos fueron exportados desde una base de datos y luego tratados en Python™ con la librería Pandas. En la Figura 1 se indica el aspecto de los datos cargados. En el recuadro azul se encuentran las variables de entrada previamente descriptas (WHP, WHT, *Choke*, WHPAO, WHTAO) y en el recuadro verde el caudal de gas que es la variable de salida o a estimar por el modelo (QG). QW y QO hacen referencia a caudales de agua y petróleo respectivamente, los cuales no se emplearon en el presente trabajo. La variable DATE (fecha del dato) solo se consideró a los efectos de graficar los resultados, manteniendo la correspondencia entre datos de entrada y salida, no como una variable dentro del modelo.

	DATE	WHP	WHT	CHOKE	WHPAO	WHTAO	QG	QW	QO
0	14-08-18 13:00	0.6875	9.7500	2.4	34.032585	13.590622	0.000000	0.00000	0.0
1	14-08-18 14:00	0.6875	9.7500	2.4	35.554649	23.703588	0.000000	0.00000	0.0
2	14-08-18 15:00	244.0625	21.4375	2.4	35.556419	28.096045	19.631997	68.71199	0.0
3	14-08-18 16:00	242.6875	26.1250	2.4	35.417034	29.920633	19.521394	68.32488	0.0
4	14-08-18 17:00	244.0625	28.2500	2.4	35.383404	31.157745	19.631997	68.71199	0.0

INPUT OUTPUT

Figura 1. Datos horarios de monitoring exportados y tratados en Python™ con la librería Pandas. En el recuadro azul se indican las variables de entrada del modelo y en verde la variable de salida o a predecir con el modelo. QW y QO hacen referencia a caudales de agua y petróleo respectivamente, los cuales no se emplearon en el presente trabajo.

FÓRMULA DEL *CHOKE*

En pozos de gas, bajo condición de fluido monofásico, se puede aplicar la fórmula del *choke* (FC). Su expresión matemática toma la forma:

$$Qg[m^3/day] = WHP[psi]*CHOKE\ [mm]^2.$$

Considerando la buena capacidad de estimación de caudales de gas con la fórmula del *choke*, especialmente en las etapas de declinación de un pozo, se incluyó en el trabajo a fin de compararla con el desempeño de los modelos con ML. Estimaciones de caudales con ML comparables con FC indicaría un grado aceptable de desempeño del modelo.

ACUMULADA DE GAS

La acumulada de gas (o gas producido) fue calculado de tres fuentes distintas:

1 - Datos de monitoreo: Caudal de gas medido a partir de datos de monitoreo y luego integrados para calcular su acumulada.
2 - Fórmula del *choke* (FC): Se aplicó la fórmula del *choke* para estimar el caudal de gas (Qg_{FC}) y luego se integró para calcular el gas acumulado (Gp_{FC}).
3 - Modelo con ML: Se aplicó el modelo ML para obtener una estimación del caudal de gas (Qg_{ML}) y luego se integró para calcular el gas acumulado (Gp_{ML}).

MÉTRICAS

El desempeño del modelo alcanzado se puede medir en términos de una gran cantidad de métricas o también denominadas scores. Antes de especificar las métricas empleadas, es necesario introducir el concepto de residuos. Residuo (e) es la diferencia entre un valor específico (dato medido) y su predicción dada por un modelo, la cual puede ser positiva o negativa. En base a los residuos, los modelos para estimar los caudales de gas se evaluaron mediante las siguientes métricas:

- Coeficiente de determinación (R2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

donde e_i es el residuo del modelo, y_i es la variable a predecir (dato medido), $\bar{y}$ es el valor medio del dato a predecir y f_i es la predicción lograda por el modelo.

- Error medio absoluto (MSE por sus siglas en inglés):

$$MSE = \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}$$

Por lo general, el R2 fue suficiente para establecer el desempeño del modelo, en pocos casos no tan claros, el desempeño se determinó a su vez con el MSE. Estas métricas se aplicaron a la estimación de caudales tanto con FC como con ML a los fines de comparar y determinar el mejor estimador de caudales.

La métrica empleada para la estimación de la acumulada de gas fue:

- Error de la acumulada porcentual (PAE [%] por sus siglas en inglés):

$$PAE = 100\% * \frac{\left|G_p - \widehat{G_p}\right|}{G_p},$$

donde G_p es el gas acumulado a partir de los datos de monitoreo y $\widehat{G_p}$ es el gas acumulado a partir de la estimación de caudales, tanto con FC (Gp_{FC}) como con ML (Gp_{ML}).

Debido a la potencial aplicación práctica de este modelo, se estableció un margen de tolerancia del 5% para el error de la acumulada porcentual (PAE).

CICLO DE TRABAJO

Los modelos con ML requieren atravesar ciertas etapas las cuales no siempre suelen seguir una metodología concreta. En la Figura 2 se esquematizan simplificadamente las distintas etapas en la creación de un modelo de ML, a modo de orientación. Sin embargo, vale aclarar que en ciertos casos es posible que se deba volver a una etapa previa, convirtiéndose en un proceso iterativo.

