Mejora de la eficiencia operativa en equipos de producción de gas: detección temprana de eventos de erosión con modelos de *Machine Learning (ML)*

Por ***Camila Sabrina Juan Suriano; Jaime Andrés Vega Becerra*** (Practia), ***Julio Sandoval; Cristian Grau Viñolo; Ignacio Mason*** (Tecpetrol).

El uso de modelos de Machine Learning permite anticipar fallas por erosión en equipos clave de producción de gas, mejorando la eficiencia operativa y reduciendo riesgos. Este enfoque reemplaza el monitoreo tradicional con soluciones predictivas basadas en datos. Una innovación clave para garantizar la seguridad y continuidad de las operaciones.

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

Planteo del problema

En la industria del petróleo y gas, el aseguramiento del flujo de los pozos y la salud de los equipos de producción son temáticas de vital importancia debido a los altos costos asociados a las pérdidas de producción y a los costos de los equipos.

A partir del año 2020, comenzaron a ocurrir fallas por erosión en los calentadores de los PAD's. Estos eventos, además de los costos asociados mencionados anteriormente, también presentaban un alto riesgo en cuanto a seguridad de las operaciones. Por ello, se decidió abordar el problema, buscando una solución innovadora que permitiera mitigar estos efectos.

Hasta ese momento, la única forma de detectar señales previas de erosión en las instalaciones era mantener un operador de sala, encargado de monitorear todas las variables físicas de los activos, aunque los indicios capturados a partir de este monitoreo eran detectables solamente si se materializaban de manera abrupta y en algunos casos con el evento de falla ya teniendo curso

Por ende, se buscó ir más allá, y explorar que información adicional se podría obtener a partir de los datos disponible. En este contexto, surgió la inclusión de la ciencia de datos como herramienta para alcanzar una posible solución.

Desarrollo técnico del trabajo

A partir de la aparición de varios eventos de fallas causados por erosión en los chokes de calentadores (como la pérdida de contención primaria de fluidos del pozo), se decidió buscar una solución para prevenir estos incidentes mediante la generación de alertas que permitan actuar de forma preventiva. Dado que se disponía de información sobre múltiples variables físicas de los equipos (por ejemplo, temperatura, presión, etc.), se optó por explorar la utilización de modelos de machine learning supervisados como posible solución.

La clasificación supervisada es una rama del aprendizaje automático que se basa en el uso de un conjunto de datos de entrenamiento etiquetados para enseñar a un algoritmo cómo clasificar nuevos datos de entrada [1]. Para poder entrenar este tipo de algoritmos, es necesario contar con datos etiquetados, es decir, se precisa tener fallas detectadas con sus tiempos respectivos. Esta particularidad es de suma importancia, dado que, si bien se contaba con la detección visual de la falla, dicha detección era efectuada de forma manual y mientras el pozo se encontraba cerrado, adjudicándose una temporalidad posterior a la ocurrencia de esta. Este factor impulsa la necesidad de detectar y aislar un comportamiento característico previo dentro de la información en el análisis de datos históricos de presión y temperatura. La primera etapa de este proceso (Figura 1) involucró la recopilación y preprocesamiento de datos, asegurando que toda la información relevante estuviera limpia y adecuadamente estructurada. Esto incluyó la eliminación de valores atípicos, la imputación de datos faltantes y la normalización de las variables para estandarizar las unidades de medida y facilitar el análisis subsecuente. Esta fase de preprocesamiento de datos en Machine Learning es esencial para garantizar la calidad y eficacia de los modelos [2].

Una vez preparados los datos, se procedió al análisis descriptivo de las series temporales de los pozos en producción. En esta etapa, el factor clave fue la evaluación minuciosa del comportamiento de las principales curvas, presión y temperatura, previos a la ocurrencia de la falla, con el fin de detectar algún cambio característico a partir del cual comenzara a ocurrir la falla. Este hito de cambio marcaría el punto de inicio mediante el cual sería posible aislar el tramo de falla, de manera de etiquetarlo para poder entrenar al modelo supervisado con el comportamiento característico de la misma. Tras el análisis

Figura 1. **Metodología empleada en la resolución del problema.**

riguroso de varios eventos de falla ocurridos, fue posible aislar un cambio notorio en las pendientes de la presión y la temperatura previos a la detección de la falla. Si bien este cambio de pendiente podría ocurrir de forma más o menos acelerado, dicho cambio fue detectado y aislado en todos los casos analizados (Figura 2).