En la primera etapa se encuentra el preprocesamiento y *scaling* de los datos, una tarea de gran relevancia para el correcto entrenamiento y desempeño de los modelos (Lee *et al.*, 2019). El preprocesamiento consiste en realizar un control de calidad de los datos y un análisis descriptivo de los mismos a fin de identificar variabilidad, presencia de outliers, entre otros. El *scaling* se refiere a la transformación de las dimensiones y/o rangos de datos de las variables, de forma tal que el rango de variación de las distintas variables sea comparable. Se encuentran varios criterios para llevar adelante esta tarea y el método a emplear puede depender de la naturaleza del dato como del modelo a emplear. Los métodos más usuales suelen ser la estandarización y la normalización.

En la segunda etapa se encuentra la selección del modelo, los cuales pueden ser: Decision tree, Random Forest, Gradient Boosting, Support Vector Machine, entre otros. Para esta instancia es necesaria la realización de pruebas con diversos modelos en base a sus características y probabilidad de éxito. Como cada modelo posee su metodología, es posible que para su implementación requiera un tipo de *scaling* diferente acorde al modelo. Luego de la realización de una comparación de rendimiento entre los diversos modelos, se elige aquel que indique el mejor desempeño.

La tercera etapa se trata del ajuste de los hiperparámetros del modelo elegido. Cada uno de los posibles modelos a seleccionar en la segunda etapa conllevan una cierta cantidad de parámetros que definen su propio funcionamiento. Estos parámetros deben adecuarse al tipo de problema que se quiere resolver, por ende, es de esperar que cada dúo problema-modelo tenga una configuración óptima de hiperparámetros. Alcanzar dicha configuración no es un procedimiento directo debido a la complejidad asociada, para intentar obtener la mejor aproximación se proponen varias metodologías como pueden ser: manual search, grid search, random search, optimización bayesiana, entre otros. En la Figura 2 se indican parte de los hiperparámetros que se pueden modificar para el modelo Random Forest.

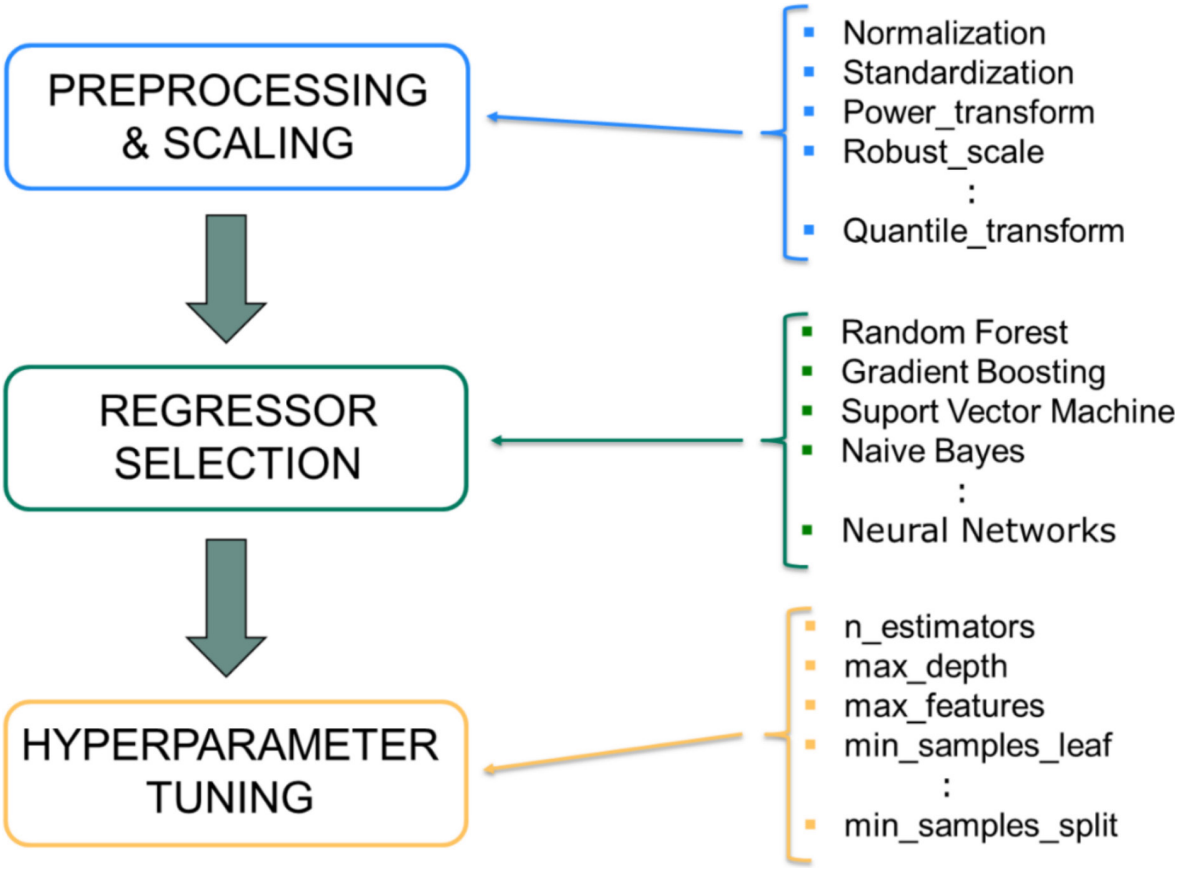

Figura 2. Diferentes etapas en la creación de un modelo con ML.

PROCEDIMIENTO

Como parte del preprocesamiento, se eliminaron valores incorrectos o atípicos para cada variable, como pueden ser flujos o presiones negativas y valores anormalmente altos correspondientes a errores de medición. También se eliminaron los datos asociados con el cierre de pozos a largo plazo, ya que implican datos no representativos del tipo de condición que se quiere modelar. El *scaling* se basó en una normalización, teniendo en cuenta el rango de cada una de las variables para la población de pozos considerados.

El modelo seleccionado fue Random Forest, un caso particular de los modelos de árboles de decisión. La elección del mismo se sustentó en su popularidad y aplicabilidad probada en diferentes campos (Nasir y Rickabaugh, 2018) con buena tasa de éxito y a su vez propias experiencias positivas con aplicaciones a problemas similares. El conjunto de datos de entrenamiento se dividió en 80-20 (como procedimiento estándar) de manera aleatoria, donde el 80 % de los datos se utilizó para el entrenamiento y el 20 % restante para la validación. El desempeño del modelo se estimó en base al Coeficiente de Determinación (R^2) y el Error Cuadrático Medio (MSE). Como R^2

y MSE presentaron resultados compatibles, generalmente solo se usó R^2 para la interpretación del desempeño; salvo en unos pocos casos en los que se necesitó un análisis más detallado, el desempeño se definió con el MSE.

La predicción siempre se realizó en un solo pozo, mientras que el entrenamiento se realizó con un único pozo o múltiples teniendo en cuenta que presentaran una tendencia de producción similar (misma fase de desarrollo), como se ha recomendado en trabajos previos (Gupta *et al.*, 2014). Los pozos con problemas operativos se descartaron para evitar entrenar un modelo con una tendencia de producción atípica. Los tiempos de producción para el entrenamiento fueron variables en función de la antigüedad del pozo a emplear, con casos de solo meses para pozos en el ciclo de apertura a incluso años para pozos en etapa de declinación.

En la instancia de ajuste de hiperparámetros se optó por una búsqueda manual mediante rangos para cada tipo de hiperparámetro en función de experiencias previas y el modelo en cuestión.

MODELO

Los modelos de árboles de decisión básicamente clasifican la información mediante sucesivos cortes de las variables de entrada, que generan una cierta cantidad de nodos (bifurcaciones) y hojas (nodos terminales), hasta lograr una división óptima dado por una función objetivo. La extensión y morfología de este árbol son función de los hiperparámetros seleccionados para el modelo. En la Figura 3 se indican dos ejemplos de árboles de decisión, uno simple a la izquierda y otro más extenso y complejo a la derecha. Múltiples versiones de este método fueron introducidas con posterioridad, de los cuales se quiere destacar Random Forest debido a su corto tiempo de procesamiento y gran poder predictivo. En este trabajo se implementó una versión en Python™ cuyo nombre específico es *Random Forest Regressor* y pertenece a una librería llamada *Scikit-learn*. Random Forest hace uso de una amplia cantidad de árboles de decisión y su resultado final es un promedio de los resultados obtenidos de todos los árboles generados previamente.

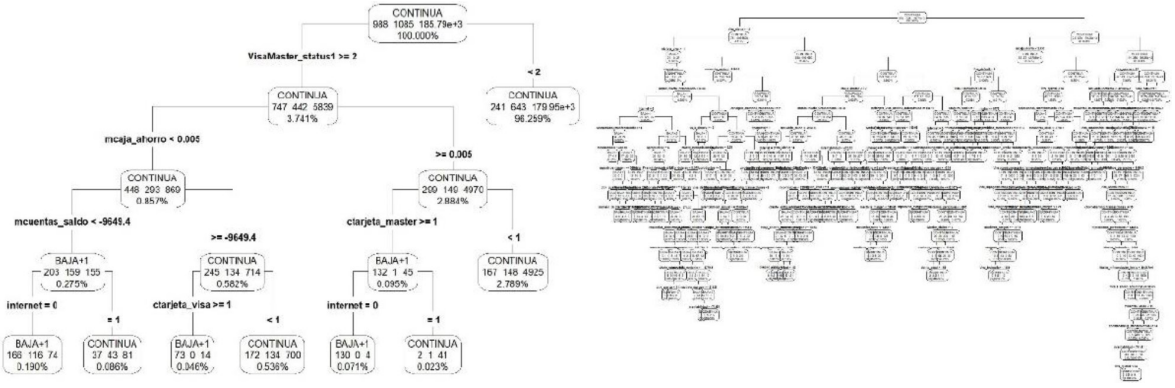

Figura 3. Ejemplos de árboles de decisión: uno simple a la izquierda y uno más extenso y complejo a la derecha.

PRUEBAS DE CONCEPTO

Con el fin de evaluar la potencial aplicación del modelo con ML, se realizaron pruebas de concepto que analizaron los siguientes aspectos:

A) GRADO DE PREDICTIBILIDAD DEL MODELO

Se aisló un grupo de 11 pozos de una misma fase de desarrollo de forma tal de estimar la producción de cada uno de ellos por separado, entrenando el modelo en cada caso con los 10 pozos restantes/complementarios.