Con el objetivo de poder identificar este cambio de forma más precisa, se emplearon técnicas provenientes del análisis de series temporales para poder obtener, mediante descomposición aditiva, la tendencia que describe a las curvas analizadas y permitiendo disminuir el ruido presente en estas mismas. Este tipo de descomposición es la aconsejada a emplear en casos donde la magnitud de la estacionalidad no aumenta con el tiempo, pues en ese caso se debería utilizar la descomposición multiplicativa [3]. A partir de la tendencia obtenida, se calcularon las derivadas primera y segunda, permitiendo detectar así los cambios visualizados, automatizando la detección del inicio de cada falla. Este análisis preliminar proporcionó el poder de distinguir entre comportamiento normal y anómalo de los sistemas bajo estudio, sentando las bases para el desarrollo del modelo predictivo. Partiendo de la detección del inicio de falla, se generaron las etiquetas de estos intervalos prolongándose desde ese inicio hasta la indicación de cierre de pozo.

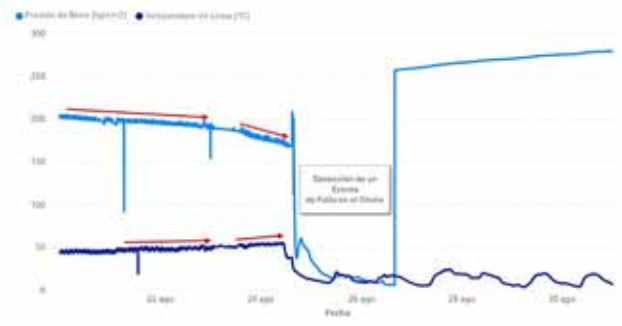

Figura 2. **Comportamiento observado previamente a la detección de la falla**

El siguiente paso en la metodología fue el desarrollo y entrenamiento del modelo de Machine Learning de clasificación. El desafío principal radicaba en el desequilibrio de datos de fallas frente a las condiciones operativas normales, lo que condujo al sobreajuste cuando se utilizaron técnicas tradicionales de aprendizaje supervisado. Se ha demostrado que los modelos de aprendizaje supervisado, como los basados en técnicas de ensamble, son robustos y capaces de manejar datos desbalanceados [4] y [5]. A pesar de las fortalezas mencionadas, los modelos sobreajustaban los datos de entrenamiento, especialmente al intentar generalizar en diferentes pozos. Con el fin de resolver esta situación, se utilizaron técnicas de remuestreo, pero no dieron resultados satisfactorios. El problema se agravó por la alta variabilidad entre diferentes pozos, lo que causaba que los modelos fallaran cuando se aplicaban a pozos no incluidos en el conjunto de entrenamiento. Para abordar estos desafíos, fue planteada la hipótesis de que la variabilidad entre los pozos estaba contribuyendo al problema del sobreajuste.

Finalmente, con el fin de mejorar la detección de fallas, se implementó una metodología en dos fases. En la primera fase, se llevó a cabo la clasificación no supervisada de los pozos. Para ello, se recolectaron los datos de todos pozos, que incluían las instancias normales. Luego, se utilizó una técnica basada en redes neuronales autoorganizativas para agrupar los pozos en un mapa bidimensional de características [6]. Posteriormente, se aplicó una técnica de clusterización sobre las neuronas para identificar clusters de pozos [7]. Este enfoque permitió agrupar pozos con características similares y desarrollar modelos personalizados para cada cluster. En la segunda fase, se llevó a cabo la clasificación supervisada para la detección de fallas. Con este fin, fueron entrenados modelos supervisados separados para cada grupo de pozos preagrupados, utilizando los datos de entrenamiento específicos de cada grupo de pozos similares. Luego se realizó la validación a partir de dejar un pozo fuera de los clusters para asegurar la robustez. Asimismo, se proba-