B) POSIBLE CALIBRACIÓN A FUTURO

Se generó una condición de mal desempeño del modelo mediante la aplicación del mismo en una situación para la que no fue entrenado. Esto se provocó entrenando el modelo con datos de producción correspondientes a una franja reducida del tiempo de producción. Tras verificar el mal desempeño del modelo, se agregaron al conjunto de datos de entrenamiento, unos pocos datos de producción equivalentes a dos mediciones puntuales de calibración. Una vez reentrenado el modelo con la incorporación de los nuevos datos, se analizaron las mejoras en la estimación de los caudales de gas con el modelo de ML.

RESULTADOS

Los resultados obtenidos con el modelo ML se muestran con una gráfica para el caudal de gas vs. tiempo y una segunda gráfica para el gas acumulado vs. tiempo. Cada gráfico contiene 3 curvas, en azul datos de monitoreo (medición FQI), en verde estimación con la fórmula del *choke* (FC) y en naranja estimación con ML (ML).

En la Figura 4 se muestra la estimación del caudal de gas vs. tiempo, donde el R^2 de ML es $R^2 = 0.968$. El modelo ML se entrenó con un pozo de la misma fase de desarrollo sin problemas operativos. La predicción de ML sigue la tendencia del pozo y su R^2 es más alto que el R^2 de FC (0,968 frente a 0,888). En la Figura 5, se muestra el gas acumulado vs. tiempo para el mismo pozo de la Figura 4. El PAE de ML está por debajo del margen de tolerancia del 5%, mientras que el PAE de FC no lo está (3,3% frente a 10,4%).

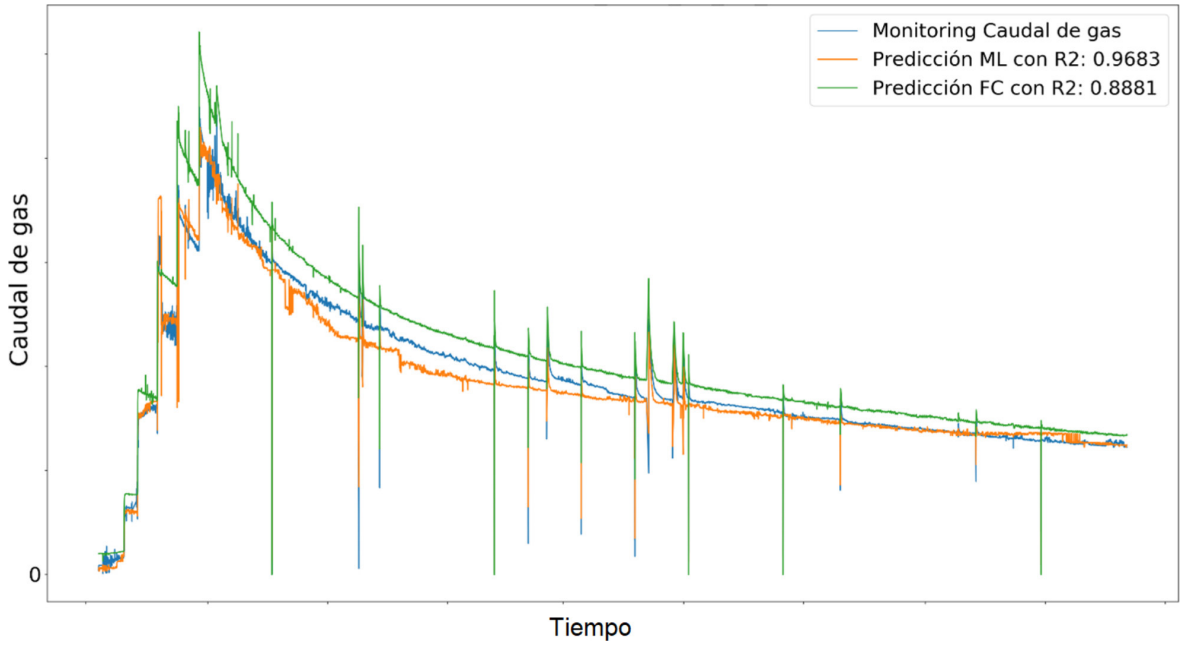

Figura 4. Caudal de gas versus tiempo. El modelo ML se entrenó con un pozo de la misma fase de desarrollo sin problemas operativos. En azul, datos de monitoring (reales de producción), en verde la fórmula del *choke* (FC) y en naranja el modelo con ML (ML). Los valores de R2 se indican para ambas estimaciones de los caudales de gas.

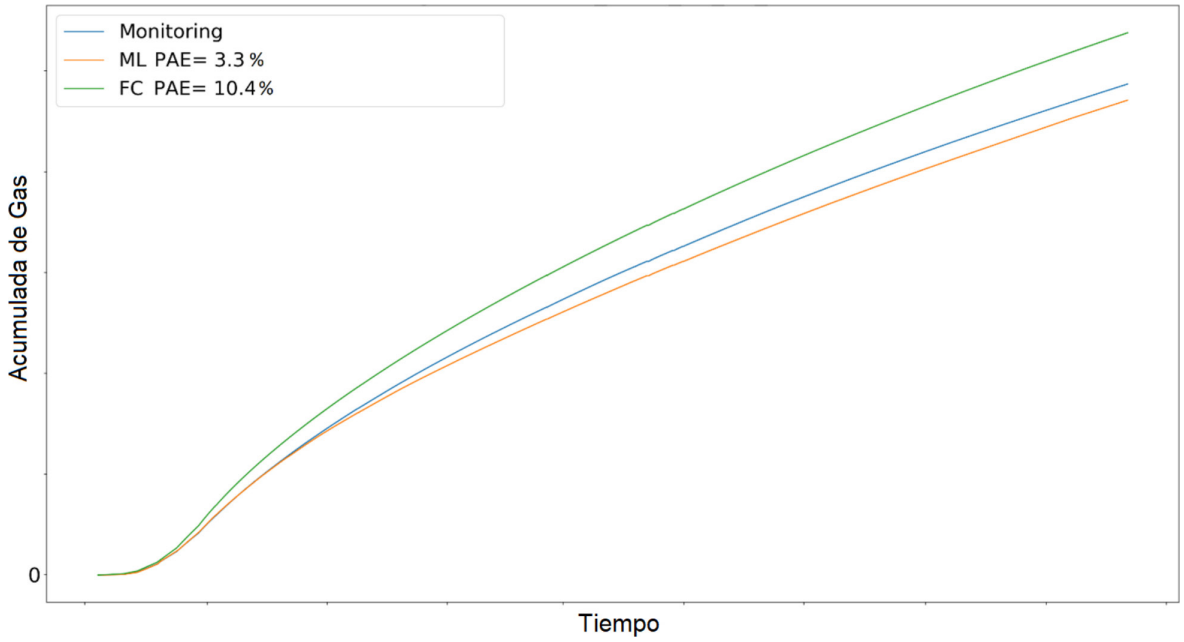

Figura 5. Acumulada de gas versus tiempo. El modelo ML se entrenó con un pozo de la misma fase de desarrollo sin problemas operativos. En azul datos de monitoring (reales de producción), en verde la fórmula del *choke* (FC) y en naranja el modelo con ML (ML). Los valores de PAE (Error de la Acumulada Porcentual) se indican para ambas estimaciones de los caudales de gas.

Tras la evaluación en el total de pozos considerados, el modelo ML presentó puntajes altos de rendimiento de predicción en el 68% de los casos. Los casos de bajo rendimiento se asociaron con pozos que registraron algún tipo de problema operativo, como *frac-hits*, restricciones de *casing*, pescas, falta de *coil tubing* (CT) *wash* o *side track*.

Las pruebas de concepto evaluaron diferentes aspectos del modelo logrado con ML y ayudó a comprender su campo de aplicabilidad:

A) GRADO DE PREDICTIBILIDAD DEL MODELO

Las pruebas relacionadas con el grado de predictibilidad indicaron que el modelo con ML debería predecir correctamente un nuevo pozo con una tendencia de producción similar como han indicado y recomendado otros autores (Zhan *et al.*, 2019), lo cual generalmente se asocia a una arquitectura similar y no ocurrencia de problemas operativos. Las evaluaciones en los 11 pozos seleccionados indicaron que en un 82% de los casos el modelo con ML tuvo mejor desempeño que FC, y en un 73% de los casos el PAE estuvo por debajo del 5% de margen de tolerancia. En la Figura 6 se indica uno de los 11 casos analizados mediante un gráfico de caudales de gas vs. tiempo, análogo al de la Figura 4. En la Figura 7 se indica la acumulada de gas a partir de ML y FC, para el mismo caso de la Figura 6, mediante un gráfico de gas acumulado vs. tiempo análogo al de la Figura 5.