Figura 3. **Señal de alerta generada por el modelo**

ron los modelos en pozos dentro del mismo cluster para evaluar la mejora en el rendimiento [4] y [5], es decir, si en un cluster había N pozos, se dejaba uno fuera y se entrenaba con los N-1 pozos restantes, probando con el pozo restante. Es importante destacar que la predicción se consideraba correcta si el modelo predecía tempranamente la falla e incorrecta si no lo hacía. Por otro lado, para evitar generar alertas cortas y recurrentes asociadas con falsos positivos del modelo, se tomó la decisión de considerar la alerta válida luego de extenderse al menos durante 2 horas, siendo este el intervalo empleado tanto en la obtención de la tendencia por descomposición aditiva como del cálculo de derivadas. Esto significa que debe contarse con al menos un histórico de 2 horas de datos para poder correr el modelo correctamente. En la Figura 3 se evidencia como es la visualización final de alerta en función del tiempo, tomando como referencia la curva de presión.

Resultados obtenidos

Dentro de los principales resultados obtenidos, se destacan los siguientes puntos:

- La implementación del modelo de Machine Learning desarrollado para la detección temprana de fallas en el choke de producción arrojó resultados prometedores basándose en la validación mediante blind tests (aplicación del modelo a un conjunto de datos históricos no empleados dentro del entrenamiento inicial). Este modelo logró identificar más del 60% de los eventos de falla antes de su ocurrencia, lo que evidencia su capacidad para predecir fallas de manera efectiva. Estos resultados no solo demuestran la viabilidad de utilizar técnicas avanzadas de análisis de datos en el monitoreo de equipos estáticos, sino que también subrayan la importancia de una gestión proactiva en la industria del petróleo y gas.
- Tras el análisis desarrollado se comprobó que las derivadas primera y segunda de las curvas de presión y temperatura fueron indicadores críticos en la identificación de cambios en las tendencias operativas, que preceden a las fallas.
- El modelo fue capaz de distinguir entre el comportamiento normal y anómalo de los sistemas, permitiendo una intervención temprana para evitar eventos que podrían ser severos desde el punto de vista de la seguridad y reducir el tiempo de inactividad.

Actualmente el modelo está siendo optimizado a partir de la mejora en la definición de las etiquetas y mediante la incorporación de nuevos casos, sumando más información al entrenamiento y mejorando su capacidad de generalización.

Conclusiones

Los hallazgos sugieren que la variabilidad entre los pozos es un factor crítico que contribuye al sobreajuste en las tareas de detección de fallas. Este desafío fue resuelto mediante el uso de la técnica de clusterización para agrupar pozos similares antes de aplicar modelos supervisados. Este enfoque aseguró que los modelos fueran más adecuados para las condiciones específicas de cada cluster, mejorando así la precisión en la detección. Además, la combinación de técnicas no supervisadas permitió una preclasificación efectiva de los pozos, facilitando el entrenamiento y la validación de modelos supervisados dentro de cada cluster.

La capacidad del modelo para distinguir entre comportamientos normales y anómalos en las curvas de presión y temperatura ha permitido intervenir tempranamente, evitando eventos potencialmente severos y reduciendo significativamente el tiempo de inactividad. Esto no solo mejora la seguridad operativa, sino que también optimiza la productividad y minimiza los costos asociados a interrupciones inesperadas en la producción. Este resultado subraya la viabilidad de aplicar técnicas avanzadas de análisis de datos en el monitoreo de equipos estáticos, promoviendo una gestión operativa más proactiva.

En resumen, la implementación del modelo de Machine Learning desarrollado para la detección temprana de fallas en el choke de producción ha demostrado ser una herramienta eficaz para la industria del petróleo y gas.

Bibliografía

[1] 6 Métodos de clasificación | Estadística y Machine Learning con R (bookdown.org)

[2] https://iartificial.blog/aplicaciones/la-importancia-del-preprocesamiento-de-datos-en-el-machine-learning/

[3] "Descomposición de Series Temporales" (https://es.planetcalc.com/7910/)

[3]Breiman, L. (2001). Random Forest. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324

[5] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley.

[6] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480. doi:10.1109/5.58325

[7] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.