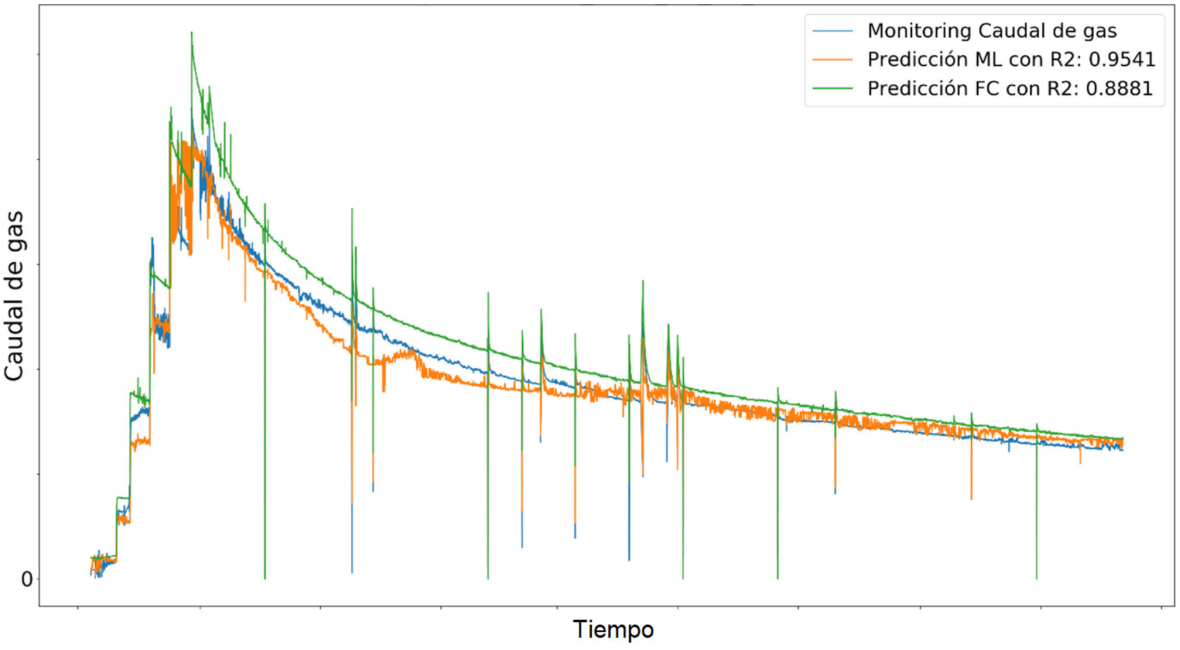

Figura 6. Caudal de gas versus tiempo. El modelo ML se entrenó con 10 pozos de la misma fase de desarrollo sin problemas operativos (entrenamiento con varios pozos). En azul datos de monitoring (reales de producción), en verde la fórmula del *choke* (FC) y en naranja el modelo con ML (ML). Los valores de R2 se indican para ambas estimaciones de los caudales de gas.

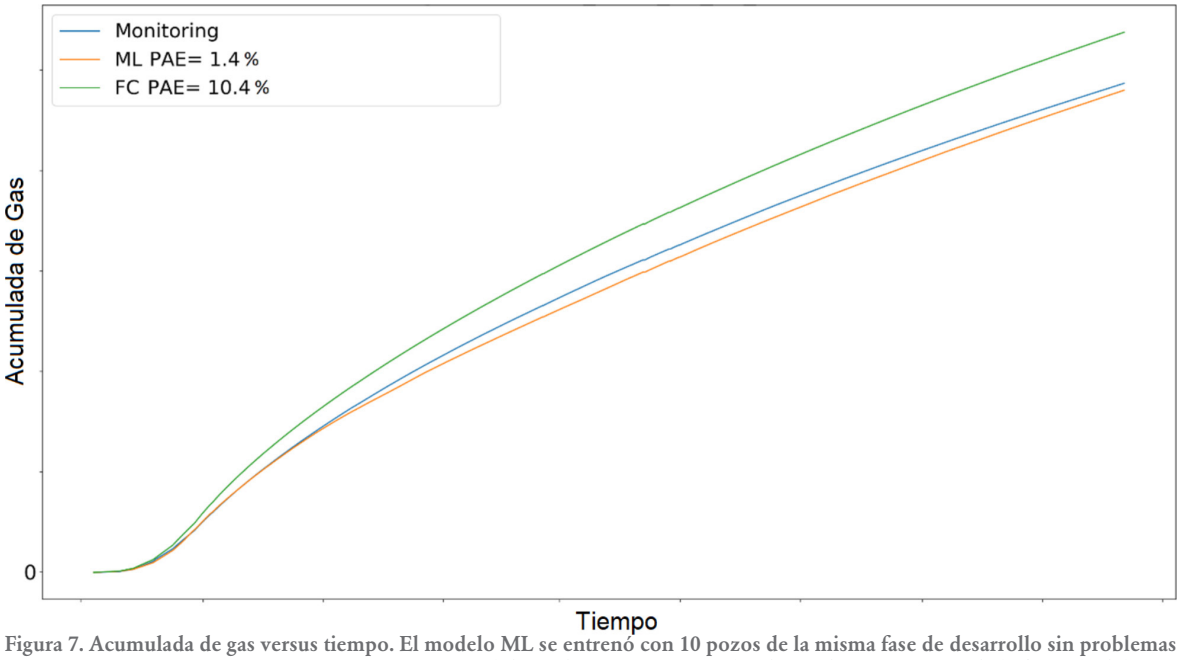

Figura 7. Acumulada de gas versus tiempo. El modelo ML se entrenó con 10 pozos de la misma fase de desarrollo sin problemas operativos (entrenamiento con varios pozos). En azul datos de monitoring (reales de producción), en verde la fórmula del *choke* (FC) y en naranja el modelo con ML (ML). Los valores de PAE (Error de la Acumulada Porcentual) se indican para ambas estimaciones de los caudales de gas.

B) POSIBLE CALIBRACIÓN A FUTURO

Las pruebas relacionadas con la calibración demostraron que, en caso de necesidad, el modelo ML podría corregirse con futuras mediciones de calibración.

En la Figura 8 se muestra el desacuerdo (remarcado con una línea punteada de color rojo) del modelo con ML generado adrede mediante el entrenamiento con datos de producción correspondiente a la primera mitad de la historia de producción de un pozo (indicado con la flecha en color rojo).

En la Figura 9 se puede observar como el modelo recobra la tendencia de producción luego de haber añadido dentro del conjunto de datos de entrenamiento, datos equivalentes a dos mediciones puntuales durante uno o dos días y especiadas tres meses entre sí (círculos en color rojo). Tras este ajuste, el modelo con ML logra un desempeño aceptable dado por un R^2 de 0,942 y sigue la tendencia de producción del pozo.

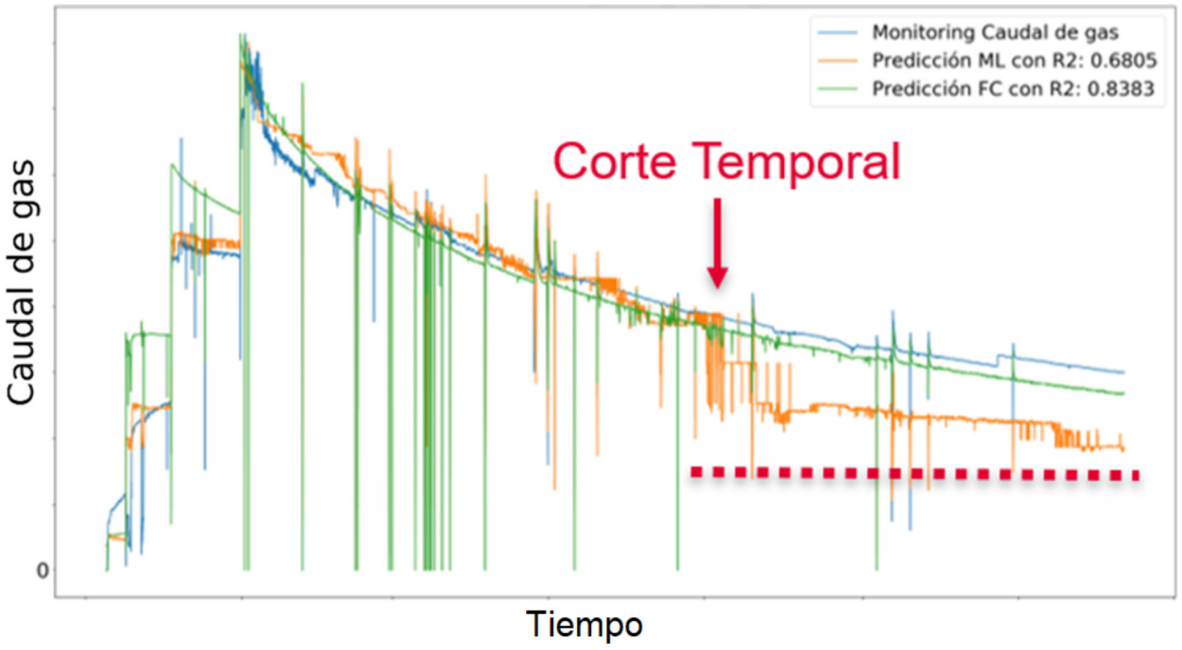

Figura 8. Caudal de gas versus tiempo. En azul datos de monitoring (reales de producción), en verde la fórmula del *choke* (FC) y en naranja el modelo con ML (ML). Los valores de R^2 se indican para ambas estimaciones de los caudales de gas. Para entrenar el modelo, se empleó un pozo de la misma fase de desarrollo sin problemas operativos. En este caso los datos de entrenamiento corresponden a aproximadamente la primera mitad de la historia de producción, como lo indica la fecha en color rojo, a partir de la cual se puede apreciar el desacuerdo del modelo por intentar estimar caudales en instancias para la cual no fue entrenado (remarcado con una línea punteada de color rojo).

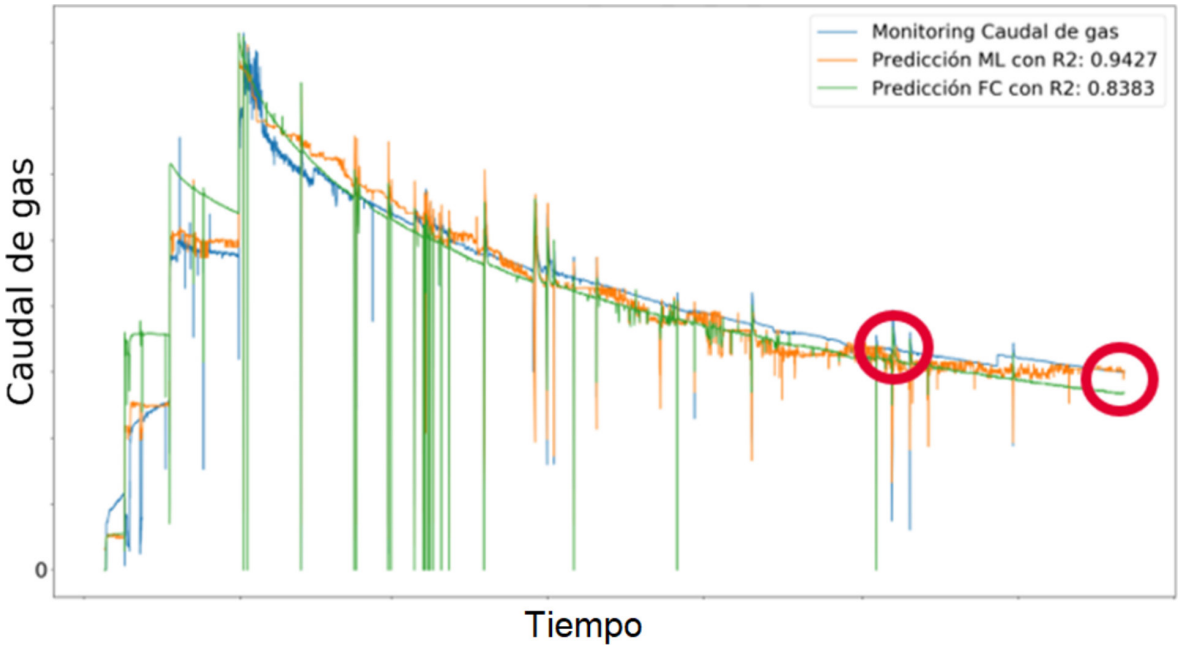

Figura 9. Caudal de gas versus tiempo. En azul datos de monitoring (reales de producción), en verde la fórmula del *choke* (FC) y en naranja el modelo con ML (ML). Los valores de R^2 se indican para ambas estimaciones de los caudales de gas. Para entrenar el modelo, se empleó un pozo de la misma fase de desarrollo sin problemas operativos. Los datos de entrenamiento corresponden al caso de la Figura 8 junto con la adición de datos equivalentes a dos mediciones puntuales cada tres meses durante uno o dos días, indicadas mediante los círculos en color rojo. A diferencia de la Figura 8, en este caso el modelo con ML logra un desempeño aceptable y sigue la tendencia de producción del pozo.

CONCLUSIONES

Se construyó un modelo ML para estimar el caudal de gas producido en pozos horizontales. El modelo ML ha demostrado estimar el caudal de gas en una amplia gama de pozos de forma satisfactoria, con un coeficiente de determinación (R^2) aceptable, que en la mayoría de los casos fue mayor que el obtenido con FC. En los casos de estimación exitosa, el gas acumulado derivado del modelo ML arrojó valores de PAE más bajos que aquellos logrados con FC. Además, en la mayoría de los casos, los valores de PAE estuvieron por debajo del margen de tolerancia del 5%.

Las pruebas de concepto arrojaron resultados positivos en cuanto al grado de predictibilidad al aplicar el modelo en un pozo aleatorio que posea una tendencia de producción similar. En cuanto a la posibilidad de poder calibrar a futuro el modelo con ML una vez puesto en funcionamiento, el mismo podría ser calibrado mediante la incorporación de nuevos datos dentro del conjunto de entrenamiento. Los datos a incorporar podrían ser mediciones puntuales de uno o dos días de duración cada tres meses a fin de que el modelo vuelva a obtener un desempeño aceptable.

Estos resultados sugieren una posible forma de explotar los datos de monitoreo para estimar los caudales de producción en los campos de gas. A su vez, permiten creer en el alto potencial de predicción de los algoritmos de IA cuando se combinan con el uso adecuado de los datos. En este sentido, mayores esfuerzos pueden conducir a un modelo más robusto para todos los pozos con el mismo tipo de fluido. Los obstáculos encontrados y las lecciones aprendidas durante este trabajo han ayudado a definir el siguiente camino y metodología para afrontar este desafío. Mientras tanto, el modelo logrado con ML podría aplicarse en campo para comprobar su posible uso en los casos de éxito ya determinados en el presente trabajo. Este último punto podría validar el modelo ML y reducir costos al evitar mediciones de monitoreo redundantes.

AGRADECIMIENTOS

Los autores agradecen a TotalEnergies y sus socios en el bloque: PanAmerican Energy, YPF y Wintershall DEA, por la autorización para presentar este trabajo.

REFERENCIAS

Gupta S., Fuehrer F. y Chelinsky Jeyachandra B., 2014. "Production Forecasting in Unconventional Resources using Data Mining and Time Series Analysis". Society of Petroleum Engineers.

Hegde, C., Wallace, S. y Gray, K., 2015. Using Trees, Bagging, and Random Forests to Predict Rate of Penetration During Drilling. Society of Petroleum Engineers.

Lee, K., Lim, J., Yoon, D. y Hyungsik, 2019. "Prediction of Shale-Gas Production at Duvernay Formation Using Deep-Learning Algorithm". SPE J. 24 (06): 2423-2437.

Liao, L., Zeng, Y., Liang, Y. y Zhang, H., 2020. Data Mining: A Novel Strategy for Production Forecast in Tight Hydrocarbon Resource in Canada by Random Forest Analysis. International Petroleum Technology Conference.

Nasir, E. y Rickabaugh, C., 2018. Optimizing Drilling Parameters Using a Random Forests ROP Model in the Permian Basin. Society of Petroleum Engineers.

Zhan C., Sankaran S., LeMoine V., Graybill J., Sher Mey, D., 2019. "Application of *Machine Learning* for Production Forecasting for Unconventional Resources". Unconventional Resources Technology Conference.